Journal of Multivariate Analysis 143 (2016) 327–344

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The Fine–Gray model under interval censored competing risks data

Chenxi Li

Department of Epidemiology and Biostatistics, Michigan State University, East lansing, MI 48824, USA

ARTICLE INFO

Article history:
Received 30 March 2014
Available online 22 October 2015

AMS subject classifications:
62G20
62E20
62N02
62P10

Keywords:
Competing risk
Cumulative incidence function
Interval censored data
Subdistribution hazard
Semiparametric efficiency
Sieve estimation

ABSTRACT

We consider semiparametric analysis of competing risks data subject to mixed case interval censoring. The Fine–Gray model (Fine and Gray, 1999) is used to model the cumulative incidence function and is coupled with sieve semiparametric maximum likelihood estimation based on univariate or multivariate likelihood. The univariate likelihood of cause-specific data enables separate estimation of cumulative incidence function for each competing risk, in contrast with the multivariate likelihood of full data which estimates cumulative incidence functions for multiple competing risks jointly. Under both likelihoods and certain regularity conditions, we show that the regression parameter estimator is asymptotically normal and semiparametrically efficient, although the spline-based sieve estimator of the baseline cumulative subdistribution hazard converges at a rate slower than root-n. The proposed method is evaluated by simulation studies regarding its finite sample performance and is illustrated by a competing risk analysis of data from a dementia cohort study.

1. Introduction

Interval censored failure time data arise widely from longitudinal studies where the failure occurrence is only detectable at periodic assessments, yielding the time to failure being known up to an interval. An example of such data is from the PAQUID project [20], where the age of onset of dementia on incident cases was only known to be between the date of the visit last seen without dementia and the date of the visit first seen with dementia. Another example of such data is from the Breastfeeding, Antiretrovirals, and Nutrition (BAN) study [5], where the time to HIV-1 infection of an infant breast-fed by its HIV-1 positive mother is only known up to being between two successive HIV tests. Besides interval censoring, competing risk is another common issue in the statistical analysis of failure time data. Competing risks data arise when subjects may fail from several dependent causes but the occurrence of failure from one cause precludes the occurrence of failure from the others or only time to the first failure occurrence is of interest. For example, in the PAQUID study, a subject may die before the onset of dementia, so death is a competing risk that precludes the occurrence of dementia. Also, in the BAN study, an infant may be weaned before getting HIV infected, which greatly reduces the risk of infection; or an infant could die without being HIV infected, which totally precludes the infection. Therefore, only the time to the first event of weaning, death and HIV infection is of clinical interest with respect to mother-to-child HIV transmission.

Competing risks data have drawn a great deal of attention since 1970s. Most of the works in the literature summarize competing risks data by the cause-specific hazard function and/or the cumulative incidence function. The former quantifies the instantaneous risk of failure from a specific cause, while the latter quantifies the cumulative risk of failure from a

E-mail address: cli@epi.msu.edu.

http://dx.doi.org/10.1016/j.jmva.2015.10.001
0047-259X/

competing risk. There have been a lot of works on the estimation of these two quantities with right censored competing risks data. The cause-specific hazard function may be estimated nonparametrically by the increment of Nelson–Aalen estimator censoring individuals who experience competing causes of failure, and the cumulative incidence function may be estimated nonparametrically by the Aalen–Johansen estimator [1]. To incorporate covariates, one can estimate the cause-specific hazard function semiparametrically by the standard Cox proportional hazards model treating failures from competing causes as independent right censoring events [23], or estimate the cumulative incidence function semiparametrically by the Fine–Gray model [7] or other semiparametric transformation models [6]. However, there are not as many published works on analyzing interval censored competing risks data, which are seen quite often in longitudinal medical studies. For the estimation of the cumulative incidence function with interval censored competing risks data, Hudgens et al. [15] derived an algorithm to compute the nonparametric maximum likelihood estimator, and Jewell et al. [17] proposed a simpler naive nonparametric maximum likelihood estimator based on reduced current status data for the cause of interest, which could be naturally extended to mixed case interval censored data. Groeneboom et al. [8] and Groeneboom et al. [9] rigorously derived the asymptotic properties of these two nonparametric cumulative incidence estimators for current status competing risks data in terms of consistency, convergence rates and limiting distributions. For the estimation of the cause-specific hazard function with interval censored competing risks data, Li and Fine [21] proposed two cause-specific hazard estimators for current status competing risks data based on smoothing the nonparametric cumulative incidence estimators and the plug-in principle, and derived their asymptotic distributions. The computation of the cause-specific hazard estimators in [21] could be naturally extended to mixed case interval censored data. Despite these works, several important research areas on interval censored competing risks data are still open. For example, the large sample properties of the nonparametric maximum likelihood estimator of the cumulative incidence function [15] and the nonparametric estimators of the cause specific hazard [21] have not been derived for mixed case interval censored data with competing risks; the semiparametric estimation of the cause-specific hazard and the cumulative incidence with interval censored competing risks data has not been rigorously studied yet.

Unlike right censored competing risks data, the likelihood of interval censored competing risks data can be formulated concisely by just cumulative incidence functions [14, Lemma 1]. Furthermore, the likelihood of reduced mixed case interval censored data for the cause of interest involves only the cumulative incidence function of that cause [14, Lemma 2]. Such properties make models that directly link the cumulative incidence to explanatory variables appealing in the analysis of interval censored competing risks data. The most popular model of this kind is the Fine–Gray model, which however has been only applied to competing risks data under right censoring so far.

In this article, we study the semiparametric inference for the Fine–Gray model with mixed case interval censored competing risks data where observation times are strictly separated. Inspired by Zhang et al. [30], we employ a monotone B-spline [25] to approximate the log baseline cumulative subdistribution hazard, reducing the dimension of the parameter space to a finite number. With this approximation, maximum likelihood estimators of the regression parameters and the B-spline coefficients are obtained by an adaptive barrier algorithm [19]. Under certain regularity conditions, we establish that the proposed estimators of the regression parameter and the baseline cumulative subdistribution hazard are consistent. Moreover, we show that the estimator of the baseline cumulative subdistribution hazard achieves the optimal convergence rate as in the nonparametric regression setting, and the estimator of the regression parameter is asymptotically normal and semiparametrically efficient. When there are more than one competing risk under consideration, one could perform a univariate analysis modeling each competing risk separately or a multivariate analysis modeling multiple competing risks jointly. Through simulation studies, we show that the multivariate analysis is more efficient than the univariate analysis for estimating the regression parameters when sample size is large.

The rest of the article is organized as follows. Section 2 specifies the Fine–Gray model and its associated likelihood for interval censored competing risks data. Section 3 describes an spline-based maximum likelihood estimation for the model. Section 4 derives the information matrix for the regression parameter in the semiparametric inference setting. Section 5 presents the asymptotic properties of the maximum likelihood estimators of baseline cumulative subdistribution hazards and regression parameters. The finite sample performance of the proposed method is then evaluated by simulation studies in Section 6, followed by a real data example in Section 7 that illustrates the application of the method. Finally, we give some discussion on our method and future research directions in Section 8. The proofs of the asymptotic results are collected in the Appendix.

2. Model specification

We assume that a Fine–Gray model holds for each of J competing risks under consideration. Specifically, the cumulative incidence function of cause k given a vector of covariates $\mathbf{Z} \in \mathbb{R}^d$, $F_k(t; \mathbf{Z})$, is modeled as:

$$F_k(t; \mathbf{Z}) = 1 - \exp\left\{-\int_0^t \lambda_{k0}(s) \exp(\mathbf{Z}^T \boldsymbol{\theta}_k) ds\right\} \quad k = 1, \ldots, J, \tag{1}$$

i.e., the subdistribution hazard of $F_k(t; \mathbf{Z})$ is $\lambda_k(t; \mathbf{Z}) = \lambda_{k0}(t) \exp(\mathbf{Z}^T \boldsymbol{\theta}_k)$. For the identifiability of the model parameters, we assume throughout:

A0. $\sum_{k=1}^J F_k(\infty; \mathbf{0}) < 1$.

Remark 1. When $J = 1$, the upper bound one is allowed to be achieved, in which case there is actually no other failure cause and the Fine–Gray model becomes the proportional hazards model. The estimation and inferential procedures for that case are discussed in [30].

Remark 2. The total number of competing risks could be equal to or larger than J. When it is equal to J, A0 means that there is a cured fraction of the population.

Define the following notations:

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_J^T)^T \in \mathbb{R}^{Jd},$$

$$\Lambda_{k0}(t) = \int_0^t \lambda_{k0}(s)ds, \qquad \phi_{k0}(t) = \log \Lambda_{k0}(t), \quad k = 1, \ldots, J,$$

and

$$\boldsymbol{\phi} = (\phi_{10}(\cdot), \ldots, \phi_{J0}(\cdot)).$$

For subject i, $i = 1, \ldots, n$, denote its underlying failure time by T_i, its failure cause by K_i and its covariate vector by $\mathbf{Z}_i$. We consider the situation where the failure time is not observed exactly but subject to 'mixed case' interval censoring as defined in [24]. Let M_i be a random positive integer denoting the number of examination times for failure on subject i and $W_i = \{W_{m,l}^{(i)} : l = 1, \ldots, m, m = 1, 2, \ldots\}$ be a triangular array of random examination times with $W_{m,l}^{(i)} < \cdots < W_{m,m}^{(i)}$. The observed examination times are $W_{M_i,1}^{(i)}, \ldots, W_{M_i,M_i}^{(i)}$. From the observed examination times, we select a pair of examination times (U_i, V_i) bracketing T_i as follows. $(U_i, V_i) = (W_{M_i,l-1}^{(i)}, W_{M_i,l}^{(i)})$ if $T_i \in (W_{M_i,l-1}^{(i)}, W_{M_i,l}^{(i)}]$ for some $l \in \{2, \ldots, M_i\}$; $(U_i, V_i) = (W_{M_i,1}^{(i)}, W_{M_i,2}^{(i)})$ if $T_i \le W_{M_i,1}^{(i)}$; $(U_i, V_i) = (W_{M_i,M_i-1}^{(i)}, W_{M_i,M_i}^{(i)})$ if $T_i > W_{M_i,M_i}^{(i)}$. Let $\Delta_{k1}^{(i)} = \mathrm{I}\{T_i \le U_i, K_i = k\}$, $\Delta_{k2}^{(i)} = \mathrm{I}\{U_i < T_i \le V_i, K_i = k\}$ and $\Delta_3^{(i)} = 1 - \sum_{k=1}^J(\Delta_{k1}^{(i)} + \Delta_{k2}^{(i)})$. The observable data related to the J competing risks from subject i consist of $(M_i, W_{M_i,1}^{(i)}, \ldots, W_{M_i,M_i}^{(i)}, \mathbf{Z}_i)$, whether any of the J types of failure has occurred by the time of each examination, and K_i if a failure has occurred by the last examination time. In light of assumption A0, we define a convention that $K_i = J + 1$ if no failure from any of the J causes occurs to subject i, which happens with a positive probability.

Under mixed case interval censoring, (M, W) are independent of (T, K) conditional on $\mathbf{Z}$ and the joint distribution of $(M, W, \mathbf{Z})$ does not depend on $(\boldsymbol{\theta}, \boldsymbol{\phi})$. Thus, by Lemma 1 and 2 in [14], the likelihood of the data related to the J competing risks, omitting the multiplicative terms that do not involve $(\boldsymbol{\theta}, \boldsymbol{\phi})$, can be written as

$$L_n(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^n \left[\prod_{k=1}^J F_k(U_i; \mathbf{Z}_i)^{\Delta_{k1}^{(i)}} \{F_k(V_i; \mathbf{Z}_i) - F_k(U_i; \mathbf{Z}_i)\}^{\Delta_{k2}^{(i)}}\right] \left\{1 - \sum_{k=1}^J F_k(V_i; \mathbf{Z}_i)\right\}^{\Delta_3^{(i)}}. \tag{2}$$

Note that the data involved in the likelihood (2) are only composed of $\mathbf{Y}_i = (\Delta_{11}^{(i)}, \Delta_{12}^{(i)}, \Delta_{21}^{(i)}, \Delta_{22}^{(i)}, \ldots, \Delta_{J1}^{(i)}, \Delta_{J2}^{(i)}, \Delta_3^{(i)}, U_i, V_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, not the full data, since $\mathbf{Y}_i$'s contain sufficient information for the model parameters. In order to facilitate deriving the large sample properties of maximum likelihood estimators for $(\boldsymbol{\theta}, \boldsymbol{\phi})$, we make a working assumption that (U, V) are independent of (T, K) conditional on $\mathbf{Z}$ and the joint distribution of $(U, V, \mathbf{Z})$ is uninformative about $(\boldsymbol{\theta}, \boldsymbol{\phi})$. This assumption was also assumed in [11,30] for mixed case interval censored data except not involving failure cause K. It is satisfied under case 2 interval censoring but may not hold in general. However, in view of the equivalence between (2) and the likelihood written in terms of the full data, we can derive the same large sample properties under the likelihood (2) and the working assumption as derived under the likelihood in terms of the full data, and the former derivation would be less complicated.

The log likelihood of (2) is

$$l_n(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^n \left[\sum_{k=1}^J \left[\Delta_{k1}^{(i)} \log F_k(U_i; \mathbf{Z}_i) + \Delta_{k2}^{(i)} \log\{F_k(V_i; \mathbf{Z}_i) - F_k(U_i; \mathbf{Z}_i)\}\right] + \Delta_3^{(i)} \log \left\{1 - \sum_{k=1}^J F_k(V_i; \mathbf{Z}_i)\right\}\right],$$

which, under model (1), is

$$\begin{aligned} l_n(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \sum_{i=1}^n \left[\sum_{k=1}^J \left[\Delta_{k1}^{(i)} \log\left[1 - \exp\left\{-e^{\mathbf{Z}_i^T\boldsymbol{\theta}_k + \phi_{k0}(U_i)}\right\}\right] + \Delta_{k2}^{(i)} \log\left[\exp\left\{-e^{\mathbf{Z}_i^T\boldsymbol{\theta}_k + \phi_{k0}(U_i)}\right\}\right.\right.\right. \\ &\quad \left.\left. - \exp\left\{-e^{\mathbf{Z}_i^T\boldsymbol{\theta}_k + \phi_{k0}(V_i)}\right\}\right]\right] + \Delta_3^{(i)} \log\left[1 - J + \sum_{k=1}^J \exp\left\{-e^{\mathbf{Z}_i^T\boldsymbol{\theta}_k + \phi_{k0}(V_i)}\right\}\right]\Bigg] \\ &\equiv \sum_{i=1}^n l(\mathbf{Y}_i; \boldsymbol{\theta}, \boldsymbol{\phi}). \end{aligned} \tag{3}$$

3. Sieve maximum likelihood estimation

To estimate the parameters $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ in model (1), we consider a spline-based semiparametric maximum likelihood estimation. The estimation procedure extends that of Zhang et al. [30] for univariate interval censored data to competing risks setting. Suppose τ_0 and τ_1 are respectively the lower and upper bounds of the bracketing examination times (U, V). Let $\tau_0 = d_{k0} < d_{k1} < \cdots < d_{k,K_{kn}} < d_{k,K_{kn}+1} = \tau_1$ $(k = 1, \ldots, J)$ be J partitions of $[\tau_0, \tau_1]$, where $K_{kn} = O(n^{\nu_k})$ for some $\nu_k > 0$ is a positive integer such that $\max_{1\le j\le K_{kn}+1} |d_j - d_{j-1}| = O(n^{-\nu_k})$. Denote the kth set of partition points by $D_{kn} = \{d_{k1}, \ldots, d_{k,K_{kn}}\}$. Let $\mathcal{S}_{kn}(D_{kn}, K_{kn}, m_k)$ be the space of polynomial splines of order $m_k \ge 1$ comprising functions s satisfying: (i) the restriction of s to $[d_{kt}, d_{k,t+1})$, $t = 0, \ldots, K_{kn}$, is a polynomial of order m_k, and (ii) for $m_k \ge 2$ and $0 \le m'_k \le m_k - 2$, s is m'_k times continuously differentiable on $[\tau_0, \tau_1]$. According to Schumaker [25, Corollary 4.10], there is a local basis $\mathcal{B}_{kn} \equiv \{b_{kt}(\cdot), 1 \le t \le q_{kn}\}$, called B-spline, for $\mathcal{S}_{kn}(D_{kn}, K_{kn}, m_k)$, where $q_{kn} \equiv K_{kn} + m_k$. These basis functions are nonnegative and sum up to one at each point in $[\tau_0, \tau_1]$, and every b_{kt} is zero outside the interval $[d_{k,t-m_k}, d_{kt}]$, where $d_{k,1-m_k} = d_{k,2-m_k} = \cdots = d_{k,0} = \tau_0$ and $d_{k,K_{kn}+1} = d_{k,K_{kn}+2} = \cdots = d_{k,K_{kn}+m_k} = \tau_1$.

Since $\phi_{k0}(t)$ is a nondecreasing function, it is desirable to restrict its estimate to be also nondecreasing. Let

$$\mathcal{M}_{kn}(D_{kn}, K_{kn}, m_k) = \left\{ \phi_{kn} : \phi_{kn}(t) = \sum_{j=1}^{q_{kn}} \beta_{kj} b_{kj}(t) \in \mathcal{S}_{kn}(D_{kn}, K_{kn}, m_k), \boldsymbol{\beta}_k \in \mathbf{B}_{kn}, t \in [\tau_0, \tau_1] \right\},$$

where $\mathbf{B}_{kn} = \{\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{k,q_{kn}})^T : \beta_{k1} \le \beta_{k2} \ldots \le \beta_{k,q_{kn}}\}$. As a consequence of the variation diminishing properties of B-spline (see, for example [25], example 4.75 and Theorem 4.76), every element of $\mathcal{M}_{kn}(D_{kn}, K_{kn}, m_k)$ is a nondecreasing function because of the monotonicity constraints on $\beta_{k1}, \ldots, \beta_{k,q_{kn}}$. Denote the feasible domain for the regression parameter $\boldsymbol{\theta}$ by $\boldsymbol{\Theta} \subset \mathbb{R}^{Jd}$. We search for $\hat{\boldsymbol{\zeta}}_n = (\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)$ that maximizes $l_n(\boldsymbol{\theta}, \boldsymbol{\phi})$ over $\boldsymbol{\Theta} \times \prod_{k=1}^J \mathcal{M}_{kn}$, which is equivalent to maximizing $l_n(\boldsymbol{\theta}, \mathcal{B}_{1n}^T\boldsymbol{\beta}_1, \ldots, \mathcal{B}_{Jn}^T\boldsymbol{\beta}_J)$ over $\boldsymbol{\Theta} \times \prod_{k=1}^J \mathbf{B}_{kn}$. In practice, we use an adaptive barrier algorithm [19] to maximize $l_n(\boldsymbol{\theta}, \mathcal{B}_{1n}^T\boldsymbol{\beta}_1, \ldots, \mathcal{B}_{Jn}^T\boldsymbol{\beta}_J)$ subject to the linear inequality constraints. It is implemented by the R function 'constrOptim'. In the subsequent simulation study and real data example, we set the initial values of the regression parameters for the optimization to be all 0 and the initial value of β_{kj} to be $\mu + (j-1)\delta_k$, where $\mu = \log\{-\log(1 - 0.15J^{-1})\}$ and $\delta_k = [\log\{-\log(1 - 0.85J^{-1})\} - \mu]/(q_{kn} - 1)$ $(j = 1, \ldots, q_{kn};\ k = 1, \ldots, J)$.

4. Information for θ

We would like to show that the estimate of the regression parameter $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_n$, is semiparametrically efficient [18, Section 3.1]. To prove that, we first derive the efficient score function and information matrix for $\boldsymbol{\theta}$ [18, Section 3.2]. Our derivation is similar to that in [12].

By some algebra, the score function for $\boldsymbol{\theta}$ is

$$\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}) = \frac{\partial l(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} [\Delta_{11}A_{11}(U,V,\mathbf{Z})\Lambda_{10}(U) - \Delta_{12}\{A_{12}(U,V,\mathbf{Z})\Lambda_{10}(U) - A_{13}(U,V,\mathbf{Z})\Lambda_{10}(V)\} \\ -\Delta_3 A_{14}(U,V,\mathbf{Z})\Lambda_{10}(V)]\, e^{\mathbf{Z}^T\boldsymbol{\theta}_1}\mathbf{Z} \\ \vdots \\ [\Delta_{J1}A_{J1}(U,V,\mathbf{Z})\Lambda_{J0}(U) - \Delta_{J2}\{A_{J2}(U,V,\mathbf{Z})\Lambda_{J0}(U) - A_{J3}(U,V,\mathbf{Z})\Lambda_{J0}(V)\} \\ -\Delta_3 A_{J4}(U,V,\mathbf{Z})\Lambda_{J0}(V)]\, e^{\mathbf{Z}^T\boldsymbol{\theta}_J}\mathbf{Z} \end{pmatrix}, \tag{4}$$

where

$$A_{k1}(U,V,\mathbf{Z}) = \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(U)\right\} \Big/ \left[1 - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(U)\right\}\right],$$

$$A_{k2}(U,V,\mathbf{Z}) = \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(U)\right\} \Big/ \left[\exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(U)\right\} - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(V)\right\}\right],$$

$$A_{k3}(U,V,\mathbf{Z}) = \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(V)\right\} \Big/ \left[\exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(U)\right\} - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(V)\right\}\right],$$

and

$$A_{k4}(U,V,\mathbf{Z}) = \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k}\Lambda_{k0}(V)\right\} \Big/ \left[1 - J + \sum_{j=1}^{J}\exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_j}\Lambda_{j0}(V)\right\}\right]$$

for $k = 1, \ldots, J$.

Let $\mathscr{P}$ be the statistical model consisting of probability measures on the sample space $\mathscr{Y}$ of $\mathbf{Y}$, and it can be expressed in the form $\mathscr{P} = \{P_{\boldsymbol{\theta},\boldsymbol{\phi}} : (\boldsymbol{\theta}, \boldsymbol{\phi}) \in \boldsymbol{\Theta} \times \prod_{k=1}^{J} \Phi_k\}$, where Φ_k is a class of functions with bounded p_kth derivative in $[\tau_0, \tau_1]$ for $p_k \geq 1$ $(k = 1, \dots, J)$. Consider its one-dimensional submodels of the form $\{P_{\boldsymbol{\theta},\boldsymbol{\phi}^{(s)}} : s \in [0, \epsilon)\}$, where the cumulative incidence function corresponding to $\boldsymbol{\phi}^{(s)}$'s kth component $\phi_{k0}^{(s)}$ $(k = 1, \dots, J)$ is $F_{k0}^{(s)}(t) = \int_0^t (1 + sa(w, k))dF_{k0}(w)$ for some function $a(w, k)$ that is a function of w with bounded variation on $\mathbb{R}_+$ when holding k fixed, and $\epsilon > 0$ depends on $a(\cdot, \cdot)$. Define a convention that

$$a(w, J+1) = \begin{cases} \left\{1 - \sum_{k=1}^{J} F_{k0}(\infty)\right\}^{-1} \left.\dfrac{\partial}{\partial s}\right|_{s=0} \left\{1 - \sum_{k=1}^{J} F_{k0}^{(s)}(\infty)\right\} & \text{if } w = \infty, \\ 0 & \text{if } w < \infty \end{cases}$$

in order for the score operator for $\boldsymbol{\phi}$, derived from the score functions of the above one-dimensional submodels, to be a linear continuous map $\dot{l}_{\boldsymbol{\phi}} : L_2^0(F_{(T,K)|\mathbf{Z}=0}) \mapsto L_2^0(P_{\boldsymbol{\theta},\boldsymbol{\phi}})$, where $F_{(T,K)|\mathbf{Z}}$ denote the conditional distribution of (T, K) given $\mathbf{Z}$, and $L_2^0(Q)$ denote the function space $\{f : \int f^2 dQ < \infty \text{ and } \int f dQ = 0\}$ for any measure Q. Specifically, for any $a \in L_2^0(F_{(T,K)|\mathbf{Z}=0})$,

$$\begin{aligned} \dot{l}_{\boldsymbol{\phi}} a(\mathbf{Y}) = \left.\frac{\partial}{\partial s}\right|_{s=0} l(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi}^{(s)}) &= \sum_{k=1}^{J} [\Delta_{k1} A_{k1}(U, V, \mathbf{Z}) h_k(U) - \Delta_{k2}\{A_{k2}(U, V, \mathbf{Z}) h_k(U) \\ &\quad - A_{k3}(U, V, \mathbf{Z}) h_k(V)\} - \Delta_3 A_{k4}(U, V, \mathbf{Z}) h_k(V)]\, e^{\mathbf{Z}^T \boldsymbol{\theta}_k} \\ &\equiv \sum_{k=1}^{J} \dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\eta_k) \\ &\equiv \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\boldsymbol{\eta}), \end{aligned} \tag{5}$$

where $h_k(t) = \frac{\partial}{\partial s}\big|_{s=0} \Lambda_{k0}^{(s)}(t) = \int_0^t a(w, k) dF_{k0}(w)/\{1 - F_{k0}(t)\}$, $\eta_k(t) = \frac{\partial}{\partial s}\big|_{s=0} \phi_{k0}^{(s)}(t) = h_k(t)/\Lambda_{k0}(t)$ $(k = 1, \dots, J)$ and $\boldsymbol{\eta} = (\eta_1(\cdot), \dots, \eta_J(\cdot))$. Let $\mathscr{N}$ be the class of functions $\eta_k(t)$.

Since $L_2^0(F_{(T,K)|\mathbf{Z}=0})$ and $L_2^0(P_{\boldsymbol{\theta},\boldsymbol{\phi}})$ are two Hilbert spaces with respective inner products $\langle \cdot, \cdot \rangle_{F_{(T,K)|\mathbf{Z}=0}}$ and $\langle \cdot, \cdot \rangle_{P_{\boldsymbol{\theta},\boldsymbol{\phi}}}$, where $\langle f, g \rangle_Q = \int fg dQ$, there is an adjoint operator $\dot{l}_{\boldsymbol{\phi}}^* : L_2^0(P_{\boldsymbol{\theta},\boldsymbol{\phi}}) \mapsto L_2^0(F_{(T,K)|\mathbf{Z}=0})$ satisfying $\langle b, \dot{l}_{\boldsymbol{\phi}} a \rangle_{P_{\boldsymbol{\theta},\boldsymbol{\phi}}} = \langle \dot{l}_{\boldsymbol{\phi}}^* b, a \rangle_{F_{(T,K)|\mathbf{Z}=0}}$ for any $b \in L_2^0(P_{\boldsymbol{\theta},\boldsymbol{\phi}})$ and $a \in L_2^0(F_{(T,K)|\mathbf{Z}=0})$. To determine the efficient score function for $\boldsymbol{\theta}$, for each of the Jd components in $\dot{l}_{\boldsymbol{\theta}}$, $\dot{l}_{\theta_{lj}}$ $(l = 1, \dots, J;\ j = 1, \dots, d)$, we need to find an $a_{lj}^* \in L_2^0(F_{(T,K)|\mathbf{Z}=0})$ such that $\langle \dot{l}_{\theta_{lj}} - \dot{l}_{\boldsymbol{\phi}} a_{lj}^*, \dot{l}_{\boldsymbol{\phi}} a \rangle_{P_{\boldsymbol{\theta},\boldsymbol{\phi}}} = 0$ for any $a \in L_2^0(F_{(T,K)|\mathbf{Z}=0})$. This amounts to solving the following normal equations:

$$\dot{l}_{\boldsymbol{\phi}}^* \dot{l}_{\boldsymbol{\phi}} a_{lj}^* = \dot{l}_{\boldsymbol{\phi}}^* \dot{l}_{\theta_{lj}}, \quad l = 1, \dots, J;\ j = 1, \dots, d. \tag{6}$$

To solve for a_{lj}^*, we need to evaluate both sides of (6). In the remainder of this section, we assume $d = 1$ to simplify the notations in the derivation of a_{lj}^*. The derivation for the general case is similar. As a result, all the notations θ_{lj} $(j = 1, \dots, d)$ can be simplified to a single scalar θ_l for $l = 1, \dots, J$, a_{lj}^* $(j = 1, \dots, d)$ can be simplified to a_l^*, and we change the notation of one-dimensional $\mathbf{Z}$ to Z.

From [3, pp. 271–272], we know that

$$\dot{l}_{\boldsymbol{\phi}}^* \dot{l}_{\boldsymbol{\phi}} a(t, k) = E[\dot{l}_{\boldsymbol{\phi}} a(Y) | T = t, K = k] = E_Z E[\dot{l}_{\boldsymbol{\phi}} a(Y) | T = t,\ K = k, Z].$$

By assumptions A3 and A7 in Section 5,

$$\begin{aligned} E[\dot{l}_{\boldsymbol{\phi}} a(Y) | T = t, K = k, Z = z] &= \int_{u=t}^{\tau_1 - \eta} \int_{v=u+\eta}^{\tau_1} e^{z\theta_k} h_k(u) A_{k1} g(u, v|z) dv du \\ &\quad - \int_{u=\tau_0}^{t} \int_{v=t}^{\tau_1} e^{z\theta_k} \{h_k(u) A_{k2} - h_k(v) A_{k3}\} g(u, v|z) \mathbf{1}[v - u \geq \eta] dv du \\ &\quad - \int_{u=\tau_0}^{t-\eta} \int_{v=u+\eta}^{t} \sum_{l=1}^{J} e^{z\theta_l} A_{l4} h_l(v) g(u, v|z) dv du, \end{aligned}$$

where $g(u, v|z)$ is the conditional density of (U, V) given Z. Let $B_{kj}(u, v) = E_Z[e^{Z\theta_k} A_{kj}(u, v, Z) g(u, v|Z)]$ $(k = 1, \dots, J;\ j = 1, 2, 3, 4)$. We obtain

$$\begin{aligned} L(t, k) \equiv \dot{l}_{\boldsymbol{\phi}}^* \dot{l}_{\boldsymbol{\phi}} a(t, k) &= \int_{u=t}^{\tau_1 - \eta} \int_{v=u+\eta}^{\tau_1} h_k(u) B_{k1}(u, v) dv du \\ &\quad - \int_{u=\tau_0}^{t} \int_{v=t}^{\tau_1} \{h_k(u) B_{k2}(u, v) - h_k(v) B_{k3}(u, v)\} \mathbf{1}[v - u \geq \eta] dv du \\ &\quad - \int_{u=\tau_0}^{t-\eta} \int_{v=u+\eta}^{t} \sum_{l=1}^{J} B_{l4}(u, v) h_l(v) dv du. \end{aligned}$$

Similar calculation yields

$$R_l(t,k) \equiv \dot{l}^*_{\boldsymbol{\phi}} \dot{l}_{\theta_l}(t,k) = \begin{cases} \int_{u=t}^{\tau_1-\eta} \int_{v=u+\eta}^{\tau_1} \Lambda_{l0}(u) C_{l1}(u,v)dvdu \\ - \int_{u=\tau_0}^{t} \int_{v=t}^{\tau_1} \{\Lambda_{l0}(u) C_{l2}(u,v) - \Lambda_{l0}(v) C_{l3}(u,v)\} \mathbf{1}[v-u \geq \eta] dvdu & \text{if } k = l, \\ - \int_{u=\tau_0}^{t-\eta} \int_{v=u+\eta}^{t} \Lambda_{l0}(v) C_{l4}(u,v)dvdu \\ - \int_{u=\tau_0}^{t-\eta} \int_{v=u+\eta}^{t} \Lambda_{l0}(v) C_{l4}(u,v)dvdu & \text{if } k \neq l, \end{cases}$$

where $C_{lj}(u,v) = E_Z[Ze^{Z\theta_l} A_{lj}(u,v,Z) g(u,v|Z)]$ $(l = 1,\dots,J;\ j = 1,2,3,4)$. Eq. (6) implies that

$$L(t,k) = R_l(t,k), \quad t \in \mathbb{R}_+,\ k = 1,\dots,J,\ l = 1,\dots,J. \tag{7}$$

Let $b_k(t) = \int_{t+\eta}^{\tau_1} B_{k1}(t,x)dx + \int_{t+\eta}^{\tau_1} B_{k2}(t,x)dx + \int_{\tau_0}^{t-\eta} B_{k3}(x,t)dx + \int_{\tau_0}^{t-\eta} B_{k4}(x,t)dx$. Then

$$\frac{\partial}{\partial t} L(t,k) = -b_k(t) h_k(t) + \int_{\tau_0}^{t-\eta} h_k(x) B_{k2}(x,t)dx + \int_{t+\eta}^{\tau_1} h_k(x) B_{k3}(t,x)dx - \sum_{m=1,m\neq k}^{J} h_m(t) \int_{\tau_0}^{t-\eta} B_{m4}(x,t)dx.$$

Let $c_l(t) = \int_{t+\eta}^{\tau_1} C_{l1}(t,x)dx + \int_{t+\eta}^{\tau_1} C_{l2}(t,x)dx + \int_{\tau_0}^{t-\eta} C_{l3}(x,t)dx + \int_{\tau_0}^{t-\eta} C_{l4}(x,t)dx$. Then

$$r_l(t,k) \equiv \frac{\partial}{\partial t} R_l(t,k) = \begin{cases} -c_l(t)\Lambda_{l0}(t) + \int_{\tau_0}^{t-\eta} \Lambda_{l0}(x) C_{l2}(x,t)dx + \int_{t+\eta}^{\tau_1} \Lambda_{l0}(x) C_{l3}(t,x)dx & \text{if } k = l, \\ -\Lambda_{l0}(t) \int_{\tau_0}^{t-\eta} C_{l4}(x,t)dx & \text{if } k \neq l. \end{cases}$$

Taking the derivative with respect to t on both sides of (7) and fixing l, we get

$$h_k(t) - \int K_k(t,x) h_k(x)dx + \frac{1}{b_k(t)} \sum_{m=1,m\neq k}^{J} h_m(t) \int_{\tau_0}^{t-\eta} B_{m4}(x,t)dx = d_l(t,k), \quad k = 1,\dots,J, \tag{8}$$

where $K_k(t,x) = [B_{k2}(x,t)\mathbf{1}[\tau_0 \leq x \leq t-\eta] + B_{k3}(t,x)\mathbf{1}[t+\eta \leq x \leq \tau_1]]/b_k(t)$ and $d_l(t,x) = -r_l(t,x)/b_k(t)$.

(8) is a system of Fredholm integral equations of the second kind. There exists a solution $\mathbf{h}^*_l(t) = (h_1^{(l),*}(t),\dots,h_J^{(l),*}(t))^T$. The corresponding a is denoted by a^*_l. Let $\mathbf{a}^* = (a^*_1,\dots,a^*_J)^T$ and $\dot{l}_{\boldsymbol{\phi}} \mathbf{a}^*(\mathbf{y}) = (\dot{l}_{\boldsymbol{\phi}} a^*_1(\mathbf{y}),\dots,\dot{l}_{\boldsymbol{\phi}} a^*_J(\mathbf{y}))^T$. Then $l^*_{\boldsymbol{\theta}}(\mathbf{y}) \equiv \dot{l}_{\boldsymbol{\theta}}(\mathbf{y}) - \dot{l}_{\boldsymbol{\phi}} \mathbf{a}^*(\mathbf{y})$ is the efficient score function for $\boldsymbol{\theta}$, and its covariance matrix $I(\boldsymbol{\theta}) = E[l^*_{\boldsymbol{\theta}}(\mathbf{Y}) l^*_{\boldsymbol{\theta}}(\mathbf{Y})^T]$ is the efficient information matrix for $\boldsymbol{\theta}$. The efficient influence function for $\boldsymbol{\theta}$ is $\tilde{l}_{\boldsymbol{\theta}}(\mathbf{y}) \equiv I(\boldsymbol{\theta})^{-1} l^*_{\boldsymbol{\theta}}(\mathbf{y})$.

5. Large sample properties

To study the large sample properties of the semiparametric maximum likelihood estimator $\hat{\boldsymbol{\zeta}}_n = (\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)$, we make the following assumptions similar to those in [30].

A1. $E(\mathbf{Z}\mathbf{Z}^T)$ is non-singular, and $\mathbf{Z}$ is bounded with probability 1.
A2. $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^{Jd}$.
A3. There exists a positive number η such that $\Pr(V - U \geq \eta) = 1$, and the union of the supports of U and V is contained in an interval $[\tau_0, \tau_1]$, where $0 < \tau_0 < \tau_1 < \infty$, and $0 < \Lambda_{k0}(\tau_0) < \Lambda_{k0}(\tau_1) < \infty$ $(k = 1,\dots,J)$.
A4. $\phi_{k0} = \log \Lambda_{k0}$ belongs to Φ_k $(k = 1,\dots,J)$, a class of functions with bounded p_kth derivative in $[\tau_0, \tau_1]$ for $p_k \geq 1$ and the first derivative of ϕ_{k0} is uniformly positive and continuous on $[\tau_0, \tau_1]$.
A5. The conditional density $g(u,v|\mathbf{z})$ of (U,V) given $\mathbf{Z}$ has bounded partial derivatives with respect to (u,v). The bounds of these partial derivatives do not depend on $(u,v,\mathbf{z})$.
A6. For some $\kappa \in (0,1)$, $x^T \operatorname{var}(\mathbf{Z}|U)x \geq \kappa x^T E(\mathbf{Z}\mathbf{Z}^T|U)x$ and $x^T \operatorname{var}(\mathbf{Z}|V)x \geq \kappa x^T E(\mathbf{Z}\mathbf{Z}^T|V)x$ a.s. for all $x \in \mathbb{R}^d$.
A7. (T,K) and (U,V) are conditionally independent given $\mathbf{Z}$, and the distribution of $(U,V,\mathbf{Z})$ does not involve $(\boldsymbol{\theta}, \boldsymbol{\phi})$.

These conditions, together with A0, guarantee the results in the forthcoming theorems.

Let $\boldsymbol{\Phi} = \prod_{k=1}^{J} \Phi_k$. For any $\boldsymbol{\phi}_1 = \left(\phi_{10}^{(1)},\dots,\phi_{J0}^{(1)}\right)$, $\boldsymbol{\phi}_2 = \left(\phi_{10}^{(2)},\dots,\phi_{J0}^{(2)}\right) \in \boldsymbol{\Phi}$, define

$$\|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|^2_{\boldsymbol{\Phi}} = \sum_{k=1}^{J} \left[E\left\{\phi_{k0}^{(1)}(U) - \phi_{k0}^{(2)}(U)\right\}^2 + E\left\{\phi_{k0}^{(1)}(V) - \phi_{k0}^{(2)}(V)\right\}^2 \right],$$

and for any $\boldsymbol{\zeta}_1 = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\phi}_1)$ and $\boldsymbol{\zeta}_2 = (\boldsymbol{\theta}^{(2)}, \boldsymbol{\phi}_2)$ in the space of $\mathcal{T} = \boldsymbol{\Theta} \times \boldsymbol{\Phi}$, define an L_2-metric:

$$d(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|_{\mathcal{T}} = \left\{\|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2 + \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2\|_{\boldsymbol{\Phi}}^2\right\}^{1/2}.$$

Theorem 1. *For $k = 1, \dots, J$, let $K_{kn} = O(n^{\nu_k})$, where ν_k satisfies the restriction $1/\{2(1 + p_k)\} < \nu_k < 1/(2p_k)$, and $m_k \geq p_k + 2$. Furthermore, suppose that the assumptions* A0–A7 *hold. Denote the true parameters of model (1) by $\boldsymbol{\zeta}_0 = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}_0)$. Then $d(\hat{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_0) = O_p(n^{-\min\{\min_k(p_k\nu_k), (1-\max_k \nu_k)/2\}})$.*

This theorem implies that if $\nu_k = 1/(1 + 2p_k)$ $(k = 1, \dots, J)$, $d(\hat{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_0) = O_p(n^{-\min_k\{p_k/(1+2p_k)\}})$ which is the optimal convergence rate as in the nonparametric regression setting.

Theorem 2. *Under the conditions of Theorem 1, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^{(0)}) \mapsto N(\mathbf{0}, I^{-1}(\boldsymbol{\theta}^{(0)}))$ in distribution, where $I(\boldsymbol{\theta}^{(0)})$ is the efficient information matrix for $\boldsymbol{\theta}^{(0)}$.*

The theorem implies that, although the estimator of the baseline cumulative subdistribution hazard converges at a rate lower than $n^{1/2}$, the estimator of the regression parameter converges to the truth at the usual root-n rate and is semiparametrically efficient. Because $I(\boldsymbol{\theta}^{(0)})$ is determined by a system of integral equations and does not have an explicit expression as shown in Section 4, direct plug-in estimator of $I(\boldsymbol{\theta}^{(0)})$ is not available. We give an approach of estimating $I(\boldsymbol{\theta}^{(0)})$ which is parallel to that of Zhang et al. [30].

For a Jd-dimensional $\boldsymbol{\theta}$, $\dot{l}_{\boldsymbol{\theta}}(\mathbf{y})$ is the vector of partial derivatives of $l(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to the components of $\boldsymbol{\theta}$. For each component of $\dot{l}_{\boldsymbol{\theta}}(\mathbf{y})$, we consider a corresponding score function for $\boldsymbol{\phi}$, $\dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\boldsymbol{\eta})$, as defined in (5). Thus the score vector for $\boldsymbol{\phi}$ corresponding to $\dot{l}_{\boldsymbol{\theta}}(\mathbf{y})$ is

$$\dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\vec{\boldsymbol{\eta}}) \equiv \{\dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\boldsymbol{\eta}^{(11)}), \dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\boldsymbol{\eta}^{(12)}), \dots, \dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\boldsymbol{\eta}^{(Jd)})\}^T, \tag{9}$$

where $\vec{\boldsymbol{\eta}} = \{\boldsymbol{\eta}^{(11)}, \boldsymbol{\eta}^{(12)}, \dots, \boldsymbol{\eta}^{(Jd)}\}$ and $\boldsymbol{\eta}^{(lj)} = \{\eta_1^{(lj)}(\cdot), \dots, \eta_J^{(lj)}(\cdot)\}$ with $\eta_k^{(lj)} \in \mathcal{N}$ $(k = 1, \dots, J;\ l = 1, \dots, J;\ j = 1, \dots, d)$. According to Bickel et al. [3, Theorem 1, pp. 70], the efficient score vector for $\boldsymbol{\theta}$ is $\dot{l}_{\boldsymbol{\theta}}(\mathbf{y}) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\vec{\boldsymbol{\eta}}_*)$, where $\vec{\boldsymbol{\eta}}_* = \{\boldsymbol{\eta}_*^{(11)}, \boldsymbol{\eta}_*^{(12)}, \dots, \boldsymbol{\eta}_*^{(Jd)}\}$ is an element of $\mathcal{N}^{J^2d}$ that minimizes

$$\rho(\vec{\boldsymbol{\eta}}) \equiv E\|\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\vec{\boldsymbol{\eta}})\|^2 \tag{10}$$

over $\mathcal{N}^{J^2d}$ and is called the least favorable direction. Then the information for $\boldsymbol{\theta}$ is

$$I(\boldsymbol{\theta}) = E[\{\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\vec{\boldsymbol{\eta}}_*)\}\{\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})(\vec{\boldsymbol{\eta}}_*)\}^T].$$

The definition of $\vec{\boldsymbol{\eta}}_*$ given by (10) leads to a least-squares estimator of $\vec{\boldsymbol{\eta}}_*$ based on the method of Huang et al. [13]. Specifically, with a random sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ and the consistent estimators $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\phi}}_n$, we can estimate the least favorable direction $\vec{\boldsymbol{\eta}}_*$ corresponding to $\boldsymbol{\theta}^{(0)}$ by the minimizer $\hat{\vec{\boldsymbol{\eta}}}_n$ of

$$\rho_n(\vec{\boldsymbol{\eta}}) \equiv n^{-1} \sum_{i=1}^n \|\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)(\vec{\boldsymbol{\eta}})\|^2 \tag{11}$$

over $\mathcal{N}^{J^2d}$. Then a natural estimator of $I(\boldsymbol{\theta}^{(0)})$ is

$$\hat{I}_n \equiv n^{-1} \sum_{i=1}^n \{\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)(\hat{\vec{\boldsymbol{\eta}}}_n)\}\{\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)(\hat{\vec{\boldsymbol{\eta}}}_n)\}^T.$$

In practice, one can compute the component of $\hat{\vec{\boldsymbol{\eta}}}_n$ corresponding to $\boldsymbol{\eta}_*^{(lj)}$ $(l = 1, \dots, J;\ j = 1, \dots, d)$ using the ordinary least-squares regression with its feasible region $\mathcal{N}^J$ approximated by the space $\prod_{k=1}^J \mathcal{N}_{kn}$, where $\mathcal{N}_{kn}$ is the linear span of the B-spline basis functions $\mathcal{B}_{kn}$ $(k = 1, \dots, J)$.

6. Numerical experiment

In this section, we present the results of a numerical investigation on the finite sample performance of our estimates of regression parameters and baseline subdistribution hazards in Model (1). In the numerical experiment, the underlying overall survival time and the observed failure cause were generated from Model (1) with $J = 2$, $d = 3$, $Z_1 \sim Bin(1, 0.5)$, $Z_2 \sim Unif(0, 5)$ and $Z_3 \sim \min\{\Gamma(3, 1), 5\}$, the true regression parameters were $(\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}) = (0.15, -0.10, -0.20, 0.39, -0.08, -0.13)$, and the true baseline cumulative incidence functions were $F_{01}(t) = 1 - \exp[-0.15\{1 - \exp(-0.25t)\}/0.25]$ and $F_{02}(t) = 1 - \exp[-0.175\{1 - \exp(-0.5t)\}/0.5]$, which are the Gompertz models considered in [16]. There were five successive inspections for each subject, and the inspection times were uniformly distributed from 0.8 to 1.2, 1.8 to 2.2, ..., and 4.8 to 5.2, respectively.

Table 1
Simulation results of the estimates of regression parameters.

n	θ	Univariate analysis					Multivariate analysis				
		Bias($\hat{\theta}$)	Var($\hat{\theta}$) ($\times 10^2$)	E(vâr) ($\times 10^2$)	MSE($\hat{\theta}$) ($\times 10^2$)	Pr ($\theta \in$ CI)	Bias($\hat{\theta}$)	Var($\hat{\theta}$) ($\times 10^2$)	E(vâr) ($\times 10^2$)	MSE($\hat{\theta}$) ($\times 10^2$)	Pr ($\theta \in$ CI)
150	0.15	−0.001	17.105	20.538	17.088	0.973	−0.026	17.878	22.484	17.930	0.976
	−0.10	−0.006	1.761	2.507	1.763	0.977	−0.004	1.803	2.717	1.803	0.977
	−0.20	−0.019	2.171	3.162	2.204	0.977	−0.009	2.170	3.373	2.176	0.983
	0.39	0.005	15.831	17.702	15.817	0.967	−0.001	13.820	17.828	13.806	0.977
	−0.08	−0.002	1.707	2.072	1.706	0.975	−0.005	1.745	2.124	1.746	0.974
	−0.13	−0.014	2.037	2.494	2.054	0.961	−0.011	2.017	2.529	2.027	0.969
450	0.15	−0.017	4.799	5.323	4.823	0.960	0.005	4.497	5.350	4.495	0.968
	−0.10	−0.010	0.624	0.645	0.634	0.945	−0.006	0.506	0.650	0.509	0.978
	−0.20	−0.023	0.744	0.793	0.798	0.962	−0.012	0.611	0.798	0.624	0.971
	0.39	0.011	5.001	4.904	5.008	0.954	−0.017	3.982	4.952	4.006	0.968
	−0.08	−0.002	0.534	0.576	0.534	0.965	−0.009	0.560	0.581	0.568	0.958
	−0.13	−0.002	0.577	0.682	0.577	0.968	−0.010	0.636	0.685	0.646	0.961
900	0.15	−0.008	2.382	2.531	2.385	0.966	−0.029	1.885	2.467	1.965	0.969
	−0.10	−0.008	0.315	0.306	0.321	0.950	0.002	0.244	0.299	0.244	0.965
	−0.20	−0.012	0.410	0.373	0.423	0.947	0.019	0.285	0.364	0.319	0.957
	0.39	0.0005	2.217	2.335	2.214	0.961	−0.061	2.114	2.298	2.489	0.947
	−0.08	0.001	0.282	0.274	0.281	0.942	0.009	0.241	0.269	0.248	0.954
	−0.13	−0.003	0.322	0.323	0.323	0.957	0.012	0.260	0.320	0.273	0.959

We computed the sieve maximum likelihood estimates using the cubic B-splines and estimated the variances of the regression parameter estimates using the least-squares method based on the same cubic B-splines as in the parameter estimation. For the B-splines for estimating $\phi_{10}(t)$ and $\phi_{20}(t)$, the knots were equally spaced, and the numbers of knots over the observation window (0.8, 5.2) were respectively chosen to be $K_{1n} = \lfloor n^{1/3} \rfloor$ and $K_{2n} = \lfloor n^{1/5} \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x and n is the sample size. For each simulated data set, we performed a multivariate analysis, which models the two competing risks jointly, and two univariate analyses, each of which models one competing risk based on the reduced interval censored data corresponding to it. We conducted the Monte Carlo simulation study with 1000 repetitions for sample size $n = 150$, 450 and 900 representing small, moderate and large sample size respectively.

The simulation results of the estimates of regression parameters are in Table 1, where Bias($\hat{\theta}$) and Var($\hat{\theta}$) are the empirical bias and variance of the parameter estimator respectively, E(vâr) is the average of the 1000 estimated asymptotic variances of the parameter estimator, MSE($\hat{\theta}$) is the empirical mean squared error of the parameter estimator, and Pr($\theta \in$ CI) is the proportion of the 1000 95% asymptotic confidence intervals that cover the true parameter. The estimation biases are negligible compared to the empirical standard errors, indicating that the asymptotic bias of $\hat{\theta}$ is on the order of $o(1/\sqrt{n})$, which is consistent with the large sample theory. When the sample size is small, the estimated asymptotic variances are larger than the empirical variances in both the univariate analysis and the multivariate analysis, leading to higher empirical coverage probabilities of the asymptotic confidence intervals. This variance overestimation lessens as sample size increases, as indicated by the coverage probabilities approaching their nominal level. In the univariate analysis, the overestimation is not worrisome when the sample size is 450 and does not exist for sample size 900. In the multivariate analysis, it is not worrisome when sample size is 900. Comparing the variance and MSE of the two types of analyses, the multivariate analysis is more efficient when the sample size is large. This is expected because the additional data in the multivariate analysis contain information about the cumulative incidence function of the competing risk estimated in the univariate analysis. However, since the multivariate analysis maximizes a likelihood in much higher dimension compared to the univariate analysis, its estimation variances could be bigger when the sample size is not large, as shown in the portions of Table 1 for $n = 150$ and 450. All the above findings suggest that compared to the univariate analysis, the multivariate one needs a larger sample size to perform well because of the higher dimension.

The simulation results of the estimates of baseline cumulative subdistribution hazards are shown in Figs. 1 and 2. Overall, the bias of the spline-based estimator is small. The estimator for the cumulative hazard of Cause 2 has a little bit smaller bias than the estimator for Cause 1. This is expected because the estimation assumed that the cumulative hazard of Cause 2 is one order more differentiable than that of Cause 1, yielding estimators with different convergence rates.

7. A time-to-dementia data analysis

7.1. Backgrounds and data

Section 7 presents an application of the foregoing methods to the data from an ongoing cohort study of dementia: the Memory and Aging Project (MAP) [2]. The MAP study recruits older individuals without dementia who agree to receive clinical and psychological evaluation each year and to donate their brain for postmortem examination. The study began in 1997 and currently includes more than 1400 participants from about 40 retirement communities and senior housing

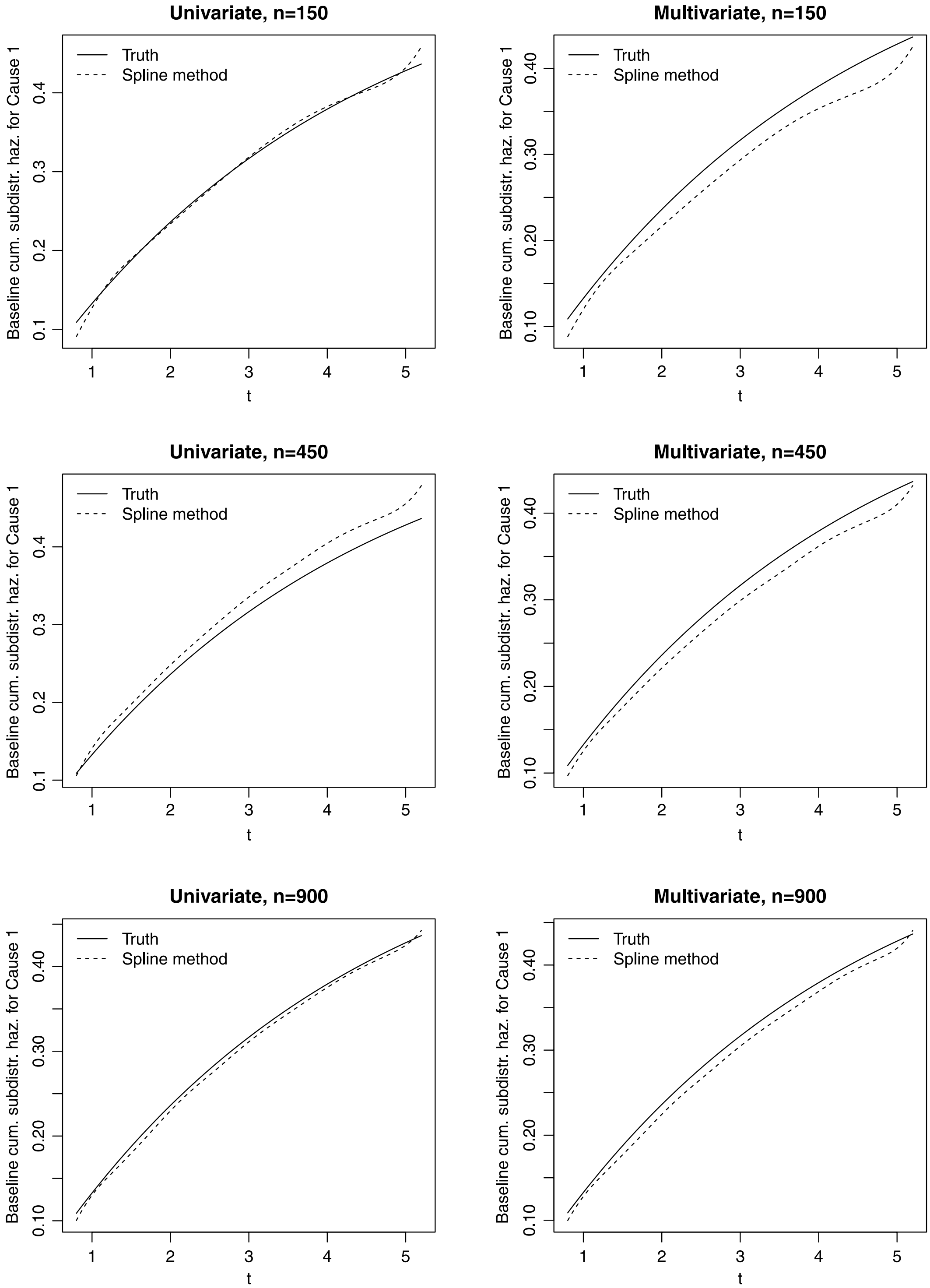

Fig. 1. Averaged estimates of the baseline cumulative subdistribution hazard functions for Cause 1 over the 1000 Monte Carlo samples.

facilities in the Chicago metropolitan area. In the study, annual visits with diagnoses of mild cognitive impairment (MCI), Alzheimer's disease (AD), and other types of dementia are scheduled for every participant. We used the MAP data to investigate the effects of years of education (categorized into two levels: ≤ 12 years and >12 years), gender, and the presence of the apolipoprotein E $\epsilon 4$ allele (ApoE4) on time from baseline study visit to incident dementia (AD or other dementia). Two

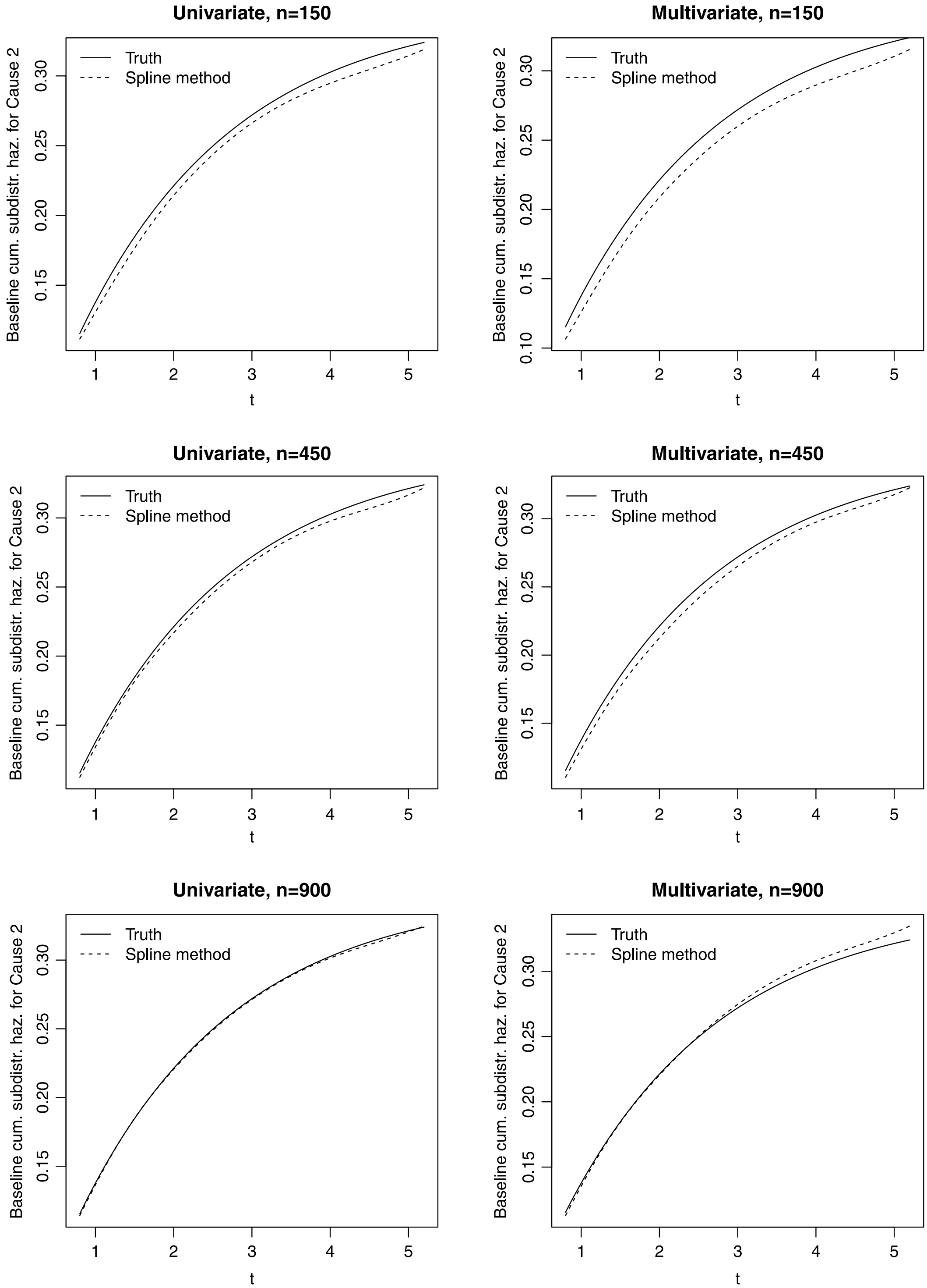

Fig. 2. Averaged estimates of the baseline cumulative subdistribution hazard functions for Cause 2 over the 1000 Monte Carlo samples.

other covariates, baseline age and the presence of baseline MCI, were also included in the analysis to further explain inter-individual differences in time to dementia. In the MAP study, the time to incident dementia is known up to between two consecutive study visits, and death may occur before dementia in older people. This yields interval censored time-to-event data with dementia and death as two competing risks.

Table 2
Characteristics of the sample for the analysis.

Characteristic	Analysis
Gender—no. (%)	
Male	255 (26.7%)
Female	701 (73.3%)
Years of education—no. (%)	
>12 (attended college)	651 (68.1%)
≤ 12 (not attended college)	305 (31.9%)
ApoE4 status—no. (%)	
Carrier	214 (22.4%)
Non-carrier	742 (77.6%)
Age at baseline (yrs)—mean (sd)	80.28 (7.16)
MCI at baseline—no. (%)	
MCI	267 (27.9%)
No MCI	689 (72.1%)
No. of study visits—mean (sd)	5.7 (2.4)
Years of follow-up—mean (sd)	4.8 (2.5)
Outcome event—no. (%)	
Dementia	183 (19.1%)
Death before dementia	181 (18.9%)

Table 3
Inference for regression parameters in the univariate analysis of cumulative incidence of dementia.

Covariate	Parameter estimate	Standard error	p-value
MCI at baseline (yes)	1.406	0.168	<0.001
Baseline age (yrs)	0.039	0.013	0.003
College education (yes)	−0.318	0.158	0.044
Gender (male)	0.015	0.158	0.926
ApoE4 (carrier)	0.521	0.163	0.001

The data set we used from the MAP study was frozen in 2010. We excluded from that data set the participants who had only baseline visit and were not known to have died as well as the ones who have missing covariates. The resulting sample for the analysis has 956 subjects. A summary of characteristics of the sample is presented in Table 2.

7.2. Univariate analysis for cumulative incidence of dementia

In this subsection, we focus on estimating the cumulative incidence function for dementia and thus model the observed data related only to dementia, essentially treating deaths prior to dementia as right censored observations. The fact that death terminates the follow-up makes mixed case interval censoring model not directly applicable to the inspection process of the MAP study. However, according to the study design, those deceased subjects in our data set would have dementia examinations approximately one year apart until 2010 had they not died ahead of that year. Therefore, it is reasonable to treat the subjects who deceased before developing dementia as being right censored at 2010, and the resulting data are the time-to-dementia data under mixed case interval censoring, for which the likelihood is of form (2) with J equal to one. We used the Fine–Gray model to model the covariates' effects on the cumulative incidence function of dementia. The sieve maximum likelihood estimation was performed using the cubic B-spline. The estimates of the regression parameters and their asymptotic standard errors are given in Table 3 along with the p-values for significance. All the covariates except gender have significant effects on the cumulative incidence for dementia at 0.05 level. Specifically, MCI at baseline, being older at baseline and ApoE4 allele all increase the cumulative risk for dementia, while having college education decreases the cumulative risk. Fig. 3, which shows the estimated cumulative incidence functions for women with different values of the significant categorical covariates, reflects the magnitude of those covariates' effects on the cumulative risk of dementia.

7.3. Multivariate analysis for cumulative incidences of dementia and deaths

In this subsection, we estimate the cumulative incidence functions of dementia and death jointly. Keep in mind that death considered here is death prior to dementia. As mentioned above, the deceased subjects in our data set would have dementia examinations approximately one year apart until 2010 had they not died ahead of that year. Thus, it is reasonable to treat the subjects who deceased before developing dementia as being interval censored between the last visit before death and the first visit after death, which did not happen but was scheduled. This scheduled visit is m years apart from the last visit before death where m is a positive integer and most often equal to one. The resulting data are interval censored

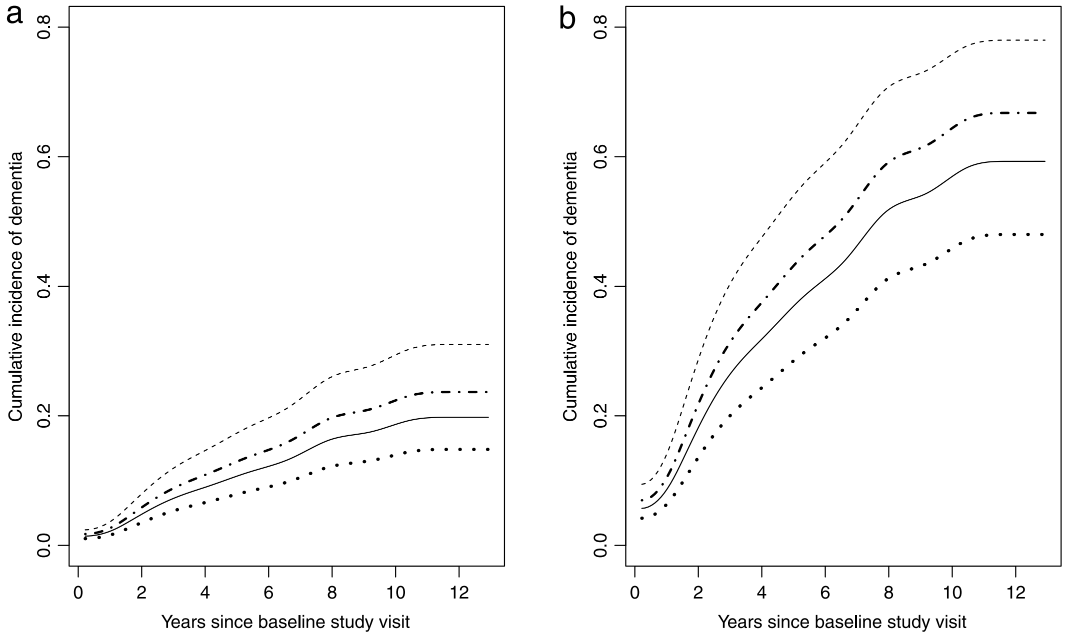

Fig. 3. Cumulative incidence functions of dementia estimated from the univariate analysis for women with baseline age 80.28 years, the mean baseline age of the sample. Plots (a) and (b) show the cumulative incidence functions for no MCI at baseline and having MCI at baseline respectively. Solid lines are for no college education and no ApoE4 allele; dashed lines are for no college education and having ApoE4 allele; dotted lines are for having college education and no ApoE4 allele; dot dash lines are for having college education and ApoE4 allele.

Table 4
Inference for regression parameters in the multivariate analysis for cumulative incidences of dementia and death.

Cause	Covariate	Parameter estimate	Standard error	p-value
Dementia	MCI at baseline (yes)	1.341	0.160	<0.001
	Baseline age (yrs)	0.063	0.014	<0.001
	College education (yes)	−0.207	0.156	0.182
	Gender (male)	0.048	0.162	0.786
	ApoE4 (carrier)	0.496	0.162	0.002
Death	MCI at baseline (yes)	−0.187	0.172	0.278
	Baseline age (yrs)	0.062	0.013	<0.001
	College education (yes)	0.058	0.161	0.716
	Gender (male)	0.418	0.162	0.010
	ApoE4 (carrier)	−0.280	0.188	0.137

time-to-event data of the competing risks, dementia and death, for which the likelihood is of the form (2) with $J = 2$. It loses statistical efficiency to treat death as being interval censored rather than use its exact time. This interval censoring treatment was performed in order to use the proposed method to estimate the cumulative incidences of dementia and death jointly. Since the time intervals bracketing death are mostly one year in length in the analytic sample (138 out of 181 death bracketing intervals), the efficiency loss is expected to be small. Annually scheduled dementia examinations also decrease the chance that dementia incidence was not caught in subjects who died after developing dementia. The analysis results of regression parameters are given in Table 4. We can see that the results of the covariate effects on cumulative incidence of dementia are similar to the univariate analysis except that the effect of college education became non-significant. For the cumulative incidence of death, baseline age has a significant increasing effect, men have significantly higher cumulative risk than women, and the other covariates are not significant. Fig. 4 is the counterpart of Fig. 3 obtained from the multivariate analysis. The two figures look similar. Fig. 5 shows the estimated cumulative incidence functions of death for people who attended college and whose baseline age is 80.28 years.

8. Discussion

In this article, we proposed a spline-based sieve maximum likelihood estimation for the Fine–Gray model to analyze interval-censored competing risks data. We showed that, under certain regularity conditions, the sieve estima-

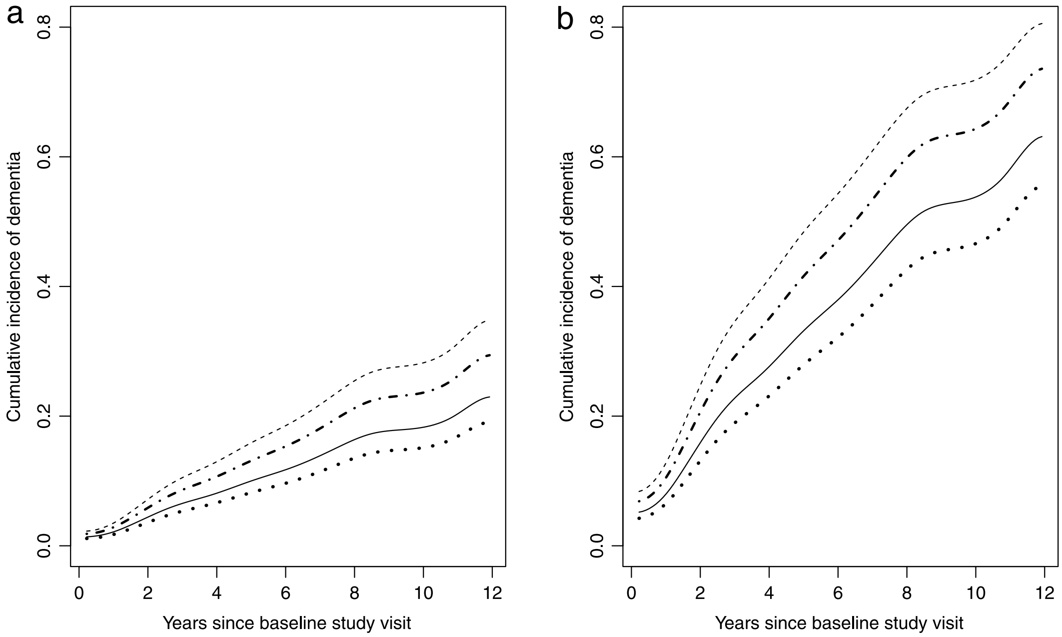

Fig. 4. Cumulative incidence functions of dementia estimated from the multivariate analysis for women with baseline age 80.28 years, the mean baseline age of the sample. Plots (a) and (b) show the cumulative incidence functions for no MCI at baseline and having MCI at baseline respectively. Solid lines are for no college education and no ApoE4 allele; dashed lines are for no college education and having ApoE4 allele; dotted lines are for having college education and no ApoE4 allele; dot dash lines are for having college education and ApoE4 allele.

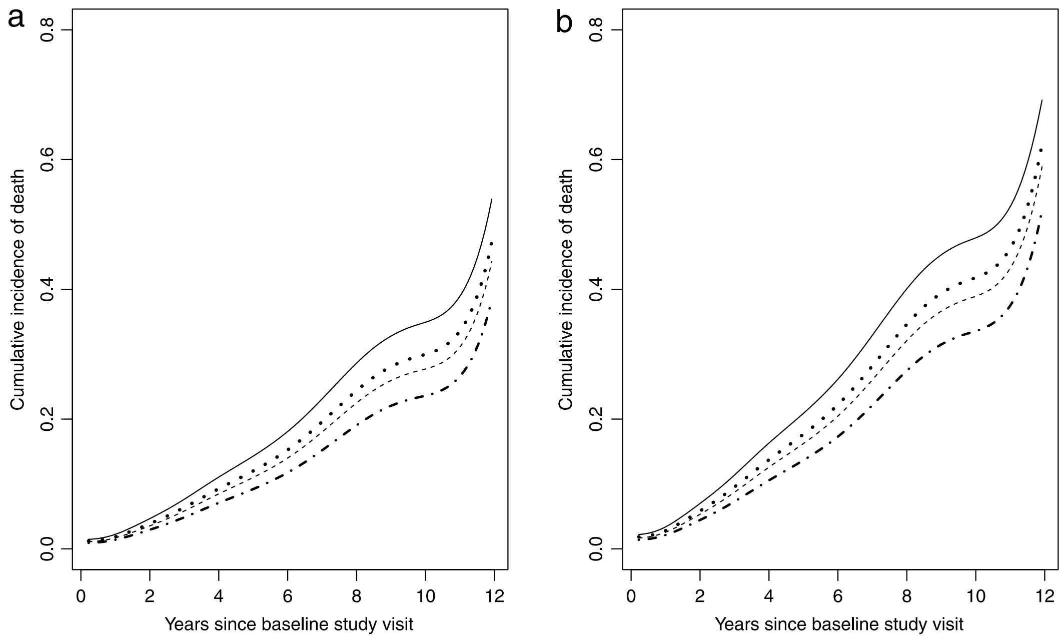

Fig. 5. Cumulative incidence functions of death estimated from the multivariate analysis for people with college education and baseline age 80.28 years, the mean baseline age of the sample. Plots (a) and (b) show the cumulative incidence functions for female and male respectively. Solid lines are for no MCI at baseline and no ApoE4 allele; dashed lines are for no MCI at baseline and having ApoE4 allele; dotted lines are for having MCI at baseline and no ApoE4 allele; dot dash lines are for having MCI at baseline and ApoE4 allele.

tor of the baseline cumulative subdistribution hazard is consistent and converges with the optimal rate as in the nonparametric regression setting, and the regression parameter estimator is asymptotically normal and semiparametrically efficient. The efficient information matrix for the regression parameter can be consistently estimated by a least-squares method.

One could also use the nonparametric maximum likelihood estimator (NPMLE) for the baseline cumulative subdistribution hazard in the Fine–Gray model under interval censored competing risks data. The NPMLE approach assumes that the baseline cumulative subdistribution hazard jumps only at the effective observation times of (U_i, V_i) $(i = 1, \ldots, n)$ as defined in Definition 1.1, page 45, of Groeneboom and Wellner [10]. In the univariate analysis, this way is equivalent to the method of Huang and Wellner [12]. But we recommend the spline-based estimator over the NPMLE for the following reasons. First, the dimension of spline-based estimator is actually smaller than that of the NPMLE, making the computation simpler. The former is of order $\sum_{k=1}^{J} n^{\nu_k}$ $(\nu_k < 1/(2p_k))$, while the latter is on the order of n. Second, if the number of knots is selected properly, e.g. $\nu_k = 1/(1 + 2p_k)$, the spline-based estimator will converge at a rate, which is $n^{p_k/(1+2p_k)}$ when $\nu_k = 1/(1 + 2p_k)$, no less than that of the NPMLE, which is always $n^{1/3}$. Third, a smooth cumulative hazard estimator is more realistic than an estimator that is a step function.

Though the joint analysis of multiple competing risks improves statistical efficiency when sample size is large, it is more vulnerable to model mis-specification because it requires that the Fine–Gray model holds for every competing risk under consideration. If a competing risk is of much more interest compared to the others, the univariate analysis of that specific competing risk is preferred to the joint analysis since it would be more robust. Take the MAP study for example, the investigators are more interested in dementia incidence in alive people compared to death prior to dementia. Thus the univariate analysis in Section 7.2 should be used to report study findings. As a result, we would conclude that college education has a decreasing effect on the cumulative risk of dementia, which was not reflected by the joint analysis in Section 7.3.

It should not take much effort to generalize the developed inferential approach to the class of semiparametric transformation models [6] for the analysis of interval censored competing risks data, since the Fine–Gray model is a member of that class. Motivated by the late entry of the MAP participants to the study for the analysis of age to dementia, extending our method to accommodate left truncation is warranted. Also, the fact that time to some competing risk such as death was exactly known in the MAP study as well as many other longitudinal studies makes it worthwhile to extend our method to partly interval censored competing risks data. Computationally, these two extensions should be straightforward, since one just needs to respectively divide the likelihood (2) by the overall survival function at the entry time and replace $F_k(V_i; \mathbf{Z}_i) - F_k(U_i; \mathbf{Z}_i)$ in (2) by the corresponding subdensity at the exactly observed occurrence time, and the resulting likelihoods can be maximized using the same model for $F_k(t; \mathbf{Z})$ as (1) and the same sieve method as in Section 3. The derivation of the relevant asymptotic theory is not that straightforward and needs further investigation.

Acknowledgments

The data for the example presented in this article were provided by a study supported by the National Institute of Aging, National Institute of Health: the Rush Memory and Aging Project (NIA Grant R01AG17917). We would like to thank the principal investigator of this project, David A. Bennett, MD, for the authorization to use the data. This work is supported in part by a grant from National Institute of Dental and Craniofacial Research (1R03DE023889).

Appendix. Proofs

The proofs of Theorems 1 and 2 followed similar arguments in [30]. Throughout the following proofs, $P_\zeta f = \int f(\mathbf{y}) dP_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{y})$, $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(\mathbf{Y}_i)$, $\lfloor x \rfloor$ denotes the largest integer below x, and we let C represent a generic constant that may vary from place to place.

Proof of Theorem 1. We first prove the consistency of $\hat{\boldsymbol{\zeta}}_n$ in the metric d. This can be accomplished by verifying the conditions of Theorem 5.7 in [27].

Let $\mathbb{M}(\boldsymbol{\zeta}) = P_{\zeta_0} l(\mathbf{Y}; \boldsymbol{\zeta}) = P_{\zeta_0} l(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})$ and $\mathbb{M}_n(\boldsymbol{\zeta}) = \mathbb{P}_n l(\mathbf{Y}; \boldsymbol{\zeta}) = \mathbb{P}_n l(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\phi})$. Therefore, for any $\boldsymbol{\zeta} \in \mathcal{T}_n = \boldsymbol{\Theta} \times \prod_{k=1}^{J} \mathcal{M}_{kn}$, $\mathbb{M}_n(\boldsymbol{\zeta}) - \mathbb{M}(\boldsymbol{\zeta}) = (\mathbb{P}_n - P_{\zeta_0}) l(\mathbf{Y}; \boldsymbol{\zeta})$.

Let $\mathcal{L}_1 = \{l(\mathbf{Y}; \boldsymbol{\zeta}) : \boldsymbol{\zeta} \in \mathcal{T}_n\}$. By the calculation of Shen and Wong [26, pp. 597], for all $\epsilon > 0$, there exist J sets of brackets $\left\{\left[\phi_{ki}^L, \phi_{ki}^U\right] : i = 1, 2, \ldots, \left\lfloor (1/\epsilon)^{Cq_{kn}} \right\rfloor\right\}$ $(k = 1, \ldots, J)$ such that for any $\boldsymbol{\phi} = (\phi_{10}(\cdot), \ldots, \phi_{J0}(\cdot)) \in \prod_{k=1}^{J} \mathcal{M}_{kn}$, one has $\phi_{ki_k}^L(t) \le \phi_{k0}(t) \le \phi_{ki_k}^U(t)$ $(k = 1, \ldots, J)$ for some $1 \le i_k \le \left\lfloor (1/\epsilon)^{Cq_{kn}} \right\rfloor$ and all $t \in [\tau_0, \tau_1]$, and $\mathbb{P}_n \left|\phi_{ki}^U(Y) - \phi_{ki}^L(Y)\right| \le \epsilon$ with $Y = U$ and V for all $1 \le i \le \left\lfloor (1/\epsilon)^{Cq_{kn}} \right\rfloor$ and $1 \le k \le J$. As $\boldsymbol{\Theta} \subset \mathbb{R}^{Jd}$ is compact, $\boldsymbol{\Theta}$ can be covered by $\left\lfloor C(1/\epsilon)^{Jd} \right\rfloor$ balls with radius ϵ and centers being denoted by $\boldsymbol{\theta}^{(s)}$ $(s = 1, \ldots, \left\lfloor C(1/\epsilon)^{Jd} \right\rfloor)$ respectively; that is for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists a $1 \le s \le \left\lfloor C(1/\epsilon)^{Jd} \right\rfloor$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(s)}\| \le \epsilon$ and hence $|\mathbf{Z}^T \boldsymbol{\theta}_k - \mathbf{Z}^T \boldsymbol{\theta}_k^{(s)}| \le C\epsilon$ for all $\mathbf{Z}$ and $1 \le k \le J$ because of A1. This implies that $\mathbf{Z}^T \boldsymbol{\theta}_k \in [\mathbf{Z}^T \boldsymbol{\theta}_k^{(s)} - C\epsilon, \mathbf{Z}^T \boldsymbol{\theta}_k^{(s)} + C\epsilon]$ for all $\mathbf{Z}$ and $1 \le k \le J$. Hence we can easily construct such a set of brackets $\{[l_{s,i_1,\ldots,i_J}^L(\mathbf{Y}), l_{s,i_1,\ldots,i_J}^U(\mathbf{Y})] : s = 1, 2, \ldots, \lfloor C(1/\epsilon)^{Jd} \rfloor;\ i_k = 1, \ldots, \lfloor (1/\epsilon)^{Cq_{kn}} \rfloor, k = 1, \ldots, J\}$ that for any $l(\mathbf{Y}; \boldsymbol{\zeta}) \in \mathcal{L}_1$, there exist a $1 \le s \le \left\lfloor C(1/\epsilon)^{Jd} \right\rfloor$ and a $1 \le i_k \le \lfloor (1/\epsilon)^{Cq_{kn}} \rfloor$ $(k = 1, \ldots, J)$ such that $l(\mathbf{Y}; \boldsymbol{\zeta}) \in [l_{s,i_1,\ldots,i_J}^L(\mathbf{Y}), l_{s,i_1,\ldots,i_J}^U(\mathbf{Y})]$ for any sample point $\mathbf{Y}$, where

$$l^L_{s,i_1,\ldots,i_J}(\mathbf{Y}) = \sum_{k=1}^J \left[\Delta_{k1} \log\left[1 - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} - C\epsilon + \phi^L_{ki_k}(U)}\right\}\right] + \Delta_{k2}\log\left[\exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} + C\epsilon + \phi^U_{ki_k}(U)}\right\} - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} - C\epsilon + \phi^L_{ki_k}(V)}\right\}\right]\right] + \Delta_3 \log\left[1 - J + \sum_{k=1}^J \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} + C\epsilon + \phi^U_{ki_k}(V)}\right\}\right]$$

and

$$l^U_{s,i_1,\ldots,i_J}(\mathbf{Y}) = \sum_{k=1}^J \left[\Delta_{k1} \log\left[1 - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} + C\epsilon + \phi^U_{ki_k}(U)}\right\}\right] + \Delta_{k2}\log\left[\exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} - C\epsilon + \phi^L_{ki_k}(U)}\right\} - \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} + C\epsilon + \phi^U_{ki_k}(V)}\right\}\right]\right] + \Delta_3 \log\left[1 - J + \sum_{k=1}^J \exp\left\{-e^{\mathbf{Z}^T\boldsymbol{\theta}_k^{(s)} - C\epsilon + \phi^L_{ki_k}(V)}\right\}\right].$$

Using Taylor expansion along with assumptions A1–A3, we can easily demonstrate that $\mathbb{P}_n|l^U_{s,i_1,\ldots,i_J}(\mathbf{Y}) - l^L_{s,i_1,\ldots,i_J}(\mathbf{Y})| \le C\epsilon$ for all $1 \le s \le \lfloor C(1/\epsilon)^{Jd}\rfloor$ and $1 \le i_k \le \lfloor (1/\epsilon)^{Cq_{kn}}\rfloor$ $(k = 1,\ldots,J)$, which leads to the conclusion that the ϵ-bracketing number for $\mathcal{L}_1$ with $L_1(\mathbb{P}_n)$-norm is bounded by $C(1/\epsilon)^{C\sum_{k=1}^J q_{kn} + Jd}$. As $N(\epsilon, \mathcal{L}_1, L_1(\mathbb{P}_n)) \le N_{[]}(\epsilon, \mathcal{L}_1, L_1(\mathbb{P}_n))$, $\mathcal{L}_1$ is Glivenko–Cantelli by van der Vaart and Wellner [28, Theorem 2.4.3]. Therefore, $\sup_{\boldsymbol{\zeta}\in\mathcal{T}_n} |\mathbb{M}_n(\boldsymbol{\zeta}) - \mathbb{M}(\boldsymbol{\zeta})| \longrightarrow_{\text{a.s.}} 0$. Let $g_k(\mathbf{Z}, t) = \exp\{\mathbf{Z}^T\boldsymbol{\theta}_k + \phi_{k0}(t)\}$ and $g_{k0}(\mathbf{Z}, t) = \exp(\mathbf{Z}^T\boldsymbol{\theta}_k^{(0)} + \phi_{k0}^{(0)}(t))$ $(k = 1,\ldots,J)$. Some algebra yields that

$$\begin{aligned}
\mathbb{M}(\boldsymbol{\zeta}_0) - \mathbb{M}(\boldsymbol{\zeta}) &= E\left[\sum_{k=1}^J \left[[1 - \exp\{-g_{k0}(\mathbf{Z}, U)\}] \log \frac{1 - \exp\{-g_{k0}(\mathbf{Z}, U)\}}{1 - \exp\{-g_k(\mathbf{Z}, U)\}} \right.\right.\\
&\quad \left. + [\exp\{-g_{k0}(\mathbf{Z}, U)\} - \exp\{-g_{k0}(\mathbf{Z}, V)\}] \log \frac{\exp\{-g_{k0}(\mathbf{Z}, U)\} - \exp\{-g_{k0}(\mathbf{Z}, V)\}}{\exp\{-g_k(\mathbf{Z}, U)\} - \exp\{-g_k(\mathbf{Z}, V)\}}\right]\\
&\quad \left. + \left[1 - J + \sum_{k=1}^J \exp\{-g_{k0}(\mathbf{Z}, V)\}\right] \log \frac{1 - J + \sum_{k=1}^J \exp\{-g_{k0}(\mathbf{Z}, V)\}}{1 - J + \sum_{k=1}^J \exp\{-g_k(\mathbf{Z}, V)\}}\right]\\
&= E\left[\sum_{k=1}^J \left[[1 - \exp\{-g_k(\mathbf{Z}, U)\}]\, w\left[\frac{1 - \exp\{-g_{k0}(\mathbf{Z}, U)\}}{1 - \exp\{-g_k(\mathbf{Z}, U)\}}\right]\right.\right.\\
&\quad \left. + [\exp\{-g_k(\mathbf{Z}, U)\} - \exp\{-g_k(\mathbf{Z}, V)\}] w\left[\frac{\exp\{-g_{k0}(\mathbf{Z}, U)\} - \exp\{-g_{k0}(\mathbf{Z}, V)\}}{\exp\{-g_k(\mathbf{Z}, U)\} - \exp\{-g_k(\mathbf{Z}, V)\}}\right]\right]\\
&\quad \left. + \left[1 - J + \sum_{k=1}^J \exp\{-g_k(\mathbf{Z}, V)\}\right] w\left[\frac{1 - J + \sum_{k=1}^J \exp\{-g_{k0}(\mathbf{Z}, V)\}}{1 - J + \sum_{k=1}^J \exp\{-g_k(\mathbf{Z}, V)\}}\right]\right],
\end{aligned}$$

where $w(x) = x\log x - x + 1 \ge (x-1)^2/4$ for $0 \le x \le 5$. Further analysis by using Taylor expansion leads to

$$\begin{aligned}
\mathbb{M}(\boldsymbol{\zeta}_0) - \mathbb{M}(\boldsymbol{\zeta}) &\ge \frac{1}{4}E\left[\sum_{k=1}^J \frac{1}{1 - \exp\{-g_k(\mathbf{Z}, U)\}}[\exp\{-g_k(\mathbf{Z}, U)\} - \exp\{-g_{k0}(\mathbf{Z}, U)\}]^2\right.\\
&\quad \left. + \frac{1}{1 - J + \sum_{k=1}^J \exp\{-g_k(\mathbf{Z}, V)\}}\left[\sum_{k=1}^J [\exp\{-g_{k0}(\mathbf{Z}, V)\} - \exp\{-g_k(\mathbf{Z}, V)\}]\right]^2\right]
\end{aligned}$$

$$= \frac{1}{4} E \left[\sum_{k=1}^{J} \frac{1}{1-\exp\{-g_k(\mathbf{Z},U)\}} \gamma^2[\xi\{\boldsymbol\theta_k^{(0)},\boldsymbol\theta_k,\mathbf{Z},\phi_{k0}^{(0)}(U),\phi_{k0}(U)\}]\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(U)\}^2 \right.$$
$$\left. + \frac{\left[\sum_{k=1}^{J} \gamma[\xi\{\boldsymbol\theta_k^{(0)},\boldsymbol\theta_k,\mathbf{Z},\phi_{k0}^{(0)}(V),\phi_{k0}(V)\}]\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(V)\}\right]^2}{1-J+\sum_{k=1}^{J}\exp\{-g_k(\mathbf{Z},V)\}} \right],$$

where $\gamma(x) = d\{\exp(-e^x)\}/dx = -\exp(-e^x+x)$, $\xi\{\boldsymbol\theta_k^{(0)},\boldsymbol\theta_k,\mathbf{Z},\phi_{k0}^{(0)}(U),\phi_{k0}(U)\}$ is some number between $\mathbf{Z}^T\boldsymbol\theta_k^{(0)}+\phi_{k0}^{(0)}(U)$ and $\mathbf{Z}^T\boldsymbol\theta_k+\phi_{k0}(U)$, and $\xi\{\boldsymbol\theta_k^{(0)},\boldsymbol\theta_k,\mathbf{Z},\phi_{k0}^{(0)}(V),\phi_{k0}(V)\}$ is some number between $\mathbf{Z}^T\boldsymbol\theta_k^{(0)}+\phi_{k0}^{(0)}(V)$ and $\mathbf{Z}^T\boldsymbol\theta_k+\phi_{k0}(V)$. Let

$$\mathbb{Z} = \begin{pmatrix} \mathbf{Z}^T & & \\ & \ddots & \\ & & \mathbf{Z}^T \end{pmatrix}_{J\times Jd}, \qquad \Gamma = \begin{pmatrix} \gamma[\xi\{\boldsymbol\theta_1^{(0)},\boldsymbol\theta_1,\mathbf{Z},\phi_{10}^{(0)}(V),\phi_{10}(V)\}] \\ \vdots \\ \gamma[\xi\{\boldsymbol\theta_J^{(0)},\boldsymbol\theta_J,\mathbf{Z},\phi_{J0}^{(0)}(V),\phi_{J0}(V)\}] \end{pmatrix}$$

and $\lambda_{\min}$ = the minimum eigenvalue of $\Gamma\Gamma^T$. Further analysis using assumptions A1–A3 leads to

$$\begin{aligned} \mathbb{M}(\boldsymbol\zeta_0)-\mathbb{M}(\boldsymbol\zeta) \ge\ & CE\left[\sum_{k=1}^{J}\left\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(U)\right\}^2\right. \\ & \left. +\ \{\mathbb{Z}(\boldsymbol\theta^{(0)}-\boldsymbol\theta)+(\boldsymbol\phi_0-\boldsymbol\phi)(V)\}^T\Gamma\Gamma^T\{\mathbb{Z}(\boldsymbol\theta^{(0)}-\boldsymbol\theta)+(\boldsymbol\phi_0-\boldsymbol\phi)(V)\}\right] \\ \ge\ & CE\left[\sum_{k=1}^{J}\left\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(U)\right\}^2\right. \\ & \left. +\ \lambda_{\min}\{\mathbb{Z}(\boldsymbol\theta^{(0)}-\boldsymbol\theta)+(\boldsymbol\phi_0-\boldsymbol\phi)(V)\}^T\{\mathbb{Z}(\boldsymbol\theta^{(0)}-\boldsymbol\theta)+(\boldsymbol\phi_0-\boldsymbol\phi)(V)\}\right] \\ \ge\ & CE\left[\sum_{k=1}^{J}\left\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(U)\right\}^2+\sum_{k=1}^{J}\left\{\mathbf{Z}^T(\boldsymbol\theta_k^{(0)}-\boldsymbol\theta_k)+(\phi_{k0}^{(0)}-\phi_{k0})(V)\right\}^2\right]. \end{aligned}$$

With assumptions A1–A4 and A6, using the same arguments as those in [29, pp. 2126–2127] leads to

$$\mathbb{M}(\boldsymbol\zeta_0)-\mathbb{M}(\boldsymbol\zeta) \ge C(\|\boldsymbol\theta^{(0)}-\boldsymbol\theta\|^2+\|\boldsymbol\phi_0-\boldsymbol\phi\|_{\boldsymbol\Phi}^2) = Cd^2(\boldsymbol\zeta_0,\boldsymbol\zeta).$$

Then it implies that $\sup_{\boldsymbol\zeta:d(\boldsymbol\zeta,\boldsymbol\zeta_0)\ge\epsilon}\mathbb{M}(\boldsymbol\zeta) \le \mathbb{M}(\boldsymbol\zeta_0)-C\epsilon^2 < \mathbb{M}(\boldsymbol\zeta_0)$.

For $\boldsymbol\phi_0 \in \boldsymbol\Phi$, Lu [22] has shown that there exists a $\phi_{k0,n} \in \mathcal{M}_{kn}$ of order $m_k \ge p_k+2$ such that $\|\phi_{k0,n}-\phi_{k0}^{(0)}\|_\infty \le Cq_{kn}^{-p_k} = O(n^{-p_k\nu_k})$ $(k=1,\dots,J)$. This also implies that $\|\phi_{k0,n}-\phi_{k0}^{(0)}\|_{\Phi_k} \le Cq_{kn}^{-p_k} = O(n^{-p_k\nu_k})$ $(k=1,\dots,J)$. Now let $\boldsymbol\phi_{0,n} = (\phi_{10,n}(\cdot),\dots,\phi_{J0,n}(\cdot))$ and $\boldsymbol\zeta_{0,n} = (\boldsymbol\theta^{(0)},\boldsymbol\phi_{0,n})$, we have

$$\begin{aligned} \mathbb{M}_n(\hat{\boldsymbol\zeta}_n)-\mathbb{M}_n(\boldsymbol\zeta_0) &= \mathbb{M}_n(\hat{\boldsymbol\zeta}_n)-\mathbb{M}_n(\boldsymbol\zeta_{0,n})+\mathbb{M}_n(\boldsymbol\zeta_{0,n})-\mathbb{M}_n(\boldsymbol\zeta_0) \\ &\ge \mathbb{P}_n l(\mathbf{Y};\boldsymbol\zeta_{0,n})-\mathbb{P}_n l(\mathbf{Y};\boldsymbol\zeta_0) \\ &= (\mathbb{P}_n-P)\{l(\mathbf{Y};\boldsymbol\zeta_{0,n})-l(\mathbf{Y};\boldsymbol\zeta_0)\}+\mathbb{M}(\boldsymbol\zeta_{0,n})-\mathbb{M}(\boldsymbol\zeta_0). \end{aligned}$$

Using the brackets for $\mathcal{M}_{kn}$ $(k=1,\dots,J)$ given before, we can similarly construct a set of brackets for the class $\mathcal{L}_2 = \{l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi)-l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi_0) : \boldsymbol\phi \in \prod_{k=1}^{J}\mathcal{M}_{kn}$ and $\|\boldsymbol\phi_{k0}-\boldsymbol\phi_{k0}^{(0)}\|_{\Phi_k} \le Cn^{-p_k\nu_k}$ for $k=1,\dots,J\}$ with the ϵ-bracketing number associated with $L_2(P)$-norm bounded by $(1/\epsilon)^{C\sum_{k=1}^{J}q_{kn}}$. This yields a finite-valued bracketing integral defined in [27, p. 270]. Hence the class $\mathcal{L}_2$ is P-Donsker. By the dominated convergence theorem, it is obvious that in this class $P\{l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi)-l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi_0)\}^2 \to 0$ as $n\to\infty$. Hence

$$(\mathbb{P}_n-P)\{l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi_{0,n})-l(\mathbf{Y};\boldsymbol\theta^{(0)},\boldsymbol\phi_0)\} = o_p(n^{-1/2})$$

by the relationship between Donsker and asymptotic equicontinuity given by van der Vaart and Wellner [28, Corollary 2.3.12]. By the dominated convergence theorem again, it is easy to see that $\mathbb{M}(\boldsymbol\zeta_{0,n})-\mathbb{M}(\boldsymbol\zeta_0) > -o(1)$ as $n\to\infty$. Therefore,

$$\mathbb{M}_n(\hat{\boldsymbol\zeta}_n)-\mathbb{M}_n(\boldsymbol\zeta_0) \ge o_p(n^{-1/2})-o(1) = -o_p(1).$$

This completes the proof of $d(\hat{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_0) \to 0$ in probability.

Next, we verify the conditions of van der Vaart and Wellner [28, Theorem 3.2.5] to derive the convergence rate. First, we have already shown in the proof of consistency that $\mathbb{M}(\boldsymbol{\zeta}_0) - \mathbb{M}(\boldsymbol{\zeta}) \geq Cd^2(\boldsymbol{\zeta}_0, \boldsymbol{\zeta})$.

Second, we further explore $\mathbb{M}_n(\hat{\boldsymbol{\zeta}}_n) - \mathbb{M}_n(\boldsymbol{\zeta}_0)$. In the proof of consistency, we know that $\mathbb{M}_n(\hat{\boldsymbol{\zeta}}_n) - \mathbb{M}_n(\boldsymbol{\zeta}_0) \geq I_{1,n} + I_{2,n}$, where $I_{1,n} = (\mathbb{P}_n - P)\{l(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}_{0,n}) - l(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}_0)\}$ and $I_{2,n} = P\{l(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}_{0,n}) - l(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}_0)\}$. By Taylor expansion, we have, for any $0 < \epsilon < 1/2 - \min_{1\leq k\leq J}\{p_k\nu_k\}$,

$$I_{1,n} = (\mathbb{P}_n - P)\left\{\sum_{k=1}^{J} \dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \tilde{\boldsymbol{\phi}})(\phi_{k0,n} - \phi_{k0}^{(0)})\right\} = \sum_{k=1}^{J} n^{-p_k\nu_k+\epsilon}(\mathbb{P}_n - P)\left\{\dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \tilde{\boldsymbol{\phi}})\frac{\phi_{k0,n} - \phi_{k0}^{(0)}}{n^{-p_k\nu_k+\epsilon}}\right\},$$

where $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_{10}, \ldots, \tilde{\phi}_{J0})$ and $\tilde{\phi}_{k0} = \phi_{k0}^{(0)} + s^*(\phi_{k0,n} - \phi_{k0}^{(0)})$ for some $s^* \in (0, 1)$ and all $1 \leq k \leq J$. Because $\|\phi_{k0,n} - \phi_{k0}^{(0)}\|_\infty = O(n^{-p_k\nu_k})$ and $\dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \tilde{\boldsymbol{\phi}})$ is uniformly bounded as a result of assumptions A1–A4, we can easily obtain that $P\left\{\dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \tilde{\boldsymbol{\phi}})\frac{\phi_{k0,n}-\phi_{k0}^{(0)}}{n^{-p_k\nu_k+\epsilon}}\right\}^2 \to 0$. As a result of $\mathcal{L}_2$ being Donsker, using van der Vaart and Wellner [28, Corollary 2.3.12] again, we can conclude that $(\mathbb{P}_n - P)\left\{\dot{l}_{\boldsymbol{\phi}}^{(k)}(\mathbf{Y}; \boldsymbol{\theta}^{(0)}, \tilde{\boldsymbol{\phi}})\frac{\phi_{k0,n}-\phi_{k0}^{(0)}}{n^{-p_k\nu_k+\epsilon}}\right\} = o_p(n^{-1/2})$ $(k = 1, \ldots, J)$. Hence,

$$I_{1,n} = \sum_{k=1}^{J} o_p(n^{-p_k\nu_k+\epsilon}n^{-1/2}) = o_p(n^{-2\min_k\{p_k\nu_k\}}),$$

owing to the selection of ϵ. Using the fact that the function $m(x) = x\log(x) - x + 1 \leq (x-1)^2$ in the neighborhood of $x = 1$, it can be easily argued that $\mathbb{M}(\boldsymbol{\zeta}_0) - \mathbb{M}(\boldsymbol{\zeta}_{0,n}) \leq C\sum_{k=1}^{J}\|\phi_{k0,n} - \phi_{k0}\|_{\Phi_k}^2 = C\sum_{k=1}^{J} O(n^{-2p_k\nu_k}) = CO(n^{-2\min_k\{p_k\nu_k\}})$, which implies that $I_{2,n} = \mathbb{M}(\boldsymbol{\zeta}_{0,n}) - \mathbb{M}(\boldsymbol{\zeta}_0) \geq -O(n^{-2\min_k\{p_k\nu_k\}})$. Thus, we conclude that

$$\mathbb{M}_n(\hat{\boldsymbol{\zeta}}_n) - \mathbb{M}_n(\boldsymbol{\zeta}_0) \geq -O_p(n^{-2\min_k\{p_k\nu_k\}}) = -O_p\left(n^{-2\min\{\min_k(p_k\nu_k),(1-\max_k\nu_k)/2\}}\right).$$

Let $\mathcal{L}_3(\eta) = \{l(\mathbf{Y}; \boldsymbol{\zeta}) - l(\mathbf{Y}; \boldsymbol{\zeta}_0) : \boldsymbol{\phi} \in \prod_{k=1}^{J}\mathcal{M}_{kn}$ and $d(\boldsymbol{\zeta}, \boldsymbol{\zeta}_0) \leq \eta\}$. Using the same arguments as in the proof of consistency, we obtain that the logarithm of the ϵ-bracketing number of $\mathcal{L}_3(\eta)$, $\log N_{[]}\{\epsilon, \mathcal{L}_3(\eta), L_2(P)\}$, is bounded by $C(\sum_{k=1}^{J} q_{kn})\log(\eta/\epsilon)$. This leads to

$$J_{[]}\{\eta, \mathcal{L}_3(\eta), L_2(P)\} = \int_0^\eta \sqrt{1 + \log N_{[]}\{\epsilon, \mathcal{L}_3(\eta), L_2(P)\}}d\epsilon \leq C\left(\sum_{k=1}^{J} q_{kn}\right)^{1/2}\eta.$$

Because assumptions A1–A3 guarantee the uniform boundedness of $l(\mathbf{Y}; \boldsymbol{\zeta})$, using van der Vaart and Wellner [28, Lemma 3.4.2], the key function $\phi_n(\eta)$ in [28, Theorem 3.2.5] is given by $\phi_n(\eta) = (\sum_{k=1}^{J} q_{kn})^{1/2}\eta + \sum_{k=1}^{J} q_{kn}/\sqrt{n}$. Note that

$$\begin{aligned} n^{2\min_k\{p_k\nu_k\}}\phi_n(1/n^{\min_k\{p_k\nu_k\}}) &= n^{\min_k\{p_k\nu_k\}}O(n^{\max_k\nu_k/2}) + n^{2\min_k\{p_k\nu_k\}}n^{-1/2}O(n^{\max_k\nu_k}) \\ &\leq Cn^{1/2}\left\{n^{\min_k(p_k\nu_k)-(1-\max_k\nu_k)/2} + n^{2\min_k(p_k\nu_k)-(1-\max_k\nu_k)}\right\}. \end{aligned}$$

Therefore, if $\min_k\{p_k\nu_k\} \leq (1 - \max_k\nu_k)/2$, $n^{2\min_k\{p_k\nu_k\}}\phi_n(1/n^{\min_k\{p_k\nu_k\}}) \leq n^{1/2}$. This implies that if we choose $r_n = n^{\min\{\min_k(p_k\nu_k),(1-\max_k\nu_k)/2\}}$, it follows that $r_n^2\phi_n(1/r_n) \leq n^{1/2}$ and $\mathbb{M}_n(\hat{\boldsymbol{\zeta}}_n) - \mathbb{M}_n(\boldsymbol{\zeta}_0) \geq -O_p(r_n^{-2})$. Hence, $r_n d(\hat{\boldsymbol{\zeta}}_n, \boldsymbol{\zeta}_0) = O_p(1)$. □

Proof of Theorem 2. To derive the asymptotic normality for $\hat{\boldsymbol{\theta}}_n$, we just need to verify the conditions (B1)–(B3) of the general theorem given in Appendix B of Zhang et al. [30]. For condition (B1), we only need to verify that $\mathbb{P}_n\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n)(\boldsymbol{\eta}^*) = o_p(n^{1/2})$ as $\mathbb{P}_n\dot{l}_{\boldsymbol{\theta}}(\mathbf{Y}; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}}_n) \equiv 0$. Because each component of $\boldsymbol{\eta}^*$ has a bounded derivative, it is also a function with bounded variation. Then it can be easily shown using the argument in [4, pp. 435–436] that there exists a $\boldsymbol{\eta}_n^* \in (\prod_{k=1}^{J}\mathcal{S}_{kn}(D_{kn}, K_{kn}, m_k))^{\otimes Jd}$ such that $\|\eta_{n,jk}^* - \eta_{jk}^*\|_{\Phi_k} = O(q_{kn}^{-1}) = O(n^{-\nu_k})$ for $j = 1, \ldots, Jd$ and $1 \leq k \leq J$ and $\mathbb{P}_n\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \hat{\boldsymbol{\zeta}}_n)(\boldsymbol{\eta}_n^*) = 0$. Therefore, we can write $\mathbb{P}_n\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \hat{\boldsymbol{\zeta}}_n)(\boldsymbol{\eta}^*) = I_{3,n} + I_{4,n}$, where

$$I_{3,n} = (\mathbb{P}_n - P)\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \hat{\boldsymbol{\zeta}}_n)(\boldsymbol{\eta}^* - \boldsymbol{\eta}_n^*)$$

and

$$I_{4,n} = P\{\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \hat{\boldsymbol{\zeta}}_n)(\boldsymbol{\eta}^* - \boldsymbol{\eta}_n^*) - \dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\zeta}_0)(\boldsymbol{\eta}^* - \boldsymbol{\eta}_n^*)\}.$$

Let $\mathcal{L}_4 = \{\dot{l}_{\boldsymbol{\phi}}(\mathbf{Y}; \boldsymbol{\zeta})(\boldsymbol{\eta}^* - \boldsymbol{\eta}) : \boldsymbol{\zeta} \in \mathcal{T}_n, \boldsymbol{\eta} \in (\prod_{k=1}^{J}\mathcal{S}_{kn}(D_{kn}, K_{kn}, m_k))^{\otimes Jd}$ and $\|\eta_{jk}^* - \eta_{jk}\|_{\Phi_k} = O(n^{-\nu_k})$ for $j = 1, \ldots, Jd$ and $1 \leq k \leq J\}$. It can be similarly argued that the ϵ-bracketing number associated with $L_2(P)$-norm is bounded by $C(1/\epsilon)^{Jd}(1/\epsilon)^{C\sum_{k=1}^{J} q_{kn}}(1/\epsilon)^{C\sum_{k=1}^{J} q_{kn}}$, which leads to $\mathcal{L}_4$ being Donsker. Furthermore, for any $r(\mathbf{Y}; \boldsymbol{\zeta}, \boldsymbol{\eta}) \in \mathcal{L}_4$, $Pr^2 \to 0$ as $n \to \infty$. Hence $I_{3,n} = o_p(n^{-1/2})$ by van der Vaart and Wellner [28, Corollary 2.3.12]. By Cauchy–Schwarz inequality and assumptions A1–A4, it can be easily shown that $\|I_{4,n}\| \le C\sum_{j=1}^{Jd}\sum_{k=1}^{J} d(\hat{\boldsymbol{\zeta}}_{kn}, \boldsymbol{\zeta}_{k0})\|\eta^*_{jk} - \eta^*_{n,jk}\|_{\Phi_k} = O_p(n^{-\min_k[\min\{(p_k+1)\nu_k,(1+\nu_k)/2\}]}) = o_p(n^{-1/2})$, where $\hat{\boldsymbol{\zeta}}_{kn}$ and $\boldsymbol{\zeta}_{k0}$ are respectively the collections of the components of $\hat{\boldsymbol{\zeta}}_n$ and $\boldsymbol{\zeta}_0$ associated with cause k. So (B1) holds. (B2) holds by similarly verifying that the class $\mathcal{L}_5(\delta) = \{l^*_{\boldsymbol{\theta}}(\mathbf{Y}; \boldsymbol{\zeta}) - l^*_{\boldsymbol{\theta}}(\mathbf{Y}; \boldsymbol{\zeta}_0) : \boldsymbol{\zeta} \in \mathcal{T}_n \text{ and } d(\boldsymbol{\zeta}, \boldsymbol{\zeta}_0) \le \delta\}$ is P-Donsker and for any $r(\mathbf{Y}; \boldsymbol{\zeta}) \in \mathcal{L}_5(\delta)$, $Pr^2 \to 0$ as $\delta \to 0$. (B3) can be easily established using Taylor expansion and the convergence rate derived in Theorem 1. Hence the proof is complete. □

References

[1] O.O. Aalen, S. Johansen, An empirical transition matrix for non-homogeneous Markov chains based on censored observations, Scand. J. Statist. 5 (1978) 141–150.
[2] D. Bennett, J. Schneider, Z. Arvanitakis, R. Wilson, Overview and findings from the religious orders study, Curr. Alzheimer Res. 9 (2012) 628–645.
[3] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, J.A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, Johns Hopkins University Press, Baltimore, 1993.
[4] P. Billingsley, Probability and Measure, John Wiley, New York, 1986.
[5] C.S. Chasela, M.G. Hudgens, D.J. Jamieson, D. Kayira, M.C. Hosseinipour, A.P. Kourtis, F. Martinson, G. Tegha, R.J. Knight, Y.I. Ahmed, D.D. Kamwendo, I.F. Hoffman, S.R. Ellington, Z. Kacheche, A. Soko, J.B. Wiener, S.A. Fiscus, P. Kazembe, I.A. Mofolo, M. Chigwenembe, D.S. Sichali, C.M. van der Horst, Maternal or infant antiretroviral drugs to reduce HIV-1 transmission, N. Engl. J. Med. 362 (2010) 2271–2281. PMID: 20554982.
[6] J.P. Fine, Regression modeling of competing crude failure probabilities, Biostatistics 2 (2001) 85–97.
[7] J.P. Fine, R.J. Gray, A proportional hazards model for the subdistribution of a competing risk, J. Amer. Statist. Assoc. 94 (1999) 496–509.
[8] P. Groeneboom, M.H. Maathuis, J.A. Wellner, Current status data with competing risks: consistency and rates of convergence of the MLE, Ann. Statist. 36 (2008) 1031–1063.
[9] P. Groeneboom, M.H. Maathuis, J.A. Wellner, Current status data with competing risks: Limiting distribution of the MLE, Ann. Statist. 36 (2008) 1064–1089.
[10] P. Groeneboom, J.A. Wellner, Information Bounds and Nonparametric Maximum Likelihood Estimation, Birkhäuser, Basel, 1992.
[11] J. Huang, A.J. Rossini, Sieve estimation for the proportional-odds failure-time regression model with interval censoring, J. Amer. Statist. Assoc. 92 (1997) 960–967.
[12] J. Huang, J.A. Wellner, Efficient estimation for the proportional hazards model with "case 2" interval censoring. Tech. Rep., Department of Statistics, University of Washington, Seattle, 1995.
[13] J. Huang, Y. Zhang, L. Hua, A least-squares approach to consistent information estimation in semiparametric models. Tech. Rep. 2008-3, Department of Biostatistics, University of Iowa, 2008.
[14] M.G. Hudgens, C. Li, J.P. Fine, Parametric likelihood inference for interval censored competing risks data, Biometrics 70 (2014) 1–9.
[15] M.G. Hudgens, G.A. Satten, I.M. Longini JR., Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation, Biometrics 57 (2001) 74–80.
[16] J. Jeong, J.P. Fine, Direct parametric inference for the cumulative incidence function, J. R. Stat. Soc. Ser. C. Appl. Stat. 55 (2006) 187–200.
[17] N.P. Jewell, M.J. van der Laan, T. Henneman, Nonparametric estimation from current status data with competing risks, Biometrika 90 (2003) 183–197.
[18] M. Kosorok, Introduction to Empirical Processes and Semiparametric Inference, Springer, New York, 2008.
[19] K. Lange, An adaptive barrier method for convex programming, Methods Appl. Anal. 1 (1994) 392–402.
[20] L. Letenneur, V. Gilleron, D. Commenges, C. Helmer, J.-M. Orgogozo, J.-F. Dartigues, Are sex and educational level independent predictors of dementia and alzheimer's disease? Incidence data from the paquid project, J. Neurol. Neurosurg. Psychiatry 66 (1999) 177–183.
[21] C. Li, J.P. Fine, Smoothed nonparametric estimation for current status competing risks data, Biometrika 100 (2013) 173–187.
[22] M. Lu, Monotone spline estimations for panel count data (Ph.D. thesis) Department of Biostatistics, University of Iowa, 2007.
[23] R.L. Prentice, J.D. Kalbfleisch, A.V. Peterson, N. Flournoy, V.T. Farewell, N.E. Breslow, The analysis of failure times in the presence of competing risks, Biometrics 34 (1978) 541–554.
[24] A. Schick, Q. Yu, Consistency of the GMLE with mixed case interval-censored data, Scand. J. Statist. 27 (2000) 45–55.
[25] L.L. Schumaker, Spline Functions: Basic Theory, John Wiley, New York, 1981.
[26] X. Shen, W.H. Wong, Convergence rate of sieve estimates, Ann. Statist. 22 (1994) 580–615.
[27] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[28] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes with Applications to Statistics, Springer-Verlag, New York, 1996.
[29] J.A. Wellner, Y. Zhang, Two likelihood-based semiparametric estimation methods for panel count data with covariates, Ann. Statist. 35 (2007) 2106–2142.
[30] Y. Zhang, L. Hua, J. Huang, A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data, Scand. J. Statist. 37 (2010) 338–354.