Published in final edited form as:

Scand Stat Theory Appl. 2016 June ; 43(2): 476–486. doi:10.1111/sjos.12186.

Fitting Cox Models with Doubly Censored Data Using Spline-Based Sieve Marginal Likelihood

Zhiguo Li and Kouros Owzar

Department of Biostatistics and Bioinformatics, Duke University

Abstract

In some applications, the failure time of interest is the time from an originating event to a failure event, while both event times are interval censored. We propose fitting Cox proportional hazards models to this type of data using a spline-based sieve maximum marginal likelihood, where the time to the originating event is integrated out in the empirical likelihood function of the failure time of interest. This greatly reduces the complexity of the objective function compared with the fully semiparametric likelihood. The dependence of the time of interest on time to the originating event is induced by including the latter as a covariate in the proportional hazards model for the failure time of interest. The use of splines results in a higher rate of convergence of the estimator of the baseline hazard function compared with the usual nonparametric estimator. The computation of the estimator is facilitated by a multiple imputation approach. Asymptotic theory is established and a simulation study is conducted to assess its finite sample performance. It is also applied to analyzing a real data set on AIDS incubation time.

Keywords

doubly censored data; multiple imputation; proportional hazards model; spline-based marginal likelihood

1 Introduction

In some clinical studies, the outcome of interest is the time from an originating event to a failure event, while the times of both of the events are censored. This type of data are called doubly censored data. In contrast, if the failure event of interest is interval censored while the time of the originating event is not censored, we call this type of data singly interval censored data. The most extensively discussed example in doubly censored data arises from an AIDS study (DeGruttola & Lagakos, 1989; Kim et al. 1993). In this study, 188 hemophiliacs who were treated with contaminated blood factor in two hospitals in France were found to be HIV infected, and 41 of them progressed to AIDS. One group of patients were heavily treated while the other group were lightly treated. It is of interest to know if the two groups differ in their AIDS incubation times, which is the time from the infection of

Zhiguo Li, Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Hock Plaza Suite 1102, Durham, NC 27710, USA. ; Email: zhiguo.li@duke.edu

Supporting information

Additional information for this article is available online, which includes proofs of Theorems 1 and 2.

HIV to the time of the onset of AIDS. Both of these two events were detected via periodic screening and thus both of the times were censored. Another example comes from victims of the Chernobyl nuclear accident who suffered from severe radiation syndrome. Some of those subjects received bone marrow transplantation while the others did not (Klymenko et al., 2011). To assess the effectiveness of the bone marrow transplantation, investigators need to know if the recovery time differs in the two groups, where the recovery time refers to the time from the blood cell count first drops below a threshold to the time the cell count rises and reaches a certain threshold. Again, both of these times were not exactly observed, since the cell counts were only measured intermittently at the patients' visits to the clinic. In both of the examples, the time to the originating event was censored in an interval. In the AIDS study, the time of the failure event (onset of AIDS) is right censored, but in the other example, the time of the failure event (blood cell count reaches the threshold again) may be censored in a finite interval.

Methods for analyzing doubly censored data have been studied by a number of authors. These include nonparametric inference of the distribution function of the time of interest (DeGruttola & Lagakos, 1989; Gomez & Lagakos, 1994) and semiparametric inference under the Cox proportional hazards model (Kim et al., 1993, Goggins et al., 1999; Sun et al., 1999; Pan, 2001; Kim, 2006). Kim et al. (1993) considered estimation in a proportional hazards model assuming discrete distributions for the time to the originating event and the failure time of interest. Goggins et al. (1999) used an EM algorithm to impute the time to the originating event to fit a proportional hazards model with right censored failure times of interest. Sun et al. (1999) proposed an estimating equation approach based on the marginal likelihood idea (Kalbfleisch & Prentice, 1973) under a proportional hazards model; Pan (2001) pointed out that the estimator in Sun et al. (1999) is asymptotically equivalent to a multiple imputation estimator; and Kim (2006) used an approximate likelihood (Geotghbeur and Ryan, 2000) and the EM algorithm to fit the proportional hazards model with doubly censored data. In most of the literature, it is assumed that the failure time of interest (e.g., the AIDS incubation time) is independent of the time to the originating event (e.g., time to HIV infection), and it is right censored. The only work we are aware of which does not impose these assumptions is Kim (2006). Kim's paper used the frailty model to account for the potential dependence between the time to the originating event and the failure time of interest. Under this framework, at first, the theoretical properties of the approximate likelihood estimator are unknown, as pointed out in Goetghbeur and Ryan (2000) for the singly interval censored data problem. Secondly, it is not convenient to perform a formal significance test for the dependence between the time to the originating event and the failure time of interest, except when the normal frailty is appropriate which may not be the case in practice. This can be an interesting hypothesis to be tested in many situations. For example, investigators may want to know if the time to HIV infection is associated with the AIDS incubation time.

The fully semiparametric likelihood approach, in which no smoothness assumptions are imposed on the baseline cumulative hazard function, to fitting the proportional hazards model with doubly censored data is very difficult if not impossible. In this article, we propose to use a spline-based marginal likelihood to fit the proportional hazards model with this type of data. This marginal likelihood is based on the semiparametric likelihood for

singly interval censored data (Zhang et al., 2010). We assume that both the time to the originating event and the failure time of interest can be interval censored. Our model allows a test for the dependence between those two times. The rate of convergence of the estimator of the unknown parameters and the asymptotic normality of the finite dimensional regression parameters are established using semiparametric theory. Moreover, we employ multiple imputation to calculate the marginal likelihood estimator, an approach that is also used in Pan (2001) when the failure event of interest is right censored. This approach reduces the complexity of the calculation because only the maximization of the likelihood for singly interval censored data as in Zhang et al. (2010) is involved. The spline-based approach is still a semiparametric approach because the dimension of the spline space increases with the sample size. However, compared with the fully semiparametric approach, the spline-based approach results in a higher rate of convergence of the estimator for the baseline hazard function when the true baseline hazard function is sufficiently smooth, although it cannot achieve the $\sqrt{n}$ rate as in parametric problems. This is because, unlike in the fully semiparametric approach, in the spline-based approach we take into account the smoothness of the baseline hazard function in its estimation. Moreover, the dimensionality of the optimization problem is greatly reduced by using the spline-based approach, especially when the sample size is large.

The article is organized as follows. We introduce the model and describe the spline-based marginal likelihood estimator in Section 2, and then study its asymptotic properties in Section 3. A multiple imputation approach to compute the estimator is described in Section 4. Section 5 presents results of a simulation study, and in Section 6 we illustrate the proposed method by analyzing the doubly censored AIDS incubation time data. We conclude with a discussion in Section 7. Proofs of asymptotic results are included in the Appendix.

2 The spline-based sieve maximum marginal likelihood estimator

Denote time to the originating event as S and time to the failure event as $T(T \geq S)$. The failure time of interest is $T^* = T - S$. In doubly censored data, both S and T are interval censored. We assume S is observed in an interval (L, R) and T is observed in (U, V) , where L, R, U and V are random observation times. Given $d - 1$ dimensional covariate vector Z and S , we assume the following proportional hazards model for T^* :

$$\Lambda(t|Z, S) = \Lambda(t) e^{\beta^T Z + \alpha S},$$

where $\Lambda(t|Z, S)$ is the cumulative hazard function of T^* given Z and S and $\Lambda(t)$ is the baseline cumulative hazard function. In this model, we include S as a covariate in the model for T^* , which enables us to test for the dependence between S and T^* . The time S usually has a practically meaningful start point and is not chosen arbitrarily. For example, in the AIDS study, the start point for S is the time of blood transfusion. However, it is worth mentioning that the choice of the start point for S is irrelevant in the above model because of the multiplicative baseline hazard function and the fact that S is linear in the exponential function.

We make the assumption that $R - U < \infty$, which guarantees that S and T are not censored in the same time interval. If S and T are censored in the same time interval, then the data contain no information about the difference $T^* = T - S$, and such subjects need to be excluded in the analysis. As usual, we also assume that the observation times L, R, U and V are independent of T^* given covariate Z . At first, for estimating the unknown parameters in the above model, if S was known, the log likelihood function for a single observation is, up to an additive constant (Zhang et al., 2010),

$$l(\beta, \alpha, \phi; Z, S) = \Delta_1 \log \left[1 - e^{-e^{\beta^T Z + \alpha S + \phi(U-S)}} \right] + \Delta_2 \log \left[e^{-e^{\beta^T Z + \alpha S + \phi(U-S)}} - e^{-e^{\beta^T Z + \alpha S + \phi(V-S)}} \right] - \Delta_3 e^{\beta^T Z + \alpha S + \phi(V-S)},$$

where $\Delta_1 = I(T < U)$, $\Delta_2 = I(U < T - V)$, $\Delta_3 = I(T > V)$, and $\phi(s) = \log \Lambda(s)$. However, in doubly censored data S is not exactly known. At first, suppose the true distribution function of S , denoted by $F_0(s)$, is known. Then we can use the following marginal likelihood (Sun et al., 1999; Kalbfleisch and Prentice, 1973) for the inference of unknown parameters:

$$m(\theta, \phi; F_0; X) = \int_L^R l(\theta, \phi; Z, s) dF_0(s),$$

where X denotes all the observed data. Let $\theta = (\beta^T, \alpha)^T$ and denote the d -dimensional parameter space for θ by Θ . Denote the parameter space for ϕ by Φ . The marginal likelihood estimator of (θ, ϕ) seeks to maximize

$$\mathbb{P}_n m(\theta, \phi; F_0; X) = \mathbb{P}_n \int_L^R l(\theta, \phi; Z, s) dF_0(s)$$

with respect to (θ, ϕ) over Θ , where $\mathbb{P}_n$ is the empirical measure, i.e., $\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$, for any function f . In practice, since $F_0(s)$ is unknown, we replace it with a consistent estimator, denoted by $\hat{F}_n(s)$. Then we maximize

$$\mathbb{P}_n m(\theta, \phi; \hat{F}_n; X) = \mathbb{P}_n \int_L^R l(\theta, \phi; Z, s) d\hat{F}_n(s),$$

over $\Theta \times \Phi$ to get an estimator for (θ, ϕ) . This maximization problem has a maximum dimension of $2n + d + 1$ and is challenging when the sample size is large. Hence, we propose to use a spline-based sieve maximum marginal likelihood estimation procedure. Zhang et al. (2010) used the spline-based sieve maximum likelihood to fit proportional hazards models with singly interval censored data. Spline-based methods have also been used by Kooperberg & Clarkson (1997) and Cai & Betensky (2003) in analyzing singly interval censored data. When the underlying baseline hazard function is sufficiently smooth, the spline-based method not only reduces the dimensionality of the optimization problem but also improves the rate of convergence of the estimator of the baseline hazard function.

Suppose a and b are the lower and upper bounds of (U_i, V_i) , $1 \leq i \leq n$. Let $a = d_0 < d_1 < \dots < d_{K_n} < d_{K_n+1} = b$ be a partition of $[a, b]$ such that $\max_{1 \leq k \leq K_n+1} |d_k - d_{k-1}| = O(n^{-\nu})$, where $K_n = O(n^\nu)$. Let $D_n = \{d_1, \dots, d_{K_n}\}$. Let $\mathcal{S}_n(D_n, K_n, m)$ be the space of polynomial splines of order m with knots D_n . According to Schumaker (1981), there exists a local basis $\mathcal{B}_n = \{b_j(t), 1 \leq j \leq q_n\}$ for $\mathcal{S}_n(D_n, K_n, m)$, where $q_n = K_n + m$. Define

$$\Phi_n = \left\{ \phi_n : \phi_n(t) = \sum_{j=1}^{q_n} \gamma_j b_j(t), \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{q_n}, t \in [a, b] \right\},$$

where the restriction $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{q_n}$ guarantees that the functions in Φ_n are monotone increasing, a consequence of the variation diminishing property of B-spline (Schumaker, 1981). For the spline-based marginal likelihood estimation of (θ, ϕ) , we maximize $\mathbb{P}_n \mathfrak{M}(\theta, \phi, X, F_n)$ over $\Theta \times \Phi_n$. That is,

$$(\hat{\theta}_n, \hat{\phi}_n) = \arg \max_{\theta \in \Theta, \phi \in \Phi_n} \mathbb{P}_n \int_L^R l(\theta, \phi; Z, s) d\hat{F}_n(s). \quad (1)$$

This is equivalent to maximization with respect to $\theta \in \Theta$ and $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{q_n}$.

Before calculating the estimator $(\hat{\theta}_n, \hat{\phi}_n)$, we need to calculate $F_n(s)$, which is based on the observed data (L_i, R_i) , $1 \leq i \leq n$, using the spline-based sieve maximum likelihood estimator as in Zhang et al. (2010). Denote $\Lambda(s)$ as the cumulative hazard function of S . Due to technical reasons, similar to Zeng (2005), we impose some restrictions on the spline space when we define the sieve maximum likelihood estimator for $\Lambda(s)$. Denote $\varphi(s) = \log\{\Lambda(s)\}$ with the truth being $\varphi_0(s) = \log\{\Lambda_0(s)\}$. Denote the parameter space for φ to be $\tilde{\Phi}$. Define $\tilde{\mathcal{S}}_n(D_n, K_n, m)$ in parallel to $\mathcal{S}_n(D_n, K_n, m)$ above but with (U_i, V_i) replaced by (L_i, R_i) , $1 \leq i \leq n$. Let $\tilde{\mathcal{B}}_n = \{\tilde{b}_j(t), 1 \leq j \leq \tilde{q}_n\}$ be a local basis for $\tilde{\mathcal{S}}_n(D_n, K_n, m)$, where $\tilde{q}_n = K_n + m$ and $K_n = O(n^\nu)$. Define

$$\tilde{\Phi}_n = \left\{ \tilde{\phi}_n : \tilde{\phi}_n(s) = \sum_{j=1}^{\tilde{q}_n} \gamma_j \tilde{b}_j(s), \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{\tilde{q}_n}, s \in [c, d], \sum_{j=1}^{\tilde{q}_n} |\gamma_j| \leq M_n \right\},$$

where c and d are the lower and upper bounds of L_i, R_i , $1 \leq i \leq n$, respectively, and $M_n = \log n$. The spline-based sieve maximum likelihood estimator of φ_0 is defined as

$$\hat{\phi}_n = \arg \max_{\tilde{\phi} \in \tilde{\Phi}_n} \mathbb{P}_n \log \left\{ e^{-e^{\tilde{\phi}(L)}} - e^{-e^{\tilde{\phi}(R)}} \right\},$$

and then we set $\hat{F}_n(s) = 1 - e^{-e^{\hat{\phi}_n(s)}}$. Compared with the nonparametric estimator of $\varphi_0(s)$, the above estimator of $\varphi_0(s)$ has a higher rate of convergence when the true function $\varphi_0(s)$ is sufficiently smooth.

3 Asymptotic properties

At first, we introduce some necessary notation. For two functions ϕ_1 and ϕ_2 in Φ , define

$$\|\phi_1 - \phi_2\|_{\Phi}^2 = E \left[\int_L^R E\{\phi_1(U-s) - \phi_2(U-s)\}^2 + E\{\phi_1(V-s) - \phi_2(V-s)\}^2 dF_0(s) \right].$$

Based on this we define the distance between two elements in $\Theta \times \Phi$ by

$$d^2((\theta_1, \phi_1), (\theta_2, \phi_2)) = \|\theta_1 - \theta_2\|^2 + \|\phi_1 - \phi_2\|_{\Phi}^2,$$

where $\|\cdot\|$ is the Euclidean norm. Define

$$\dot{m}_1(\theta, \phi, F; X) = \int_L^R \frac{\partial l(\theta, \phi; Z, s)}{\partial \theta} dF(s).$$

Let φ_t be a parametric submodel for φ with $\varphi_0 = \varphi$ and let $g = \varphi_t - \varphi|_{t=0}$. Denote $\mathcal{G}$ to be the collection of all such functions g . Define

$$\dot{m}_2(\theta, \phi, F; X)[g] = \int_L^R \frac{\partial l(\theta, \phi_t; Z, s)}{\partial t} \Big|_{t=0} dF(s),$$

and

$$\dot{m}_2(\theta, \phi, F; X)[\mathbf{g}] = (\dot{m}_2(\theta, \phi, F; X)[g_1], \dots, \dot{m}_2(\theta, \phi, F; X)[g_d])^T,$$

where $\mathbf{g} = (g_1, \dots, g_d)^T$. Let

$$m^*(\theta, \phi, F; X) = \dot{m}_1(\theta, \phi, F; X) - \dot{m}_2(\theta, \phi, F; X)[\mathbf{g}^*],$$

where $\mathbf{g}^* = (g_1^*, \dots, g_d^*)^T$ minimizes $E\|\dot{m}_1(\theta, \phi, F; X) - \dot{m}_2(\theta, \phi, F; X)[\mathbf{g}]\|^2$ over $\mathbf{g} \in \mathcal{G}^d$. Let

$$I_F(\theta, \phi) = E\{m^*(\theta, \phi, F; X)m^{*T}(\theta, \phi, F; X)\},$$

which is an analogue of the semiparametric information matrix for θ (Bickel et al., 1993) (or pseudo semiparametric information matrix).

The following regularity conditions are imposed to obtain the asymptotic properties.

- A1** (a) $E\{(Z, S)^T(Z, S)\}$ is nonsingular; and (b) There exists constants z_0 and s_0 such that $P(|Z| < z_0) = 1$ and $P(|S| < s_0) = 1$.
- A2** Θ is a compact subset of R^d .
- A3** (a) There exists a positive number η such that $P(V - U \leq \eta) = 1$, the union of the supports of U and V is a finite interval $[a, b]$, and $0 < \Lambda_0(a) < \Lambda_0(b) < \infty$; and (b) There exists a positive number η such that $P(R - L \leq \eta) = 1$, the union of the supports of L and R is a finite interval $[c, d]$, and $0 < \Lambda_0(c) < \Lambda_0(d) < \infty$.
- A4** (a) The parameter space Φ is a class of functions with bounded p th derivative in $[a, b]$ for $p \geq 1$, and $\phi'_0(s)$ is strictly positive and continuous on $[a, b]$; and (b) The parameter space Φ is a class of functions with bounded q th derivative in $[c, d]$ for $q \geq 1$, and $\tilde{\phi}'_0(s)$ is strictly positive and continuous on $[c, d]$.
- A5** (a) The conditional density of (U, V) given Z has bounded partial derivative with respect to (u, v) . The bounds of these partial derivatives do not depend on (u, v, z) ; and (b) The density functions of L and R are bounded above and bounded away from 0 in their supports.
- A6** For some $\kappa \in (0, 1)$, $\mu^T \text{var}((Z^T, S)^T | U) \mu \leq \kappa \mu^T E((Z^T, S)^T (Z^T, S) | U) \mu$ and $\mu^T \text{var}((Z^T, S)^T | V) \mu \leq \kappa \mu^T E((Z^T, S)^T (Z^T, S) | V) \mu$, a.s., for all $\mu \in R^d$.
- A7** The distribution of (L, R, U, V, Z, S) does not depend on (θ, Λ) , S and (L, R) are independent, and T and (U, V) are independent given (Z, S) .

These conditions are mostly the same as those in Zhang et al. (2010), with some additional assumptions for our particular problem. The condition $P(|S| < s_0)$ appears to be restrictive. However, since our analysis is conditional on the fact that the first event is observed (otherwise both events are right censored and such a subject must be excluded as mentioned above), that condition is satisfied under Assumption A3. Under these assumptions we derive the rate of convergence for $(\hat{\theta}_n, \hat{\phi}_n)$ and prove the asymptotic normality of $\hat{\theta}_n$ as stated in Theorems 1 and 2 below, respectively.

Theorem 1

Under assumptions A1–A7, we have

$$d((\hat{\theta}_n, \hat{\phi}_n), (\theta_0, \phi_0)) = O_p(r_n^{-1}), \text{ as } n \rightarrow \infty,$$

where $r_n = n^\alpha (\log n)^{-2/5}$, $\alpha = \min(\beta, p\nu, (1 - \nu)/2)$ and $\beta = (3q - 4)/(5(2q + 1))$.

The factor $\log n$ appears in the rate of convergence because of the restriction on the spline space $\tilde{\Phi}_n$ and the choice $M_n = \log n$. Here M_n is chosen for technical reasons to restrict the spline space to guarantee the rate of convergence for the derivative $\hat{F}'_n(s)$. If $\nu = 1/(2p+1)$, then $\min(p\nu, (1 - \nu)/2) = p/(2p+1)$. Since $p/(2p+1) > (3q - 4)/(5(2q + 1))$ for any $p, q \geq 1$, this theorem indicates that, when $\nu = 1/(2p + 1)$, the overall rate of convergence for the parameter estimator can achieve $n^\beta (\log n)^{-1/4}$. From the proof of the theorem, this is a rate of

convergence for $\hat{F}'_n(s)$. From the proof, it is also clear that, if $F_0(s)$ was known, the rate of convergence for the parameter estimator would be $n^{\min(p\nu, (1-\nu)/2)}$, which is the same as the rate of convergence for the spline-based maximum likelihood estimator based on singly interval censored data (Zhang et al., 2010). This theorem shows that, since $F_0(s)$ has to be estimated and $\hat{F}'_n(s)$ is involved in the objective function, the rate of convergence for $(\hat{\theta}_n, \hat{\varphi}_n)$ is controlled by that for $\hat{F}'_n(s)$. On the other hand, if, as in Sun et al. (1999), we assume that S has finite supporting points, then the density of S is strongly consistent and has a $\sqrt{n}$ rate of convergence (Yu et al., 1998). In that case, it is easy to show that the rate of convergence for $(\hat{\theta}_n, \hat{\varphi}_n)$ is the same as in Zhang et al. (2010). Moreover, Groeneboom & Wellner (1992) showed that the nonparametric estimator for the cumulative hazard function based on singly interval censored data has a $n^{1/3}$ rate of convergence and its asymptotic distribution is not normal. Therefore, it is expected that the asymptotic distribution of $\hat{\varphi}_n$ is not normal, either, but the determination of this asymptotic distribution is out of the scope of this paper. The rate of convergence given in Theorem 1 is essentially that for $\hat{\varphi}_n$ because the following theorem shows that the rate of convergence for $\hat{\theta}_n$ is $\sqrt{n}$ and it is asymptotically normally distributed.

Theorem 2

Under assumptions A1–A7, it holds that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I_{F_0}^{-1}(\theta_0, \phi_0)),$$

as $n \rightarrow \infty$.

Here the matrix $I_{F_0}(\theta_0, \phi_0)$ is the pseudo semiparametric information matrix for θ evaluated at the true parameter values. This matrix can be consistently estimated by a least squares estimator similarly as in Zhang et al. (2010) as follows. Define

$$\rho_n(\mathbf{g}) = \frac{1}{n} \sum_{i=1}^n \|\dot{m}_1(\hat{\theta}_n, \hat{\phi}_n, \hat{F}_n; X_i) - \dot{m}_2(\hat{\theta}_n, \hat{\phi}_n, \hat{F}_n; X_i)[\mathbf{g}]\|^2.$$

Suppose $\hat{\mathbf{g}}_n$ is a minimizer of $\rho_n(\mathbf{g})$ over $\mathbf{g} \in \mathcal{G}^d$, then we estimate $I_{F_0}(\theta_0, \phi_0)$ by $\rho_n(\hat{\mathbf{g}}_n)$. Since we propose the multiple imputation method in the next section, this method for variance estimation is not used and hence we do not include more details which can be found in Huang et al. (2008).

4 A multiple imputation procedure for computing the estimator

For convenience of computation, we use a multiple imputation approach to approximate $\hat{\theta}_n$ and its standard error. Utilizing this approach reduces the maximization problem (1) to the less complicated maximization problem for singly interval censored data and hence the computation program for singly interval censored data can be used to calculate the estimator

(1) and its standard error. A similar approach has been used in Pan (2001) to approximate the estimator in Sun et al. (1999).

As pointed out by Pan (2001), the marginal likelihood can be treated as an expectation of the usual likelihood given the observed data. Specific to our situation, we can approximate this expectation by an empirical mean as

$$\mathbb{P}_n^m(\theta, \phi, \hat{F}_n; X) \approx \frac{1}{M} \sum_{m=1}^M \mathbb{P}_n l(\theta, \phi; X, S^{(m)}),$$

where $S^{(m)}, 1 \leq j \leq M$, is a random sample drawn from the distribution $F_n^{\wedge}(s)$ conditional on the observed data. We can maximize the right-hand side of the above display to obtain an approximate maximum marginal likelihood estimator of (θ_0, φ_0) . Alternatively, one can maximize each of the summands and obtain M estimates and then summarize them to form a final estimate. This is the multiple imputation procedure described in detail in the following.

For each $1 \leq m \leq M$, we maximize $\mathbb{P}_n l(\theta, \varphi; Z, S^{(m)})$ over $\Theta \times \mathcal{M}_n$ to obtain $\hat{\theta}_n^{(m)}$. This is the maximum likelihood estimator of the unknown parameters based on singly interval censored data, and a generalized version of the Rosen algorithm in Zhang et al. (2010) can be used for the computation. The variance of $\hat{\theta}_n^{(m)}$, which is the inverse of the information matrix for θ , can be estimated using the least squares method in Zhang et al. (2010), too.

We denote the variance estimate for $\hat{\theta}_n^{(m)}$ as $\rho_n(\hat{\mathbf{g}}_n^{(m)})/n, 1 \leq m \leq M$. Finally, we set

$$\bar{\theta}_n = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_n^{(m)}$$

to be the estimator for θ_0 with a variance estimate

$$\hat{\Sigma} = \left(1 + \frac{1}{M}\right) \frac{\sum_{m=1}^M (\hat{\theta}_n^{(m)} - \bar{\theta}_n)^T (\hat{\theta}_n^{(m)} - \bar{\theta}_n)}{M-1} + \frac{1}{M} \sum_{m=1}^M \rho_n(\hat{\mathbf{g}}_n^{(m)}).$$

By Lemma 1 of Wang and Robbins (1998), it can be shown that $\bar{\theta}_n$ is asymptotically equivalent to $\hat{\theta}_n$ in the sense that $\sqrt{n}(\bar{\theta}_n - \hat{\theta}_n) = o_p(1)$.

As pointed out by Pan (2001), for the multiple imputation, a small number M usually suffices, although it is not an issue to use a large M with the current computing power. It is advisable to increase M gradually to see how the results change. An M is considered to be good enough when the results do not change much compared with those with a smaller M .

5 Simulation

A small simulation study is carried out to assess the performance of the proposed methods in the previous sections. In this simulation, we generate time to the originating event S from an exponential distribution with mean 2. The failure time of interest T^* is also generated from an exponential distribution with mean $1/e^{\beta_0^T Z + \alpha_0 S}$, given S and a covariate vector Z , where $\beta_0 = (0.5, -0.5)^T$, $\alpha_0 = 1$, $Z = (Z_1, Z_2)^T$, $Z_1 \sim \text{Bernoulli}(0.5)$, and $Z_2 \sim N(0, 1)$. After this we set $T = S + T^*$. Then we generate a series of visit times for each subject. The first visit time is generated from a uniform distribution in $(0, 1)$. The lengths of the time intervals between two adjacent visit times are generated independently from a uniform distribution in $(0.5, 1)$. Any visit times greater than 5 are discarded. According to these data, the time intervals in which S and T are censored are determined, and the cases in which they are censored in the same time interval are discarded. Under this simulation setting, the percentages of the event of interest being left censored, censored between two (nonzero) finite observation times, and right censored, are all between about 20% and 40%.

The sieve maximum likelihood estimator $F_n(\hat{s})$ is computed using the cubic B-spline by the algorithm described in Zhang et al. (2010). After this, we use the multiple imputation procedure described in Section 4 to calculate the estimate θ_n and its standard error. In calculating the spline estimates for $F_n(\hat{s})$ and $\varphi_n(\hat{t})$, the numbers of knots are chosen to be $L_n = \lceil N_1^{1/3} \rceil$ and $K_n = \lceil N_2^{1/3} \rceil$, respectively, where N_1 and N_2 are the numbers of distinct observation times for S and T , respectively. The knots are put at the L_n (or K_n) quantiles of the distinct observation times for S (or T) in calculating $F_n(\hat{s})$ (or $\varphi_n(\hat{t})$). A total of 1000 simulation replications are repeated, over which the biases, empirical standard errors, the average of estimated standard errors, and the empirical coverage rate of the 95% confidence intervals of the regression parameters are computed.

The simulation results are presented in Table 1 for sample sizes $n = 100, 200, 500$ and 800 . The results show that for smaller sample sizes, there can be considerable biases, and the biases for the β s are larger than that for α , but the biases decrease when the sample size becomes larger. In addition to the bias of θ_n when $F_0(s)$ is known, the bias for estimating $F_0(s)$ adds to the bias of θ_n . For smaller sample sizes, the variances are overestimated which results in coverage probabilities of the 95% confidence intervals larger than the nominal level, but for larger sample sizes this becomes less a problem.

6 An application

A common situation in which doubly censored data arise is the study of AIDS incubation time. In these studies, it is of interest to make inference about the distribution of the AIDS incubation time, which is defined as the time between the infection of human immunodeficiency virus (HIV) and the onset of AIDS. Both the HIV infection and AIDS onset are observed at periodic screenings, and the exact times of the events are unknown but can only be determined to belong to two adjacent examination times: the time of last negative test and the time of the first positive test.

In this section we apply our proposed method to analyzing the AIDS incubation time in the example described in Section 1. All the data are tabulated in Table 1 in Sun et al. (1999). Overall, 188 patients were infected with HIV, and 41 of them progressed to AIDS. Ninety six patients were heavily treated (received $> 1000 \mu\text{g/kg}$ of blood factor) and the other 92 patients were lightly treated. We denote $Z = 1$ if a patient was heavily treated and $Z = 0$ otherwise. The goal is to investigate the association between treatment (Z) and AIDS incubation time (T^*), and between the time to HIV infection (S) and AIDS incubation time. We fit a proportional hazards model for T^* with Z and S as covariates. The estimated log hazard ratio for Z is 0.68 with a standard error 0.32, which gives a two-sided p value 0.03. For time to HIV infection, the estimated log hazard ratio is -0.11 with a standard error 0.09, which results in a two-sided p value 0.22. The conclusion is that the heavily treated group had shorter incubation time than the lightly treated group. Longer time to HIV infection seems to be associated with longer incubation time but this association is not significant under the 0.05 significance level.

7 Discussion

For semiparametric regression analysis of doubly censored failure time data, we proposed a spline-based marginal likelihood method with rigorously established theoretical properties and reasonable finite sample performance. The implementation of the method is facilitated by a multiple imputation approach, which reduces the problem to one for singly interval censored data. In addition to testing for the effects of covariates, the method allows for a straightforward test for the association between the time to the originating event and the failure time of interest.

As noted in Meyer (2008) and Lu (2015), the monotonicity of the coefficients of the spline basis functions is a sufficient but not necessary condition for the monotonicity of the spline functions we use. Consequently this spline space is more restrictive than necessary, although this is not an issue as we can achieve the desired precision in approximating smooth functions using this space of spline functions by Lemma A1 in Lu et al. (2007).

Some authors considered data-adaptive selection of the number of knots. For example, Xue & Liang (2010), Lu & Loomis (2013) and Lu (2015) selected the number of knots by minimizing the Akaike information criterion. However, data-adaptive knot selection methods are not fully developed and can introduce theoretical as well as practical difficulties, and hence it is not further explored in this paper.

For future research, the proposed method can be applied to other semiparametric regression models with doubly censored data, such as the proportional odds regression (Huang & Rossini, 1997; Shen, 1998), partial linear regression (Xue et al., 2004), and additive hazards model (Lin et al., 1998; Martinussen & Scheike, 2002). The theoretical properties can be established in a similar way as those for singly interval censored data. The multiple imputation procedure described above can also be used to reduce the computation to that for singly interval censored data for these models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by a grant from the National Institutes of Health with award number P01CA142538. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank the editor, associate editor and two anonymous referees for their helpful comments.

References

- Bickel, PJ.; Klaassen, CA.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins University Press; 1993.
- Cai T, Betensky R. Hazard regression for interval-censored data with penalized spline. *Biometrics*. 2003; 59:570–579. [PubMed: 14601758]
- DeGruttola V, Lagakos SW. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*. 1989; 45:1–12. [PubMed: 2497809]
- Goetghebuer E, Ryan L. Semiparametric regression analysis of interval-censored data. *Biometrics*. 2000; 56:1139–1144.
- Goggins WB, Feinkelstein DM, Zaslavsky AM. Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine*. 1999; 18:2737–2747. [PubMed: 10521863]
- Gomez G, Lagakos SW. Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*. 1994; 50:204–212. [PubMed: 8086603]
- Groeneboom, P.; Wellner, JA. DMV Seminar, Band 19. Birkhauser; New York: 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation.
- Huang J, Rossini A. Sieve estimation of the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*. 1997; 92:960–967.
- Huang, J.; Zhang, Y.; Hua, L. Technical Report 2008-3. University of Iowa, Department of Biostatistics; 2008. A least-squares approach to consistent information estimation in semiparametric models.
- Kalbfleisch JD, Prentice RL. Marginal likelihood based on Cox regression and life model. *Biometrika*. 1973; 60(2):267–278.
- Kim MY, DeGruttola V, Lagakos SW. Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*. 1993; 49:13–22. [PubMed: 8513098]
- Kim YJ. Regression analysis of doubly censored failure time data with frailty. *Biometrics*. 2006; 62:458–464. [PubMed: 16918909]
- Klymenko SV, Belyi DA, Ross JR, Owzar K, Jiang C, Li Z, Minchenko NJ, Kovalenko NA, Bebesko VG, Chao JN. Hematopoietic cell infusion for the treatment of nuclear disaster victims: new data from the Chernobyl accident. *International Journal of Radiation Biology*. 2011; 87(8):846–850. [PubMed: 21406047]
- Kooperberg C, Clarkson D. Hazard regression with interval-censored data. *Biometrics*. 1997; 53:1485–1494. [PubMed: 9423263]
- Lin D, Oakes D, Ying Z. Additive hazard regression with current status data. *Biometrika*. 1998; 85:289–298.
- Lu M. Spline estimation of generalized monotonic regression. *Journal of Nonparametric Statistics*. 2015; 27:19–39.
- Lu M, Loomis D. Spline-based semiparametric estimation of partially linear Poisson regression with single-index models. *Journal of Nonparametric Statistics*. 2013; 25:905–922.
- Lu M, Zhang Y, Huang J. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*. 2007; 94:705–718.

- Martinussen T, Scheike T. Efficient estimation in additive hazards regression with current status data. *Biometrika*. 2002; 89:649–658.
- Meyer MC. Inference using shape-restricted regression splines. *The Annals of Applied Statistics*. 2008; 2:1013–1033.
- Pan W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics*. 2001; 57:1245–1250. [PubMed: 11764266]
- Schumaker, L. Spline function: basic theory. New York: John Wiley; 1981.
- Shen X. Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*. 1998; 85:165–177.
- Sun J, Liao Q, Pagano M. Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics*. 1999; 55:909–194. [PubMed: 11315027]
- Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika*. 1998; 85:935–948.
- Xue H, Lam K, Li G. Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*. 2004; 99:346–356.
- Xue L, Liang H. Polynomial spline estimation for a generalized additive coefficient model. *Scandinavian Journal of Statistics*. 2010; 37:26–46.
- Yu Q, Schick A, Li L, Wong GYC. Estimation of a distribution function with case 2 interval-censored data. *Statistics and Probability Letters*. 1998; 37:223–228.
- Zeng D. Likelihood approach for marginal proportional hazards regression in the presence of dependent censoring. *Annals of Statistics*. 2005; 33(2):501–521.
- Zhang Y, Hua L, Huang J. Spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*. 2010; 37:338–354.

Table 1

Simulation results for the spline-based sieve maximum likelihood estimator

n	β_1				β_2				α			
	100	200	500	800	100	200	500	800	100	200	500	800
Bias	0.103	0.064	0.028	0.011	0.132	0.077	0.040	0.025	0.088	0.057	0.033	0.021
SE	0.719	0.389	0.196	0.091	0.375	0.168	0.069	0.033	0.295	0.133	0.053	0.021
SD	0.625	0.320	0.174	0.082	0.307	0.145	0.063	0.029	0.283	0.130	0.049	0.018
95% CP	0.978	0.969	0.961	0.958	0.974	0.973	0.965	0.953	0.976	0.969	0.963	0.955

SE: average estimated SE; SD: empirical SD; CP: coverage probability.