An Adaptive Spline-Based Sieve Semiparametric Maximum Likelihood Estimation for the Cox Model with Interval-Censored Data

By YING ZHANG

Department of Biostatistics, The University of Iowa

ying-j-zhang@uiowa.edu

Jiang Huang

Department of Statistics and Actuarial Science, The University of Iowa

jian-huang@uiowa.edu

AND Lei Hua

Department of Biostatistics, The University of Iowa

lei-hua@uiowa.edu

Summary

We propose to analyze interval-censored data with Cox model using a spline-based sieve semiparametric maximum likelihood approach in which the baseline cumulative hazard function is approximated by a monotone B-splines function. We apply the generalized Rosen algorithm, used in Zhang & Jamshidian (2004), for computing the maximum likelihood estimate. We show that the the estimator of regression parameter is asymptotically normal and semiparametrically efficient. We also develop an easy-to-implement method to consistently estimate the standard error of the regression parameter estimate that facilitates an adaptive

inference procedure for semiparametric likelihood analysis for interval censored-data with the Cox model. The method is evaluated by simulation studies regarding its finite sample performance and is illustrated using the data from breast cosmesis study.

Some key words: B-splines; Convergence rate; Counting process; Current status data; Empirical processes; Efficient estimation; Monotone polynomial splines; Proportional odds model; Semiparametric model

1. Introduction

Interval censoring refers to a censoring mechanism where an event time can not be directly observed but only is known to lie between two adjacent examination times in a sequence of examinations or follow-up visits. An important application of the analysis of interval-censored data is in HIV/AIDS studies. Examples include Goggins & Finkelstein (2000), Betensky et al. (2001), Seaman & Bird (2001), Gómez et al. (2003), Song & Ma (2008) and Hsu et al. (2007), among others. Recently, the analysis of interval-censored data has appeared in many other biomedical and epidemiological studies. For example, Kim & Xue (2002) analyze interval-censored data for an ongoing clinical trial for systemic lupus erythematosus, Bogaerts et al. (2002) analyze multivariate interval-censored dental data, Bellamy et al. (2004) develop a parametric frailty model for clustered and interval-censored data with application to the East Boston Asthma study, and Sparling et al. (2006) study the risk of progression of diabetic retinopathy with parametric survival models for interval-censored data.

The development of regression analysis of interval-censored data has been very active in the last decade. While the likelihood-based approach for the Weibull parametric models with interval-censored data has been implemented, exemplified by Bellamy et al. (2004) and Sparling et al. (2006), most work has been focusing on semiparametric models. Imputation-based approach was proposed by Satten et al. (1998), Song & Ma (2008), Zhang et al. (2008) in which interval-censored event times are imputed and then some well-known semiparametric regression methods such as Cox model (1972) for right censored data can be handily utilized. However, the imputation methods in general produce biased estimates for the regression parameter. The semiparametric accelerated failure time model for interval-censored data was considered by Rabinowitz et al. (1995), Li & Zhang (1998), Betensky et al. (2001), Li & Pu (2003), and Gómez et al. (2003).

The Cox model, the most popular semiparametric model in the regression analysis with right censored data, has also been considered in the analysis for interval-censored data. However, for the likelihood analysis of this model with interval-censored data, the baseline hazard function can not be eliminated using the partial likelihood approach as in the case of right censored data. One has to estimate the regression parameter and the baseline hazard jointly. This turns out to be a challenging task both numerically and theoretically. Finkelstein (1986) appears to be the first to propose the Cox model for interval-censored data with discrete hazard assumption. With this set up, the semiparametric regression problem is essentially converted to a parametric regression problem. This approach has been adopted by Goggins & Finkelstein (2000), Seaman & Bird (2001) and Kim & Xue (2002). The

fully semiparametric maximum likelihood analysis for current status data is developed by Huang (1996) in which he showed that despite the nonparametric estimator of the baseline cumulative hazard function converges slower than the standard rate $n^{1/2}$, the maximum likelihood estimator of regression parameter can still be asymptotically normal and achieves the semiparametric efficiency bound defined in Bicke et al. (1993). Although Huang & Wellner (1997) discussed the possible extension of Huang (1996) to interval-censored data, to the best of our knowledge, the theory and numerical implementation of semiparametric maximum likelihood analysis of interval-censored data have not been fully developed in statistical literatures.

In this article, we address the theoretical and numerical challenges in the semiparametric estimation of the Cox proportional hazards model with interval-censored data. We propose an adaptive spline-based sieve semiparametric likelihood estimation procedure, in which the log baseline cumulative hazard function is approximated by monotone B-splines (Schumaker, 1981). The generalized Rosen's algorithm, proposed by Zhang & Jamshidian (2004) for computing the nonparametric maximum likelihood estimator with linear inequality constraints, is implemented to compute the sieve semiparametric maximum likelihood estimate. We show that the proposed estimator of the regression parameter is asymptotically normal and semiparametrically efficient, and the spline-based sieve estimator of the baseline hazard function can converge faster than that based on the ordinary semiparametric maximum likelihood analysis described in Huang & Wellner (1997). We also develop an easy-to-implement method to consistently estimate the standard error based on the ordinary

least-squares approach, in order to make statistical inference using the asymptotic results. The proposed method facilitates an easy-to-implement semiparametric likelihood inference procedure for analyzing interval-censored data with the Cox model and will be promising in general semiparametric inference problem.

The rest of the paper is organized as follows: Section 2 describes the model and likelihood for interval-censored data; Section 3 introduces the spline-based sieve semiparametric maximum likelihood approach and the generalized Rosen algorithm for computing the estimate; Section 4 presents the asymptotic results of the estimator; Section 5 provides numerical results consisting of simulation studies and application to an illustrating example in breast cosmesis study; Section 6 discusses the further application of the proposed method; Finally the technical details are outlined in Appendices.

2. Model and Likelihood

Consider the Cox proportional hazards model, in which the conditional hazard of T given a covariate vector $Z \in R^d$ is proportional to the baseline hazard (the hazard for $Z = 0$). In terms of the cumulative hazard function, this model is

$$\Lambda(t|z) = \Lambda_0(t) \exp(\theta_0' z), \tag{1}$$

where θ_0 is a d-dimensional regression parameter and Λ_0 is the unspecified baseline cumulative hazard function.

Let (U, V) be the pair of examination times bracketing the event time T. That is, U is

the last examination time before and V is the first examination time after the event. Let G_z be the joint distribution function of (U, V) given covariate $Z = z$ with $P(U \leq V|z) = 1$ for any $z \in R^d$ and $H(z)$ be the distribution of Z. Let $\delta_1 = 1_{[T \leq U]}, \delta_2 = 1_{[U<T \leq V]}$ and $\delta_3 = 1 - \delta_1 - \delta_2$ and denote the observation from a single subject by $X = (\delta_1, \delta_2, \delta_3, U, V, Z)$. Under the common assumption that conditional on Z, T is independent of (U, V), the joint density of X is given by

$$p(x) = F(u|z)^{\delta_1} \{F(v|z) - F(u|z)\}^{\delta_2} \{1 - F(u|z)\}^{\delta_3} g_z(u, v)h(z),$$

where $F(\cdot|z)$ is the conditional distribution function of the event time and $g_z(u, v)$ and $h(z)$ are the density functions of G_z and H, respectively. Further assume that the distribution of (U, V) is noninformative of T, then under the Cox model, the log-likelihood of an identically and independently distributed sample $X_i = (\delta_{1i}, \delta_{2i}, \delta_{3i}, U_i, V_i, Z_i)$ for $i = 1, 2, \ldots, n$ is given by

$$\begin{aligned} l_n(\theta, \Lambda; \cdot) &= \sum_{i=1}^{n} \Big(\delta_{1i} \log \left[1 - \exp \left\{ -\Lambda(u_i) e^{\theta' z_i} \right\} \right] \\ &\quad + \delta_{2i} \log \left[\exp \left\{ -\Lambda(u_i) e^{\theta' z_i} \right\} - \exp \left\{ -\Lambda(v_i) e^{\theta' z_i} \right\} \right] - \delta_{3i} \Lambda(v_i) e^{\theta' z_i} \Big), \end{aligned}$$

omitting the additive terms that do not involve (θ, Λ). Let $\phi = \log \Lambda$, the resulting log-likelihood in terms of (θ, ϕ) is

$$\begin{aligned} l_n(\theta, \Lambda; \cdot) &= \sum_{i=1}^{n} \Big(\delta_{1i} \log \left[1 - \exp \left\{ -e^{\theta' z_i + \phi(u_i)} \right\} \right] \\ &\quad + \delta_{2i} \log \left[\exp \left\{ -e^{\theta' z_i + \phi(u_i)} \right\} - \exp \left\{ -e^{\theta' z_i + \phi(v_i)} \right\} \right] - \delta_{3i} e^{\theta' z_i + \phi(v_i)} \Big). \quad (2) \end{aligned}$$

3. Spline-Based Sieve Maximum Likelihood Estimation

Suppose $0 = t_0 < t_1 < t_2 < \cdots < t_m < \infty$ are the distinct time points in the collection of $\{U_i, V_i : \quad i = 1, 2, \cdots, n\}$. The value of the log likelihood function (2) is completely determined by the values of ϕ at these points and θ. Conventionally, the semiparametric maximum likelihood estimator is sought by maximizing (2) with respect to θ and $\phi(t_i)$, for $i = 1, 2, \cdots, m$. The upper bound of m is $2n$ if there is no tie among $\{U_i, V_i\}$, $i = 1, 2, \cdots, n$. However, as Huang & Wellner (1997) pointed out that this optimization problem is hard to solve, particularly when θ is a multidimensional vector and sample size is large.

To ease the numerical difficulty in nonparametric estimation problem, Geman & Hwang (1982) proposed a sieve maximum likelihood estimation procedure for which the unknown function in the log likelihood is approximated by a linear span of some known basis functions to form a sieve log likelihood. Then maximizing the log likelihood with respect to the unknown function converts to maximizing the sieve log likelihood with respect to the unknown coefficients in the linear span. This, in turn, reduces the dimensionality of the optimization problem significantly since the number of basis functions required to reasonably approximate the unknown function grows a lot slower as same size increases.

Spline technique has been well recognized in statistical literature as an efficient tool in dimension reduction for nonparametric estimation since the theoretical development on spline estimation by Stone (1985,1986). Therefore, it is natural to consider spline-based sieve maximum likelihood estimation in the context of regression models with interval-censored

data. Some further theoretical results of spline-based sieve estimator has been obtained by Shen & Wong (1994). Shen (1998) has also applied the spline-based sieve maximum likelihood estimation to proportional odds model with censored data. Other applications of splines in analyzing interval-censored data can be found in Koopetberg & Clarkson (1997) and Cai & Betensky (2003).

We now describe the spline-based sieve semiparametric maximum likelihood estimation for the Cox model with interval-censored data. Suppose a and b are the lower and upper bounds of censoring times $\{(U_i, V_i) : \quad i = 1, 2, \cdots, n\}$. Let $a = d_0 < d_1 < \cdots < d_{K_n} < d_{K_n+1} = b$ be a partition of $[a, b]$ into $K_n + 1$ subintervals $I_{Kt} = [d_t, d_{t+1}), t = 0, \ldots, K$, where $K \equiv K_n \approx n^v$ is a positive integer such that $\max_{1 \le k \le K+1} |d_k - d_{k-1}| = O(n^{-v})$. Denote the set of partition points by $D_n = \{d_1, \ldots, d_{K_n}\}$. Let $\mathcal{S}_n(D_n, K_n, m)$ be the space of polynomial splines of order $m \ge 1$ consisting of functions s satisfying: (i) the restriction of s to I_{Kt} is a polynomial of order m for $m \le K$; (ii) for $m \ge 2$ and $0 \le m' \le m - 2$, s is m' times continuously differentiable on $[a, b]$. This definition is phrased after Stone (1985), which is a descriptive version of Schumaker (1981), page 108, Definition 4.1. According to Schumaker (1981), page 117, Corollary 4.10, there exists a *local* basis $\mathcal{B}_n \equiv \{\mathbf{b}_t, 1 \le t \le q_n\}$, so called B-splines, for $\mathcal{S}_n(D_n, K_n, m)$, where $q_n \equiv K_n + m$. These basis functions are nonnegative and sum up to one at each point in $[a, b]$, and each $\mathbf{b}_t$ is zero outside the interval $[d_t, d_{t+m}]$.

Because ϕ in (2) is a nondecreasing function, it is desirable to restrict its estimate to be

nondecreasing as well. Let

$$\mathcal{M}_n(D_n, K_n, m) = \left\{ \phi_n : \phi_n(t) = \sum_{j=1}^{q_n} \beta_j \mathbf{b}_j(t) \in \mathcal{S}_n(D_n, K_n, m),\ \beta \in B_n,\ \ t \in [a, b] \right\}.$$

where $B_n = \{\beta : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{q_n}\}$. Each element of $\mathcal{M}_n(D_n, K_n, m)$ is a nondecreasing function because of the monotonicity constraints on $\beta_1, \ldots, \beta_{q_n}$. This fact is a consequence of the *variation diminishing properties* of B-splines. See for instance, Schumaker (1981), Example 4.75 and Theorem 4.76, pages 177-178. Denote $\Theta \in R^d$ the feasible domain for the regression parameter and abbreviate $\mathcal{M}_n(D_n, K_n, m)$ by $\mathcal{M}_n$. We look for $\hat{\tau}_n = (\hat{\theta}_n, \hat{\phi}_n)$ that maximizes $l_n(\theta, \phi; \cdot)$ over $\Theta \times \mathcal{M}_n$. This is equivalent to maximizing $l_n(\theta, \mathcal{B}_n'\beta; \cdot)$ over $\Theta \times B_n$. No restriction will be imposed on Θ in the optimization.

For restricted parametric maximum likelihood estimation problems, Jamshidian (2004) generalized the gradient projection algorithm originally proposed by Rosen (1960) using the generalized Euclidean metric $\|x\| = x^T W x$, where W is a positive definite matrix and possibly varying from iteration to iteration. Zhang & Jamshidian (2004) applied the algorithm to large-scale nonparametric maximum likelihood estimation problems by choosing $W = D_H$, the matrix containing only the diagonal elements of the negative Hessian matrix H, in order to avoid the storage problem in updating H. However, this will increase the number of iterations and thereby the computing time. In this article, we use $W = -H$ directly because the dimension of unknown parameter space is largely reduced due to the use of splines. The numerical advantage of this algorithm over the iterative convex minorant algorithm studied by Jongbloed (1998) for nonparametric maximum likelihood estimation with monotone constraints has been demonstrated by Lu et al. (2007). In the following, we describe the

algorithm for computing the proposed spline-based sieve semiparametric estimate.

Let $\dot{\ell}(\tau)$ and W be the gradient and negative Hessian matrix of the log likelihood given by (2) with respect to $\tau = (\theta, \beta)$, respectively. Let $\mathcal{A} = \{i_1, i_2, \cdots, i_r\}$ denote the index set of active constraints, i.e. $\alpha_{i_j} = \alpha_{i_j+1}$, for $j = 1, 2, \cdots, r$, during the numerical computation. We define a working matrix corresponding to this set,

$$A = \begin{bmatrix} 0 & \cdots & -1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & -1 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 & \cdots & 0 \end{bmatrix}_{r \times (q_n + d)}.$$

The generalized Rosen algorithm is implemented in the following steps:

S0: (**Computing the feasible search direction**)

$$\underline{d} = \left\{ I - W^{-1} A^T \left(A W^{-1} A^T \right)^{-1} A \right\} W^{-1} \dot{\ell}(\tau).$$

S1: (**Forcing the updated τ fulfill the constraints**) If the resulted direction $\underline{d}$ is not nondecreasing in its components, compute

$$\gamma = \min_{i \notin \mathcal{A} \text{ and } d_i > d_{i+1}} \left(-\frac{\alpha_{i+1} - \alpha_i}{d_{i+1} - d_i} \right).$$

Doing so guarantees that $\alpha_{i+1} + \gamma d_{i+1} \geq \alpha_i + \gamma d_i$, for $i = 1, 2, \cdots, q_n$.

S2: (**Step-Halving line search**) Looking for a smallest integer k starting from 0 such that

$$\ell \left\{ \tau + (1/2)^k \, \underline{d} \right\} > \ell(\tau).$$

S3: (**Updating the solution**) If $\gamma > (1/2)^k$, replace τ by $\tilde{\tau} = \tau + (1/2)^k \underline{d}$ and check the stopping criterion (S5).

S4: (**Updating the active constraint set**) If $\gamma \leq (1/2)^k$, in addition to replace τ by $\tilde{\tau} = \tau + \gamma \underline{d}$, modify $\mathcal{A}$ by adding indexes of all the newly active constraints to $\mathcal{A}$ and accordingly modify the working matrix A.

S5: (**Checking the stopping criterion**) If $\|\underline{d}\| \geq \varepsilon$ for a small $\varepsilon > 0$, go to S0. Otherwise, compute $\lambda = \left(AW^{-1}A^T\right)^{-1} AW^{-1}\dot{\ell}(\tau)$.

i. If $\lambda_i \leq 0$ for all $i \in \mathcal{A}$, set $\hat{\tau} = \tau$ and stop.

ii. If at least one $\lambda_i > 0$ for $i \in \mathcal{A}$, remove the index corresponding to the largest λ_i from $\mathcal{A}$, and update A and go to S0.

4. Asymptotic Properties

In this section, we describe the asymptotic results of the estimator. For any $\phi_1, \phi_2 \in \Phi$, define

$$\|\phi_1 - \phi_2\|_\Phi^2 = E\{\phi_1(U) - \phi_2(U)\}^2 + E\{\phi_1(V) - \phi_2(V)\}^2.$$

and for any $\tau_1 = (\theta_1, \phi_1)$ and $\tau_2 = (\theta_2, \phi_2)$ in the space of $\mathcal{T} = \Theta \times \Phi$, define

$$d(\tau_1, \tau_2) = \|\tau_1 - \tau_2\|_\mathcal{T} = \left\{\|\theta_1 - \theta_2\|^2 + \|\phi_1 - \phi_2\|_\Phi^2\right\}^{1/2}.$$

As usual, the study of asymptotic properties of the semiparametric maximum likelihood estimator requires some regularity conditions to be satisfied. The following conditions sufficiently guarantee the results in the forthcoming theorems.

(C1) (a) $E(ZZ')$ is nonsingular; (b) Z is bounded, that is, there exists $z_0 > 0$ such that $P(\|Z\| \leq z_0) = 1$.

(C2) Θ is a compact subset of R^d.

(C3) (a) There exists a positive number η such that $P(V - U \geq \eta) = 1$; (b) the union of the supports of U and V is contained in an interval $[a, b]$, where $0 < a < b < \infty$, and $0 < \Lambda_0(a) < \Lambda_0(b) < \infty$.

(C4) $\phi_0 = \log \Lambda_0$ belongs to Φ, a class of functions with bounded pth derivative in $[a, b]$ for $p \geq 1$ and the first derivative of ϕ_0 is strictly positive and continuous on $[a, b]$.

(C5) The conditional density $g(u, v|z)$ of (U, V) given Z has bounded partial derivatives with respect to (u, v). The bounds of these partial derivatives do not depend on (u, v, z).

(C6) For some $\kappa \in (0, 1)$, $a^T var(Z|U)a \geq \kappa a^T E(ZZ^T|U)a$ and $a^T var(Z|V)a \geq \kappa a^T E(ZZ^T|V)a$ a.s. for all $a \in R^d$.

These conditions are usually satisfied in practice. Although some of these conditions may be stronger than needed and could be weakened, it will make the proof considerably more difficult.

THEOREM 1. *Let $K_n = O(n^\nu)$, where ν satisfies the restriction $\frac{1}{2(1+p)} < \nu < \frac{1}{2p}$. Suppose that T and (U, V) are conditionally independent given Z and that the distribution of (U, V, Z) does not involve (θ, Λ). Furthermore, suppose that conditions (C1)–(C6) hold. Then*

$$d\left(\widehat{\tau}_n, \tau_0\right) = O_p\left\{n^{-\min(p\nu, (1-\nu)/2)}\right\}.$$

This theorem implies that if $\nu = 1/(1+2p)$, $d(\widehat{\tau}_n, \tau_0) = O_p\left\{n^{-p/(1+2p)}\right\}$ which is the optimal convergence rate in the nonparametric regression setting. So if the baseline hazard function is smooth, the proposed estimator can achieve a better convergence rate than the conventional semiparametric estimator considered in Huang & Wellner (1997). Although the overall convergence rate is lower than $n^{1/2}$, the proposed estimator of the regression parameter is still asymptotically normal and can be shown semiparametrically efficient.

THEOREM 2. *Suppose the conditions given in Theorem 1 hold, then*

$$n^{1/2}\left(\hat{\theta}_n - \theta_0\right) \to N\left\{0, I^{-1}(\theta_0)\right\}$$

in distribution.

In Theorem 2, $I(\theta_0)$ is the information matrix evaluated at θ_0 based on the general semiparametric information theory described by Bicke et al. (1993). The theorem implies that the estimator of regression parameter, $\hat{\theta}$ converges to the true parameter at the usual root-n rate and achieves the semiparametric efficiency bound despite the lower convergence rate of the nonparametric component. Below we describe an approach for estimating $I(\theta_0)$.

Let

$$\begin{aligned} l(\theta, \phi; x) &= \delta_1 \log\left[1 - \exp\left\{-e^{\theta' z + \phi(u)}\right\}\right] + \delta_2\left[\exp\left\{-e^{\theta' z + \phi(u)}\right\} - \exp\left\{-e^{\theta' z + \phi(v)}\right\}\right] \\ &\quad -\delta_3 e^{\theta' z + \phi(v)} \end{aligned}$$

be the log-likelihood for a sample of size one. Consider a parametric smooth submodel with parameter $(\theta, \phi_{(s)})$, where $\phi_{(0)} = \phi$ and

$$\left.\frac{\partial \phi_{(s)}}{\partial s}\right|_{s=0} = h.$$

Let $\mathcal{H}$ be the class of functions h defined by this equation. The score operator for ϕ is

$$\dot{l}_2(\tau; x)(h) = \left.\frac{\partial}{\partial s} l(\theta, \phi_{(s)}; x)\right|_{s=0}. \tag{3}$$

For a d-dimensional θ, $\dot{l}_1(\tau; x)$ is the vector of partial derivatives of $l(\tau; x)$ with respect to the components of θ. For each component of $\dot{l}_1$, a score operator for ϕ is defined as in (3). So the score operator for ϕ corresponding to $\dot{l}_1$ is

$$\dot{l}_2(\tau; x)(\mathbf{h}) \equiv \{\dot{l}_2(\tau; x)(h_1), \ldots, \dot{l}_2(\tau; x)(h_d)\}', \tag{4}$$

where $\mathbf{h} \equiv (h_1, \ldots, h_d)'$ with $h_k \in \mathcal{H}, 1 \leq k \leq d$.

According to Bicke et al. (1993), Theorem 1, page 70, the efficient score vector for θ is $\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(\xi_0)$, where ξ_0 is an element of $\mathcal{H}^d$ that minimizes

$$\rho(\mathbf{h}) \equiv E\|\dot{l}_1(\tau; X) - \dot{l}_2(\tau; X)(\mathbf{h})\|^2 \tag{5}$$

over $\mathcal{H}^d$. The minimizer $\xi_0 = (\xi_{01}, \xi_{02}, \ldots, \xi_{0d})'$ is called the *least favorable direction*. Denote the efficient score by $l^*(\tau; x) \equiv \dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(\xi_0)$. Then the information for θ is

$$I(\theta) = E\|l^*(\tau; X)\|^2 = E\|\dot{l}_1(\tau; X) - \dot{l}_2(\tau; X)(\xi_0)\|^2. \tag{6}$$

With interval-censored data, the least favorable direction $\xi_0(t)$ has no explicit solution and in fact, it is the solution of a Fredholm integral equation of the second kind,

$$\xi_0(t) - \int K(t,x)\xi_0(x)dx = d(t)$$

with two complicate functions $K(t,x)$ and $d(t)$ described in Huang & Wellner (1997). Apparently, a direct estimation of $\xi_0(t)$ for the information matrix is impossible. Nevertheless, the definition of $\xi_0(t)$ given by (5) leads us to consider a least-squares estimator of the information matrix. The detailed development of the least-squares method for consistent variance estimation in semiparametric models is given by Huang et al. (2008). Specifically, with the random sample $X_1, \ldots, X_n$ and the consistent estimator $\widehat{\tau}_n$, we can estimate $I(\theta)$ by the minimum value of

$$\rho_n(\mathbf{h}) \equiv n^{-1} \sum_{i=1}^{n} \|\dot{l}_1(\widehat{\tau}_n; X_i) - \dot{l}_2(\widehat{\tau}_n; X_i)(\mathbf{h})\|^2 \tag{7}$$

over $\mathcal{H}^d$. That is, if $\widehat{\xi}_n$ is a minimizer of ρ_n over $\mathcal{H}^d$, then a natural estimator of $I(\theta_0)$ is $\widehat{\mathcal{I}}_n \equiv \rho_n(\widehat{\xi}_n)$.

In practice, one can easily estimate the components of optimal $\hat{\xi}_n$ using the ordinary least-squares regression with the Hilbert space $\mathcal{H}_n$ linearly spanned by the B-splines basis functions $\mathcal{B}_n$. This estimation is implemented in the subsequent simulation studies and application.

5. Numerical Results

5.1 *Simulation Studies*

Simulation studies are carried out to evaluate the finite sample performance of the proposed method. Interval-censored data are generated as follows: for each subject, we independently generate $X_i = (\delta_{i,1}\delta_{i,2}, \delta_{i,3}, U_i, V_i, Z_i)$, for $i = 1, 2, \cdots, n$, where the event time is generated according to the Cox model $\Lambda(t|Z) = t^{1/2}\exp(\theta_0^T Z)$ for which the true parameters are $\theta_0 = (-1.0, 0.5, 1.5)^T$ and $\log\Lambda_0(t) = 0.5\log t$, the covariate vector $Z_i = (Z_{1,i}, Z_{2,i}, Z_{3,i})^T$ is simulated by $Z_{1,i} \sim$ Uniform$(0, 1)$, $Z_{2,i} \sim$ Normal$(0, 1)$, and $Z_{3,i} \sim$ Bernoulli(0.5); a series of examination times are produced by the partial sum of inter-arrival times that are independently and identically distributed according to Exp(0.5), U_i is the last examination time before 5 at which the event has not occurred yet and V_i is the first observation time before 5 at which the event has occurred.

We perform the sieve semiparamteric maximum likelihood analysis using cubic B-splines and estimate the standard error of the regression parameter estimates using the least-squares method based on cubic B-splines as well. For the B-splines, the number of knots is chosen to be $K_n = [N^{1/3}]$, the largest integer below $N^{1/3}$, where N is the number of distinct observation time points of the collection $\{(U_i, V_i) : \ \ i = 1, 2, \cdots, n\}$, and the knots are placed at the K_n quantiles of the N distinct observation times.

We consider three different sample sizes: n =50, 100, and 200. In each case, the Monte-Carlo simulations with 1000 repetitions are conducted. Table 1 displays the estimation bias

Table 1

Simulation Results of the Monte-Carlo study for the sieve semiparametric maximum likelihood analysis of θ_0 with 1000 repetitions

	$\theta_{1,0}$			$\theta_{2,0}$			$\theta_{3,0}$		
	n=50	n=100	n=200	n=50	n=100	n=200	n=50	n=100	n=200
Bias	-0.1280	-0.0754	-0.0335	0.1055	0.0304	0.0132	0.2172	0.0917	0.0498
M-C sd	0.7613	0.4662	0.3160	0.4352	0.2716	0.1908	0.4967	0.2834	0.2056
ASE	0.8514	0.5086	0.3310	0.5002	0.3010	0.1944	0.5663	0.3283	0.2098
95%-CP	97.6%	97.1%	96.2%	98.7%	97.2%	95.0%	0.98.2%	0.97.7%	0.95.6%

(Bias), Monte Carlo standard deviation (MC s.d), the average of standard errors (ASE) based on the asymptotic result given in Theorem 2, and the coverage probability of 95% Wald-confidence interval for $\hat{\theta}_n$. The results show that this adaptive spline-based seive semiparametric maximum likelihood estimation method performs quite well: the bias is negligible compared to the standard error and the estimated standard error decreases as sample size increases. The least-squares method for estimating the standard error overestimates the true standard error slightly, but the overestimation lessens as sample size increases and it provides a reasonable estimate of the standard error when sample size reaches 200. As the result of overestimation, the coverage probability of 95% confidence interval exceeds the nominal value but approaches to 95% when sample size increases to 200. In addition, we

also plot in Fig. 1 the averages of the B-spline sieve estimates of the true log cumulative hazard function for the case $0.5 \log t$. It shows that estimation bias is relatively large when the sample size is small ($n = 50$) but drops significantly when the sample size increases to 200.

5.2 *Breast Cosmesis Study*

The breast cosmesis study is a clinical trial for comparing radiotherapy alone with primary radiotherapy with adjuvant chemotherapy in terms of subsequent cosmetic deterioration of the breast following tumorectomy. Subjects (46 assigned to radiotherapy alone and 48 to radiotherapy plus chemotherapy) were followed for up to 60 months, with pre-scheduled follow-up visits for every 4-6 months. In this paper, we use the Cox model to analyze the difference of the hazard for the time until the appearance of breast retraction between the two treatments,

$$\Lambda(t|Z) = \Lambda_0(t) \exp(\theta_0 Z), \tag{8}$$

where Λ_0 is the baseline hazard (the cumulative hazard for radiotherapy alone) and Z is the indicator for the treatment of radiotherapy plus chemotherapy. Using the method proposed in this paper, the cubic B-splines sieve semiparametric maximum likelihood estimate of θ_0 is $\widehat{\theta}_n = 0.8948$ with asymptotic standard error given by 0.2926. The Wald test statistic is $Z = 3.0582$ with p-value=0.0011. This indicates that the treatment of radiotherapy with adjuvant chemotherapy significantly increases the risk of the breast retraction and the result is comparable with what has been concluded in Finkelstein & Wolf (1985).

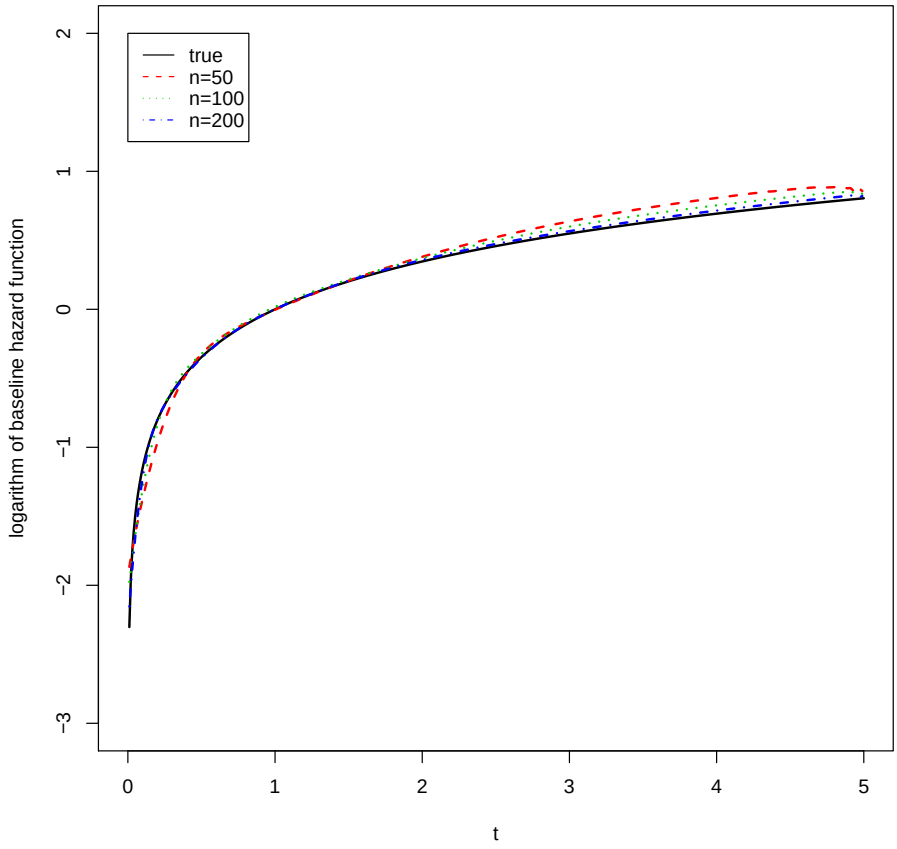

Figure 1. The average of the B-splines sieve maximum likelihood estimates of the log baseline cumulative hazard function with 1000 repetitions

6. Final Remarks

In this article, we proposed an adaptive spline-based sieve semiparametric maximum likelihood method. This method reduces the dimensionality of the estimation problem using the splines and therefore releases the numerical burden of the computation without interfering the asymptotic properties of the regression parameter estimates. An adaptive spline method for consistently estimating the standard error is also developed in order to make inference of the regression parameter. Although the spline function is used to approximate the baseline log cumulative hazard function for the sieve likelihood, our simulation experiments indicate that the number and placement of the spline knots have very little impact on the inference made for the regression parameter using the proposed method and hence this method facilitates a practical and easy-to-implement semiparametric likelihood inference procedure for analyzing interval-censored data with the Cox model which is often viewed as a challenging task in the liteartures.

It should be a straightforward task to apply the method presented here to other semiparametric regression models for interval-censored data such as the partial linear regression proposed by Xue et al. (2004), the proportional odds regression studied by Huang & Rossini (1997) and Shen (1998), and the additive hazard model studied by Lin et al. (1998) and Martinussen & Scheike (2002). In principle, our proposed method can be applied to any semiparametric maximum likelihood estimation problems in which the maximum likelihood estimator of finite-dimensional parameter can be shown asymptotically efficient and the nu-

merical computation for infinite-dimensional nuisance parameter is a burden,

References

Bellamy, S., Li, Y., Ryan, L., Lipsitz, S., Canner, M. & Wright, R. (2004). Analysis of clustered and interval censored data from a community-based study in asthma. *Statist. Med.* 23, 3607–21.

Betensky, R., Rabinowitz, D. & Tsiatis, A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* 88, 703–11.

Bicke, P., Klaassen, C., Ritov, Y. & Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Billingsley, P. (1986). *Probability and Measure.* John Wiley, New York.

Bogaerts, K., Leroy, R., Lesaffre, E. & Declerck, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statist. Med.* 21, 3775–87.

Cai, T. & Betensky, R. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* 59, 570–9.

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* 42, 845–54.

Finkelstein, D. & Wolf, R. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 41, 933–45.

Geman, A. & Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* 10, 401–14.

Goggins, W. & Finkelstein, D. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* 56, 940–3.

Gómez, G., Espinal, A. & Lagakos, S. (2003). Inference for a linear regression model with an interval-censored covariate. *Statist. Med.* 22, 409–25.

Hsu, C., Taylor, J., Murray, S. & Commenges, D. (2007). Multiple imputation for interval-censored data with auxiliary variables. *Statist. Med.* 26, 769–81.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* 24, 540–68.

Huang, J. & Rossini, A. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Statist. Assoc.* 92, 960–7.

Huang, J. & Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the first Seattle Symposium in Biostatistics: Survival Analysis*. New York: Springer-Verlag, 123–69.

Huang, J., Zhang, Y. & Hua, L. (2008). A least-squares approach to consistent information estimation in semiparametric models. *Technical Report, Department of Biostatistics, The University of Iowa* .

Jamshidian, M. (2004). On algorithms for restricted maximum likelihood estimation. *Comput. Statist. and Data Anal.* 45, 137–57.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comput. Graph. Statist.* 7, 310–21.

Kim, M. & Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statist. Med.* 21, 3715–26.

Koopetberg, C. & Clarkson, D. (1997). Hazard regression with interval-censored data. *Biometrics* 53, 1485–94.

Li, G. & Zhang, C. (1998). Linear regression with interval censored data. *Ann. Statist.* 26, 1306–27.

Li, L. & Pu, Z. (2003). Rank estimation of log-linear regression with interval-censored data. *Lifetime Data Analysis* 9, 57–70.

Lin, D., Oakes, D. & Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* 85, 289–98.

Lu, M. (2007). *Monotone Spline Estimations for Panel Count Data.* Ph.D Dissertation, Department of Biostatistics, The University of Iowa.

Lu, M., Zhang, Y. & Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* 94, 705–18.

Martinussen, T. & Scheike, T. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* 89, 649–58.

Rabinowitz, D., Tsiatis, A. & Aragon, J. (1995). Regression with interval-censored data. *Biometrika* 82, 501–13.

Rosen, J. (1960). The gradient projection method for nonlinear programming. *J. Soc. Indust. Appl. Math.* 8, 181–217.

Satten, G., Datta, S. & Williamson, J. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *J. Am. Statist. Assoc.* 93, 318–27.

Schumaker, L. (1981). *Spline Function: Basic Theory.* New York: Wiley.

Seaman, S. & Bird, S. (2001). Proportional hazards model for interval-censored failure times and time-dependent covariates: application to hazard of hiv infection of injecting drug users in prison. *Statist. Med.* 20, 1855–70.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* 85, 165–77.

Shen, X. & Wong, W. (1994). Convergence rate of sieve estimates. *Ann. Statist* 22, 580–615.

Song, X. & Ma, S. (2008). Multiple augmentation for interval-censored data with measurement error. *Statist. Med.* 27, to appear.

Sparling, Y., Younes, N. & Lachin, J. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* 7, 599–614.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Application to Statistics*. New York: Springer-Verlag.

Wellner, J. & Zhang, Y. (2007). Likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* 28, 2106–42.

Xue, H., Lam, K. & Li, G. (2004). Sieve maximum likelihood estimatior for semiparametric regression models with current status data. *J. Am. Statist. Assoc.* 99, 346–56.

Zhang, W., Zhang, Y., Chaloner, K. & Stapleton, J. (2008). Imputation methods for doubly censored hiv data. *J. Statist. Comput. Simul.* to appear.

Zhang, Y. & Jamshidian, M. (2004). On algorithms for npmle of the failure function with censored data. *J. Comput. Graph. Statist.* 13, 123–40.

7. Appendices

7.1 *Appendix A-Proofs*

This section contains the sketch of the proofs for Theorems 1 and 2. Some empirical process theorems developed in van der Vaart (1998) and van der Vaart & Wellner (1996) will be heavily involved. Throughout the following proofs, we denote $Pf = \int f(x)dP(x)$ and $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$, the empirical process indexed by function $f(X)$ and we let C represent a generic constant that may vary from place to place.

The proof of Theorem 1:

Before deriving the convergence rate, we need to show that the sieve semiparametric maximum likelihood estimator $\widehat{\tau}_n$ is consistent in the metric d. This can be accomplished by verifying the conditions of Theorem 5.7 in van der Vaart (1998). Let $\mathbb{M}(\tau) = Pl(\tau; X) = Pl(\theta, \phi; X)$ and $\mathbb{M}_n(\tau) = \mathbb{P}_n l(\tau; X) = \mathbb{P}_n l(\theta, \phi; X)$. Hence for any $\tau = \mathcal{T}_n = \Theta \times \mathcal{M}_n$, $\mathbb{M}_n(\tau) - \mathbb{M}(\tau) = (\mathbb{P}_n - P)l(\tau; X)$.

Let $\mathcal{L}_1 = \{l(\tau; X) : \quad \tau \in \mathcal{T}_n\}$. By the calculation of Shen & Wong (1994), page 597, $\forall \epsilon > 0$, there exists a set of brackets $\left\{[\phi_i^L, \phi_i^U] : \quad i = 1, 2, \cdots, \left[(1/\epsilon)^{Cq_n}\right]\right\}$ such that for any $\phi \in \mathcal{M}_n$, one has $\phi_i^L(u) \leq \phi(u) \leq \phi_i^U(u)$ for some $1 \leq i \leq \left[(1/\epsilon)^{Cq_n}\right]$ and all $u \in [a, b]$, and $P|\phi_i^U(X) - \phi_i^L(X)| \leq \epsilon$. Since $\Theta \subset R^d$ is compact, Θ can be covered by $\left[C(1/\epsilon)^d\right]$ balls with radius ϵ; that is, for any $\theta \in \Theta$, there exists an $1 \leq s \leq \left[C(1/\epsilon)^d\right]$ such that $|\theta - \theta_s| \leq \epsilon$ and hence $|\theta' z - \theta_s' z| \leq C\epsilon$ for any z because of (C1). This implies

that $\theta' z \in [\theta_s' z - C\epsilon, \theta_s' z + C\epsilon]$ for all z. Hence we can easily construct a set of brackets $\left\{[l_{s,i}^L(X), l_{s,i}^U(X)] : \quad s = 1, 2, \cdots, \left[C(1/\epsilon)^d\right]; \quad i = 1, 2, \cdots, \left[(1/\epsilon)^{Cq_n}\right]\right\}$ that for any $l(\tau; X) \in \mathcal{L}_1$, there exist an $s \leq \left[C(1/\epsilon)^d\right]$ and an $i \leq \left[(1/\epsilon)^{Cq_n}\right]$ such that $l(\tau; X) \in [l_{s,i}^L(X), l_{s,i}^U(X)]$ for any sample point X, where

$$\begin{aligned} l_{s,i}^L(X) &= \delta_1 \log\left[1 - \exp\left\{-e^{\theta_s' z + \phi_i^L(u) - C\epsilon}\right\}\right] \\ &\quad + \delta_2 \log\left[\exp\left\{-e^{\theta_s' z + \phi_i^U(u) + C\epsilon}\right\} - \exp\left\{-e^{\theta_s' z + \phi_i^L(v) - C\epsilon}\right\}\right] - \delta_3 e^{\theta_s' z + \phi_i^U(v) + C\epsilon} \end{aligned}$$

and

$$\begin{aligned} l_{s,i}^U(X) &= \delta_1 \log\left[1 - \exp\left\{-e^{\theta_s' z + \phi_i^U(u) + C\epsilon}\right\}\right] \\ &\quad + \delta_2 \log\left[\exp\left\{-e^{\theta_s' z + \phi_i^L(u) - C\epsilon}\right\} - \exp\left\{-e^{\theta_s' z + \phi_i^U(v) + C\epsilon}\right\}\right] - \delta_3 e^{\theta_s' z + \phi_i^L(v) - C\epsilon}. \end{aligned}$$

Using Taylor expansion along with Conditions (C1)-(C3), we can easily demonstrate that $P|l_{s,i}^U(X) - l_{s,i}^L(X)| \leq C\epsilon$ for all $1 \leq s \leq \left[C(1/\epsilon)^d\right]$ and $1 \leq i \leq \left[(1/\epsilon)^{Cq_n}\right]$ which leads to the conclusion that the ϵ-bracketing number for $\mathcal{L}_1$ with $L_1(P)$-norm is bounded by $C(1/\epsilon)^{Cq_n + d}$. Hence $\mathcal{L}_1$ is Glivenko-Cantelli by Theorem 2.4.1 of van der Vaart & Wellner (1996). Therefore, $\sup_{\tau \in \mathcal{T}_n} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau)| \to_p 0$. Let $g(z, t) = \exp\{\theta' z + \phi(t)\}$ and $g_0(z, t) =$

$\exp\{\theta_0' z + \phi_0(t)\}$. Some algebra yields that

$$\begin{aligned}
\mathbb{M}(\tau_0) - \mathbb{M}(\tau) = E\Bigg(& [1-\exp\{-g_0(Z,U)\}]\log\frac{1-\exp\{-g_0(Z,U)\}}{1-\exp\{-g(Z,U)\}} \\
& +[\exp\{-g_0(Z,U)\}-\exp\{-g_0(Z,V)\}]\log\frac{\exp\{-g_0(Z,U)\}-\exp\{-g_0(Z,V)\}}{\exp\{-g(Z,U)\}-\exp\{-g(Z,V)\}} \\
& +\exp\{-g_0(Z,V)\}\log\frac{\exp\{-g_0(Z,V)\}}{\exp\{-g(Z,V)\}}\Bigg) \\
= E\Bigg(& [1-\exp\{-g(Z,U)\}]m\left[\frac{1-\exp\{-g_0(Z,U)\}}{1-\exp\{-g(Z,U)\}}\right] \\
& +[\exp\{-g(Z,U)\}-\exp\{-g(Z,V)\}]m\left[\frac{\exp\{-g_0(Z,U)\}-\exp\{-g_0(Z,V)\}}{\exp\{-g(Z,U)\}-\exp\{-g(Z,V)\}}\right] \\
& +\exp\{-g(Z,V)\}m\left[\frac{\exp\{-g_0(Z,V)\}}{\exp\{-g(Z,V)\}}\right]\Bigg),
\end{aligned}$$

where $m(x) = x\log x - x + 1 \geq (x-1)^2/4$ for $0 \leq x \leq 5$. Further analysis by using Taylor expansion and Conditions (C1)-(C3) leads to

$$\begin{aligned}
\mathbb{M}(\tau_0) - \mathbb{M}(\tau) &\geq CE\Bigg(\frac{1}{1-\exp\{-g(Z,U)\}}\left[\exp\{-g_0(Z,U)\}-\exp\{-g(Z,U)\}\right]^2 \\
&\quad +\frac{1}{\exp\{-g(Z,V)\}}\left[\exp\{-g_0(Z,V)\}-\exp\{-g(Z,V)\}\right]^2\Bigg) \\
&\geq CE\left[\{(\theta_0-\theta)'Z+(\phi_0-\phi)(U)\}^2+\{(\theta_0-\theta)'Z+(\phi_0-\phi)(V)\}^2\right].
\end{aligned}$$

With Conditions (C1)-(C6), using the same arguments as those in Wellner & Zhang (2007), page 2126-2127 leads to

$$\mathbb{M}(\tau_0) - \mathbb{M}(\tau) \geq C\left(\|\theta-\theta_0\|^2 + \|\phi-\phi_0\|_\Phi^2\right) = Cd^2(\tau_0,\tau).$$

Then it implies that $\sup_{\tau: d(\tau,\tau_0)\geq\epsilon}\mathbb{M}(\tau) \leq \mathbb{M}(\tau_0) - C\epsilon^2 < \mathbb{M}(\tau_0)$.

For $\phi_0 \in \Phi$, Lu (2007) has shown that there exists a $\phi_{0,n} \in \mathcal{M}_n$ of order $m \geq p+2$ such that

$$\|\phi_{0,n} - \phi_0\|_\infty \leq Cq_n^{-p} = O\left(n^{-p\nu}\right).$$

This also implies that $\|\phi_{0,n} - \phi_0\|_\Phi \leq C q_n^{-p} = O\left(n^{-p\nu}\right)$. Now let $\tau_{0,n} = (\theta_0, \phi_{0,n})$, we have

$$\begin{aligned}
\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) &= \mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_{0,n}) + \mathbb{M}_n(\tau_{0,n}) - \mathbb{M}_n(\tau_0) \\
&\geq \mathbb{P}_n l(\tau_{0,n}; X) - \mathbb{P}_n l(\tau_0; X) \\
&= (\mathbb{P}_n - P)\left\{l(\tau_{0,n}; X) - l(\tau_0; X)\right\} + \mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0).
\end{aligned}$$

Using the brackets for $\mathcal{M}_n$ given above, we can similarly construct a set of brackets for the class $\mathcal{L}_2 = \{l(\theta_0, \phi; x) - l(\theta_0, \phi_0; x) : \ \phi \in \mathcal{M}_n \ \text{ and } \ \|\phi - \phi_0\|_\Phi \leq C n^{-p\nu}\}$ with the ϵ-bracketing number associated $L_2(P)$-norm bounded by $(1/\epsilon)^{Cq_n}$. This yields a finite-valued bracketing integral defined in van der Vaart (1998), page 270. Hence the class $\mathcal{L}_2$ is P-Donsker by Theorem 19.5 of van der Vaart (1998). By the Dominated Convergence Theorem, it is obvious that in this class $P\left\{l(\theta_0, \phi; X) - l(\theta_0, \phi_0; X)\right\}^2 \to 0$ as $n \to \infty$. Hence

$$(\mathbb{P}_n - P)\left\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\right\} = o_p(n^{-1/2})$$

by the relationship between P-Donsker and asymptotic equicontinuity given by Corollary 2.3.12 of van der Vaart & Wellner (1996). By the Dominated Convergence Theorem again, it is easy to see that $\mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0) > -o(1)$ as $n \to \infty$. Therefore,

$$\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq o_p(n^{-1/2}) - o(1) = -o_p(1).$$

This completes the proof of $d(\widehat{\tau}_n, \tau_0) \to 0$ in probability.

Next, we verify the conditions of Theorem 3.2.5 of van der Vaart & Wellner (1996) in order to derive the convergence rate. First, we have already shown in the proof of consistency that $\mathbb{M}(\tau_0) - \mathbb{M}(\tau) \geq C d^2(\tau_0, \tau)$.

Second, we further explore $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0)$. In the proof of consistency, we know that $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq I_{1,n} + I_{2,n}$, where $I_{1,n} = (\mathbb{P}_n - P)\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\}$ and $I_{2,n} = P\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\}$. By Taylor expansion, we have

$$I_{1,n} = (\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)(\phi_{0,n} - \phi_0)\right\} = n^{-p\nu+\epsilon}(\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)\frac{\phi_{0,n} - \phi_0}{n^{-p\nu+\epsilon}}\right\}$$

for any $0 < \epsilon < 1/2 - p\nu$. Because $\|\phi_{0,n} - \phi_0\|_\infty = O(n^{-p\nu})$ and $\dot{l}_2(\theta_0, \tilde{\phi}; X)$ is uniformly bounded due to Conditions (C1)-(C4), we can easily obtain that $P\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)\frac{\phi_{0,n}-\phi_0}{n^{-p\nu+\epsilon}}\right\}^2 \to 0$. Due to $\mathcal{L}_2$ being P-Donsker, using Corollary 2.3.12 of van der Vaart & Wellner (1996) again, we can conclude that $(\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)\frac{\phi_{0,n}-\phi_0}{n^{-p\nu+\epsilon}}\right\} = o_p(n^{-1/2})$. Hence

$$I_{1,n} = o_p(n^{-p\nu+\epsilon}n^{-1/2}) = o_p(n^{-2p\nu}),$$

due to the selection of ν. Using the fact that the function $m(x) = x\log x - x + 1 \leq (x-1)^2$ in the neighborhood of $x = 1$, it can be easily argued that $\mathbb{M}(\tau_0) - \mathbb{M}(\tau_{0,n}) \leq C\|\phi_{0,n} - \phi_0\|_\Phi^2 = O(n^{-2p\nu})$, which implies that $I_{2,n} = \mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0) \geq -O(n^{-2p\nu})$. Thus we conclude that $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq -O_p(n^{-2p\nu}) = -O_p\left(n^{-2\min(p\nu,(1-\nu)/2)}\right)$.

Let $\mathcal{L}_3(\eta) = \{l(\tau; x) - l(\tau_0; x) : \ \phi \in \mathcal{M}_n \ \text{ and } \ d(\tau, \tau_0) \leq \eta\}$. Using the same argument as that in the proof of consistency, we obtain that the logarithm of the ϵ-bracketing number of $\mathcal{L}_3(\eta)$, $\log N_{[\,]}\{\epsilon, \mathcal{L}_3(\eta), L_2(P)\}$ is bounded by $Cq_n\log(\eta/\epsilon)$. This leads to

$$J_{[\,]}\{\eta, \mathcal{L}_3(\eta), L_2(P)\} = \int_0^\eta \sqrt{1 + \log N_{[\,]}\{\epsilon, \mathcal{L}_3(\eta), L_2(P)\}}d\epsilon \leq Cq_n^{1/2}\eta.$$

Because Conditions (C1) and (C3) guarantee the uniform boundedness of $l(\tau; x)$, using Theorem 3.4.1 of van der Vaart & Wellner (1996), the key function $\phi_n(\eta)$ in Theorem 3.2.5

of van der Vaart & Wellner (1996) is given by $\phi_n(\eta) = q_n^{1/2}\eta + q_n/n^{1/2}$. Note that

$$n^{2p\nu}\phi_n(1/n^{p\nu}) = n^{p\nu}n^{\nu/2} + n^{2p\nu}n^{\nu} + n^{2p\nu}n^{\nu}/n^{1/2} = n^{1/2}\left\{n^{p\nu-(1-\nu)/2} + n^{2p\nu-(1-\nu)}\right\}.$$

Therefore, if $p\nu \leq (1-\nu)/2$, $n^{2p\nu}\phi_n(1/n^{p\nu}) \leq n^{1/2}$. This implies that if we choose $r_n = \min(p\nu, (1-\nu)/2)$, it follows that $r_n^2\phi_n(1/r_n) \leq n^{1/2}$ and $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq -O_p(r_n^{-2})$. Hence $r_n d(\widehat{\tau}_n, \tau_0) = O_p(1)$.

The proof of Theorem 2:

To derive the asymptotic normality for $\hat{\theta}_n$, we just need to verify the conditions of the general theorem given in Appendix B. For Condition (B1), we only need to verify that $\mathbb{P}_n\dot{l}_2(\widehat{\theta}_n, \widehat{\phi}_n; X)(\xi_0) = o_p(n^{-1/2})$ since $\mathbb{P}_n\dot{l}_1(\widehat{\theta}_n, \widehat{\phi}_n; X) \equiv 0$. Because ξ_0 has a bounded derivative, it is also a function with bounded variation. Then it can be easily shown using the argument in Billingsley (1986), page 435-436, that there exist a $\xi_{0,n} \in S_n(D_n, K_n, m)$ such that $\|\xi_{0,n} - \xi_0\|_\Phi = O(q_n^{-1}) = O(n^{-\nu})$ and $\mathbb{P}_n\dot{l}_2(\widehat{\tau}_n; X)(\xi_{0,n}) = 0$. Therefore we can write $\mathbb{P}_n\dot{l}_2(\widehat{\tau}_n; X)(\xi_0) = I_{3,n} + I_{4,n}$, where

$$I_{3,n} = (\mathbb{P}_n - P)\dot{l}_2(\widehat{\tau}_n; X)(\xi_0 - \xi_{0,n})$$

and

$$I_{4,n} = P\left\{\dot{l}_2(\widehat{\tau}_n; X)(\xi_0 - \xi_{0,n}) - \dot{l}_2(\tau_0; X)(\xi_0 - \xi_{0,n})\right\}.$$

Let $\mathcal{L}_4 = \{\dot{l}_2(\tau; x)(\xi_0 - \xi) : \ \tau \in \mathcal{T}_n, \ \ \xi \in S_n(D_n, K_n, m) \ \text{ and } \ \|\xi_0 - \xi\|_\Phi \leq n^{-\nu}\}$. It can be similarly argued that the ϵ-bracketing number associated with $L_2(P)$-norm is bounded by $C\,(1/\epsilon)^d\,(1/\epsilon)^{Cq_n}\,(1/\epsilon)^{Cq_n}$ which leads $\mathcal{L}_4$ being a P-Donsker due to Theorem 19.5 of

van der Vaart (1998). Furthermore, for any $r(\tau, \xi; x) \in \mathcal{L}_4$, $Pr^2 \to 0$ as $n \to \infty$. Hence $I_{3,n} = o_p(n^{-1/2})$ by Corollary 2.3.12 of van der Vaart & Wellner (1996). By Cauchy-Schwatz inequality and regularity conditions (C1)-(C4), it can be easily shown that

$$\begin{aligned} I_{4,n} &\leq Cd(\widehat{\tau}_n, \tau_0)\|\xi_0 - \xi_{0,n}\|_\Phi = O_p\left(n^{-\min(p\nu,(1-\nu)/2)}n^{-\nu}\right) = O_p\left(n^{-\min(\nu(p+1),(1+\nu)/2)}\right) \\ &= o_p(n^{-1/2}). \end{aligned}$$

So (B1) holds. (B2) holds by similarly verifying that the class $\mathcal{L}_5(\eta) = \{l^*(\tau; x) - l^*(\tau_0; x) : \tau \in \mathcal{T}_n \text{ and } d(\tau, \tau_0) \leq \eta\}$ is P-Donsker and for any $r(\tau; x) \in \mathcal{L}_5(\eta)$, $Pr^2 \to 0$ as $\eta \to 0$. (B3) can be easily established using Taylor expansion and the convergence rate derived in Theorem 1. Hence the proof is complete.

7.2 *Appendix B-A General Theorem*

This section presents a general theorem for asymptotic normality of the maximum likelihood estimator of the finite-dimensional parameter in a setting of semiparametric maximum likelihood estimation when the infinite-dimensional parameter is treated as a nuisance parameter. This theorem is the simplified version of the general theorem given in Huang (1996). The following conditions will be assumed.

(B1): $\mathbb{P}_n \dot{l}_1(\widehat{\theta}_n, \widehat{\phi}_n; X) = o_p(n^{-1/2})$ and $\mathbb{P}_n \dot{l}_2(\widehat{\theta}_n, \widehat{\phi}_n; X)(\xi_0) = o_p(n^{-1/2})$

(B2): $(\mathbb{P}_n - P)\left\{l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X)\right\} = o_p(n^{-1/2})$

(B3): $P\left\{l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X)\right\} = -I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2})$

THEOREM 3. *Suppose (B1)-(B3) are satisfied, and suppose that $I(\theta_0)$ is nonsingular. Then*

$$n^{1/2}(\widehat{\theta}_n - \theta_0) = n^{1/2} I^{-1}(\theta_0) \sum_{i=1}^{n} l^*(\theta_0, \phi_0; X_i) + o_p(1) \to_d N\left\{0, I^{-1}(\theta_0)\right\}.$$

Proof: Combining (B2) and (B3), we have

$$\mathbb{P}_n \left\{ l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X) \right\} = -I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2}).$$

By (B1), it follows that

$$\mathbb{P}_n l^*(\theta_0, \phi_0; X) = I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2})$$

Because $I(\theta_0)$ is nonsingular, and $\mathbb{P}_n l^*(\theta_0, \phi_0; X) = O_p(n^{-1/2})$ due to the ordinary large sample theory, one has $\|\widehat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$. Thus $o_p(\|\widehat{\theta}_n - \theta_0\|) = o_p(n^{-1/2})$ and therefore

$$\mathbb{P}_n l^*(\theta_0, \phi_0; X) = I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(n^{-1/2}).$$

The result follows.