Published in final edited form as:

J Am Stat Assoc. 2011 December 1; 106(496): 1434–1449. doi:10.1198/jasa.2011.tm10156.

Maximum Likelihood Estimations and EM Algorithms with Length-biased Data

Jing Qin,

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, NIH
Bethesda, Maryland 20892, USA, Phone: 301-451-2436

Jing Ning,

Division of Biostatistics, The University of Texas, Health Science Center at Houston, School of
Public Health, Houston, Texas 77030, USA

Hao Liu, and

Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston,
Texas 77030, USA

Yu Shen

Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston,
Texas 77030, USA

Jing Qin: jingqin@niaid.nih.gov

SUMMARY

Length-biased sampling has been well recognized in economics, industrial reliability, etiology applications, epidemiological, genetic and cancer screening studies. Length-biased right-censored data have a unique data structure different from traditional survival data. The nonparametric and semiparametric estimations and inference methods for traditional survival data are not directly applicable for length-biased right-censored data. We propose new expectation-maximization algorithms for estimations based on full likelihoods involving infinite dimensional parameters under three settings for length-biased data: estimating nonparametric distribution function, estimating nonparametric hazard function under an increasing failure rate constraint, and jointly estimating baseline hazards function and the covariate coefficients under the Cox proportional hazards model. Extensive empirical simulation studies show that the maximum likelihood estimators perform well with moderate sample sizes and lead to more efficient estimators compared to the estimating equation approaches. The proposed estimates are also more robust to various right-censoring mechanisms. We prove the strong consistency properties of the estimators, and establish the asymptotic normality of the semi-parametric maximum likelihood estimators under the Cox model using modern empirical processes theory. We apply the proposed methods to a prevalent cohort medical study. Supplemental materials are available online.

Jing Qin is Mathematical Statistician, National Institution of Allergy and Infectious Diseases, Bethesda, MD 20817 (jingqin@niaid.nih.gov); Jing Ning is Assistant Professor, Division of Biostatistics, The University of Texas, Health Science Center at Houston, TX 77030 (jing.ning@uth.tmc.edu); Hao Liu is Assistant Professor, Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas 77030 (haol@bcm.edu); and Yu Shen is Professor, Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX 77030 (yshen@mdanderson.org).

B Supplemental Materials

Jumps for NPMLE of baseline cumulative hazard function for length-biased data: A detail description on why the NPMLE has jumps at both censored and uncensored times for length-biased data.

Keywords

Cox regression model; EM algorithm; Increasing failure rate; Non-parametric likelihood; Profile likelihood; Right-censored data

1. INTRODUCTION

When the observed failure times are not randomly selected from the target population of interest but with probability proportional to their underlying length, we have length-biased time-to-event data. Length-biased data are naturally encountered in applications of renewal processes (Cox and Miller, 1977; Vardi, 1982; Dewanji and Kalbfleisch, 1987; Vardi, 1989), industrial applications (Kvam, 2008), etiologic studies (Simon, 1980), genome-wide linkage studies (Terwilliger et al., 1997), epidemiologic cohort studies (Keiding, 1991; Gail and Benichou, 2000; Gordis, 2000; Sansgiry and Akman, 2000; Scheike and Keiding, 2006), cancer prevention trials (Zelen and Feinleib, 1969; Zelen, 2004), and studies of labor economy (Lancaster, 1990; McClean and Devine, 1995; De Uña Álvarez et al., 2003). In observational studies, a prevalent cohort design that draws samples from individuals with a condition or disease at the time of enrollment is generally more efficient and practical. The recruited patients who have already experienced an initiating event are followed prospectively for the failure event (e.g. disease progression or death) or are right censored. Under this sampling design, individuals with longer survival times measured from the onset of the disease are more likely to be included in the cohort, whereas those with shorter survival times are selectively excluded. Length-biased sampling thereby manifests in the observations, because the “observed” time intervals from initiation to failure within the prevalent cohort tend to be longer than those arising from the underlying distribution of the general population. How to properly adjust for potential selection bias in analyzing length-biased data has been a longstanding statistical problem. Although we use a prevalent cohort study in medical applications here to illustrate length-biased data, it is apparent that the issues caused by biased sampling are common in many potential applications and sampling designs.

In a seminal paper, Vardi (1989) described the multiplicative censorship model, which connected four well-investigated statistical problems: A. Estimating a non-parametric distribution function under multiplicative censoring, B. Estimating the underlying distribution in renewal processes, C. Solving a nonparametric deconvolution problem, and D. Estimating a monotone decreasing density function. Vardi (1989) presented problems A and C, which have a natural connection with the measurement error problem and inverse problem discussed by van der Vaart and Wellner (1992) and Bickel and Ritov (1994). Most importantly, Vardi (1989) and Wang (1991) showed that the nonparametric maximum likelihood estimation (NPMLE) of the survival distribution under multiplicative censoring (problem B) is equivalent to the nonparametric estimation for survival distribution of the observed length-biased data. The large sample properties of the corresponding NPMLE is established in Asgharian et al. (2002), and the asymptotic efficiency of the NPMLE follows from Asgharian and Wolfson (2005) and van der Vaart (1998, Theorem 25.47). In this paper we explore the potential to extend the approach of Vardi (1989) to nonparametric estimations in more general settings and to semiparametric regression models.

The Cox proportional hazards model, the most popular semiparametric model for regression analysis of traditional survival data, assumes a nonparametric baseline hazard function and a regression function of the covariates (Cox, 1972, 1975). Only limited literature exists that describes modeling risk factors on the distribution of the underlying population when observed failure times are subject to length bias. Recently, Tsai (2009) generalized the

pseudo-partial likelihood approach of Wang (1996) to model right-censored length-biased data. Qin and Shen (2010) proposed inverse weighted estimating equation approaches for right-censored length-biased data under the proportional hazards model. These approaches do not provide a straightforward way to analyze length-biased data if the censoring time depends on the covariates, and may not yield efficient estimators. For traditional survival data, Zeng and Lin (2007) demonstrated that estimating equations approaches under either the semiparametric Cox model or the transformation models are less efficient than the profile maximum likelihood estimation approach. (For related works on the profile likelihood for traditional survival data, see Nielsen et al. (1992); Klein (1992); Murphy (1994, 1995); Murphy and van der Vaart (2000); Zeng et al. (2005); Zeng and Lin (2007).) For right-censored length-biased data, we expect the similar efficiency advantage of the maximum likelihood estimation (MLE) method, and the robustness of the method to various assumptions of the censoring distribution.

Implementing the profile likelihood method is much more challenging when working with right-censored length-biased data compared to traditional survival data. One significant difference is that the full profile likelihoods have positive support on both censored and failure time points for length-biased data in contrast to the MLEs for traditional survival data and the conditional likelihood estimates for length-biased data. We propose new expectation-maximization (EM) algorithms for the maximum likelihood estimation of the nonparametric and semiparametric Cox regression models for right-censored length-biased data. One new aspect of our method is that we derive the likelihood for the unobserved (i.e. left-truncated) subpopulation given the observed length-biased data in the full likelihood, which serves as the missing data mechanism in the EM algorithm. In contrast to the EM algorithm of Vardi (1989), which estimates the underlying distribution function via estimation of the biased distribution, our EM algorithm directly estimates the target unbiased distribution function. As a result, any model and parameter constraints for the target distribution function can be directly imposed.

The rest of the paper is organized as follows. In Section 2, we introduce a new EM algorithm for the nonparametric estimation of the target distribution given length-biased data. In Section 3, we apply the new EM algorithm to estimate a distribution function with an increasing failure rate constraint. In Section 4, we propose the maximum semiparametric likelihood estimation under the Cox proportional hazards model and derive the large sample properties for length-biased data. We provide a convenient profile estimation approach based on the EM algorithm, with which the standard software for Cox regression can be adapted for right-censored length-biased data. We describe our simulation studies in Section 5 and the application of our method to a data example in Section 6. Section 7 contains some concluding remarks.

2. A NEW EM ALGORITHM FOR ESTIMATING NONPARAMETRIC SURVIVAL FUNCTION

Consider a prevalent cohort study in which the subjects are diagnosed with a disease and are at risk for a failure event. Let T , be the duration from the disease onset to failure with the unbiased density function $f(t) = dF(t)/dt$ and survival function $S(t)$. The observed data include the backward recurrence time A (from disease onset to the study entry), forward recurrence time V (from the study entry to failure), and length-biased time $T = A+V$. Based on the renewal theory (Vardi, 1982, 1989; Lancaster, 1990, Chapter 3), the joint distribution of (A, V) is

$$\frac{f(a+v)}{\mu}, a, v > 0, \text{ where } \mu = \int t f(t) dt.$$

When the prevalent cohort is followed prospectively, V is subject to right censoring. The censoring time, denoted by C , is measured from the study entry. Let $\delta = I(V < C)$ be the censoring indicator and assume that (A, V) is independent of C . Let $X = \min(A + V, A + C)$. Denote the observed data as (X_i, A_i, δ_i) , $i = 1, 2, \dots, n$. The density function of the observed biased T is defined as $g(y) = dG(y)/dy$, where $dG(y) = ydF(y)/\mu$, and the survival function of $\tilde{T}$ is

$$S(t) = \int_t^\infty dF(u) = \mu \int_t^\infty u^{-1} dG(u).$$

Therefore, the likelihood for the observed data (X_i, A_i, δ_i) is proportional to

$$\prod_{i=1}^n \frac{f^{\delta_i}(X_i) S^{1-\delta_i}(X_i)}{\mu} \propto \prod_{i=1}^n [dG(X_i)]^{\delta_i} \prod_{i=1}^n \left[\int_{x \geq X_i} x^{-1} dG(x) \right]^{1-\delta_i}. \quad (1)$$

Vardi (1989) proposed an EM-algorithm for the NPMLE of G . Using the relationship between G and F , $dF(t) = t^{-1}dG(t)/\int t^{-1}dG(t)$, the NPMLE for F can be derived. However, it is often difficult to impose constraints on F when F is estimated from the NPMLE of G , because the constraints on F may not be easily translated to the constraints on G .

As demonstrated in Vardi (1989), to maximize (1) it is sufficient to consider the discrete version of distribution F , i.e., $p(T = t_i) = p_i$, nonparametrically on the point masses at

$$t_1 < t_2 < \dots < t_k,$$

where $t_1, \dots, t_k$ are the ordered unique failure and censoring times for $\{X_1, \dots, X_n\}$, $k \leq n$. In principle, the length-biased observations (A, T) can be equivalently generated from a truncation model with

$$A \sim U(0, \hat{\tau}), \quad \tilde{T} \sim F, \text{ on } (0, \hat{\tau}), \quad (2)$$

where $\hat{\tau} = t_{k \wedge} A$ and $\tilde{T}$ are independent, $dF(t_i) = p_i$ and $\sum_{i=1}^k p_i = 1$, and (A, T) is observed if and only if $T \geq A$. The probability of observing a length-biased observation under this setting is $\pi = P(\tilde{T} \geq A) = E(\tilde{T})/\hat{\tau}$.

We propose an EM algorithm with a different missing mechanism to directly estimate the target distribution, F . For a cohort subject to left truncation, the biased samples on n subjects denoted by $\mathbf{O} = \{(X_1, \delta_1, A_1), \dots, (X_n, \delta_n, A_n), A_i \leq X_i, i = 1, \dots, n\}$, are observed, whereas the data on m subjects are left truncated. Here the latent left-truncated data are denoted by $\mathbf{O}^* = \{(T_1^*, A_1^*), \dots, (T_m^*, A_m^*), A_i^* > T_i^*, i = 1, 2, \dots, m\}$. The random integer m then follows a negative binomial distribution with parameter π . The probability mass function of m is

$$\binom{m+n-1}{m} (1-\pi)^m \pi^n, m=0, 1, 2, \dots \text{ and } E(m|\mathbf{O})=n(1-\pi)/\pi.$$

Following the principle of the EM algorithm, we think of $\{\mathbf{O}, \mathbf{O}^*\}$ as the ‘complete data’, and consider that pseudo missing data also referred to as “ghosts” data in Turnbull (1976)

are $\mathbf{O}^* = \{(T_1^*, A_1^*), \dots, (T_m^*, A_m^*), m\}$ and the observed ‘incomplete data’ are $\mathbf{O}$. We derive the full likelihood including the component of the truncated observations. The log-likelihood based on the complete data $\{\mathbf{O}, \mathbf{O}^*\}$ is

$$\sum_{j=1}^k \left[\sum_{i=1}^n I(T_i=t_j) + \sum_{i=1}^m I(T_i^*=t_j) \right] \log p_j, \quad (3)$$

where $T_i \geq A_i, i = 1, 2, \dots, n$ and $T_l^* < A_l^*, l = 1, \dots, m$. Then conditional on the observed data,

$$\begin{aligned} E \left[\sum_{i=1}^n I(T_i=t_j) | \mathbf{O} \right] &= \sum_{i=1}^n \left\{ \delta_i I(T_i=t_j) + (1-\delta_i) P(T_i=t_j | T_i \geq A_i, T_i \geq X_i) \right\} \\ &= \sum_{i=1}^n \left[\delta_i I(X_i=t_j) + (1-\delta_i) \frac{I(X_i \leq t_j) p_j}{\int_{s \geq X_i} f(s) ds} \right], \end{aligned}$$

because $(1-\delta_i) P(T_i=t_j | T_i \geq A_i, T_i \geq X_i) = (1-\delta_i) P(T=t_j | T \geq X_i) = (1-\delta_i) I(X_i \leq t_j) p_j / S(X_i)$. Conditional on the observed data $\mathbf{O}$, the expectation for the missing left-truncated data can be expressed as

$$E \left\{ E \left[\sum_{i=1}^m I(T_i^*=t_j) | m \right] | \mathbf{O}, T^* < A^* \right\}.$$

Under the truncation model specified in (2),

$$EI(T^*=t_j | \mathbf{O}, T^* < A^*) = Pr(T^*=t_j, A^* > t_j) / Pr(T^* < A^*) = \frac{p_j(1-t_j/\tau)}{1-\pi}.$$

This together with $E(m | \mathbf{O}) = n(1-\pi)/\pi$,

$$E \left\{ E \left[\sum_{i=1}^m I(T_i^*=t_j) | m \right] | \mathbf{O}, T^* < A^* \right\} = \frac{n(1-\pi)}{\pi} \frac{(1-t_j/\tau) p_j}{1-\pi} = \frac{n}{\pi} (1-t_j/\tau) p_j.$$

Subject to $\sum_{i=1}^k p_i = 1$ and $p_i \geq 0$, we maximize the expected complete-data log-likelihood conditional on the observed data via the EM algorithm,

$$\ell_E(\mathbf{p}) = \sum_{j=1}^k w_j \log p_j, \quad (4)$$

where $\mathbf{p} = (p_1, \dots, p_k)$, and

$$w_j = \sum_{i=1}^n \left[\delta_i I(X_i = t_j) + (1 - \delta_i) \frac{p_j I(X_i \leq t_j)}{\sum_{j=1}^k p_j I(X_i \leq t_j)} \right] + \frac{n}{\pi} (1 - t_j / \tau) p_j.$$

By simple algebra, $\sum_{j=1}^k w_j = n + n(1 - \pi)/\pi = n/\pi$. The following iterative EM algorithm can be used to solve $\hat{p}_j$ for $j = 1, \dots, k$.

Step 1

Select an arbitrary $p_j^{(0)}$ satisfying $\sum_{j=1}^k p_j^{(0)} = 1$, $p_j^{(0)} \geq 0$.

Step 2

Solve $p_j^{(1)}$ by maximizing (4), so that we replace $p_j^{(0)}$ with

$$\hat{p}_j^{(1)} = \frac{\hat{\pi}^{(0)}}{n} \left\{ \sum_{i=1}^n \left[\delta_i I(X_i = t_j) + (1 - \delta_i) \frac{\hat{p}_j^{(0)} I(X_i \leq t_j)}{\sum_{j=1}^k \hat{p}_j^{(0)} I(X_i \leq t_j)} \right] + \frac{n}{\hat{\pi}^{(0)}} (1 - \frac{t_j}{\tau}) \hat{p}_j^{(0)} \right\}, \quad (5)$$

$$\text{where } \hat{\pi}^{(0)} = \sum_{j=1}^k t_j \hat{p}_j^{(0)} / \tau.$$

With a given convergence criterion, we can solve p_j iteratively. Let $\hat{p}_j$ denote the MLE of p_j ,

$j = 1, \dots, k$, the NPMLE $\hat{F}(t) = \sum_{j=1}^k \hat{p}_j I(t_j \leq t)$, $\hat{\pi} = \int t d\hat{F}(t) / \tau$, $\hat{\pi} = \int t d\hat{F}(t) / \tau$, and

$$Q_1^n(t) = \frac{1}{n_1} \sum_{i=1}^n \delta_i I(X_i \leq t), \quad Q_0^n(t) = \frac{1}{n_0} \sum_{i=1}^n (1 - \delta_i) I(X_i \leq t),$$

where $n_1 = \sum_{i=1}^n \delta_i$ and $n_0 = \sum_{i=1}^n (1 - \delta_i)$. Thus, the limiting form of (5) is

$$d\hat{F}(t) = \hat{\pi} n_1 n^{-1} dQ_1^n(t) + \hat{\pi} n_0 n^{-1} d\hat{F}(t) \int_0^t \frac{dQ_0^n(s)}{1 - \hat{F}(s)} + (1 - t/\tau) d\hat{F}(t). \quad (6)$$

Remark 1

In contrast to the NPMLE for traditional survival analyses, which has jumps only at the observed failure time points, the proposed NPMLE for length-biased data has jumps at all observed but unique points including censored times, similar to that of Vardi (1989).

Remark 2

Equation (6) for the constructed EM algorithm with the unbiased distribution function F is equivalent to that for Vardi's EM algorithm based on a 'multiplicative-censorship' model

with the biased distribution function G . Denoting $d\hat{G}(t) = t d\hat{F}(t)/\hat{\mu}$, where $\hat{\mu} = \hat{\pi}\hat{\tau}$, we re-express equation (6) as an equation of $\hat{G}$,

$$d\hat{G}(t) = n_1 n^{-1} dQ_1^n(t) + n_0 n^{-1} \frac{d\hat{G}(t)}{t} \int_0^t \left[\int_{r \geq s} r^{-1} d\hat{G}(r) \right]^{-1} dQ_0^n(s),$$

which is the same equation derived by Vardi (1989) and Vardi and Zhang (1992). The advantage of the new EM algorithm is that it directly estimates the target distribution function of the unbiased data, which allows one to directly impose constraints on F . This advantage will be further elucidated in the next two sections.

Remark 3

The ‘missing’ data (i.e. left-truncated failure times), $\{T_1^*, \dots, T_m^*\}$ are assumed not subject to right censoring. It is clear that whether T^* is subject to right censoring or not is irrelevant in the derivation of the above EM algorithm.

Remark 4

The development of the methods and large sample properties is focused on $[0, \tau]$ throughout the paper, where τ is a finite upper bound of the support of the population survival times, and $\Lambda(\tau) < \infty$. In practice, τ can be estimated by $t_k = \max_{i=1}^n X_i$. We prove (in Appendix A.2) the following lemma that $\hat{\tau} \equiv t_k \rightarrow \tau$ in probability, and that the convergence rate is faster than $n^{1/2}$.

Lemma 1—Suppose that $E(C) > 0$ and $\tau < \infty$. Then for $1 > \eta > 1/2$, $n^\eta(\hat{\tau} - \tau) = o_p(1)$.

3. NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION WITH INCREASING FAILURE RATE

In some applications, it is known or assumed that the survival function for the target population has an increasing failure rate (Barlow and Proschan, 1975; Padgett and Wei, 1980; Tsai, 1988). The maximum likelihood estimation of a distribution function with an increasing failure rate was derived for traditional right-censored data by Padgett and Wei (1980), and for left-truncated and right-censored data by Tsai (1988). Using the same notation as in Section 2, the observed right-censored length-biased data are denoted by (X, A, δ) . Let $\lambda(t)$ denote the hazard function for the target cumulative density function F . Let $z_1 < \dots < z_{k^*}$ denote the distinct ordered failure times $\{X_1, \dots, X_n\}$. Let the size of the risk set at time x be denoted by $R(x) = \sum_{i=1}^n I(A_i \leq x \leq X_i)$ and the size of failure at time x be denoted by $d(x) = \sum_{i=1}^n I(X_i = x, \delta_i = 1)$. Under the increasing failure rate constraint, Tsai (1988) proposed a maximum conditional likelihood estimator of λ , conditional on the truncation time A ,

$$\hat{\lambda}(y) = \begin{cases} 0 & y < z_1, \\ \hat{\lambda}_j & z_j \leq y < z_{j+1}; j = 1, 2, \dots, k^* - 1, \\ \hat{\lambda}_{k^*} & z_{k^*} \leq y \end{cases}$$

where

$$\widehat{\lambda}_j = \max_{1 \leq r \leq j} \min_{j \leq s \leq k-1} \left\{ \sum_{i=r}^s d(z_i) / \sum_{i=r}^s R(z_i) (z_{i+1} - z_i) \right\}.$$

By applying the new EM algorithm, we consider a full likelihood estimation of the hazard function for the target population. Define $\lambda(t_j) = \lambda_j$; p_j can be expressed as $\lambda_j \exp(-\int_0^{t_j} \lambda(s) ds)$; thus the expected complete-data log-likelihood function in (4) is

$$\ell_E(\lambda) = \sum_{i=1}^k w_i \log \lambda_i - \sum_{i=1}^k w_i \sum_{j=1}^i \int_{t_{j-1}}^{t_j} \lambda(t) dt, \quad (7)$$

where $\lambda = (\lambda_1, \dots, \lambda_k)$, and $t_1 < \dots < t_k$ is defined in §2. Because the hazard function $\lambda(\cdot)$ increases with time,

$$\ell_E(\lambda) \leq \sum_{i=1}^k w_i \log \lambda_i - \sum_{j=1}^k \left[\sum_{i=j}^k w_i (t_j - t_{j-1}) \right] \lambda_{j-1},$$

where $\lambda_0 = 0$. Taking a partial derivative of λ_j to the right side of the above inequality, we have

$$\frac{w_j}{\lambda_j} - \sum_{i=j+1}^k w_i (t_{j+1} - t_j) = 0. \quad (8)$$

Using arguments similar to those of Marshall and Proschan (1965) and Padgett and Wei (1980), the solution to equation (8) also maximizes the expected log-likelihood $\ell_E(\lambda)$ defined in (7),

$$\lambda_j = \frac{w_j}{\sum_{i=j+1}^k w_i (t_{j+1} - t_j)}.$$

Applying the pool-adjacent-violators algorithm, we can then achieve monotonicity for the NPMLE of $\lambda(\cdot)$,

$$\widehat{\lambda}(x) = \begin{cases} 0 & x < t_1 \\ \widehat{\lambda}_j & t_j \leq x < t_{j+1}; \quad j=1, 2, \dots, k-1, \\ \widehat{\lambda}_k & x = t_k, \end{cases}$$

where $\widehat{\lambda}_j = \max_{1 \leq r \leq j} \min_{j \leq s \leq k-1} \left\{ \sum_{i=r}^s w_i / \sum_{i=r}^s [\sum_{l=i+1}^k w_l] (t_{i+1} - t_i) \right\}$. Although the formula for the proposed NPMLE of the monotone hazard function bears some similarity to that of Tsai (1988), the full likelihood approach is essentially different from the conditional likelihood approach of Tsai (1988), where the estimated function has jumps only at distinct

failure time points. By using the information of the left-truncated data in the full likelihood function, the NPMLLE is expected to be more efficient and smoother than the maximum conditional likelihood estimate. We will further compare the two approaches in empirical studies.

As the hazard function $\lambda(t)$ is increasing on $t \in [0, \tau]$, the corresponding cumulative hazard function $\Lambda(t)$ is convex. Let $\hat{\Lambda}_n(\cdot)$ denote the estimator obtained by the EM algorithm together with the pool-adjacent-violators algorithm. Then $\hat{\Lambda}_n(\cdot)$ is the greatest convex minorant of $\Lambda_n(\cdot)$, where $\Lambda_n(\cdot)$ is the NPMLLE $\Lambda(\cdot)$ when there is no constraint on its shape. The strong consistent results of $\hat{\Lambda}_n(\cdot)$ uniformly on $[0, \tau]$ can be easily derived from the uniform consistency of its survival function, established by Asgharian and Wolfson (2005). The consistency of $\hat{\Lambda}_n$ can be inferred, because the pool-adjacent-violator algorithm leads to a continuous map for $\Lambda_n(\cdot)$ to $\hat{\Lambda}_n(\cdot)$. The technical details are provided in the Appendix A.3.

4. MLE UNDER COX REGRESSION MODEL

4.1 Full Likelihood and Score Functions

Since the cornerstone work of Cox (1972, 1975), the proportional hazards model has become the standard regression model for analyzing traditional right-censored survival data. Specifically, the covariate-specific hazard function is specified as

$$\lambda_{\mathbf{Z}}(t) = \lambda(t) \exp(\boldsymbol{\beta}' \mathbf{Z})$$

where $\mathbf{Z}$ is a covariate vector and the baseline hazard function $\lambda(t)$ is not specified parametrically. Breslow (1972) showed that by inserting an estimator (Breslow's estimator) of the hazard function with a fixed $\boldsymbol{\beta}$ into the full likelihood, the profile likelihood for $\boldsymbol{\beta}$ is reduced to Cox's partial likelihood for $\boldsymbol{\beta}$. Later, Kalbfleisch and Prentice (1973) and Andersen et al. (1992) proved that the rank-based likelihood method is also equivalent to the partial likelihood method. When survival data are subject to biased sampling, neither the Cox's partial likelihood approach nor Kalbfleisch and Prentice's rank-based likelihood method can be directly applied. This is because the observed biased data do not follow the proportional hazards model that is assumed for unbiased data from the target population; and because the rank-based likelihood method is not applicable, due to the dependency of the length-biased data on the magnitude of the length.

The density function of an unbiased $\tilde{T}$ given $\mathbf{Z}$ is denoted by $f(t | \mathbf{Z})$ and the corresponding survival function by $S(t | \mathbf{Z})$. For random but length-biased samples of n subjects, the observed data consist of $\{\boldsymbol{\phi}_i \equiv (A_i, X_i, \delta_i, \mathbf{Z}_i), i = 1, \dots, n\}$, which are n i.i.d. copies of $\boldsymbol{\phi} \equiv (A, X, \delta, \mathbf{Z})$. The full likelihood function of the observed data is proportional to

$$L_n = \prod_{i=1}^n \frac{f^{\delta_i}(X_i | \mathbf{Z}_i) S^{(1-\delta_i)}(X_i | \mathbf{Z}_i)}{\mu_{\boldsymbol{\beta}, \Lambda}(\mathbf{Z}_i)} \quad (9)$$

where $\mu_{\boldsymbol{\beta}, \Lambda}(\mathbf{Z}_i) = \int_0^\infty t f(t | \mathbf{Z}_i) dt = \int_0^\infty S(t | \mathbf{Z}_i) dt$. The identifiability of the model can be established, similar to the case for the Cox model under traditional survival data, where the identifiability has been established (Elbers and Ridder, 1982). By decomposing the full likelihood to the product of the conditional likelihood of X given A and the marginal likelihood of A , we have

$$L_n = \left[\prod_{i=1}^n \frac{f^{\delta_i}(X_i|Z_i) S^{1-\delta_i}(X_i|Z_i)}{S(A_i|Z_i)} \right] \left[\prod_{i=1}^n \frac{S(A_i|Z_i)}{\mu_{\beta, \Lambda}(Z_i)} \right].$$

Although the estimating equation derived from the likelihood conditional on A (the first component in L_n) shares the same advantage of Cox's partial likelihood by canceling the baseline hazard function (Wang et al., 1993; Kalbfleisch and Lawless, 1991), the conditional likelihood approach is generally less efficient than the full likelihood approach.

Using the notation of counting process, we denote $N_i(t) = I(X_i \leq t)\delta_i$, $Y_i(t) = I(X_i \geq t)$ for $i = 1, \dots, n$. The log-likelihood function of (9) can be expressed as

$$\ell_n(\beta, \Lambda) = n^{-1} \sum_{i=1}^n \left\{ \int_0^{\tau} (\beta' Z_i + \log \lambda(t)) dN_i(t) - \Lambda_{Z_i}(X_i) - \log \mu_{\beta, \Lambda}(Z_i) \right\}, \quad (10)$$

where $\Lambda(t) = \int_0^t \lambda(s) ds$, and $\Lambda_{Z_i}(t) = \int_0^t \exp(\beta' Z_i) d\Lambda(s)$. The estimation for MLE of β and the infinite dimensional parameter Λ can be computationally intractable if directly maximizing (10) or solving its score equations. We thereby propose an alternative computational approach, which is a generalization of the EM algorithm for NPMLE discussed in Section 2 under the Cox proportional hazards model.

The semiparametric MLE for the baseline hazard function Λ is obtained by maximizing the likelihood over the set of piece-wise constant functions. Of note, the estimator can have jumps at both censored and uncensored times by observing that the likelihood function achieves its maximum for the hazard function with jumps on $\{t_1, \dots, t_k\}$, where $t_1 < \dots < t_k$ denotes distinct failure and censored time points. Similar to the argument in (Vardi, 1989, page 754), for any estimator of $\Lambda(t)$ that jumps outside of the event times $\{t_1, \dots, t_k\}$ one can find a greater likelihood with jumps on $\{t_1, \dots, t_k\}$ only. A detail explanation is in Supplemental Materials available online.

4.2. MLE and EM Algorithm

For $i = 1, \dots, n$, let T_{ij}^* , $j = 1, 2, \dots, m_i$ be the truncated latent data corresponding to covariate Z_i . We develop the EM algorithm based on the discretized version of $\Lambda(u) = \sum_{u \geq t_j} \lambda_j$, where λ_j is the positive jump at time t_j for $j = 1, \dots, k$, and $\lambda = (\lambda_1, \dots, \lambda_k)$. For notational convenience, denote $f_i(t) = dF(t | Z_i)$. The log-likelihood based on the complete data is then

$$\sum_{j=1}^k \sum_{i=1}^n \left[I(T_i = t_j) + \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \right] \log f_i(t_j)$$

Conditional on the observed data relative to the i th subject, $\mathcal{O}_i = \{X_i, A_i, \delta_i, Z_i\}$, we obtain the expectation that

$$\begin{aligned} w_{ij} &= E \left[I(T_i = t_j) + \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \middle| \mathcal{O}_i \right] \\ &= \delta_i I(X_i = t_j) + (1 - \delta_i) \frac{p_{ij} I(X_i \leq t_j)}{\sum_{j=1}^k p_{ij} I(X_i \leq t_j)} + \frac{\widehat{\tau}}{\mu_i} (1 - t_j / \widehat{\tau}) p_{ij} \end{aligned} \quad (11)$$

where

$$p_{ij} = \lambda_j \exp(\beta' \mathbf{Z}_i) \exp \left\{ - \sum_{l=1}^j \lambda_l \exp(\beta' \mathbf{Z}_i) \right\}, \text{ and } \mu_i = \sum_{j=1}^k t_j p_{ij}.$$

Thus, the expected complete-data log-likelihood function conditional on the observed data is as follows:

$$\begin{aligned} \ell_E(\beta, \lambda) &= \sum_{i=1}^n \sum_{j=1}^k w_{ij} \log f_i(t_j) \\ &= \sum_{j=1}^k w_{+j} \log \lambda_j + \sum_{i=1}^n w_{i+} \beta' \mathbf{Z}_i - \sum_{l=1}^k \sum_{j=l}^k \sum_{i=1}^n w_{ij} \exp(\beta' \mathbf{Z}_i) \lambda_l, \end{aligned}$$

where $w_{+j} = \sum_{i=1}^n w_{ij}$, and $w_{i+} = \sum_{j=1}^k w_{ij}$. In the M-step, we maximize the expected complete-data log-likelihood function conditional on the observed data with respect to the baseline hazard function at t_j , for $j = 1, \dots, k$,

$$\frac{\partial \ell_E(\beta, \lambda)}{\partial \lambda_j} = \frac{w_{+j}}{\lambda_j} - \sum_{l=j}^k \sum_{i=1}^n w_{il} \exp(\beta' \mathbf{Z}_i) = 0,$$

which leads to a closed form of λ_j as a function of β , denoted by

$$\lambda_j(\beta) = \frac{w_{+j}}{\sum_{l=j}^k \sum_{i=1}^n w_{il} \exp(\beta' \mathbf{Z}_i)}. \quad (12)$$

Here, λ_j is the maximizer of the M-step. Next, we maximize the expected complete-data log-likelihood function with respect to β

$$\frac{\partial \ell_E(\beta, \lambda)}{\partial \beta} = \sum_{i=1}^n w_{i+} \mathbf{Z}_i - \sum_{l=1}^k \sum_{j=l}^k \sum_{i=1}^n w_{ij} \mathbf{Z}_i \exp(\beta' \mathbf{Z}_i) \lambda_l. \quad (13)$$

By inserting $\lambda_j(\beta)$ of (12) into the equation (13), β can be solved from the following equation,

$$\sum_{i=1}^n w_{i+} \mathbf{Z}_i - \sum_{l=1}^k w_{+l} \left\{ \frac{\sum_{i=1}^n \sum_{j=l}^k w_{ij} \mathbf{Z}_i \exp(\beta' \mathbf{Z}_i)}{\sum_{i=1}^n \sum_{j=l}^k w_{ij} \exp(\beta' \mathbf{Z}_i)} \right\} = 0, \quad (14)$$

which is equivalent to maximizing the complete-data profile likelihood function for λ . With the estimated λ_j ($j = 1, \dots, k$) and β , one can update the expectation of the likelihood via w_{ij} in (11) and repeat the M-step until the estimators of β and λ_j ($j = 1, \dots, k$) converge.

At the M-step, the estimating equation (14) reveals that we may use the existing software for conventional right-censored data to estimate the covariate coefficient β under the Cox

proportional hazards model. To simplify the description, consider a model with one covariate $\mathbf{Z}$. First we need to create a vector with a length of nk for the weight function defined by $\mathbf{W}_{nk} = (w_{11}, \dots, w_{1k}, w_{21}, \dots, w_{2k}, \dots, w_{n1}, \dots, w_{nk})$, which is estimated at the E-step. The corresponding failure time data and covariate vectors are constructed with the same length as $\mathbf{W}_{nk}$, $\mathbf{T}_{nk} = (t_1, \dots, t_k, \dots, t_1, \dots, t_k)$ and $\mathbf{Z}_{nk} = (\mathbf{Z}_1, \dots, \mathbf{Z}_1, \dots, \mathbf{Z}_n, \dots, \mathbf{Z}_n)$, respectively. By using the function “coxph” in S-PLUS (or R) with the “weights” option, we obtain the estimator of β at the M-step from

$$> \text{coxph}(\text{Surv}(\mathbf{T}_{nk}, \Delta) \sim \mathbf{Z}_{nk}, \text{weights}=\mathbf{W}_{nk}),$$

where the censoring indicator, $\Delta = (1, \dots, 1)$, is an identity vector of length nk .

Note that the algorithm computes β and λ iteratively through the EM steps. The value of the complete log-likelihood $\ell_E(\beta, \lambda)$ increases with each EM step. More specifically, our EM algorithm falls in the general scheme of the ECM algorithm, a variation of EM methods proposed by Meng and Rubin (1993). The convergence of our EM algorithm to the local maximizer is guaranteed by the same conditions that ensure the convergence of the ECM algorithm, as proved in Meng and Rubin (1993). The uniqueness of the NPMLE is guaranteed by the Assumption 5 in Appendix A.1.

4.3. Asymptotic Properties

In this section, we establish the strong consistency and asymptotic normality of the MLE under the regularity conditions list in Appendix A.1. For asymptotic proof, we denote the MLE by $(\hat{\beta}_n, \hat{\Lambda}_n)$, and let (β_0, Λ_0) be the true value. In Appendix A.4, we prove the strong consistency by the classical Kullback-Leibler information approach, which has been successfully applied for NPMLE in traditional survival analysis (Murphy, 1994; Parner, 1998).

Theorem 1—Under the regularity conditions listed in Appendix A.1, the MLE $(\hat{\beta}_n, \hat{\Lambda}_n)$ are consistent: $\hat{\beta}_n$ converges to β_0 , and $\hat{\Lambda}_n(t)$ converges to $\Lambda_0(t)$ almost surely and uniformly in t for $t \in [0, \tau]$ as $n \rightarrow \infty$.

The computation for the MLE of Λ is based on the discretized version $\hat{\Lambda}_n(t) = \sum_{t \geq t_j} \hat{\lambda}_j$. The existence and uniqueness of the NPMLE can be proved based on the log-likelihood function $\ell_n(\beta, \lambda)$, in terms of $\{\beta, \lambda\}$ where $\lambda \equiv \{\lambda_1, \dots, \lambda_k\}$, and

$$\ell_n(\beta, \lambda) = \sum_{i=1}^n \left[\int_0^\tau \beta' Z_i dN_i(t) + \sum_{l=1}^k \delta_i \log \lambda_l \right] - \sum_{l=1}^k \lambda_l \sum_{i=1}^n e^{\beta' Z_i} 1(t_l \leq t_i) - \sum_{i=1}^n \log \left[\int_0^\tau \exp \left(-e^{\beta' Z_i} \sum_{l=1}^k \lambda_l 1(t_l \leq s) \right) ds \right].$$

Let $\hat{\lambda}(\cdot, \beta)$ be the maximizer of $\ell_n(\beta, \lambda)$ for given β . The existence and uniqueness of NPMLE are guaranteed by Assumption 5: the information matrix of the profile likelihood evaluated at the true value β_0 is positive definite.

Next, we will apply the Z-theorem for the infinite-dimensional estimating equations to prove the weak convergence of the estimators (van der Vaart and Wellner, 1996, Theorem 3.3.1, p. 310). The score equation of β is

$$U_{1n}(\beta, \Lambda) \equiv \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^\tau Z_i dN_i(u) - \int_0^\tau Y_i(u) Z_i e^{\beta' Z_i} d\Lambda(u) + \int_0^\tau Z_i \left(\int_u^\tau S(v|Z_i) dv \right) e^{\beta' Z_i} \Lambda(u) / \mu_{\beta, \Lambda}(Z_i) \right\} \quad (16)$$

To obtain the MLE of $\Lambda(\cdot)$, consider a submodel defined by $d\Lambda_\eta = (1 + \eta h)d\Lambda$, where h is a bounded and integrable function. Taking the derivative of $\ell_n(\beta, \Lambda_\eta)$ with respect to η , evaluating it at $\eta = 0$, and setting $h(\cdot) = 1(\cdot \leq t)$, we have the score equation for Λ

$$U_{2n}(t, \beta, \Lambda) \equiv \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^t dN_i(u) - \int_0^t Y_i(u) e^{\beta' Z_i} d\Lambda(u) + \int_0^t \left(\int_u^\tau S(v|Z_i) dv \right) e^{\beta' Z_i} d\Lambda(u) / \mu_{\beta, \Lambda}(Z_i) \right\}. \quad (17)$$

Denote the vector for score functions by $U_n(\cdot, \beta, \Lambda) \equiv \{U_{1n}(\beta, \Lambda), U_{2n}(t, \beta, \Lambda)\}$, and its expectation E_0 under the true values (β_0, Λ_0) by

$$U_0(\cdot, \beta, \Lambda) \equiv \{U_{10}(\beta, \Lambda), U_{20}(\cdot, \beta, \Lambda)\},$$

where $U_{10}(\beta, \Lambda) = E_0\{U_{1n}(\beta, \Lambda)\}$ and $U_{20}(t, \beta, \Lambda) = E_0\{U_{2n}(t, \beta, \Lambda)\}$. Both the score function U_n and its expectation U_0 are defined on the parameter set $\mathcal{B} \times \mathcal{A}$, where the set $\mathcal{B}$ is assumed to be compact in $\mathbb{R}^p$, and the set $\mathcal{A}$ consists of nondecreasing functions in the space of functions with bounded variation. Let $\hat{\psi}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$, $\psi = (\beta, \Lambda)$ and $\psi_0 = (\beta_0, \Lambda_0)$.

By the definition of MLE, $U_n(\cdot, \hat{\psi}_n) = 0$. As the true parameter ψ_0 satisfies $U_0(\cdot, \psi_0) = 0$, $\sqrt{n}\{U_0(\cdot, \hat{\psi}_n) - U_0(\cdot, \psi_0)\} = \sqrt{n}\{U_0(\cdot, \hat{\psi}_n) - U_n(\cdot, \hat{\psi}_n)\}$. In the Appendix A.5, we prove a stochastic approximation $|\sqrt{n}\{U_0(\cdot, \hat{\psi}_n) - U_n(\cdot, \hat{\psi}_n)\} - \sqrt{n}\{U_n(\cdot, \hat{\psi}_0) - U_0(\cdot, \psi_0)\}| = o_p(1)$. Denote that $\dot{U}_{\psi_0}$ is the Fréchet derivative of the map $U_0(\cdot, \psi)$ evaluated at ψ_0 . By the definition of the Fréchet derivative, $\dot{U}_{\psi_0}\{\sqrt{n}(\hat{\psi}_n - \psi_0)\} = -\sqrt{n}\{U_n(\cdot, \psi_0) - U_0(\cdot, \psi_0)\} + o_p(1)$.

The estimating function evaluated at ψ_0 , $\sqrt{n}U_n(\psi_0) = \sqrt{n}\{U_n(\cdot, \psi_0) - U_0(\cdot, \psi_0)\}$, is a sum of i.i.d. random quantities. We will prove by the empirical processes theory that $\sqrt{n}U_n(\psi_0)$ converges weakly to $\mathbb{W} = (\mathbb{W}_1, \mathbb{W}_2)$, where $\mathbb{W}_1$ is a Gaussian random vector and $\mathbb{W}_2$ is a tight Gaussian process. The covariance matrix for $\mathbb{W}_1$ is $\Sigma_{11} = E_0\{U_{1n}(\beta_0, \Lambda_0)^{\otimes 2}\}$, and the covariance between $\mathbb{W}_2(s)$ and $\mathbb{W}_2(t)$ is $\Sigma_{22}(s, t) = E_0\{U_{2n}(s, \beta_0, \Lambda_0)U_{2n}(t, \beta_0, \Lambda_0)\}$. By the Z-theorem for the infinite-dimensional estimating equations (van der Vaart and Wellner, 1996), we have

Theorem 2—Under the regularity conditions listed in the Appendix A.1, $\sqrt{n}(\hat{\psi}_n - \psi_0)$ converges weakly to a tight mean zero Gaussian process $-\dot{U}_{\psi_0}^{-1}(\mathbb{W})$.

Note that the asymptotic distribution of the sequence $\sqrt{n}(\hat{\psi}_n - \psi_0)$ is completely determined by the tightness of $\dot{U}_{\psi_0}^{-1}(\mathbb{W})$ and its marginal covariance function. We characterize the Fréchet derivative $\dot{U}_{\psi_0}$, viewed as an operator on the parameter space of $\mathcal{B} \times \mathcal{A}$. Define for $l = 0, 1, 2$,

$$\begin{aligned} \mathbf{K}_1^{(l)}(u) &= E_0 \left\{ \mathbf{Z}^{\otimes l} e^{\beta_0' \mathbf{Z}} S_0(u|\mathbf{Z}) \mu_0^{-1}(\mathbf{Z}) \int_0^u S_c(s|\mathbf{Z}) ds \right\} \\ \mathbf{K}_2^{(l)}(t, u) &= \int_u^\tau E_0 \left\{ \mathbf{Z}^{\otimes l} e^{2\beta_0' \mathbf{Z}} \mu_0^{-1}(\mathbf{Z}) S_0(v|\mathbf{Z}) \left(\Lambda_0(t \wedge v) - \Phi_0(t|\mathbf{Z}) \mu_0^{-1}(\mathbf{Z}) \right) \right\} dv, \end{aligned}$$

where $S_C(u|\mathbf{Z}) = P(C \geq u | \mathbf{Z})$, $\mu_0(\mathbf{Z}) = \mu_{\beta_0, \Lambda_0}(\mathbf{Z})$, $S_0(t|\mathbf{Z}) = \exp\{-\exp(\beta_0' \mathbf{Z}) \Lambda_0(t)\}$, and

$$\Phi_0(t|\mathbf{Z}) = \int_0^t \left(\int_u^\tau S_0(v|\mathbf{Z}) dv \right) d\Lambda_0(u).$$

By Assumption 5, the Fisher information of β for known Λ_0 is positive definite

$$\mathbf{J}_0 \equiv \left\{ \int_0^\tau \mathbf{K}_1^{(2)}(u) d\Lambda_0(u) + \int_0^\tau \mathbf{K}_2^{(2)}(\tau, u) d\Lambda_0(u) \right\}. \quad (18)$$

Then the Fréchet derivative $\dot{U}_{\psi_0}$ can be written in the following form:

$$\dot{U}_{\psi_0}(\beta, \Lambda) \equiv - \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \beta \\ \Lambda \end{pmatrix} \equiv (\sigma_{11}(\beta) + \sigma_{12}(\Lambda), \sigma_{21}(\beta)(\cdot) + \sigma_{22}(\Lambda)(\cdot)), \quad (19)$$

where

$$\begin{aligned} \sigma_{11}(\beta) &= \left\{ \int_0^\tau \mathbf{K}_1^{(2)}(u) d\Lambda_0(u) + \int_0^\tau \mathbf{K}_2^{(2)}(\tau, u) d\Lambda_0(u) \right\}' \beta \\ \sigma_{12}(\Lambda) &= \int_0^\tau \mathbf{K}_1^{(1)}(u) d\Lambda(u) + \int_0^\tau \mathbf{K}_2^{(1)}(\tau, u) d\Lambda(u) \\ \sigma_{21}(\beta)(t) &= \left\{ \int_0^t \mathbf{K}_1^{(1)}(u) d\Lambda_0(u) + \int_0^t \mathbf{K}_2^{(1)}(t, u) d\Lambda_0(u) \right\}' \beta \\ \sigma_{22}(\Lambda)(t) &= \int_0^t \mathbf{K}_1^{(0)}(u) d\Lambda(u) + \int_0^t \mathbf{K}_2^{(0)}(t, u) d\Lambda(u). \end{aligned}$$

We show the invertibility of $\dot{U}_{\psi_0}$ by translating the operator into the Fredholm integral equations of the second kind (Tricomi, 1985, Chapter 2). We prove in the appendix A.5 that the inverse of $\dot{U}_{\psi_0}$ exists, and the inverse is continuous with the following form

$$\dot{U}_{\psi_0}^{-1}(\beta, \Lambda) \equiv \begin{pmatrix} \sigma_{11}^{-1} + \sigma_{11}^{-1} \sigma_{12} \Phi^{-1} \sigma_{21} \sigma_{11}^{-1} & \sigma_{11}^{-1} \sigma_{12} \Phi^{-1} \\ \Phi^{-1} \sigma_{21} \sigma_{11}^{-1} & \Phi^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \Lambda \end{pmatrix}, \quad (20)$$

where the functional $\Phi = \sigma_{22} - \sigma_{21} \sigma_{11}^{-1} \sigma_{12}$. We shown in the appendix that Φ has an inverse $\Lambda \rightarrow \Phi^{-1}(\Lambda)$ expressed in the following form as a function of t

$$\Phi^{-1}(\Lambda)(t) = \int_0^t \frac{d\Lambda(u)}{K_1^{(0)}(u)} - \int_0^\tau \left(\int_0^t \frac{H(u, v)}{K_1^{(0)}(u)} du \right) d\Lambda(v), \quad (21)$$

where $H(u, v)$ is the solution of the following integral equation

$$H(u, v) = \frac{Q(u, v)}{K_1^{(0)}(v)} + \int_0^\tau H(u, s) \frac{Q(s, v)}{K_1^{(0)}(v)} ds, \quad (22)$$

$$Q(t, u) = \left(K_1^{(1)}(t) \lambda_0(t) + \int_0^\tau \dot{K}_2^{(1)}(t, v) d\Lambda_0(v) \right)' J_0^{-1} \left(K_1^{(1)}(u) + K_2^{(1)}(\tau, u) \right) - \dot{K}_2^{(0)}(t, u), \quad (23)$$

and with the notation $\dot{K}_2^{(l)}(t, u) = \partial K_2^{(l)}(t, u) / \partial t$ for $l = 0, 1$,

$$\dot{K}_2^{(l)}(t, u) = \int_u^\tau E_0 \left\{ Z^{\otimes l} e^{2\beta'_0 Z} \frac{S_0(v|Z)}{\mu_0^{-1}(Z)} \left(\lambda_0(t \wedge v) - \lambda_0(t) \int_t^\tau \frac{S_0(s|Z)}{\mu_0^{-1}(Z)} ds \right) \right\} dv.$$

By Theorem 2, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to a mean zero normal random vector characterized by

$$\sigma_{11}^{-1}(\mathbb{W}_1) + \sigma_{11}^{-1} \sigma_{12} \Phi^{-1} \sigma_{21} \sigma_{11}^{-1}(\mathbb{W}_1) - \sigma_{11}^{-1} \sigma_{12} \Phi^{-1}(\mathbb{W}_2), \quad (24)$$

where the Gaussian process $\Phi^{-1}(\mathbb{W}_2)$ has the following form

$$\Phi^{-1}(\mathbb{W}_2)(t) = \int_0^t \frac{d\mathbb{W}_2(u)}{K_1^{(0)}(u)} - \int_0^\tau \left(\int_0^t \frac{H(u, v)}{K_1^{(0)}(u)} du \right) d\mathbb{W}_2(v).$$

Note that the stochastic integral is well defined via the formula of integration by the parts because the functions $u \rightarrow 1/K_1^{(0)}(u)$ and $v \rightarrow \int_0^t H(u, v)/K_1^{(0)}(u) du$ are of bounded variation on $[0, \tau]$.

Additionally, the process $\sqrt{n}(\widehat{\Lambda}_n - \Lambda_0)$ converges weakly to a tight Gaussian process

$$\Phi^{-1} \sigma_{21} \sigma_{11}^{-1}(\mathbb{W}_1) + \Phi^{-1}(\mathbb{W}_2), \quad (25)$$

where the processes

$$\sigma_{21} \sigma_{11}^{-1}(\mathbb{W}_1)(t) = \left\{ \int_0^t K_1^{(1)}(u) d\Lambda_0(u) + \int_0^\tau K_2^{(1)}(t, u) d\Lambda_0(u) \right\}' J_0^{-1} \mathbb{W}_1.$$

If the baseline function Λ_0 is known, then $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to a Gaussian random variate $\sigma_{11}^{-1}(\mathbb{W}_1)$ with mean zero and the sandwich covariance matrix $J_0^{-1} \sum_{11} J_0^{-1}$. Because of the variation associated with the profile-likelihood estimator $\widehat{\Lambda}_n$, the asymptotic variance of $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ is more complicated, with the extra terms indicated in (24). The variance-covariance matrix may be estimated by its empirical plug-in version,

but the computation can be extremely complicated as it needs to solve the integral equation (22). We describe alternative methods for this computation in the following section.

4.4. Variance Estimation

Unlike the estimation of the regression coefficients, variances of the estimator of the MLE, $\hat{\beta}_n$ cannot be obtained directly from existing software such as R or SAS, because they cannot incorporate the variations caused by the profile likelihood estimator. Instead, we can use bootstrapping techniques or use the information matrix to estimate the variance of the estimators. When working with the observed full likelihood with unknown parameters (β, Λ) , the total number of parameters has the same order as the number of observed distinct times, which often yields an information matrix of a high dimension. Murphy and van der Vaart (1999) showed that the inverse of the information matrix for the profile likelihood provides valid variance estimates for the finite-dimensional parameters of interest, i.e., $\hat{\beta}_n$ under the semiparametric models. There is no general analytical formula for calculating the profile information matrix; thus we describe a numerical EM-aided differentiation approach (Murphy and van der Vaart, 1999; Chen and Little, 1999) to approximate the profile information matrix. Chen and Little (1999) also proved that the score function of the profile likelihood for the observed data is the same as the expected complete-data score function conditional on the observed data at the convergent point. Therefore, the second derivative of the log profile likelihood evaluated at the MLE $\hat{\beta}_n$ can be approximated by $\partial^2 \ell_E(\beta, \lambda(\beta)) / \partial \beta^2|_{\beta=\hat{\beta}_n}$, which is the first derivative of the expected complete-data score function conditional on the observed data and profiled over $\lambda(\beta) = (\lambda_1(\beta), \dots, \lambda_k(\beta))$ given by (12). By perturbation around the profile MLE $\hat{\beta}_n$, the information matrix for β can then be calculated as follows:

1. Perturb the l th component of $\hat{\beta}_n = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ by a small value $\varepsilon = 1/n$ in the neighborhood of $\hat{\beta}_l$ (in both directions). The perturbed estimator is denoted by $\hat{\beta}_{\varepsilon,l} = (\beta_1, \dots, \beta_l \pm \varepsilon, \dots, \beta_k)$, where $l = 1, \dots, p$.
2. Approximate the l th row of the information matrix of β by

$$\frac{1}{\varepsilon} \left\{ \frac{\partial \ell_E(\beta, \lambda(\beta))}{\partial \beta} \Big|_{\beta=\hat{\beta}_{\varepsilon,l}} \right\},$$

where the hazard function $\lambda(\beta)$ is obtained from (12) using the M-step described in Section 4.2.

When estimating the variance of $\hat{\lambda}_n$, the bootstrap approach can be used. This re-sampling approach is valid, given our established asymptotic normality for the MLE of $\hat{\beta}_n$ and $\hat{\lambda}_n$. In this case, we should obtain the variances of both $\hat{\beta}_n$ and $\hat{\lambda}_n$.

5. SIMULATIONS

We performed simulation studies to evaluate the proposed methods and the corresponding EM algorithms for two settings: nonparametric MLE with an increasing failure rate, and the profile likelihood estimators under the Cox proportional hazards model for length-biased data. We aimed to assess the small sample accuracy and precision of our estimators, and to compare their performance with those of the existing methods under each setting. Each study comprised 1000 repetitions. Sample sizes of 200 and 400 were used.

5.1. Estimating a Distribution Function with an Increasing Failure Rate

We generated independent pairs of $(A, \tilde{T})$ with failure times from a Weibull distribution $(F(t) = 1 - \exp\{-(t/\alpha_2)^{\alpha_1}\})$ with $\alpha_1 = 2$ and $\alpha_2 = 1$ and truncation times from a uniform distribution to ensure the stationarity assumption. Here, the specified Weibull distribution has an increasing failure rate. The censoring variables measured from the examination time were independently generated from uniform distributions.

Table 1 compares the performance of our proposed estimator and that of Tsai's estimator (Tsai, 1988), denoted by F_c . When $F(t)$ is greater than 0.5, both estimators achieved outstanding accuracy. In contrast, when $F(t)$ is small, both had downward biases, and the bias decreased with increasing sample sizes. As expected, the empirical standard deviations of our estimator were as much as 25% lower than those of Tsai's estimator.

5.2. Estimating Regression Coefficients Under the Cox Model

We generated unbiased failure times $(\tilde{T})$ from the proportional hazards model with two covariates, where $\beta = (\beta_1, \beta_2) = (0.5, 1)$, the binary covariate $Z_1 \sim \text{Bernoulli}(1, 0.5)$, the continuous covariate $Z_2 \sim \text{uniform}(-0.5, 0.5)$, and the baseline hazard function is t . The censoring times C were independently generated either from uniform distributions or from the specified covariate dependent distributions (see Table 2).

For a light (15%) or moderate (30%) censoring percentage, the mean estimates of the coefficients agreed well with the true parameter no matter whether the censoring distribution was dependent on or independent of the covariates. Even with heavy censoring (50%), the inferences associated with the proposed method were fairly accurate for all of the scenarios investigated: the means of the estimated standard errors were close to the empirical standard errors and the coverage of the 95% confidence intervals were reasonable, ranged from 92% to 96%.

In Table 2, we also show a comparison of the performance between the proposed MLE and the existing estimation methods for length-biased data. The estimating equation approaches of Qin and Shen (2010) are as follow:

$$\begin{aligned} EE - I: \quad & \sum_{i=1}^n \delta_i \left[Z_i - \frac{\sum_{j=1}^n I(X_j \geq X_i) \delta_j \{X_j S_c(X_j - A_j)\}^{-1} Z_j \exp(\beta' Z_j)}{\sum_{j=1}^n I(X_j \geq X_i) \delta_j \{X_j S_c(X_j - A_j)\}^{-1} \exp(\beta' Z_j)} \right] = 0, \\ EE - II: \quad & \sum_{i=1}^n \delta_i \left[Z_i - \frac{\sum_{j=1}^n I(X_j \geq X_i) \delta_j \{W_c(X_j)\}^{-1} Z_j \exp(\beta' Z_j)}{\sum_{j=1}^n I(X_j \geq X_i) \delta_j \{W_c(X_j)\}^{-1} \exp(\beta' Z_j)} \right] = 0, \end{aligned}$$

where $S_c(\cdot)$ is the survival function of the censoring time C , and $W_c(t) = \int_0^t S_c(s) ds$. Tsai (2009) proposed the pseudo-partial likelihood method with the following estimating equations based on the score statistics,

$$EE - PL: \quad \sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^n Z_j Y_j(t) \exp(\beta' Z_j) W(t, X_j)}{\sum_{j=1}^n Y_j(t) \exp(\beta' Z_j) W(t, X_j)} \right\} dN_i(t) = 0,$$

where $W(t, X_j) = \delta_j \frac{W_c(X_j) - W_c(X_j - t)}{W_c(X_j)} + (1 - \delta_j) \frac{S_c(X_j) - S_c(X_j - t)}{S_c(X_j)}$.

When the censoring was independent of the covariates, all the types of estimators had a small bias with light censoring (15%). With moderate (30%) or heavy censoring (50%), the biases associated with the MLE were much smaller than those associated with EE I. The

MLE method always exhibited clearly superior efficiency, with smaller empirical standard errors. For instance, the standard errors associated with the estimating equations were 1.12 to 1.62 times greater, and the standard errors associated with the pseudo-partial likelihood method were 1.09 to 1.50 times greater than those associated with the MLE method based on a sample size of 200. When the censoring distribution was dependent on the covariates, the estimators obtained by EE-I, EE-II and EE-PL were biased compared with those obtained by the MLE method. In summary, the MLE approach is the most efficient one among the four methods, and is also the most robust to various censoring mechanisms.

6. A REAL DATA EXAMPLE

Dementia is a progressive degenerative medical condition and is one of the leading causes of all deaths in the United States and Canada. The Canadian Study of Health and Aging was a multicenter epidemiologic study of dementia, in which 14,026 subjects 65 years or older were randomly chosen throughout Canada to receive an invitation for a health survey. A total of 10,263 subjects agreed to participate in the study (Wolfson et al., 2001). The participants were then screened for dementia, and 1132 people were identified as having the disease. The individuals with dementia were followed until their deaths or last follow-up dates in 1996, and their dates of dementia onset were ascertained from their medical records.

After excluding subjects with missing data regarding the date of disease onset or classification of dementia subtype, a total of 818 patients remained {393 with probable Alzheimer's disease, 252 with possible Alzheimer's disease and 173 with vascular dementia. Other study variables included the approximate date of dementia onset, date of screening for dementia, date of death or censoring and death indicator variable. Given the prevalent cases ascertained cross-sectionally, Asgharian et al. (2006) validated the stationarity assumption, which was defined as that the incidence of dementia did not change over the period of the study.

At the end of the study, 638 out of 818 patients had died and the others were right censored. Within this elderly cohort, it seems reasonable to assume that the overall death rate increases with age. Applying the NPMLE approaches described in Sections 2 and 3, we estimated the hazard function for each subtype of dementia, and plotted the survival of patients with probable Alzheimer's disease, possible Alzheimer's disease and vascular dementia with and without the constraint of the increasing risk of death (see Figure 1). It is not surprising that the estimated survival curves with the constraint, i.e., additional information, have narrower confidence intervals than their corresponding survival curves without the constraint. As pointed out by one referee, a monotone hazards constraint may not hold for death due to dementia, particularly for patients with vascular dementia, because of the uncertainty associated with the cause of death.

Using the diagnosis subtype of possible Alzheimer's disease as the baseline cohort, we defined two indicator variables for the other two subtypes of dementia for the Cox proportional hazards model. Applying the proposed method in Section 4, the estimated covariate Effects of two subtypes of dementia and their standard errors are listed in Table 3. The results showed that the long-term survival distributions were statistically significantly different between the group with vascular dementia and the group with possible Alzheimer's dementia, and marginally different between the group with probable Alzheimer's dementia and the group with possible Alzheimer's dementia. We also analyzed the same data set using the estimating equation methods (EE-I and EE-II) by Qin and Shen (2010) and the pseudo-partial likelihood method by Tsai (2009). The estimated coefficients with the associated standard errors obtained by both EE-II and EE-PL indicated no statistically significant survival differences between the three subtypes of dementia. The results from EE-I

suggested there existed a statistically significant survival difference between the group with vascular dementia and the group with possible Alzheimer's dementia, but no statistically significant survival differences between the group with vascular dementia and the group with probable Alzheimer's dementia. The discrepancy in the inferences is most likely caused by the loss in efficiency when using the estimating equation method or the pseudo-partial likelihood method compared to using the MLE method.

7. CONCLUDING REMARKS

We have proposed new EM algorithms for length-biased data to obtain full likelihood maximum estimators under three settings, and the missing data mechanism in the EM-algorithm is the left truncation for the length-biased data. In contrast to Vardi's (1989) EM algorithm for estimating nonparametric survival distributions, the advantage of the new EM algorithm is that one can directly estimate the non-parametric survival distribution or hazard function of the unbiased failure time T .

One major challenge to maximum likelihood estimation when it involves infinite dimensional parameters is computational intractability. We have implemented the new EM algorithm together with the profile likelihood method for jointly estimating the baseline hazard function and the covariate coefficients under the Cox regression model for length-biased data. Commercially available statistics software for the Cox model can be adapted for easy computation. The EM algorithm is not computationally intensive even with continuous covariates, since p_{ij} and λ_j can be obtained easily from the closed form expressions.

Similar to the NPMLE for traditional survival data, the method we have proposed requires the observation of at least one failure time to ensure that the large sample properties hold in the settings for length-biased data. As shown in our empirical studies, the proposed computational algorithms to solve MLE perform well in terms of accuracy, and are more efficient compared to the existing estimating equation approaches, which are more efficient than the conditional approach (Wang et al., 1993). Without assuming a known parametric distribution of Z as in Bergeron et al. (2008), maximizing the likelihood function (1) is equally efficient to maximizing the full likelihood including the marginal distribution of Z .

Parallel to the observation of Zeng and Lin (2007) for traditional survival data, estimators obtained from MLE are much more robust and efficient than those from the estimating equation approaches, and well suited for the proportional hazards regression model and other nonparametric estimations for length-biased right-censored data. The proposed EM algorithm may be further generalized to other semiparametric models, and the tools for model checking should be developed for length-biased right-censored data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by National Institute of Health grant R01-CA079466.

We thank one Associate Editor and two Referees for their very constructive comments. We also thank Professor Masoud Asgharian and investigators of the Canadian Study of Health and Aging (CHSA) for providing us the dementia data from CHSA. The data reported in the example were collected as part of the CHSA. The core study was funded by the Seniors' Independence Research Program, through the National Health Research and Development Program of Health Canada (Project no.6606-3954-MC(S)). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP Project 6603-1417-302(R), Bayer Incorporated, and the British

Columbia Health Research Foundation Projects 38 (93-2) and 34 (96-1). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

References

- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. Springer; New York: 1992.
- Asgharian M, M'LAN CE, Wolfson DB. Length-biased Sampling with Right Censoring: an Unconditional Approach. *Journal of the American Statistical Association*. 2002; 97:201–209.
- Asgharian M, Wolfson DB. Asymptotic Behavior of the Unconditional NPMLE of the Length-biased Survivor Function From Right Censored Prevalent Cohort Data. *The Annals of Statistics*. 2005; 33:2109–2131.
- Asgharian M, Wolfson DB, Zhang X. Checking Stationarity of the Incidence Rate Using Prevalent Cohort Survival Data. *Statistics in Medicine*. 2006; 25:1751–1767. [PubMed: 16220462]
- Barlow, RE.; Proschan, F. Statistical Theory of Reliability. New York: Holt, Rinehart & Winston; 1975.
- Bergeron P-J, Asgharian M, Wolfson DB. Covariate Bias Induced by Length-Biased Sampling of Failure Times. *Journal of the American Statistical Association*. 2008; 103:737–742.
- Bickel PJ, Ritov Y. Efficient Estimation Using both Direct and Indirect Observations. *Theory of Probability and its Applications*. 1994; 38:194–213.
- Breslow N. Contribution to the Discussion of the Paper by D. R. Cox. *Journal of the Royal Statistical Society B*. 1972; 34:187–220.
- Chen HY, Little RJA. Proportional Hazards Regression with Missing Covariates. *Journal of the American Statistical Association*. 1999; 94:896–908.
- Cox DR. Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
- Cox DR. Partial Likelihood. *Biometrika*. 1975; 62:269–276.
- Cox, DR.; Miller, HD. The Theory of Stochastic Processes. London: New York; 1977.
- De Uña Álvarez J, Otero-Giraldez MS, Alvarez-Llorente G. Estimation Under Length-bias and Right-censoring: An Application to Unemployment Duration Analysis for Married Women. *Journal of Applied Statistics*. 2003; 30:283–291.
- Dewanji A, Kalbfleisch JD. Estimation of Sojourn Time Distributions for Cyclic sSemi-markov Processes in Equilibrium. *Biometrika*. 1987; 74:281–288.
- Elbers C, Ridder G. True and Spurious Duration Dependence: the Identifiability of the Proportional Hazard Model. *The Review of Economic Studies*. 1982; 49:403–409.
- Ferguson, TS. A Course in Large Sample Theory. 1. London: Chapman & Hall; 1996.
- Gail, MH.; Benichou, J. Encyclopedia of Epidemiologic Methods. Wiley; 2000.
- Gordis, L. Epidemiology. Philadelphia, PA: W. B. Saunders Company; 2000.
- Huang J, Wellner JA. Estimation of a Monotone Density or Monotone Hazard under Random Censoring. *Scandinavian Journal of Statistics*. 1995; 22:3–33.
- Kalbfleisch JD, Lawless JF. Regression Models for Right Truncated Data with Applications to AIDS Incubation Times and Reporting Lags. *Statistica Sinica*. 1991; 1:19–32.
- Kalbfleisch JD, Prentice RL. Marginal Likelihoods Based on Cox's Regression and Life Model. *Biometrika*. 1973; 60:267–278.
- Keiding N. Age-specific Incidence and Prevalence: a Statistical Perspective (with Discussion). *Journal of the Royal Statistical Society, Series A*. 1991; 154:371–412.
- Klein J. Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics*. 1992; 48:795–806. [PubMed: 1420842]
- Kvam P. Length Bias in the Measurements of Carbon Nanotubes. *Technometrics*. 2008; 50:462–467.
- Lancaster, T. The Econometric Analysis of Transition Data. Cambridge: University Press; 1990.
- Marshall AW, Proschan F. Maximum Likelihood Estimation for Distributions with Monotone Failure Rate. *The Annals of Mathematical Statistics*. 1965; 36:69–77.

- McClellan S, Devine C. A Nonparametric Maximum Likelihood Estimator for Incomplete Renewal Data. *Biometrika*. 1995; 82:791–803.
- Meng X-L, Rubin DB. Maximum Likelihood Estimation via the ECM Algorithm: a General Framework. *Biometrika*. 1993; 80:267–278.
- Murphy SA. Consistency in a Proportional Hazards Model Incorporating a Random Effect. *The Annals of Statistics*. 1994; 22:712–731.
- Murphy SA. Asymptotic theory for the frailty model. *The Annals of Statistics*. 1995; 23:182–198.
- Murphy SA, van der Vaart AW. Observed Information in Semi-parametric Models. *Bernoulli*. 1999; 5:381–412.
- Murphy SA, van der Vaart AW. On Profile Likelihood (with Discussion). *Journal of the American Statistical Association*. 2000; 95:449–485.
- Nielsen GG, Gill RD, Andersen PK, Sorensen TIA. A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics*. 1992; 19:25–43.
- Padgett WJ, Wei LJ. Maximum Likelihood Estimation of a Distribution Function with Increasing Failure Rate Based on Censored Observations. *Biometrika*. 1980; 67:470–474.
- Parner E. Asymptotic Theory for the Correlated Gamma-frailty Model. *The Annals of Statistics*. 1998; 26:183–214.
- Qin J, Shen Y. Statistical Methods for Analyzing Right-censored Length-biased Data under Cox Model. *Biometrics*. 2010; 66:382–392. [PubMed: 19522872]
- Rothenberg TJ. Identification in Parametric Models. *Econometrica*. 1971; 39:577–591.
- Sansgiry P, Akman O. Transformations of the Lognormal Distribution as a Selection Model. *The American Statistician*. 2000; 54:307–309.
- Scheike TH, Keiding N. Design and Analysis of Time-to-pregnancy. *Statistical Methods in Medical Research*. 2006; 15:127–140. [PubMed: 16615653]
- Simon R. Length-biased Sampling in Etiologic Studies. *American Journal of Epidemiology*. 1980; 111:444–452. [PubMed: 7377187]
- Terwilliger J, Shannon W, Lathrop G, Nolan J, Goldin L, Chase G, Weeks D. True and False Positive Peaks in Genomewide Scans: Applications of Length-biased Sampling to Linkage Mapping. *American Journal of Human Genetics*. 1997; 61:430–438. [PubMed: 9311749]
- Tricomi, FG. *Integral Equations*. New York: Dover Publications; 1985.
- Tsai WY. Estimation of the Survival Function with Increasing Failure Rate Based on Left Truncated and Right Censored Data. *Biometrika*. 1988; 75:319–324.
- Tsai WY. Pseudo-partial Likelihood for Proportional Hazards Models with Biased-sampling Data. *Biometrika*. 2009; 96:601–615.
- Turnbull BW. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society Series B Methodological*. 1976; 38:290–295.
- van der Vaart, A. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, UK: Cambridge University Press; 1998. *Asymptotic Statistics*.
- van der Vaart AW, Wellner JA. Existence and Consistency of Maximum Likelihood in Upgrade Mixture Models. *Journal of Multivariate Analysis*. 1992; 43:133–146.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag; 1996. with applications to statistics
- Vardi Y. Nonparametric Estimation in the Presence of Length Bias. *The Annals of Statistics*. 1982; 10:616–620.
- Vardi Y. Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation. *Biometrika*. 1989; 76:751–761.
- Vardi Y, Zhang CH. Large Sample Study of Empirical Distributions in a Random-Multiplicative Censoring Model. *The Annals of Statistics*. 1992; 20:1022–1039.
- Wang MC. Nonparametric Estimation From Cross-Sectional Survival Data. *Journal of the American Statistical Association*. 1991; 86:130–143.
- Wang MC. Hazards Regression Analysis for Length-biased Data. *Biometrika*. 1996; 83:343–354.

- Wang MC, Brookmeyer R, Jewell NP. Statistical Models for Prevalent Cohort Data. *Biometrics*. 1993; 49:1–11. [PubMed: 8513095]
- Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB. the Clinical Progression of Dementia Study Group. A Reevaluation of the Duration of Survival after the Onset of Dementia. *The New England Journal of Medicine*. 2001; 344:1111–1116. [PubMed: 11297701]
- Zeidler, E. *Applied Mathematical Sciences*. Vol. 109. New York: Springer-Verlag; 1995. *Applied Functional Analysis: Main Principles and Their Applications*.
- Zelen M. Forward and Backward Recurrence Times and Length Biased Sampling: Age Specific Models. *Lifetime Data Analysis*. 2004; 10:325–334. [PubMed: 15690988]
- Zelen M, Feinleib M. On the Theory of Screening for Chronic Diseases. *Biometrika*. 1969; 56:601–614.
- Zeng D, Lin DY. Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data. *Journal of the Royal Statistical Society, Series B*. 2007; 69:507–564.
- Zeng D, Lin DY, Yin G. Maximum Likelihood Estimation for the Proportional Odds Model with Random Effects. *Journal of the American Statistical Association*. 2005; 100:470–483.

A APPENDIX

A.1 Assumptions

Denote the Euclidean norm by $|\cdot|$. For a vector $\mathbf{z} = (z_1, \dots, z_p)'$, $|\mathbf{z}| \equiv (|z_1|^2 + \dots + |z_p|^2)^{1/2}$. To avoid measurability issues, the probability is understood as the outer probability (van der Vaart and Wellner, 1996). We adopt the convention that $0/0 = 0$. We assume the following regularity assumptions:

1. The true value of the hazard function $\lambda_0(\cdot)$ is continuously differentiable. In addition, the upper bound τ of the support of the cumulative hazard function Λ_0 is finite.
2. The parameter $\boldsymbol{\beta}$ is in a compact set $\mathcal{B}$ that contains $\boldsymbol{\beta}_0$. The parameter set $\mathcal{A}$ for the baseline function contains all nondecreasing functions Λ satisfying $\Lambda(0) = 0$ and $\Lambda(\tau) < \infty$.
3. The residual censoring time C satisfies $P(C > V) > 0$. Its survival function $S_C(\cdot)$ is continuous.
4. For the covariate vector $\mathbf{Z}$, the terms $E_0|\mathbf{Z}|^2$ and $E_0|e^{\boldsymbol{\beta}'\mathbf{Z}}|$ are bounded.
5. The information matrix $-\partial^2 E[\ell_n(\boldsymbol{\beta}, \hat{\lambda}(\cdot, \boldsymbol{\beta}))]/\partial \boldsymbol{\beta}^2$ evaluated at the true values $\boldsymbol{\beta}_0$ is positive definite.
6. If $P(\mathbf{b}'\mathbf{Z} = c_0) = 1$ for some constant c_0 , then $\mathbf{b} = 0$.

Assumption 1 and 3 imply that the bivariate function $Q(t, u)$ defined in (23) is continuous on $[0, \tau] \times [0, \tau]$. Assumption 3 implies that the censoring may occur after V . Assumption 5 is a classical condition that appears in the study of the Cox model for traditional survival data (Andersen et al., 1992, Condition VII2.1.(e), page 497). It implies that the matrix $\mathbf{J}_0$ defined in (18) is positive definite, which is the information matrix for $\boldsymbol{\beta}$ when the baseline function Λ_0 is known. The positive definiteness of $\mathbf{J}_0$ can be also implied by the fact the model is identifiable (Rothenberg, 1971). Assumption 6 means that there is no covariate collinearity, which ensure the model identifiability.

A.2 Proof of the lemma on the convergence of $\hat{\tau} = t_k$ to τ

Recall that $X_i = \min(A_i + V_i, A_i + C_i)$, and $t_k = \max\{X_1, \dots, X_n\}$. For any arbitrary small $\varepsilon > 0$,

$$\begin{aligned}
P(|\max_{1 \leq i \leq n} X_i - \tau| > \varepsilon) &= [P(X_1 < \tau - \varepsilon)]^n = \{1 - E[I(A_1 + V_1 > \tau - \varepsilon)S_c(\tau - \varepsilon - A_1)]\}^n \\
&= \left\{1 - \int_{\tau - \varepsilon}^{\tau} w(y)f(y)dy/\mu\right\}^n \\
&= \{1 - w(\xi)f(\xi)\varepsilon/\mu\}^n,
\end{aligned}$$

where $w(y) = \int_0^y S_c(\tau - \varepsilon - a)da > 0$, $\tau - \varepsilon \leq \xi \leq \tau$, $f(\xi) > 0$, and the last equality uses the mean value theorem. Given the above equation, for any arbitrary small $\eta > 0$,

$$P(n^\eta |\max_{1 \leq i \leq n} X_i - \tau| > \varepsilon) = \{1 - w(\xi)f(\xi)\varepsilon n^{-\eta}/\mu\}^n,$$

where $\tau - \varepsilon n^{-\eta} \leq \xi \leq \tau$. Note that when $n \rightarrow \infty$,

$$\begin{aligned}
w(\xi) &= \int_0^{\xi} S_c(\tau - \varepsilon n^{-\eta} - a)da = \int_{\tau - \xi - \varepsilon n^{-\eta}}^{\tau - \varepsilon n^{-\eta}} S_c(s)ds \rightarrow \int_0^{\tau} S_c(s)ds, \\
\int_0^{\tau} S_c(s)ds &= E[\int_0^{\tau} I(C > s)ds] = E[\min(C, \tau)].
\end{aligned}$$

Thus, as long as $E[C] > 0$ which is inferred by Assumption 3, we have $w(\xi) \rightarrow w(\tau) > 0$ and $f(\xi) \rightarrow f(\tau) > 0$, for $n \rightarrow \infty$, since τ is the upper bound for the support of population time and $\Lambda(\tau) < \infty$. Therefore, when $n \rightarrow \infty$ we have that

$$P(n^\eta |\widehat{\tau} - \tau| > \varepsilon) \approx \exp(-w(\tau)f(\tau)\varepsilon n^{1-\eta}/\mu),$$

which implies that $\widehat{\tau} \rightarrow \tau$ in a rate higher than $n^{1/2}$ but lower than n (i.e. when $1/2 < \eta < 1$, the above probability converges to zero).

Note that by the Borel-Cantelli Lemma, $\widehat{\tau}_n \rightarrow \tau$ in probability implies that for every subsequence, there is a sub-subsequence $\{n'\}$ such that $\widehat{\tau}_{n'} \rightarrow \tau$ almost surely (Ferguson, 1996, page 8). This combined with the fact that τ can be consistently estimated by $\widehat{\tau}_n$ complete the proof for strong consistency of $\widehat{\Lambda}_n$. As $\widehat{\tau}_n \leq \tau$, these facts also infer that the mean μ of the population failure time $\tilde{T}$ can be consistently estimated by $\int_0^{\widehat{\tau}} \exp(-\widehat{\Lambda}_n(s))ds$.

A.3 Consistency of $\widehat{\Lambda}_n$ with increasing failure rate

Let $\|\cdot\|_\tau$ denote the supremum norm over $[0, \tau]$. We have the following strong consistency result for the NPMLE $\widehat{\Lambda}_n(\cdot)$ proposed in Section 3:

Proposition 1

Suppose Λ is convex on its support $[0, \tau]$. Under the regularity conditions,

$$\|\widehat{\Lambda}_n - \Lambda\|_\tau \leq \|\Lambda_n - \Lambda\|_\tau \rightarrow_{a.s.} 0.$$

We adapt the proof in Huang and Wellner (1995). Let $\varepsilon_n = \|\Lambda_n - \Lambda\|_\tau$. As argued in Asgharian and Wolfson (2005), $\varepsilon_n \rightarrow 0$ almost surely when $n \rightarrow \infty$. Since Λ is convex on

$[0, \tau]$, it must be continuous on $[0, \tau]$. The function $\Lambda - \varepsilon_n$ is convex and is a minorant of Λ_n , i.e., $\Lambda(s) - \varepsilon_n \leq \Lambda_n(s)$ for all $0 \leq s \leq \tau$. By the definition of $\hat{\Lambda}_n$, we have that for all $0 \leq s \leq \tau$,

$$\Lambda(s) - \varepsilon_n \leq \hat{\Lambda}_n(s) \leq \Lambda_n(s).$$

It follows that $-\varepsilon_n \leq \hat{\Lambda}_n(s) - \Lambda(s) \leq \Lambda_n(s) - \Lambda(s) \leq \varepsilon_n$ for all $0 \leq s \leq \tau$. The conclusion of the proposition follows as $\|\hat{\Lambda}_n - \Lambda\|_\tau \leq \varepsilon_n$ goes to 0 almost surely.

A.4 Consistency: Proof of Theorem 1

Note that the log-likelihood function $\ell_n(\beta, \lambda)$ is strictly concave in λ as each function of λ in $\ell_n(\beta, \lambda)$ is concave or strictly concave and the summation of concave functions is concave. Hence, for each β in a compact set $\mathcal{B}$, we can find a unique maximizer of $\hat{\lambda}(\cdot, \beta)$ of the likelihood function $\ell_n(\beta, \lambda)$. The existence of the NPMLE for $\{\beta, \lambda\}$ follows by the compactness of $\mathcal{B}$ for the profile likelihood $\ell_n(\beta, \hat{\lambda}(\cdot, \beta))$, which is continuous in β . The uniqueness of the NPMLE is guaranteed by Assumption 5 for large samples.

The technical details of the consistency proof are similar to those of Murphy (1995) or Parner (1998). We provide only a sketch of the proof. As the MLE $(\hat{\beta}_n, \hat{\Lambda}_n)$ maximizes the log-likelihood function ℓ_n , the method is to use the empirical Kullback-Leibler distance $\ell_n(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_n(\beta_0, \Lambda_0)$, which must always be nonnegative. If $(\hat{\beta}_n, \hat{\Lambda}_n)$ converges at all, say, to (β^*, Λ^*) , then $\ell_n(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_n(\beta_0, \Lambda_0)$ must converge to the negative Kullback-Leibler distance between P_{β^*, Λ^*} and P_{β_0, Λ_0} by the strong law of large numbers, where $P_{\beta, \Lambda}$ is the probability measure under the parameter (β, Λ) . The Kullback-Leibler distance between P_{β^*, Λ^*} and P_{β_0, Λ_0} therefore must be zero, and we conclude that $P_{\beta^*, \Lambda^*} = P_{\beta_0, \Lambda_0}$ almost surely. It then follows by model identifiability that $\beta^* = \beta_0$ and $\Lambda^* = \Lambda_0$.

We need to find, for any subsequence of $(\hat{\beta}_n, \hat{\Lambda}_n)$, a further convergent subsequence. The first step is to show that $(\hat{\beta}_n, \hat{\Lambda}_n)$ stays bounded. As $\hat{\beta}_n$ is in a compact set, it must stay bounded. Because $(\hat{\beta}_n, \hat{\Lambda}_n)$ maximizes the likelihood function, $\ell_n(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_n(\beta, \Lambda) \geq 0$ for each (β, Λ) in the parameter set. Recall that $\hat{\tau} = t_k$. We show that $\hat{\Lambda}_n(\hat{\tau})$ stays bounded, $\overline{\lim}_n \hat{\Lambda}_n < \infty$. We use the method of contradiction. Suppose that $\hat{\Lambda}_n(\hat{\tau})$ diverges. Then we can construct some sequence $\{\beta_n, \Lambda_n\}$ such that the empirical Kullback-Leibler distance $\ell_n(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_n(\beta_n, \Lambda_n)$ would become negative infinity. This is a contradiction as the Kullback-Leibler distance is always nonnegative. The construction of the contradiction is along the lines as in Murphy (1994). Briefly, we choose $\beta_n = \beta_0$ and define Λ_n to be

$$\bar{\Lambda}_n(t) = \int_0^t \left\{ \sum_{i=1}^n W_i(u, \beta_0, \Lambda_0) \right\}^{-1} d \left\{ \sum_{j=1}^n N_j(u) \right\}, \quad (26)$$

where $W_i(u, \beta, \Lambda) = e^{\beta' Z_i} (Y_i(u) - \int_u^\tau S(v|Z_i) dv) / \int_0^\tau S(v|Z_i) dv$, and note that

$$E W_i(u, \beta, \Lambda) - E \left[e^{\beta' Z_i} \{u - E_c(u - C)^+\} S(u|Z) / \mu(Z) \right] \geq 0.$$

It can be shown easily that $\hat{\Lambda}_n$ converges to Λ_0 almost surely and uniformly in t . By a technical argument similar to that of Murphy (1995), we can show that $\ell_n(\hat{\beta}_n, \hat{\Lambda}_n) - \ell_n(\beta_n, \hat{\Lambda}_n) \rightarrow -\infty$ as $n \rightarrow \infty$. This is impossible so $\hat{\Lambda}_n$ must stay bounded.

As $\hat{\Lambda}_n$ stays bounded, we can apply Helly's selection principle to find a convergent subsequence of $(\hat{\beta}_{n_k}, \hat{\Lambda}_{n_k})$ for an arbitrary subsequence from the sequence indexed by $\{1, \dots, n\}$. By the strong law of large numbers, such convergent subsequence must converge to (β_0, Λ_0) using the classical Kullback-Leibler information approach. For any given subsequence $\{n_k\}$, we can identify a further subsequence of $(\hat{\beta}_{n_k}, \hat{\Lambda}_{n_k})$ that converges to (β_0, Λ_0) . Helly's selection theorem implies that the entire sequence $(\hat{\beta}_n, \hat{\Lambda}_n(t))$ must converge to $(\beta_0, \Lambda_0(t))$ for each $t \in [0, \tau]$. By the assumption that $\Lambda_0(\cdot)$ is monotone and continuous, the convergence of $\hat{\Lambda}_n(t) \in [0, \tau]$ at each t is also uniformly in t . The convergence is also almost surely a true convergence as the proof is carried out for a fixed ω in the underlying probability space Ω , where we use the law of large numbers, except that it is countable many times.

A.5 Asymptotic Normality: Proof of Theorem 2

We prove the asymptotic normality by the Z-theorem for the infinite-dimensional estimating equations (van der Vaart and Wellner, 1996, Theorem 3.3.1, page 310). This approach has been successfully applied by (Murphy, 1995, Theorem 1) and (Parner, 1998, Theorem 2) among many others. The proof requires the confirmation of the three main conditions of the Z-theorem: Fréchet differentiability, weak convergence of $\sqrt{n}U_n(\beta_0, \Lambda_0)$ and a stochastic approximation of the estimating equations, which we outline below.

Fréchet Derivative and its Invertibility

Let $\Lambda_Z(t) = \int_0^t e^{\beta' Z} d\Lambda(u)$. We first show that the population estimating equation U_0 is Fréchet differentiable and its Fréchet derivative is continuously invertible. We have defined $U_0(\cdot, \beta, \Lambda) = (U_{10}(\beta, \Lambda), U_{20}(\cdot, \beta, \Lambda))$ where

$$\begin{aligned} U_{10}(\beta, \Lambda) &= \int_0^\tau E_0 \left\{ Z dN(t) - Y(u) Z d\Lambda_Z(u) + Z \left(\int_u^\tau S(v|Z) dv \right) d\Lambda_Z(u) / \mu_{\beta, \Lambda}(Z) \right\}, \\ U_{20}(t, \beta, \Lambda) &= \int_0^t E_0 \left\{ dN(u) - Y(u) d\Lambda_Z(u) + \left(\int_u^\tau S(v|Z) dv \right) d\Lambda_Z(u) / \mu_{\beta, \Lambda}(Z) \right\}. \end{aligned}$$

The Fréchet derivative can be calculated from the Gâteaux variations of $U_0(\beta, \Lambda)$ at (β_0, Λ_0) . That is done to differentiate $U_0(\beta_\eta, \Lambda_\eta)$ with respect to η and evaluated at $\eta = 0$, where the submodels are $\beta_\eta = \beta_0 + \eta\beta$ and $\Lambda_\eta = \Lambda_0 + \eta\Lambda$.

The Gâteaux derivative of $U_{20}(t, \beta, \Lambda)$ evaluated at (β_0, Λ_0) is

$$-\{\sigma_{21}(\beta) + \sigma_{22}(\Lambda)\},$$

where $\sigma_{21}(\beta) \equiv \partial / \partial \eta U_{20}(t, \beta_\eta, \Lambda_0)|_{\eta=0} = \beta' \left\{ \int_0^t K_1^{(1)}(u) d\Lambda_0(u) + \int_0^\tau K_2^{(1)}(t, u) d\Lambda_0(u) \right\}$, and $\sigma_{22}(\Lambda) \equiv \partial / \partial \eta U_{20}(t, \beta_0, \Lambda_\eta)|_{\eta=0} = \int_0^t K_1^{(0)}(u) d\Lambda(u) + \int_0^\tau K_2^{(0)}(t, u) d\Lambda(u)$. The Gâteaux derivative of $U_{10}(t, \beta, \Lambda)$ evaluated at (β_0, Λ_0) is

$$-\{\sigma_{11}(\beta) + \sigma_{12}(\Lambda)\},$$

where $\sigma_{11}(\beta) \equiv \partial/\partial\eta U_{10}(\beta_\eta, \Lambda_0)|_{\beta=\beta_0} = \{\int_0^\tau \mathbf{K}_1^{(2)}(u)d\Lambda_0(u) + \int_0^\tau \mathbf{K}_2^{(2)}(\tau, u)d\Lambda_0(u)\}'\beta$, and $\sigma_{12}(\Lambda) \equiv \partial/\partial\eta U_{10}(\beta_0, \Lambda_\eta)|_{\eta=0} = \int_0^\tau \mathbf{K}_1^{(1)}(u)d\Lambda(u) + \int_0^\tau \mathbf{K}_2^{(1)}(\tau, u)d\Lambda(u)$.

To obtain the results on weak convergence, we need to strengthen the Gateaux differentiability to Fréchet differentiability, essentially for the proof of tightness (van der Vaart and Wellner, 1996, page 310). The Fréchet differentiability of $U_0(\beta, \Lambda)$ can be confirmed by definition; its derivative has the form in (19). Note that the operator $\dot{U}_{\psi_0}$ is a linear continuous operator defined on parameter space in the product space of $\mathbb{R}^p$ and the

Banach space $L_2[0, \tau]$. If the inverse operator $\dot{U}_{\psi_0}^{-1}$ exists, then it must be continuous by Banach's continuous inverse theorem (Zeidler, 1995, page 179). Hence, to prove the continuous invertibility of U_{ψ_0} , we only need to show the existence of the inverse operator $\dot{U}_{\psi_0}^{-1}$.

To show the existence of the inverse operator $\dot{U}_{\psi_0}^{-1}$, we need to show by the formula in (20) that σ_{11} and $\Phi = \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}$ have inverse. The operator $\sigma_{11}(\beta) = \mathbf{J}_0'\beta$ is a linear operator, where the matrix $\mathbf{J}_0$ defined in (18) is the Fisher information for β for known Λ_0 . By Assumption 5, the matrix $\mathbf{J}_0$ has an inverse, and hence σ_{11} is invertible. The operator Φ has the following form:

$$\Phi(\Lambda) = - \left[\int_0^\tau \mathbf{K}_1^{(0)}(u)d\Lambda(u) + \int_0^\tau \left\{ \mathbf{K}_2^{(0)}(t, u) - \left(\int_0^\tau \mathbf{K}_1^{(1)}(v)d\Lambda_0(v) + \int_0^\tau \mathbf{K}_2^{(1)}(t, v)d\Lambda_0(v) \right)' \mathbf{J}_0^{-1} \cdot \left(\mathbf{K}_1^{(1)}(u) + \mathbf{K}_2^{(1)}(\tau, u) \right) \right\} d\Lambda(u) \right].$$

The invertibility of Φ is equivalent to show that there exists a unique solution to the equation $\Phi(\Lambda) = \tilde{\Lambda}$ for a function $\tilde{\Lambda}$ of bounded variation. Taking the derivative with respect to t on both sides of the equation,

$$-d\tilde{\Lambda}(t) = \mathbf{K}_1^{(0)}(t)d\Lambda(t) - \int_0^\tau Q(t, u)d\Lambda(u), \quad (27)$$

where $Q(t, u)$ is defined previously in (23). We observe that the integral equation (27) is the Fredholm equation of the second type. By Assumptions 1 and 3, the bivariate function $Q(t, u)$ defined in (23) is continuous on $[0, \tau] \times [0, \tau]$. Also by assumption 3, the function

$1/\mathbf{K}_1^{(0)}(t)$ is continuous and bounded away from 0 for $t > 0$. By the classical theory for integral equation (Tricomi, 1985, Chapter 2), there is a unique solution $d\Lambda(t)$ to the Fredholm integral equation (27), characterized by

$$d\Lambda(t) = - \frac{d\tilde{\Lambda}(t)}{\mathbf{K}_1^{(0)}(t)} + \int_0^\tau \frac{H(t, u)}{\mathbf{K}_1^{(0)}(t)} d\tilde{\Lambda}(u),$$

where $H(t, u)$ satisfies the equation (22). Finally, the invertibility of the functional Φ follows and its inverse operator $\Phi^{-1}(\Lambda)$ has a form expressed in (21).

Weak Convergence of $\sqrt{n}U_n(\beta_0, \Lambda_0)$

As the true value (β_0, Λ_0) satisfies $U_0(\beta_0, \Lambda_0) = 0$, we have

$$\sqrt{n}U_n(\beta_0, \Lambda_0) = \sqrt{n}\{U_{1n}(\psi_0) - U_{10}(\psi_0), U_{2n}(t, \psi_0) - U_{20}(t, \psi_0)\}.$$

By the multivariate central limit theorem for the summation of independently and identically distributed (i.i.d.) random vectors, $\sqrt{n}\{U_{1n}(\psi_0) - U_{10}(\psi_0)\}$ converges in law to $\mathbb{W}_1$, provided that the second moment is finite. The process $\sqrt{n}\{U_{2n}(t, \psi_0) - U_{20}(t, \psi_0)\}$ is a sum of i.i.d. processes of bounded variation on $[0, \tau]$. By a lemma for the central limit theorem for processes of bounded variation (van der Vaart and Wellner, 1996, Example 2.11.16), $\sqrt{n}\{U_{2n}(t, \psi_0) - U_{20}(t, \psi_0)\}$ converges to a tight Gaussian process, say $\mathbb{W}_2$, provided that the second moment is finite. The weak convergence of $\sqrt{n}U_n(\beta_0, \Lambda_0)$ follows by the continuous mapping theorem.

Stochastic Approximation

To apply the Z-theorem for the infinite dimensional estimating equations (van der Vaart and Wellner, 1996, Theorem 3.3.1), we need to confirm the following stochastic approximation:

$$\sqrt{n}\{(U_n - U_0)(\hat{\psi}_n) - (U_n - U_0)(\psi_0)\} \equiv \sqrt{n}\{U_n(\cdot, \hat{\psi}_n) - U_0(\cdot, \hat{\psi}_n) - [U_n(\cdot, \psi_0) - U_0(\cdot, \psi_0)]\} = o_p(1).$$

The function is defined on $\mathcal{B} \times \mathcal{A}$, where by Assumption 2, the set $\mathcal{B}$ is compact that contains β_0 , and $\mathcal{A}$ is a set of nondecreasing function such that for each $\Lambda \in \mathcal{A}$, $\Lambda(0) = 0$ and $\Lambda(\tau) < \infty$. Hence $\mathcal{A}$ contains Λ_0 . To apply the Z-theorem, we need Λ in the closed linear subspace $\bar{\mathcal{A}}$ generated by the set $\mathcal{A}$. The subspace $\bar{\mathcal{A}}$ is viewed in the space of functions of bounded variation on $[0, \tau]$ endowed with the variation norm by $\|\cdot\|_v$, defined by the total variation of Λ on $[0, \tau]$, i.e.,

$$\|\Lambda\|_v \equiv \sup \sum_{k=1}^{m'} |\Lambda(s_k) - \Lambda(s_{k-1})|,$$

where the supremum is taken over all finite partitions of $[0, \tau]$, $\{0 = s_0 < s_1 < \dots < s_{m'} = \tau\}$.

We first derive the score functions for β and $\Lambda(\cdot)$ for a single observation $\mathcal{O}$. We keep $\mathcal{O}$ in the notations for the score functions to emphasize its dependence on the data. By straightforward calculation, the score function for β is

$$\dot{\ell}_1(\mathcal{O}, \beta, \Lambda) = \int_0^\tau Z dN(u) - \int_0^\tau Y(u) Z e^{\beta' Z} d\Lambda(u) + \int_0^\tau Z \left(\int_u^\tau S(v|Z) dv \right) e^{\beta' Z} d\Lambda(u) / \mu_{\beta, \Lambda}(Z). \quad (28)$$

For the infinite dimensional parameter $\Lambda(\cdot)$, consider a submodel defined by $d\Lambda_\eta = (1 + \eta h) d\Lambda$, where h is a bounded and integrable function. By taking the derivative of $\ell_n(\beta, \Lambda_\eta)$ with respect to η and evaluating it at $\eta = 0$, we have the score operator for Λ

$$\dot{\ell}_2(\mathcal{O}, \beta, \Lambda)(h) = \int_0^\tau h(v) dN(v) - \int_0^\tau Y(v) e^{\beta' Z} h(v) d\Lambda(v) + \int_0^\tau \left\{ \int_u^\tau S(v|Z) dv \right\} h(u) e^{\beta' Z} d\Lambda(u) / \mu_{\beta, \Lambda}(Z). \quad (29)$$

Taking $h(\cdot) = 1(\cdot \leq t)$ in (29), we have the score function for Λ

$$\dot{\ell}_2(t; O, \beta, \Lambda) = \int_0^t dN(v) - \int_0^t Y(v) e^{\beta' Z} h(v) d\Lambda(v) + \int_0^t \left\{ \int_u^t S(v|Z) dv \right\} e^{\beta' Z} d\Lambda(u) / \mu_{\beta, \Lambda}(Z). \quad (30)$$

We can write $\dot{\ell}_1(O, \beta, \Lambda) = \int_0^\tau Z d\dot{\ell}_2(t; O, \beta, \Lambda)$.

Let $\dot{\ell}(t, O, \psi) = \{\dot{\ell}_1(O, \psi), \dot{\ell}_2(t, O, \psi)\}$, where $\psi = (\beta, \Lambda)$. We first introduce some notations from the empirical processes theory. Let $\mathbb{P}_n$ be the empirical probability measure. Then

$$U_n(\beta, \Lambda) = \mathbb{P}_n \dot{\ell}(\cdot, O, \psi) = n^{-1} \sum_{i=1}^n \dot{\ell}(\cdot, O_i, \psi).$$

Denote the empirical process by $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - P_0 f)$, where P_0 is the expectation under ψ_0 . Note that $\sqrt{n}(U_n - U_0)(\psi) = \mathbb{G}_n \dot{\ell}(\cdot, O, \psi)$ is the empirical process indexed by the class of functions

$$\{\dot{\ell}(t; O, \psi), \psi \in \mathcal{B} \times \bar{\mathcal{A}}, t \in [0, \tau]\}.$$

Let the norm $\|\cdot\|_{\mathcal{H}}$ on $\mathcal{H} = \mathcal{B} \times \bar{\mathcal{A}}$ be defined as $\|(\beta, \Lambda)\|_{\mathcal{H}} = \|\beta\| + \|\Lambda\|_{\mathcal{V}}$. Then the stochastic condition that we want to confirm is

$$\left\| \mathbb{G}_n \dot{\ell}(t, O, \widehat{\psi}_n) - \mathbb{G}_n \dot{\ell}(t, O, \psi_0) \right\|_{\mathcal{H}} = o_p(1).$$

To apply the functional central limit theory, we show that the class of functions $\{\dot{\ell}(t, O, \psi) - \dot{\ell}(t, O, \psi_0) : \|\psi\|_{\mathcal{H}} < \delta, t \in [0, \tau]\}$ is P_0 -Donsker. As $\dot{\ell}_1(O, \psi) = \int_0^\tau Z d\dot{\ell}_2(t, O, \psi)$, we only need to show that $\{\dot{\ell}_2(t, O, \psi) - \dot{\ell}_2(t, O, \psi_0) : \|\psi\|_{\mathcal{H}} < \delta, t \in [0, \tau]\}$ is P_0 -Donsker, where

$$\begin{aligned} \dot{\ell}_2(t, O, \psi) - \dot{\ell}_2(t, O, \psi_0) &= \int_0^t Y(u) e^{\beta_0' Z} d\Lambda_0(u) - \int_0^t Y(u) e^{\beta' Z} d\Lambda(u) \\ &+ \int_0^t \left\{ \int_u^\tau S(v|Z) dv \right\} e^{\beta' Z} d\Lambda(u) / \mu_Z(\psi) - \int_0^t \left\{ \int_u^\tau S_0(v|Z) dv \right\} e^{\beta_0' Z} d\Lambda_0(u) / \mu_Z(\psi_0). \end{aligned}$$

First, the class of functions $\exp(\beta' Z)$ for β in the compact set $\mathcal{B}$ is P_0 -Donsker as it is of finite dimension. The class of functions of bounded variation on $[0, \tau]$ is P_0 -Donsker (van der Vaart, 1998, page 273). As the function $\Lambda(t)$ is of bounded variation on $[0, \tau]$, the set of functions $\{\int_0^t Y(u) e^{\beta' Z} d\Lambda, t \in [0, \tau], \beta, \Lambda\}$ is P_0 -Donsker. Similarly, we can show that the sets of functions $\{\int_0^t [\int_u^\tau S(v|Z) dv] e^{\beta' Z} d\Lambda(u), t \in [0, \tau], \beta \in \mathcal{B}, \Lambda \in \bar{\mathcal{A}}\}$ and $\{\mu_Z(\beta, \Lambda), \beta \in \mathcal{B}, \Lambda \in \bar{\mathcal{A}}\}$ are P_0 -Donsker. Note that their envelope functions are $(\tau - u) e^{\beta' Z} \Lambda(\tau)$ and $\tau < \infty$, respectively. By Assumption 4, $(\tau - u)^2 \Lambda^2(\tau) E_0 e^{2\beta' Z} < \infty$ for all β and $\Lambda \in \bar{\mathcal{A}}$. By the permanent property of Donsker classes, we proceed to confirm the Donsker property by using the fact that the sums, production and Lipschitz transformations of simple P_0 -Donsker classes are still P_0 -Donsker.

Furthermore, $\|\psi - \psi_0\|_{\mathcal{H}} \rightarrow 0$, $\ell_2(t, \mathcal{O}, \psi)$ converges to $\ell_2(t, \mathcal{O}, \psi_0)$ for each t . The convergence also holds in the square moment by the dominated convergence theorem. It follows that

$$\sup_{t \in [0, \tau]} E_0 \left\| \dot{\ell}(t, \mathcal{O}, \psi) - \dot{\ell}(t, \mathcal{O}, \psi_0) \right\|_{\mathcal{H}}^2 \rightarrow 0.$$

The stochastic approximation of $\sqrt{n}(\mathbf{U}_n - \mathbf{U}_0)(\widehat{\beta}_n, \widehat{\Lambda}_n)$ to $\sqrt{n}(\mathbf{U}_n - \mathbf{U}_0)(\beta_0, \Lambda_0)$ now follows by a technical lemma of van der Vaart and Wellner (1996, Lemma 3.3.5, p. 311).

Figure 1.
Estimated survival curves according to subtypes of dementia, with and without the constraint of increasing risk of death with age

Table 1
Summary Statistics of Simulations for the Estimated Distribution Function with Increasing Failure Rate

Sample Size	C%	F(t)	$\hat{F}_p(t)$		$\hat{F}_c(t)$	
			Est.	ESD	Est.	ESD
200	10%	0.10	0.060	0.024	0.066	0.032
		0.25	0.211	0.044	0.211	0.045
		0.50	0.486	0.045	0.469	0.046
		0.75	0.745	0.030	0.732	0.034
200	30%	0.90	0.898	0.018	0.892	0.020
		0.10	0.068	0.028	0.065	0.033
		0.25	0.222	0.045	0.210	0.046
		0.50	0.490	0.046	0.468	0.049
400	10%	0.75	0.748	0.031	0.732	0.036
		0.90	0.901	0.019	0.892	0.022
		0.10	0.065	0.019	0.075	0.025
		0.25	0.217	0.031	0.221	0.033
400	30%	0.50	0.495	0.033	0.478	0.033
		0.75	0.748	0.021	0.738	0.023
		0.90	0.899	0.012	0.895	0.014
		0.10	0.080	0.021	0.075	0.025
400	30%	0.25	0.236	0.033	0.221	0.033
		0.50	0.501	0.033	0.477	0.034
		0.75	0.751	0.022	0.737	0.025
		0.90	0.901	0.013	0.895	0.016

Note: $\hat{F}_p(t)$ is the proposed estimator; $\hat{F}_c$ is Tsai's conditional estimator; C%=censoring percentage; Est.= average of estimates; ESD = empirical standard deviation; ESMSE = empirical square root of the mean squared error = $(\text{bias}^2 + \text{ESD}^2)^{1/2}$.

Table 2

Summary Statistics of Simulations for Estimating Regression Coefficients under Cox Model with $\beta = (\beta_1, \beta_2) = (0.5, 1)$. Mean SE is the mean of the estimated standard errors

Cohort Size	C%	Proposed approach			EE - 1		EE - 2		Tsai's Method	
		Est.	ESD	Mean SE	95 % CP	Est.	ESD	Est.	ESD	ESD
200	15%	(.49,.98)	(.11,.20)	(.11,.19)	(.96,.95)	(.50,1.01)	(.14,.25)	(.51,1.04)	(.13,.24)	(.51,1.01)
	30%	(.48,.94)	(.11,.21)	(.11,.20)	(.94,.93)	(.47,.91)	(.19,.33)	(.51,1.01)	(.16,.28)	(.50,1.01)
	50%	(.46,.93)	(.12,.21)	(.12,.20)	(.93,.94)	(.42,.85)	(.23,.40)	(.51,1.02)	(.19,.34)	(.49,1.01)
400	15%	(.49,.98)	(.08,.14)	(.08,.14)	(.95,.95)	(.50,.99)	(.10,.18)	(.51,1.02)	(.09,.17)	(.52,1.01)
	30%	(.48,.97)	(.08,.15)	(.08,.14)	(.93,.93)	(.46,.94)	(.14,.24)	(.50,1.01)	(.11,.21)	(.50,1.00)
	50%	(.48,.94)	(.08,.15)	(.08,.15)	(.94,.92)	(.42,.84)	(.18,.31)	(.51,1.02)	(.14,.24)	(.49,1.01)
$\lambda_c = t \exp(0.5Z_1 + 0.5Z_2)$										
200	30%	(.48,.95)	(.11,.20)	(.11,.20)	(.94,.93)	(.39,.89)	(.18,.31)	(.46,.97)	(.15,.27)	(.58,1.07)
400	30%	(.48,.97)	(.08,.15)	(.08,.14)	(.96,.96)	(.39,.91)	(.13,.23)	(.45,.97)	(.11,.20)	(.58,1.08)
$\lambda_c = t \exp(Z_2)$										
200	30%	(.48,.96)	(.11,.20)	(.11,.20)	(.95,.94)	(.52,.82)	(.18,.31)	(.53,.92)	(.15,.28)	(.49,1.15)
400	30%	(.48,.96)	(.08,.15)	(.08,.14)	(.93,.93)	(.50,.80)	(.12,.22)	(.51,.91)	(.11,.19)	(.50,1.15)
$C \sim Z_1 U(0, 1) + (1 - Z_1) U(0, 1.8)$										
200	30%	(.49,.96)	(.12,.21)	(.12,.20)	(.95,.93)	(.82,.92)	(.20,.33)	(.71,1.00)	(.17,.28)	(.41,1.02)
400	30%	(.48,.96)	(.09,.14)	(.08,.14)	(.92,.95)	(.82,.89)	(.11,.22)	(.69,.97)	(.12,.19)	(.70,.96)

Table 3

Estimates (Standard Errors) of Regression Coefficients Using Length-biased Adjusted Methods for Dementia Data.

	MLE	EE-I	EE-II	EE-PL
Probable Alzheimer	0.125 (0.062)	0.109 (0.092)	0.134 (0.091)	0.064(0.081)
Vascular Dementia	0.185 (0.077)	0.245 (0.110)	0.208 (0.110)	0.164(0.111)