Scandinavian Journal of Statistics, Vol. 37: 338–354, 2010
doi: 10.1111/j.1467-9469.2009.00680.x

A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data

YING ZHANG and LEI HUA

Department of Biostatistics, University of Iowa

JIAN HUANG

Department of Statistics and Actuarial Science and Department of Biostatistics, University of Iowa

ABSTRACT. We propose a spline-based semiparametric maximum likelihood approach to analysing the Cox model with interval-censored data. With this approach, the baseline cumulative hazard function is approximated by a monotone B-spline function. We extend the generalized Rosen algorithm to compute the maximum likelihood estimate. We show that the estimator of the regression parameter is asymptotically normal and semiparametrically efficient, although the estimator of the baseline cumulative hazard function converges at a rate slower than root-*n*. We also develop an easy-to-implement method for consistently estimating the standard error of the estimated regression parameter, which facilitates the proposed inference procedure for the Cox model with interval-censored data. The proposed method is evaluated by simulation studies regarding its finite sample performance and is illustrated using data from a breast cosmesis study.

Key words: consistent variance estimation, convergence rate, efficient estimation, empirical processes, monotonicity constraints, sieve semiparametric model

1. Introduction

Interval censoring refers to a censoring mechanism where an event time cannot be directly observed but is only 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), Hsu *et al.* (2007) and Song & Ma (2008). Recently, the analysis of interval-censored data has appeared in many other biomedical and epidemiological studies. For example, Kim & Xue (2002) analysed interval-censored data for an ongoing clinical trial for systemic lupus erythematosus, Bogaerts *et al.* (2002) analysed multivariate interval-censored dental data, Bellamy *et al.* (2004) developed a parametric frailty model for clustered and interval-censored data with application to the East Boston Asthma study and Sparling *et al.* (2006) studied 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. Although 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, notably the Cox (1972) regression model. Imputation-based approach was proposed by Satten *et al.* (1998), Song & Ma (2008) and Zhang *et al.* (2009) in which interval-censored event times are imputed and then some well-known semiparametric regression methods such as the Cox model for right-censored data can be handily utilized. However, the imputation method in general produces biased estimates for the regression parameter. As an alternative to the Cox model, 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), among others.

For the likelihood analysis of the Cox model with interval-censored data, the baseline hazard function cannot be eliminated using the partial likelihood approach as in the case of right-censored data. This makes the problem of analysing interval-censored data more difficult. Finkelstein (1986) considered the Cox model for interval-censored data with a discrete hazard assumption. This approach has also been used by Goggins & Finkelstein (2000), Seaman & Bird (2001) and Kim & Xue (2002). The fully semiparametric maximum likelihood analysis of current status data, a special case of interval-censored data, is developed by Huang (1996) in which the baseline cumulative hazard function is estimated by a monotone step function whose number of jumps increases with sample size. He showed that despite the non-parametric estimator of the baseline cumulative hazard function converges at a rate slower than the standard rate $n^{1/2}$, the maximum likelihood estimator of regression parameter is still asymptotically normal and achieves the semiparametric efficiency bound defined in Bickel *et al.* (1993). Although Huang & Wellner (1997) discussed some possible extensions of Huang (1996) to interval-censored data, the statistical theory for the semiparametric maximum likelihood analysis of interval-censored data has not been fully developed. Pan (1999) extended the iterative convex minorant (ICM) algorithm developed by Jongbloed (1998) for computing the non-parametric maximum likelihood estimates to the Cox model with interval-censored data. However, Pan's algorithm may not be efficient when sample size is large, as only the diagonal elements of the Hessian matrix are used during the numerical iterations in the extended ICM algorithm. In addition, Pan (1999) adopted the bootstrap inference for the regression parameter, which is computationally intensive. Although the information matrix and the asymptotic normality have been described by Huang & Wellner (1997), there is no discussion in the literature on how to consistently estimate the asymptotic variance of the estimated regression parameter for the Cox model with interval-censored data.

In this article, we propose a spline-based sieve semiparametric likelihood estimation procedure for the Cox model with interval-censored data, in which the log baseline cumulative hazard function is estimated by a monotone B-spline (Schumaker, 1981). The generalized Rosen's (GR) algorithm, proposed by Zhang & Jamshidian (2004) for computing non-parametric maximum likelihood estimates with linear inequality constraints, is extended 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 converges at the optimal non-parametric rate. We also develop an easy-to-implement method to consistently estimate the standard error of the estimated regression parameter based on the ordinary least-squares approach, to make statistical inference using the asymptotic results. The proposed method facilitates an easy-to-implement semiparametric likelihood inference procedure for analysing interval-censored data with the Cox model.

The rest of the article is organized as follows. Section 2 states the model and likelihood for interval-censored data, section 3 describes the spline-based sieve semiparametric maximum likelihood approach and the extended GR algorithm for computing the estimate, section 4 presents the asymptotic results of the estimator, section 5 includes simulation studies that evaluate the finite sample performance of the proposed estimator and an application to data from a breast cosmesis study, while the concluding remarks are given in section 6. Finally, the technical details are given 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

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

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

Let (U, V) be a 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 the covariate $Z=z$, with $P(U \le V \mid z)=1$ for any $z \in R^d$, and let $H(z)$ be the distribution of Z. Let $\delta_1 = 1_{[T \le U]}$, $\delta_2 = 1_{[U < T \le V]}$ and $\delta_3 = 1 - \delta_1 - \delta_2$ and denote the observation from a single subject by $X = (\delta_1, \delta_2, \delta_3, U, V, Z)$. Assume that conditional on Z, T is independent of (U, V). Then the density of X is given by

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

where $F(\cdot \mid z)$ 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 non-informative to T. Then under the Cox model, the log-likelihood of an i.i.d. sample $X_i = (\delta_{1i}, \delta_{2i}, \delta_{3i}, U_i, V_i, Z_i)$ for $i = 1, 2, \ldots, n$ is given by

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

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

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

3. Spline-based sieve semiparametric maximum likelihood estimation

Suppose $0 = t_0 < t_1 < t_2 < \cdots < t_m < \infty$ are the distinct time points in the collection of $\{U_i, V_i : i = 1, 2, \ldots, n\}$. The value of the log-likelihood (2) is completely determined by the values of ϕ at these points and θ. Conventionally, the fully semiparametric maximum likelihood estimator is sought by maximizing (2) with respect to θ and $\phi(t_i)$, for $i = 1, 2, \ldots, m$. A trivial upper bound of m is $2n$ if there are no ties among $\{U_i, V_i\}, i = 1, 2, \ldots, n$. This high-dimensional optimization problem is challenging particularly when θ is a multidimensional vector and the sample size is large.

To ease the computational difficulty in fully non-parametric estimation problems, Geman & Hwang (1982) proposed a sieve maximum likelihood estimation procedure for which the unknown function in a 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 as the number of basis functions required to reasonably approximate the unknown function grows a lot slower as sample size increases.

Spline technique has been well recognized in the statistical literature as a useful tool in non-parametric estimation (Stone, 1985, 1986). Therefore, it is natural to consider spline-based sieve semiparametric maximum likelihood estimation in the context of the Cox model with interval-censored data. Some further theoretical results of the spline-based sieve estimation have 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 spline in analysing interval-censored data can be found in Kooperberg & 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 the observation times $\{(U_i, V_i): i=1,2,\dots,n\}$. 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 $I_{Kt}=[d_t, d_{t+1}), t=0,\dots,K$, where $K\equiv K_n\approx n^{\upsilon}$ is a positive integer such that $\max_{1\le k\le K+1}|d_k-d_{k-1}|=O(n^{-\upsilon})$. Denote the set of partition points by $D_n=\{d_1,\dots,d_{K_n}\}$. 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 $m\le K$; and (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, definition 4.1). According to Schumaker (1981, corollary 4.10), there exists a *local* basis $\mathcal{B}_n\equiv\{\mathbf{b}_t, 1\le t\le q_n\}$, so-called B-spline, for $\mathcal{S}_n(D_n,K_n,m)$, where $q_n\equiv K_n+m$. These basis functions are non-negative 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 non-decreasing function, it is desirable to restrict its estimate to be non-decreasing 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 $B_n=\{\beta: \beta_1\le\beta_2\le\cdots\le\beta_{q_n}\}$. Each element of $\mathcal{M}_n(D_n,K_n,m)$ is a non-decreasing function because of the monotonicity constraints on $\beta_1,\dots,\beta_{q_n}$. This fact is a consequence of the *variation diminishing properties* of B-spline; see, for instance, Schumaker (1981, example 4.75 and theorem 4.76). Denote by $\Theta\in R^d$ 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 $\|x\|=x^{\mathrm{T}}Wx$, where W is a positive definite matrix and possibly varying from iteration to iteration. Zhang & Jamshidian (2004) applied the algorithm to some large-scale non-parametric maximum likelihood estimation problems by choosing $W=D_H$, the matrix containing only the diagonal elements of the negative Hessian matrix H, to avoid the storage problem in updating H. However, this will increase the number of iterations and thereby the computing time. Lu *et al.* (2007) adopted Zhang–Jamshidian's algorithm with $W=-H$ to compute the monotone mean function estimates in the context of spline-based sieve non-parametric maximum likelihood analysis of panel count data and demonstrated the numerical advantage of this algorithm over the ICM algorithm used by Wellner & Zhang (2000). For the spline-based sieve semiparametric maximum likelihood estimation, we will jointly update the estimate of (θ,β) by generalizing the algorithm developed by Lu *et al.* (2007) for which the Hessian matrix H contains both the second derivatives $\partial^2 l_n/\partial\theta^2$ and $\partial^2 l_n/\partial\beta^2$, and the mixed derivative $\partial^2 l_n/\partial\theta\partial\beta$. However, the monotonicity constraints only need to be imposed on the last q_n coefficients ($\beta' s$). In the following, we describe the algorithm for computing the proposed spline-based sieve semiparametric estimate. The detailed explanation of the algorithm can be found in Zhang & Jamshidian (2004).

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 $\mathcal{A} = \{i_1, i_2, \ldots, i_r\}$ denote the index set of active constraints, that is, $\tau_{i_j} = \tau_{i_j+1}$, for $i_1 > d$ and $j = 1, 2, \ldots, r$, during the numerical computation. We define an r by $d + q_n$ working matrix corresponding to this set

$$A = \begin{bmatrix} 0 & \cdots & -1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & -1 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 & \cdots & 0 \end{bmatrix},$$

where the first d columns are always set to be zero, because parameter θ is free of constraints. The working matrix A is defined in such a way that $A\tau = 0$ at the active constraints which is the essential part of the GR algorithm described by Zhang & Jamshidian (2004).

The GR algorithm is implemented in the following steps:

S0 (computing the feasible search direction):

$$\underline{d} = \{I - W^{-1}A^{\mathrm{T}}(AW^{-1}A^{\mathrm{T}})^{-1}A\}W^{-1}\dot{\ell}(\tau),$$

where $\underline{d}$ is the projection of the weighted gradient onto a subspace made by the active constraints from the preceding iteration and computes the feasible direction for the next iteration.

S1 (forcing the updated τ fulfil the constraints): If the resulted direction $\underline{d}$ is not non-decreasing in its components, compute

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

Doing so guarantees that $\tau_{i+1} + \gamma d_{i+1} \geq \tau_i + \gamma d_i$, for $i = d+1, d+2, \ldots, d+q_n-1$ and hence the updated value taken as $\tau + \gamma^* \underline{d}$ for $\gamma^* \leq \gamma$ satisfies the monotonicity constraints on βs.

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 replacing 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 $\lambda = (AW^{-1}A^{\mathrm{T}})^{-1}AW^{-1}\dot{\ell}(\tau)$.

(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

It is worth noting that the B-spline approach simplifies the fully semiparametric model to a weakly parametric problem with the advantage of controlling the expansion of the dimensionality of the model when the sample size increases. However, as the number of basis functions increases with sample size, the ordinary parametric likelihood theory does not apply to the present model. In this section, we describe some asymptotic properties of the proposed 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 an L_2-metric:

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

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

- (C1) (a) $E(ZZ')$ is non-singular; and (b) Z is bounded, that is, there exists $z_0 > 0$ such that $P(\|Z\| \le z_0) = 1$.
- (C2) Θ is a compact subset of R^d.
- (C3) (a) There exists a positive number η such that $P(V - U \ge \eta) = 1$; and (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 $p \ge 1$ and the first derivative of ϕ_0 is strictly positive and continuous on $[a, b]$.
- (C5) The conditional density $g(u, v \mid z)$ 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 $\kappa \in (0, 1)$, $a^\mathrm{T}\mathrm{var}(Z \mid U)a \ge \kappa a^\mathrm{T} E(ZZ^\mathrm{T} \mid U)a$ and $a^\mathrm{T}\mathrm{var}(Z \mid V)a \ge \kappa a^\mathrm{T} E(ZZ^\mathrm{T} \mid V)a$ 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, relaxing them will make the proofs considerably more difficult.

Theorem 1

Let $K_n = O(n^v)$, where v satisfies the restriction $1/2(1+p) < v < 1/2p$. 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*

$$d(\hat{\tau}_n, \tau_0) = O_p\{n^{-\min(pv, (1-v)/2)}\}.$$

This theorem implies that if $v = 1/(1+2p)$, $d(\hat{\tau}_n, \tau_0) = O_p\{n^{-p/(1+2p)}\}$ which is the optimal convergence rate in the non-parametric regression setting. So if the baseline hazard function is second-order differentiable, the proposed estimator can achieve a better convergence rate than the fully semiparametric estimator considered in Huang & Wellner (1997).

Theorem 2

Suppose the conditions given in theorem 1 *hold, then*

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

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 Bickel *et al.* (1993). The theorem implies that, although the estimator of the baseline cumulative hazard function converges at a rate slower than $n^{1/2}$, the estimator of the regression parameter, $\hat{\theta}_n$, converges to the true parameter at the usual root-n rate and achieves the semiparametric efficiency bound. However, in general, $I(\theta_0)$ is determined by an integral equation and does not have an explicit expression. Direct estimation of $I(\theta_0)$ is not feasible. Next, we describe an approach for estimating $I(\theta_0)$.

Let

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

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 $\dot{l}_1(\tau;x)$, a score operator for ϕ is defined as in (3). So the score operator for ϕ corresponding to $\dot{l}_1(\tau;x)$ 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 $h_k\in\mathcal{H}, 1\le k\le d$.

According to Bickel *et al.* (1993, theorem 1, p. 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 favourable 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 favourable direction $\xi_0(t)$ is the solution of a Fredholm integral equation of the second kind,

$$\xi_0(t)-\int K(t,x)\xi_0(x)\,\mathrm{d}x=d(t)$$

with two complicated 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 not straightforward. 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 a random sample $X_1,\ldots,X_n$ and the consistent estimator $\hat{\tau}_n$, we can estimate $I(\theta)$ by the minimum value of

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

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

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-spline basis functions $\mathcal{B}_n$. This estimation is implemented in subsequent simulation studies and application.

5. Numerical results

5.1. Simulation studies

A random sample of interval-censored data, $X_i = (\delta_{i,1}\delta_{i,2}, \delta_{i,3}, U_i, V_i, Z_i)$, for $i = 1, 2, \ldots, n$, is generated as follows. For each subject, the event time is generated according to the Cox model $\Lambda(t \mid Z) = t^{1/2}\exp(\theta_0^{\mathrm{T}} Z)$ for which the true parameters are $\theta_0 = (-1.0, 0.5, 1.5)^{\mathrm{T}}$ and $\log \Lambda_0(t) = 0.5 \log t$, the covariate vector $Z = (Z_1, Z_2, Z_3)^{\mathrm{T}}$ is simulated according to $Z_1 \sim$ uniform(0, 1), $Z_2 \sim$ normal(0, 1) and $Z_3 \sim$ Bernoulli(0.5); a series of examination times are generated by the partial sum of inter-arrival times that are i.i.d. according to exp(0.5), U is the last examination time before 5 at which the event has not occurred yet and V is the first observation time before 5 at which the event has occurred.

We compute the sieve semiparametric maximum likelihood estimate using the cubic B-spline and estimate the standard error of the estimated regression parameter using the least-squares method based on the cubic B-spline as well. We compare the proposed method (spline) with the fully semiparametric maximum likelihood estimation method (conventional) studied by Huang & Wellner (1997) and Pan (1999). For the B-spline, 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 $\{(U_i, V_i): i = 1, 2, \ldots, n\}$, and the knots are placed at the K_n quantiles of the N distinct observation times. The proposed algorithm described in section 3 is used to compute the spline estimate. For the fully semiparametric maximum likelihood estimation method, the extended ICM algorithm developed by Pan (1999) is adopted for computation. Table 1 presents the bias, standard deviation and average computing time in a Monte-Carlo study with 1000 repetitions for sample size $n = 50, 100$ and 200, respectively. The results show that the B-spline approach not only has the advantage in computing

Table 1. *Comparison of the B-splines and the conventional method in a Monte-Carlo simulation study with* 1000 *repetitions*

	Bias			SD			Computing time (seconds)
	$\theta_{1,0}$	$\theta_{2,0}$	$\theta_{3,0}$	$\theta_{1,0}$	$\theta_{2,0}$	$\theta_{3,0}$	
Sample size = 50							
Spline	−0.154	0.082	0.230	0.777	0.474	0.484	5.2
Conventional	−0.196	0.105	0.366	0.826	0.503	0.901	14.6
Sample size = 100							
Spline	−0.088	0.033	0.112	0.484	0.273	0.317	12.5
Conventional	−0.1001	0.045	0.138	0.490	0.276	0.332	44.9
Sample size = 200							
Spline	−0.013	0.017	0.045	0.323	0.190	0.197	35.1
Conventional	−0.021	0.023	0.058	0.325	0.192	0.198	157.9

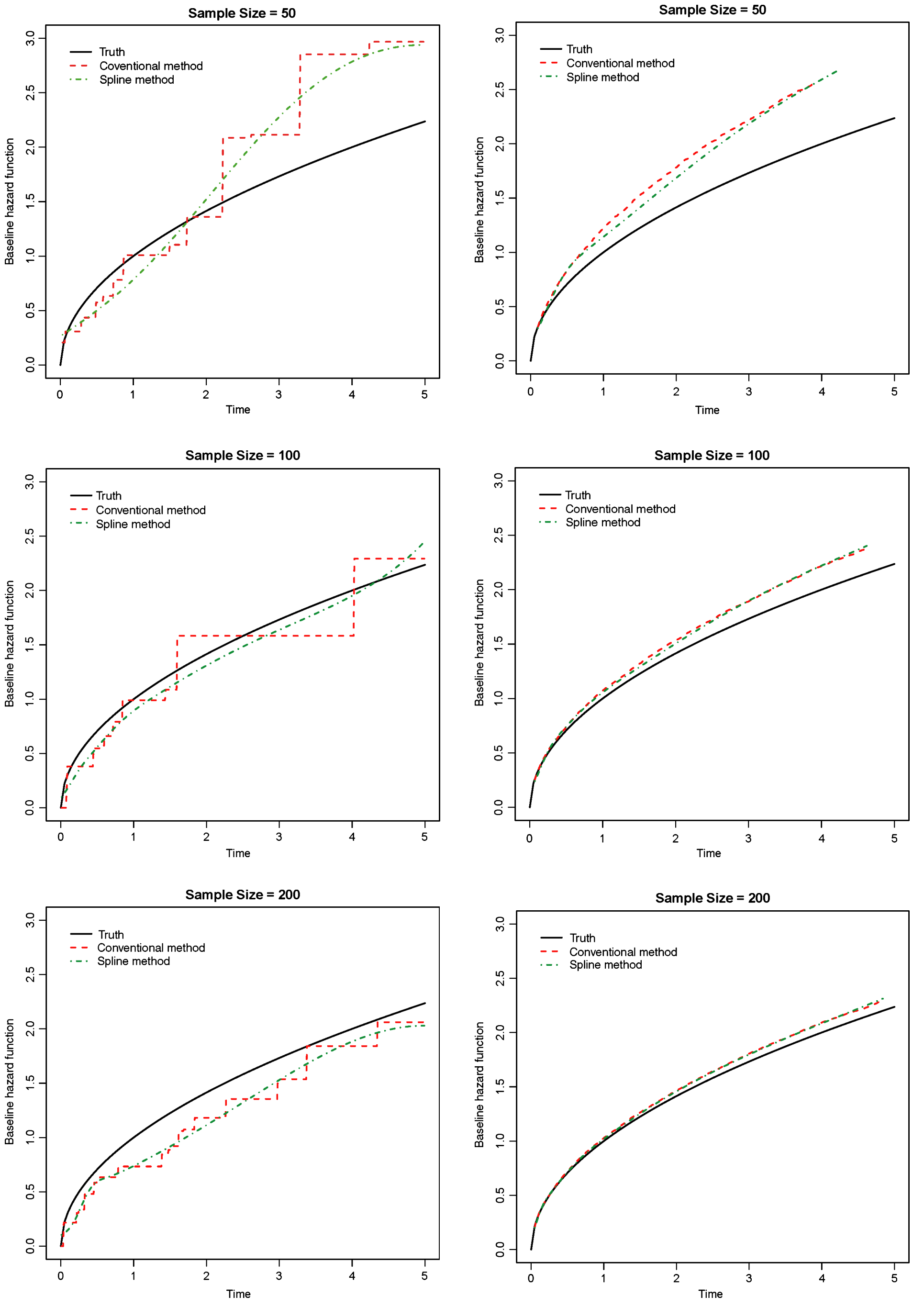

Fig. 1. Estimates of the baseline hazard function. See text for details.

time over the fully semiparametric approach, but also has smaller bias and standard deviation when sample size is relatively small. However, the Monte-Carlo standard deviations for the same parameters under both methods tend to have same values as the sample size increases. This is expected as both estimators have the same asymptotic variance.

Table 2. *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	100	200	500	50	100	200	500	50	100	200	500
Bias	−0.128	−0.075	−0.034	−0.015	0.106	0.030	0.013	0.008	0.217	0.092	0.050	0.019
M-C SD	0.761	0.466	0.316	0.191	0.435	0.272	0.191	0.113	0.497	0.283	0.206	0.122
ASE	0.851	0.509	0.331	0.198	0.500	0.301	0.194	0.115	0.566	0.328	0.210	0.124
95% CP	0.976	0.971	0.962	0.955	0.987	0.972	0.950	0.946	0.982	0.977	0.956	0.952

M-C, Monte Carlo; ASE, average of standard errors; CP, coverage probability; SD, standard deviation.

Figure 1 displays the estimates of baseline hazard function: the left panel shows the estimates based on a single Monte-Carlo sample and the right panel presents the average of estimates with 1000 repeated Monte-Carlo samples. The B-spline estimator of $\Lambda_0(t)$ smooths out the step function estimator from the conventional approach (left panel), but both estimators converge to the true function with very little bias (right panel) when the sample size is 200.

We have also conducted simulations with sample size up to 400 (results not shown here). The computing time of the conventional method grows exponentially but for the B-spline method it only grows linearly. For $n=400$, the average computing time with 10 repetitions for the conventional method is about eight times larger than the proposed B-spline method. For the conventional approach, the bootstrap method described in Pan (1999) is the only inference procedure available in the existing literature, but it is computationally demanding. Next, we will focus on evaluating the finite sample performance of the B-spline estimator.

We further conduct Monte-Carlo simulation study with 1000 repetitions for sample sizes n=50, 100, 200 and 500. For each repetition, the standard error of the estimated regression parameter is estimated using the least-squares method based on the cubic B-spline. Table 2 presents the estimation bias (bias), Monte Carlo standard deviations (M-C SD), the average of standard errors (ASE) based on the asymptotic result given in theorem 2, and the coverage probability of 95 per cent Wald confidence interval for $\hat{\theta}_n$.

The results show that the spline-based sieve semiparametric maximum likelihood estimator performs quite well. The bias is small compared with the standard error. The standard error decreases as the sample size increases. The least-squares method for estimating the standard error overestimates the true standard error slightly, but the overestimation lessens as the sample size increases. It provides a reasonable estimate of the standard error when the sample size is over 200. As the result of overestimation, the coverage probability of 95 per cent confidence interval exceeds 0.95 a little bit but approaches the nominal value when the sample size is larger than 200. In addition, we also plot in Fig. 2 the averages of the B-spline sieve estimates of the true log cumulative hazard function, $\phi(t)=0.5\log t$. 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, conducted by Beadle *et al.* (1984), is a clinical trial comparing radiotherapy alone with primary radiotherapy plus adjuvant chemotherapy in terms of subsequent cosmetic deterioration of the breast following tumourectomy. Subjects (46 assigned to radiotherapy alone and 48 to radiotherapy plus chemotherapy) were followed for up to 60 months, with prescheduled follow-up visits for every 4–6 months. In this article, we use the Cox model to analyse the difference of the hazards for time until the appearance of breast retraction between the two treatments,

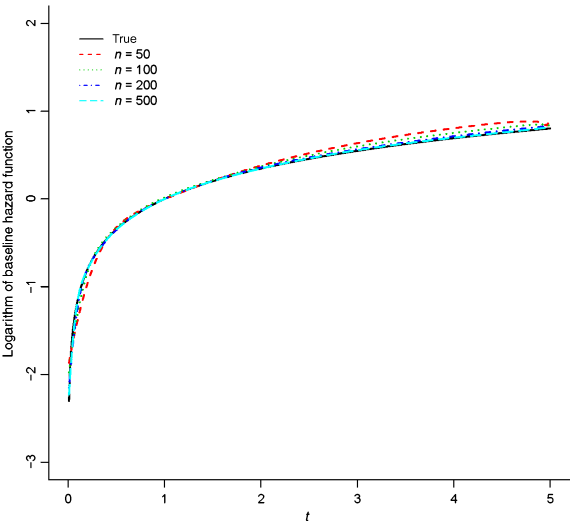

Fig. 2. The average of the B-spline sieve maximum likelihood estimates of the log baseline cumulative hazard function with 1000 repetitions.

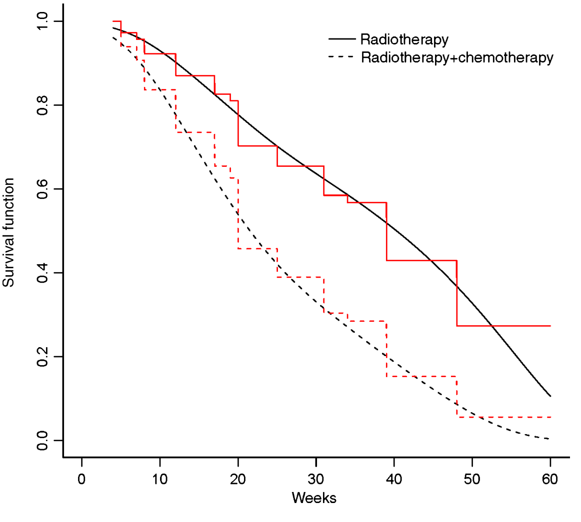

Fig. 3. The two estimates of the survival functions: the conventional method (step functions) and the B-spline method (smooth curves).

$$\Lambda(t \mid Z) = \Lambda_0(t) \exp(\theta_0 Z),$$

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 article, the cubic B-spline sieve semiparametric maximum likelihood estimate of θ_0 is $\hat{\theta}_n = 0.895$ with asymptotic standard error given by 0.293. The Wald test statistic is $Z = 3.058$ with p-value $= 0.001$. 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). For the comparison purpose, we also perform the analysis using the fully semiparametric likelihood method. It gives $\hat{\theta}_n = 0.797$ and the bootstrap s.e. $(\hat{\theta}_n) = 0.336$ with 500 bootstrap samples which results in $Z = 2.375$ with p-value$= 0.009$. Given the estimates of (θ_0, Λ_0), one can directly obtain the plug-in estimates of the survival function for the two groups. Both the B-spline and conventional estimates of the survival function shown in Fig. 3 imply that the event time is stochastically longer for the group treated with radiotherapy only.

6. Concluding remarks

In this article, we proposed a spline-based sieve semiparametric maximum likelihood method for the Cox model with interval-censored data. This method reduces the dimensionality of the estimation problem using the B-spline and therefore relieves the computation burden without interfering the asymptotic properties of the estimated regression parameter. A least-squares method for consistently estimating the standard error of the estimated regression parameter is also developed. Our proposed method facilitates a practical and easy-to-implement semiparametric likelihood inference procedure for analysing the Cox model with interval-censored data.

It should be a straightforward task to apply the method presented here to other semiparametric regression models with 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 numerical computation for infinite-dimensional nuisance parameter is a burden.

Acknowledgements

Huang's research is partially supported by grants from the US National Institutes of Health and National Science Foundation. The authors thank the editor, associate editor and two anonymous referees for their helpful comments that led to improvements in the paper.

References

Beadle, G., Harris, J., Comes, S., Henderson, C., Silver, B. & Hellman, S. (1984). The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early breast cancer: a preliminary analysis. *Int. J. Radiat. Oncol., Biol. and Phys.* **10**, 2131–2137.

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. *Stat. Med.* **23**, 3607–3621.

Betensky, R., Rabinowitz, D. & Tsiatis, A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88**, 703–711.

Bickel, 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. *Stat. Med.* **21**, 3775–3787.

Cai, T. & Betensky, R. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570–579.

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

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–854.

Finkelstein, D. & Wolf, R. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933–945.

Geman, A. & Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10**, 401–414.

Goggins, W. & Finkelstein, D. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56**, 940–943.

Gómez, G., Espinal, A. & Lagakos, S. (2003). Inference for a linear regression model with an interval-censored covariate. *Stat. Med.* **22**, 409–425.

Hsu, C., Taylor, J., Murray, S. & Commenges, D. (2007). Multiple imputation for interval-censored data with auxiliary variables. *Stat. Med.* **26**, 769–781.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–568.

Huang, J. & Rossini, A. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Amer. Statist. Assoc.* **92**, 960–967.

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* (eds D. Y. Lin & T. R. Fleming), 123–169. Springer-Verlag, New York.

Huang, J., Zhang, Y. & Hua, L. (2008). A least-squares approach to consistent information estimation in semiparametric models. *Technical Report 2008–3*, Department of Biostatistics, University of Iowa.

Jamshidian, M. (2004). On algorithms for restricted maximum likelihood estimation. *Comput. Statist. Data Anal.* **45**, 137–157.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comput. Graph. Statist.* **7**, 310–321.

Kim, M. & Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Stat. Med.* **21**, 3715–3726.

Kooperberg, C. & Clarkson, D. (1997). Hazard regression with interval-censored data. *Biometrics* **53**, 1485–1494.

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

Li, G. & Zhang, C. (1998). Linear regression with interval censored data. *Ann. Statist.* **26**, 1306–1327.

Lin, D., Oakes, D. & Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–298.

Lu, M. (2007). *Monotone spline estimations for panel count data*. PhD Dissertation, Department of Biostatistics, 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**, 705–718.

Martinussen, T. & Scheike, T. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* **89**, 649–658.

Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *J. Comput. Graph. Statist.* **8**, 109–120.

Rabinowitz, D., Tsiatis, A. & Aragon, J. (1995). Regression with interval-censored data. *Biometrika* **82**, 501–513.

Rosen, J. (1960). The gradient projection method for nonlinear programming. *J. Soc. Ind. 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. Amer. Statist. Assoc.* **93**, 318–327.

Schumaker, L. (1981). *Spline function: basic theory*. John Wiley, New York.

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. *Stat. Med.* **20**, 1855–1870.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165–177.

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. *Stat. Med.* **27**, 3178–3190.

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.

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

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.
van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes with application to statistics*. Springer-Verlag, New York.
Wellner, J. & Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779–814.
Wellner, J. & Zhang, Y. (2007). Likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35**, 2106–2142.
Xue, H., Lam, K. & Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Amer. Statist. Assoc.* **99**, 346–356.
Zhang, Y. & Jamshidian, M. (2004). On algorithms for the nonparametric maximum likelihood estimator of the failure function with censored data. *J. Comput. Graph. Statist.* **13**, 123–140.
Zhang, W., Zhang, Y., Chaloner, K. & Stapleton, J. (2009). Imputation methods for doubly censored HIV data. *J. Statist. Comput. Simulation* **79**, 1245–1257.

Received February 2009, in final form August 2009

Ying Zhang, Department of Biostatistics, University of Iowa, 200 Hawkins Drives, C22 GH, Iowa City, IA 52242, USA.
E-mail: ying-j-zhang@uiowa.edu

Appendices

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 (1988) and van der Vaart & Wellner (1996) will be heavily involved. Throughout the following proofs, we denote $Pf=\int f(x)\,\mathrm{d}P(x)$ and $\mathbb{P}_nf=\frac{1}{n}\sum_{i=1}^{n}f(X_i)$, the empirical process indexed by function $f(X)$, and we let C represent a generic constant that may vary from place to place.

Proof of theorem 1. Before deriving the convergence rate, we need to show that the sieve semiparametric maximum likelihood estimator $\hat{\tau}_n$ 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 $\mathbb{M}_n(\tau)=\mathbb{P}_nl(\tau;X)=\mathbb{P}_nl(\theta,\phi;X)$. Hence for any $\tau\in\mathcal{T}_n=\Theta\times\mathcal{M}_n$, $\mathbb{M}_n(\tau)-\mathbb{M}(\tau)=(\mathbb{P}_n-P)l(\tau;X)$.

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

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

and

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

Using Taylor expansion along with conditions (C1)–(C3), we can easily demonstrate that $\mathbb{P}_n|l^U_{s,i}(X)-l^L_{s,i}(X)|\le C\epsilon$ for all $1\le s\le[C(1/\epsilon)^d]$ and $1\le i\le[(1/\epsilon)^{Cq_n}]$ which leads to the conclusion that the ϵ-bracketing number for $\mathcal{L}_1$ with $L_1(\mathbb{P}_n)$-norm is bounded by $C(1/\epsilon)^{Cq_n+d}$.*** As $N(\epsilon,\mathcal{L}_1,L_1(\mathbb{P}_n))\le N_{[\,]}(2\epsilon,\mathcal{L}_1,L_1(\mathbb{P}_n))$, $\mathcal{L}_1$ is Glivenko–Cantelli by theorem 2.4.3 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 $m(x)=x\log x-x+1\ge(x-1)^2/4$ for $0\le x\le 5$. Further analysis by using Taylor expansion and conditions (C1)–(C3) leads to

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

With conditions (C1)–(C6), using the same arguments as those in Wellner & Zhang (2007, pp. 2126–2127), leads to

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

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

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

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

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

$$\begin{aligned}
\mathbb{M}_n(\hat\tau_n)-\mathbb{M}_n(\tau_0)&=\mathbb{M}_n(\hat\tau_n)-\mathbb{M}_n(\tau_{0,n})+\mathbb{M}_n(\tau_{0,n})-\mathbb{M}_n(\tau_0)\\
&\ge\mathbb{P}_nl(\tau_{0,n};X)-\mathbb{P}_nl(\tau_0;X)\\
&=(\mathbb{P}_n-P)\{l(\tau_{0,n};X)-l(\tau_0;X)\}+\mathbb{M}(\tau_{0,n})-\mathbb{M}(\tau_0).
\end{aligned}$$

Using the brackets for $\mathcal{M}_n$ given before, we can similarly construct a set of brackets for the class $\mathcal{L}_2=\{l(\theta_0,\phi;x)-l(\theta_0,\phi_0;x)\colon\ \phi\in\mathcal{M}_n$ and $\|\phi-\phi_0\|_\Phi\le Cn^{-p\nu}\}$ 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, p. 270). Hence the class $\mathcal{L}_2$ is P-Donsker. By the dominated convergence theorem, it is obvious that in this class $P\{l(\theta_0,\phi;X)-l(\theta_0,\phi_0;$

$X)\}^2 \to 0$ as $n \to \infty$. Hence

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

by the relationship between 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,

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

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

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

Second, we further explore $\mathbb{M}_n(\hat{\tau}_n) - \mathbb{M}_n(\tau_0)$. In the proof of consistency, we know that $\mathbb{M}_n(\hat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq I_{1,n} + I_{2,n}$, 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

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

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

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

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

Let $\mathcal{L}_3(\eta) = \{l(\tau; x) - l(\tau_0; x)\colon \phi \in \mathcal{M}_n \text{ and } d(\tau, \tau_0) \leq \eta\}$. 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

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

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

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 $\mathbb{P}_n\dot{l}_2(\hat{\theta}_n, \hat{\phi}_n; X)(\xi_0) = o_p(n^{-1/2})$ as $\mathbb{P}_n\dot{l}_1(\hat{\theta}_n, \hat{\phi}_n; X) \equiv 0$. 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, pp. 435–436) that there exists a $\xi_{0,n} \in S_n(D_n, K_n, m)$ such that

$\|\xi_{0,n}-\xi_0\|_\Phi = O(q_n^{-1}) = O(n^{-v})$ and $\mathbb{P}_n \dot{l}_2(\hat{\tau}_n; X)(\xi_{0,n}) = 0$. Therefore, we can write $\mathbb{P}_n \dot{l}_2(\hat{\tau}_n; X)(\xi_0) = I_{3,n} + I_{4,n}$, where

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

and

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

Let $\mathcal{L}_4 = \{\dot{l}_2(\tau; x)(\xi_0 - \xi) : \tau \in \mathcal{T}_n, \quad \xi \in S_n(D_n, K_n, m)$ and $\|\xi_0 - \xi\|_\Phi \le n^{-v}\}$. 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 to $\mathcal{L}_4$ being Donsker. 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

$$I_{4,n} \le Cd(\hat{\tau}_n, \tau_0)\|\xi_0 - \xi_{0,n}\|_\Phi = O_p(n - \min(pv, (1-v)/2)n^{-v}) = O_p(n - \min(v(p+1), (1+v)/2)) = o_p(n^{-1/2}).$$

So (B1) holds. (B2) holds by similarly verifying that the class $\mathcal{L}_5(\eta) = \{l^*(\tau; x) - l^*(\tau_0; x) : \tau \in \mathcal{T}_n$ and $d(\tau, \tau_0) \le \eta\}$ 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.

Appendix B: A general theorem

This appendix 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 a simplified version of the general theorem given in Huang (1996). The following conditions will be assumed.

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

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

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

Theorem 3

Suppose (B1)–(B3) *are satisfied, and suppose that* $I(\theta_0)$ *is non-singular. Then*

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

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

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

By (B1), it follows that

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

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

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

The result follows.