Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data

By DONGLIN ZENG, FEI GAO and D. Y. LIN

Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

dzeng@bios.unc.edu fgao@live.unc.edu lin@bios.unc.edu

Summary

Interval-censored multivariate failure time data arise when there are multiple types of failure or there is clustering of study subjects and each failure time is known only to lie in a certain interval. We investigate the effects of possibly time-dependent covariates on multivariate failure times by considering a broad class of semiparametric transformation models with random effects, and we study nonparametric maximum likelihood estimation under general interval-censoring schemes. We show that the proposed estimators for the finite-dimensional parameters are consistent and asymptotically normal, with a limiting covariance matrix that attains the semiparametric efficiency bound and can be consistently estimated through profile likelihood. In addition, we develop an EM algorithm that converges stably for arbitrary datasets. Finally, we assess the performance of the proposed methods in extensive simulation studies and illustrate their application using data derived from the Atherosclerosis Risk in Communities Study.

Some key words: Current-status data; EM algorithm; Multivariate failure time data; Nonparametric likelihood; Profile likelihood; Proportional hazards; Proportional odds; Random effects.

1. Introduction

Multivariate failure time data arise when each study subject may experience multiple events or when study subjects are sampled in clusters such that the failure times are potentially correlated (Kalbfleisch & Prentice, 2002, Ch. 10). The failure times are interval-censored if the events or failures can only be determined through periodic examination. In the special case of one examination per subject, the observations are called current-status data (Huang, 1996). An example of interval-censored multiple-event data is an HIV/AIDS study where laboratory tests were performed periodically on each patient to detect the presence of cytomegalovirus in the blood and urine (Goggins & Finkelstein, 2000). An example of interval-censored clustered data is a study of pandemic H1N1 influenza where blood samples of family members were collected at different time-points to determine whether there is infection with the influenza virus (Kor et al., 2013). Such data allow characterization of the dependence of related events and evaluation of the effects of covariates on the multivariate outcome. The fact that failure times are never exactly observed, together with their dependence, makes the analysis theoretically and computationally challenging.

Several methods for regression analysis of interval-censored multiple-event data have been proposed. Specifically, Goggins & Finkelstein (2000), Kim & Xue (2002), Chen et al. (2007), Tong et al. (2008) and Chen et al. (2013) constructed estimating equations for marginal models by assuming that all subjects are examined at a common set of time-points. Chen et al. (2009) and Chen et al. (2014) considered a frailty proportional hazards model for current-status data and interval-censored data, respectively. The former assumed a piecewise-constant baseline hazard function, while the latter assumed a common set of examination times for all subjects. All the aforementioned work avoids the difficult task of nonparametric estimation by parameterizing the failure time distribution or estimating the survival probabilities at fixed time-points. Wang et al. (2008) studied sieve estimation of a copula proportional hazards model for bivariate current-status data with univariate examination time, which was parameterized by a proportional hazards model. Wen & Chen (2013) established asymptotic theory for the nonparametric maximum likelihood estimation of a gamma-frailty proportional hazards model for bivariate interval-censored data and constructed a self-consistency equation, which involves an artificial tuning constant and may have multiple solutions. Wang et al. (2015) developed an EM algorithm for spline-based sieve estimation of the same model, but for bivariate current-status data.

The literature on interval-censored clustered data is relatively limited. Cook & Tolusso (2009) and Kor et al. (2013) constructed estimating functions for a copula proportional hazards model with a piecewise-constant baseline hazard function for current-status and interval-censored data, respectively. Chang et al. (2007) established a profile likelihood theory for a gamma-frailty proportional hazards model with current-status family data, and Wen & Chen (2011) developed a self-consistency algorithm similar to that in Wen & Chen (2013).

In this paper, we provide efficient estimation methods for a broad class of semiparametric transformation models with random effects for general interval-censored multivariate failure time data. Our work advances the study of multivariate interval-censored data in several directions. First, we deal with the most general form of interval censoring, allowing each subject to have an arbitrary sequence of examination times, and we do not model the examination times. Second, our models accommodate time-dependent covariates and include both proportional and non-proportional hazards structures. Third, our models allow multiple random effects and treat multiple events and clustered data in a unified framework. Fourth, we estimate the failure time distribution in a completely nonparametric manner and avoid any tuning parameters, which are required by sieve methods. Fifth, we establish a rigorous asymptotic theory for the nonparametric maximum likelihood estimators under mild conditions. Finally, we devise an EM algorithm that involves only low-dimensional parameters in each iteration and performs well in a wide variety of situations.

The present paper also substantially extends our recent work on univariate interval-censored data (Zeng et al., 2016). We expand our previous numerical algorithm to handle unobserved random effects and multiple baseline hazard functions. We address new theoretical challenges generated by the presence of random effects, especially in proving the Donsker property of relevant functions in the form of integration over random effects. In addition, the asymptotic theory of Zeng et al. (2016) hinges on the assumption that a subset of study subjects is examined at the study endpoint; here we remove that restrictive assumption and formulate new arguments to prove the consistency of the estimators. Finally, we show that the covariance matrix for the finite-dimensional parameters can be estimated consistently by the inverse empirical covariance matrix of the individual contributions to the gradient of the profile loglikelihood function. This estimator is always positive semidefinite and is numerically more stable than the Hessian matrix used by Zeng et al. (2016) and others.

2. Data, model and likelihood

We consider a general framework for modelling multivariate failure time data that encompasses both multiple events and clustered data. Suppose that there are n independent clusters with J_i subjects in the ith cluster and that each subject can potentially experience K types of events. It is assumed that J_i is small relative to n. For $i = 1, \ldots, n$, $j = 1, \ldots, J_i$ and $k = 1, \ldots, K$, let T_{ijk} denote the kth failure time for the jth subject of the ith cluster, and let $X_{ijk}(\cdot)$ denote the corresponding p-vector of possibly time-dependent covariates. We specify that the cumulative hazard function of T_{ijk} takes the form

$$\Lambda_{ijk}(t) = G_k\left[\int_0^t \exp\{\beta^{\mathrm{T}} X_{ijk}(s) + b_i^{\mathrm{T}} Z_{ijk}(s)\}\, \mathrm{d}\Lambda_k(s)\right], \tag{1}$$

where Z_{ijk} contains 1 and covariates that may be part of X_{ijk}, b_i is a d_i-vector of random effects from the multivariate normal distribution with mean zero and covariance matrix $\Sigma_i(\gamma)$ indexed by unknown parameters γ, β is a set of unknown regression parameters, $\Lambda_k(\cdot)$ is an arbitrary increasing function with $\Lambda_k(0) = 0$, and $G_k(x)$ is a specific transformation function. It is assumed that T_{ijk} $(j = 1, \ldots, J_i;\ k = 1, \ldots, K)$ are independent conditional on b_i. By letting X_{ijk} and Z_{ijk} depend on k, model (1) allows the regression parameters and random effects to be different among the K types of events; see Lin (1994). In addition, the dependence of Z_{ijk} on j allows for subject-specific random effects. Often $\Sigma_i(\gamma)$ does not depend on i, and then γ consists of the upper diagonal elements of the common covariance matrix Σ. An example in which $\Sigma_i(\gamma)$ depends on i is given in the Supplementary Material.

A variety of transformations can be generated through the log-Laplace transform

$$G_k(x) = -\log \int_0^\infty \exp(-xt) f_k(t)\, \mathrm{d}t, \tag{2}$$

where $f_k(t)$ is a density function with support on $[0, \infty)$. The choice of the gamma density with mean 1 and variance r_k for $f_k(t)$ yields the class of logarithmic transformations $G_k(x) = r_k^{-1}\log(1 + r_k x)$ with $r_k > 0$ (Chen et al., 2002), which includes the proportional odds model, $r_k = 1$, and can be extended to include the proportional hazards model by letting $r_k = 0$.

Suppose that T_{ijk} is monitored at a sequence of positive time-points $U_{ijk1} < \cdots < U_{ijk,M_{ijk}}$. We assume that $\{U_{ijkl} : l = 1, \ldots, M_{ijk};\ j = 1, \ldots, J_i;\ k = 1, \ldots, K\}$ are independent of $\{T_{ijk} : j = 1, \ldots, J_i;\ k = 1, \ldots, K\}$ and b_i conditional on $\{X_{ijk}(\cdot) : j = 1, \ldots, J_i;\ k = 1, \ldots, K\}$. Let $(L_{ijk}, R_{ijk}]$ be the shortest time interval that brackets T_{ijk}, i.e., $L_{ijk} = \max\{U_{ijkl} : U_{ijkl} < T_{ijk},\ l = 0, \ldots, M_{ijk}\}$ and $R_{ijk} = \min\{U_{ijkl} : U_{ijkl} \geqslant T_{ijk},\ l = 1, \ldots, M_{ijk} + 1\}$, where $U_{ijk0} = 0$ and $U_{ijk,M_{ijk}+1} = \infty$. Then the likelihood concerning the parameters $\theta = (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$ and $\mathcal{A} = (\Lambda_1, \ldots, \Lambda_K)$ is

$$\begin{aligned} L_n(\theta, \mathcal{A}) = \prod_{i=1}^n \int \prod_{j=1}^{J_i} \prod_{k=1}^K \Bigg\{ & \exp\left(-G_k\left[\int_0^{L_{ijk}} \exp\{\beta^{\mathrm{T}} X_{ijk}(s) + b_i^{\mathrm{T}} Z_{ijk}(s)\}\, \mathrm{d}\Lambda_k(s)\right]\right) \\ & - \exp\left(-G_k\left[\int_0^{R_{ijk}} \exp\{\beta^{\mathrm{T}} X_{ijk}(s) + b_i^{\mathrm{T}} Z_{ijk}(s)\}\, \mathrm{d}\Lambda_k(s)\right]\right)\Bigg\} \\ & \times (2\pi)^{-d_i/2} |\Sigma_i(\gamma)|^{-1/2} \exp\left\{-\frac{b_i^{\mathrm{T}} \Sigma_i(\gamma)^{-1} b_i}{2}\right\} \mathrm{d}b_i, \end{aligned} \tag{3}$$

in which $\exp(-G_k[\int_0^{R_{ijk}} \exp\{\beta^{\mathrm{T}} X_{ijk}(s) + b_i^{\mathrm{T}} Z_{ijk}(s)\}\, \mathrm{d}\Lambda_k(s)]) = 0$ if $R_{ijk} = \infty$.

3. Nonparametric maximum likelihood estimation

We adopt the nonparametric maximum likelihood estimation approach. For each $k = 1, \ldots, K$, let $0 = t_{k0} < t_{k1} < \cdots < t_{km_k} < \infty$ be the ordered sequence of all L_{ijk} and R_{ijk} with $R_{ijk} < \infty$. The estimator for Λ_k is a step function which jumps only at those time-points with respective jump sizes of $\lambda_{k0} = 0, \lambda_{k1}, \ldots, \lambda_{km_k}$. We introduce a latent variable ξ_{ijk} with density $f_k(t)$ as given in (2). Then (3) can be written as

$$L_n(\theta, \mathcal{A}) = \prod_{i=1}^{n} \int \prod_{j=1}^{J_i} \prod_{k=1}^{K} \int \left[\exp\left\{ -\xi_{ijk} \sum_{t_{kq} \leqslant L_{ijk}} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) \lambda_{kq} \right\} \right.$$

$$\left. - I(R_{ijk} < \infty) \exp\left\{ -\xi_{ijk} \sum_{t_{kq} \leqslant R_{ijk}} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) \lambda_{kq} \right\} \right] f_k(\xi_{ijk}) \, \mathrm{d}\xi_{ijk}$$

$$\times (2\pi)^{-d_i/2} |\Sigma_i(\gamma)|^{-1/2} \exp\left\{ -\frac{b_i^{\mathrm{T}} \Sigma_i(\gamma)^{-1} b_i}{2} \right\} \mathrm{d}b_i, \tag{4}$$

where $X_{ijkq} = X_{ijk}(t_{kq})$ and $Z_{ijkq} = Z_{ijk}(t_{kq})$.

To make the maximization of the likelihood more tractable, we introduce independent Poisson random variables W_{ijkq} $(q = 1, \ldots, m_k)$ with means $\lambda_{kq} \xi_{ijk} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq})$. Let $A_{ijk} = \sum_{t_{kq} \leqslant L_{ijk}} W_{ijkq}$ and $B_{ijk} = I(R_{ijk} < \infty) \sum_{L_{ijk} < t_{kq} \leqslant R_{ijk}} W_{ijkq}$. Because the joint probability of $A_{ijk} = 0$ and $B_{ijk} > 0$ given ξ_{ijk} and b_i is $\exp\{-\xi_{ijk} \sum_{t_{kq} \leqslant L_{ijk}} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) \lambda_{kq}\} - I(R_{ijk} < \infty) \exp\{-\xi_{ijk} \sum_{t_{kq} \leqslant R_{ijk}} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) \lambda_{kq}\}$, the likelihood arising from the observations $(A_{ijk} = 0, B_{ijk} > 0 : i = 1, \ldots, n; j = 1, \ldots, J_i; k = 1, \ldots, K)$ is the same as (4). Therefore, we develop an EM algorithm to maximize (4) by treating W_{ijkq} $(t_{kq} \leqslant R^*_{ijk})$, ξ_{ijk} and b_i as complete data, where $R^*_{ijk} = L_{ijk} I(R_{ijk} = \infty) + R_{ijk} I(R_{ijk} < \infty)$.

Remark 1. Conditional on ξ_{ijk} and b_i, the failure time T_{ijk} follows a proportional hazards model. Let $N_{ijk}(t)$ be a Poisson process with value 0 at $t = 0$ and intensity function the same as the hazard function of T_{ijk}. Clearly, T_{ijk} is the first time $N_{ijk}(t)$ jumps from 0 to 1, such that T_{ijk} falling in the interval $(L_{ijk}, R_{ijk}]$ is equivalent to $N_{ijk}(t)$ taking no jump before L_{ijk} but at least one jump in $(L_{ijk}, R_{ijk}]$. Thus, A_{ijk} and B_{ijk} are indeed the counts of $N_{ijk}(t)$ before L_{ijk} and between L_{ijk} and R_{ijk}, respectively.

The complete-data loglikelihood is

$$\sum_{i=1}^{n} \left\{ \sum_{j=1}^{J_i} \sum_{k=1}^{K} \left(\sum_{q=1}^{m_k} I(t_{kq} \leqslant R^*_{ijk}) \Big[W_{ijkq} \log\big\{ \lambda_{kq} \xi_{ijk} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) \big\} \right. \right.$$

$$\left. \left. - \lambda_{kq} \xi_{ijk} \exp(\beta^{\mathrm{T}} X_{ijkq} + b_i^{\mathrm{T}} Z_{ijkq}) - \log(W_{ijkq}!) \Big] + \log f_k(\xi_{ijk}) \right) \right.$$

$$\left. - \frac{d_i}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i(\gamma)| - \frac{b_i^{\mathrm{T}} \Sigma_i(\gamma)^{-1} b_i}{2} \right\}. \tag{5}$$

In the M-step, we solve the following equation for β using the one-step Newton–Raphson method:

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K}\sum_{q=1}^{m_k} I(t_{kq} \leqslant R^*_{ijk})\hat{E}(W_{ijkq}) \times \left[X_{ijkq} - \frac{\sum_{i'=1}^{n}\sum_{j'=1}^{J_{i'}} I(t_{kq} \leqslant R^*_{i'j'k})X_{i'j'kq}\hat{E}\{\xi_{i'j'k}\exp(\beta^{\mathrm{T}}X_{i'j'kq} + b_{i'}^{\mathrm{T}}Z_{i'j'kq})\}}{\sum_{i'=1}^{n}\sum_{j'=1}^{J_{i'}} I(t_{kq} \leqslant R^*_{i'j'k})\hat{E}\{\xi_{i'j'k}\exp(\beta^{\mathrm{T}}X_{i'j'kq} + b_{i'}^{\mathrm{T}}Z_{i'j'kq})\}} \right] = 0,$$

where $\hat{E}(\cdot)$ denotes the conditional expectation given the observed data. We then calculate

$$\lambda_{kq} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J_i} I(t_{kq} \leqslant R^*_{ijk})\hat{E}(W_{ijkq})}{\sum_{i=1}^{n}\sum_{j=1}^{J_i} I(t_{kq} \leqslant R^*_{ijk})\hat{E}\{\xi_{ijk}\exp(\beta^{\mathrm{T}}X_{ijkq} + b_i^{\mathrm{T}}Z_{ijkq})\}}$$

for $q = 1, \ldots, m_k$ and $k = 1, \ldots, K$, and maximize $-\log|\Sigma_i(\gamma)| - \hat{E}\{b_i^{\mathrm{T}}\Sigma_i^{-1}(\gamma)b_i\}$ to estimate γ. If the Σ_i are the same and nonparametric, then the latter becomes $\Sigma = n^{-1}\sum_{i=1}^{n}\hat{E}(b_i^{\otimes 2})$, where $a^{\otimes 2} = aa^{\mathrm{T}}$.

In the E-step, we evaluate the conditional expectations involved in the M-step. We use the fact that the joint density of ξ_{ijk} $(j = 1, \ldots, J_i;\ k = 1, \ldots, K)$ and b_i given the observed data is proportional to

$$\prod_{j=1}^{J_i}\prod_{k=1}^{K}\left[\exp\left\{-\xi_{ijk}\sum_{t_{kq}\leqslant L_{ijk}}\exp(\beta^{\mathrm{T}}X_{ijkq} + b_i^{\mathrm{T}}Z_{ijkq})\lambda_{kq}\right\} - I(R_{ijk} < \infty)\exp\left\{-\xi_{ijk}\sum_{t_{kq}\leqslant R_{ijk}}\exp(\beta^{\mathrm{T}}X_{ijkq} + b_i^{\mathrm{T}}Z_{ijkq})\lambda_{kq}\right\}\right] \times f_k(\xi_{ijk})(2\pi)^{-d_i/2}|\Sigma_i(\gamma)|^{-1/2}\exp\left\{-\frac{b_i^{\mathrm{T}}\Sigma_i(\gamma)^{-1}b_i}{2}\right\}.$$

In addition, the conditional mean of W_{ijkq} for $t_{kq} \leqslant R^*_{ijk}$ given ξ_{ijk} $(j = 1, \ldots, J_i;\ k = 1, \ldots, K)$, b_i and the observed data is

$$I(L_{ijk} < t_{kq} \leqslant R_{ijk} < \infty)\frac{\lambda_{kq}\xi_{ijk}\exp(\beta^{\mathrm{T}}X_{ijkq} + b_i^{\mathrm{T}}Z_{ijkq})}{1 - \exp\{-\sum_{L_{ijk}<t_{kq'}\leqslant R_{ijk}}\lambda_{kq'}\xi_{ijk}\exp(\beta^{\mathrm{T}}X_{ijkq'} + b_i^{\mathrm{T}}Z_{ijkq'})\}}.$$

We use Gaussian quadrature to approximate integrals over ξ_{ijk} and b_i.

Starting with $\beta = 0$, $\lambda_{kq} = 1/m_k$ and Σ_i as the identity matrix, we iterate between the E-step and the M-step until convergence to obtain the nonparametric maximum likelihood estimators $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\Lambda}_k$ $(k = 1, \ldots, K)$. The high-dimensional parameters λ_{kq} are calculated explicitly in the M-step. We show in the Supplementary Material that each iteration of the algorithm guarantees an increase in the likelihood. Due to the presence of random effects, the conditional expectations in this EM algorithm are more tedious to evaluate than those in Zeng et al. (2016).

4. Asymptotic properties

Let $\hat{\theta} = (\hat{\beta}^{\mathrm{T}}, \hat{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\mathcal{A}} = (\hat{\Lambda}_1, \ldots, \hat{\Lambda}_K)$. We establish the asymptotic properties of $(\hat{\theta}, \hat{\mathcal{A}})$ under the following regularity conditions, wherein we omit the subscript i when referring to a random variable for a cluster and use the notation $Q_{jk}(t, b; \beta, \Lambda_k) = \exp(-G_k[\int_0^t \exp\{\beta^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\}\, \mathrm{d}\Lambda_k(s)])$ and $\phi(b; \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp(-b^{\mathrm{T}} \Sigma^{-1} b/2)$.

Condition 1. The true value of θ, denoted by $\theta_0 = (\beta_0^{\mathrm{T}}, \gamma_0^{\mathrm{T}})^{\mathrm{T}}$, lies in the interior of a known compact set $\Theta = \{(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}} : \beta \in \mathcal{B},\ \gamma \in \mathcal{C}\}$, where $\mathcal{B}$ is a compact set in $\mathbb{R}^p$ and $\mathcal{C}$ is a compact set in the domain of γ such that $\Sigma(\gamma)$ is a positive-definite matrix with eigenvalues bounded away from zero and ∞. The true value of Λ_k, denoted by Λ_{0k}, is continuously differentiable with positive derivatives in $[0, \tau_k]$, which is the union of the supports of U_{jkl} $(l = 1, \ldots, M_{jk};\ j = 1, \ldots, J)$.

Condition 2. With probability one, $X_{jk}(\cdot)$ has bounded total variation in $[0, \tau_k]$. If there exists a deterministic function $a_1(t)$ and a constant vector a_2 such that $a_1(t) + a_2^{\mathrm{T}} X_{jk}(t) = 0$ with probability 1, then $a_1(t) = 0$ for $t \in [0, \tau_k]$ and $a_2 = 0$.

Condition 3. With probability one, $Z_{jk}(\cdot)$ has bounded total variation in $[0, \tau_k]$.

Condition 4. The cluster size J is bounded by a positive constant and is independent of $\{T_{jk} : j = 1, \ldots, J;\ k = 1, \ldots, K\}$, $\{U_{jkl} : l = 1, \ldots, M_{jk};\ j = 1, \ldots, J;\ k = 1, \ldots, K\}$ and b conditional on (X_{jk}, Z_{jk}) $(j = 1, \ldots, J;\ k = 1, \ldots, K)$.

Condition 5. For any $j = 1, \ldots, J$ and $k = 1, \ldots, K$, the number of examination times M_{jk} is positive with $E(M_{jk}) < \infty$. The conditional densities of $(U_{jkl}, U_{jk,l+1})$ given (J, M_{jk}, X_{jk}), denoted by $g_{jkl}(u, v)$ $(l = 0, \ldots, M_{jk})$, have continuous second-order partial derivatives with respect to u and v when $v - u \geqslant \eta$ for some positive constant η, and are continuously differentiable functionals with respect to X_{jk} and Z_{jk}. In addition, $\mathrm{pr}\{\min_{0 \leqslant l < M_{jk}} (U_{jk,l+1} - U_{jkl}) \geqslant \eta \mid J, M_{jk}, X_{jk}\} = 1$.

Condition 6. The transformation function G_k is twice continuously differentiable on $[0, \infty)$ with $G_k(0) = 0$, $G_k'(x) > 0$ and $G_k(\infty) = \infty$ for $k = 1, \ldots, K$, where $G_k'(x) = \mathrm{d}G_k(x)/\mathrm{d}x$. In addition, $G_k'(x) \exp\{-G_k(x)\}$ is uniformly bounded in $x \geqslant 0$ and there exists a positive constant r_{k0} such that $\exp\{-G_k(x)\} = O(x^{-1/r_{k0}})$ as $x \to \infty$.

Condition 7. For a pair of parameters $(\theta_1, \mathcal{A}_1)$ and $(\theta_2, \mathcal{A}_2)$, if

$$\int \left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta_1, \Lambda_{1k'}) \right\} \phi\{b; \Sigma(\gamma_1)\}\, \mathrm{d}b$$
$$= \int \left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta_2, \Lambda_{2k'}) \right\} \phi\{b; \Sigma(\gamma_2)\}\, \mathrm{d}b$$

with probability 1 for any $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$ and $t_{j'k'} \in [0, \tau_{k'}]$ with $j' \in \{1, \ldots, j\}$ and $k' \in \{1, \ldots, k\}$, then $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$ and $\Lambda_{1k}(t) = \Lambda_{2k}(t)$ for $t \in [0, \tau_k]$ and $k \in \{1, \ldots, K\}$.

Condition 8. If there exists a vector v and functions $a_{jk}(t;b)$ $(j=1,\ldots,J;\ k=1,\ldots,K)$ such that

$$\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k}Q_{j'k'}(t_{j'k'},b;\beta_0,\Lambda_{0k'})\right\}\left\{\sum_{j'=1}^{j}\sum_{k'=1}^{k}a_{j'k'}(t_{j'k'},b)+\frac{v^{\mathrm{T}}\phi'_{\gamma}(b;\Sigma_0)}{\phi(b;\Sigma_0)}\right\}\phi(b;\Sigma_0)\,\mathrm{d}b=0$$

with probability one for any $j\in\{1,\ldots,J\}$, $k\in\{1,\ldots,K\}$ and $t_{j'k'}\in[0,\tau_{k'}]$ with $j'\in\{1,\ldots,j\}$ and $k'\in\{1,\ldots,k\}$, where $\Sigma_0=\Sigma(\gamma_0)$ and ϕ'_{γ} is the derivative of $\phi\{b;\Sigma(\gamma)\}$ with respect to γ, then $v=0$ and $a_{jk}(t,b)=0$ for $j=1,\ldots,J$, $t\in[0,\tau_k]$ and $k\in\{1,\ldots,K\}$.

Remark 2. Conditions 1–4 are standard conditions for multivariate failure time regression. Condition 5 requires that two adjacent examination times be separated by at least η; otherwise, the data may contain exact observations, which need a different treatment. This condition also requires smoothness of the joint density of the examination times. Unlike Zeng et al. (2016), we do not require a subset of study subjects to be examined at the end of the study. Condition 6 holds for both the logarithmic family $G_r(x)=r^{-1}\log(1+rx)$ $(r\geqslant 0)$ and the Box–Cox family $G_\rho(x)=\rho^{-1}\{(1+x)^{\rho}-1\}$ $(\rho\geqslant 0)$, where $G_r(x)=x$ if $r=0$ and $G_\rho(x)=\log(1+x)$ if $\rho=0$. Condition 7 pertains to parameter identifiability, and Condition 8 says that the Fisher information along any submodel at the true parameter values should be nonsingular. If X_{jk} and Z_{jk} are time-independent and $G_k(x)=x$, then the equations in Conditions 7 and 8 become

$$\begin{aligned}&\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k}\exp(\tilde{\beta}_1^{\mathrm{T}}[1,X_{j'k'}^{\mathrm{T}}]^{\mathrm{T}}+b^{\mathrm{T}}Z_{j'k'})\right\}\phi\{b;\Sigma(\gamma_1)\}\,\mathrm{d}b\\&\qquad=\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k}\exp(\tilde{\beta}_2^{\mathrm{T}}[1,X_{j'k'}^{\mathrm{T}}]^{\mathrm{T}}+b^{\mathrm{T}}Z_{j'k'})\right\}\phi\{b;\Sigma(\gamma_2)\}\,\mathrm{d}b,\\&\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k}\exp(\beta_0^{\mathrm{T}}X_{j'k'}+b^{\mathrm{T}}Z_{j'k'})\right\}\\&\qquad\times\left\{\sum_{j'=1}^{j}\sum_{k'=1}^{k}v_1^{\mathrm{T}}[1,X_{j'k'}^{\mathrm{T}}]^{\mathrm{T}}+\frac{v_2^{\mathrm{T}}\phi'_{\gamma}(b;\Sigma_0)}{\phi(b;\Sigma_0)}\right\}\phi(b;\Sigma_0)\,\mathrm{d}b=0,\end{aligned}$$

respectively. It can be shown that the above equations hold if Z_{jk} is linearly independent; that is, any symmetric matrix C satisfying $Z_{jk}^{\mathrm{T}}CZ_{jk}=0$ with probability one must be a zero matrix.

We state the strong consistency and weak convergence of the nonparametric maximum likelihood estimators in Theorems 1 and 2, respectively.

THEOREM 1. *Under Conditions 1–7,* $\|\hat{\theta}-\theta_0\|+\sum_{k=1}^{K}\sup_{t\in[0,\tau_k]}|\hat{\Lambda}_k(t)-\Lambda_{0k}(t)|\to 0$ *almost surely, where* $\|\cdot\|$ *is the Euclidean norm.*

THEOREM 2. *Under Conditions 1–8,* $n^{1/2}(\hat{\theta}-\theta_0)$ *converges in distribution to a zero-mean multivariate normal vector whose covariance matrix attains the semiparametric efficiency bound.*

Remark 3. The proofs of the theorems are given in the Appendix. In the proof of Theorem 1, a major challenge is to show uniform boundedness of $\hat{\Lambda}_k$ without assuming that there is a positive

probability of $R_{jk} = \tau_k$. To address this challenge, we first obtain a sequence of $\hat{\Lambda}_k$ that converges for any interior compact sets of $[0, \tau_k)$. We then show that the limit of the sequence is the true parameter value by deriving the covering number for the loglikelihood function. In the proof of Theorem 2, we use the bounded inverse theorem to establish the convergence rates of the $\hat{\Lambda}_k$ in terms of n and the Euclidean distance of the other parameter estimators, and we show that the rates obtained are sufficient for the asymptotic normality and efficiency of the estimators.

Let $\mathrm{pl}_n(\theta) = \max_{\mathcal{A}} \log L_n(\theta, \mathcal{A})$, which is obtained by using the above EM algorithm but updating only $(\Lambda_1, \ldots, \Lambda_K)$ in the M-step. One may estimate the covariance matrix of $\hat{\theta}$ by the negative inverse of the Hessian matrix of $\mathrm{pl}_n(\theta)$ at $\hat{\theta}$, which is determined by the numerical differences of second order and a perturbation constant of the order of $n^{-1/2}$ (Murphy & van der Vaart, 2000; Zeng et al., 2016). The estimated matrix may be negative definite, especially in small samples. We propose to estimate the covariance matrix of $\hat{\theta}$ by $(n\hat{V}_n)^{-1}$ with

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \left[\left\{ \frac{\partial}{\partial \theta} l_i(\theta, \hat{\mathcal{A}}_\theta) \Big|_{\theta=\hat{\theta}} \right\}^{\otimes 2} \right],$$

where $\hat{\mathcal{A}}_\theta = \arg\max_{\mathcal{A}} \log L_n(\theta, \mathcal{A})$ for $\theta \in \Theta$ and $l_i(\theta, \mathcal{A})$ is the loglikelihood function for the ith cluster. Thus, we estimate the information matrix for θ_0 by the empirical covariance matrix of the gradient of $l_i(\theta, \hat{\mathcal{A}}_\theta)$. We approximate this gradient by a first-order numerical difference, which is quicker to calculate than its second-order counterpart. The resulting covariance matrix estimator is guaranteed to be positive semidefinite and turns out to be more robust with respect to choice of the perturbation constant than the estimator based on the second-order numerical difference. The consistency of this covariance estimator is stated in the following theorem.

THEOREM 3. *Under Conditions 1–8, $\hat{V}_n^{-1}$ is a consistent estimator for the limiting covariance matrix of $n^{1/2}(\hat{\theta} - \theta_0)$.*

5. SIMULATION STUDIES

To evaluate the performance of the proposed methods, we conducted two series of simulation studies. The first series pertained to clustered data, the cluster sizes being 1, 2 and 3 with probabilities 0·2, 0·7 and 0·1, respectively. We considered model (1) with $K = 1$ and $\Lambda(t) = \log(1+0{\cdot}5t)$. We generated two independent cluster-level covariates, the first being Ber(0·5) and the second Un(0, 1). We set the corresponding regression parameters β_1 and β_2 to 0·5 and $-0{\cdot}5$, respectively. We adopted the class of logarithmic transformations indexed by parameter r and obtained the random effect b from $N(0, \sigma^2)$ where $\sigma^2 = 0{\cdot}5$. We generated five potential examination times for each subject, with the first being Un(0, 1) and the gap between any two successive examination times being $0{\cdot}1 + \mathrm{Un}(0, 1)$. We assumed that the study ended at time 5, beyond which no examinations occurred. We simulated 10 000 replicates.

Table 1 summarizes the results on the estimation of $\beta = (\beta_1, \beta_2)^{\mathrm{T}}$ and σ^2 for various values of n and r, and Fig. 1 displays the corresponding results for the estimation of $\Lambda(t)$. The biases for all parameter estimators are small and decrease as n increases. The variance estimator for $\hat{\beta}$ is accurate, and the variance of $\hat{\sigma}^2$ tends to be overestimated. The confidence intervals for both β and σ^2 have proper coverage probabilities. Additional studies revealed that the variance estimator for $\hat{\sigma}^2$ and the confidence intervals for σ^2 become more accurate as σ^2 increases.

The second series of studies was concerned with multiple events. We considered model (1) with $K = 2$, $J = 1$, $\Lambda_1(t) = \log(1 + 0{\cdot}5t)$ and $\Lambda_2(t) = 0{\cdot}5t$. We focused on the logarithmic

Table 1. *Parameter estimation results for simulation studies with clustered data*

		$n = 100$				$n = 200$				$n = 400$			
r		Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	$\beta_1 = 0{\cdot}5$	0·014	0·263	0·258	94	0·005	0·182	0·180	95	0·002	0·127	0·126	95
	$\beta_2 = -0{\cdot}5$	−0·008	0·404	0·399	95	−0·005	0·278	0·277	95	−0·003	0·194	0·194	95
	$\sigma^2 = 0{\cdot}5$	−0·024	0·369	0·384	96	−0·009	0·244	0·259	97	−0·001	0·166	0·177	97
0·5	$\beta_1 = 0{\cdot}5$	0·014	0·302	0·299	95	0·004	0·210	0·208	95	0·002	0·147	0·146	95
	$\beta_2 = -0{\cdot}5$	−0·010	0·483	0·479	95	−0·007	0·333	0·331	95	−0·004	0·233	0·232	95
	$\sigma^2 = 0{\cdot}5$	−0·027	0·457	0·486	96	−0·010	0·309	0·330	96	0·001	0·214	0·228	96
1	$\beta_1 = 0{\cdot}5$	0·015	0·341	0·341	95	0·004	0·237	0·235	95	0·002	0·166	0·165	95
	$\beta_2 = -0{\cdot}5$	−0·012	0·558	0·552	95	−0·008	0·382	0·381	95	−0·005	0·268	0·266	95
	$\sigma^2 = 0{\cdot}5$	−0·036	0·558	0·607	95	−0·018	0·380	0·412	95	−0·001	0·265	0·286	95

SE, empirical standard error; SEE, mean standard error estimator; CP, empirical coverage percentage of 95% confidence interval. For σ^2, Bias and SEE are based on the median instead of the mean, and the confidence interval is based on the log transformation. Each entry is based on 10 000 replicates.

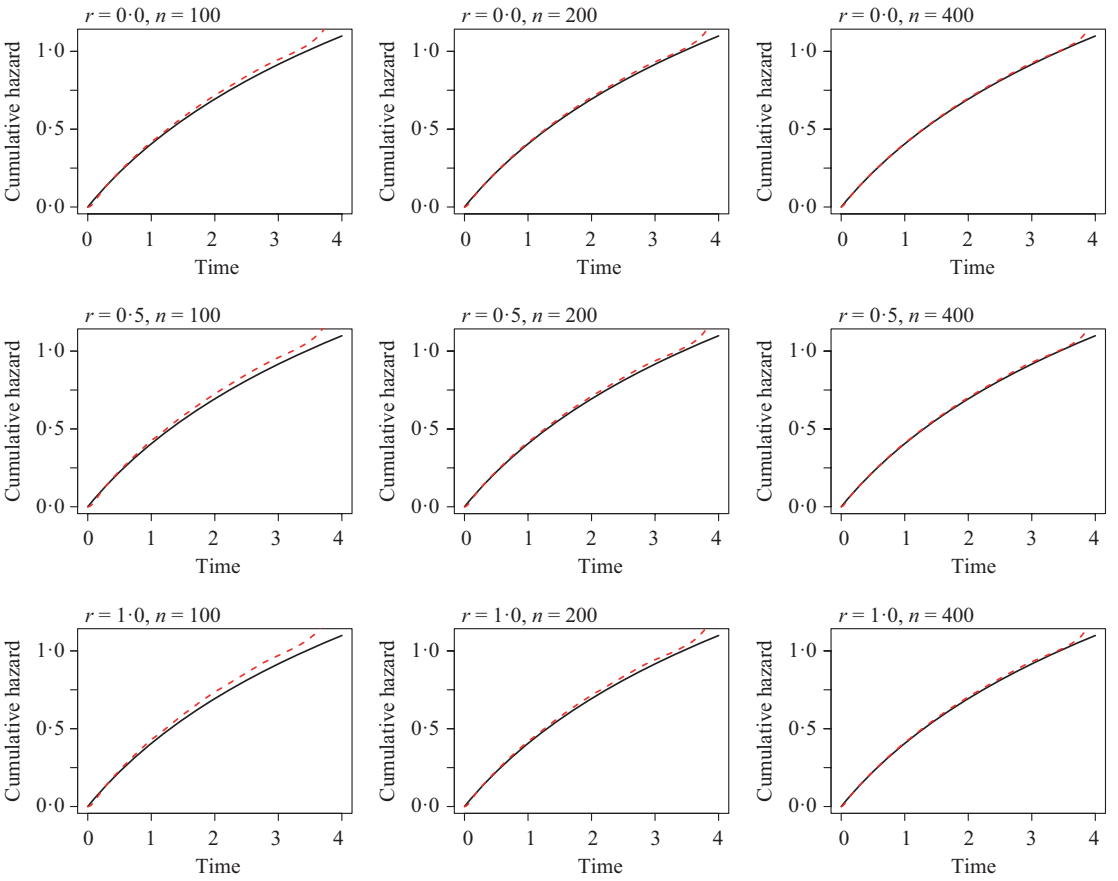

Fig. 1. Estimation of $\Lambda(t)$ for clustered data: the solid and dashed curves show the true values and averaged estimates, respectively, where each estimate is based on 10 000 replicates.

families indexed by r_1 and r_2. For each subject, we generated covariates and random effects from the same distributions as in the first series of studies. We set the regression parameters for the first event, (β_{11}, β_{12}), to $(0{\cdot}5, -0{\cdot}5)$ and those of the second event, (β_{21}, β_{22}), to $(0{\cdot}4, 0{\cdot}2)$. We generated examination times for each subject in the same manner as in the first series of studies. The results for the second series of studies are presented in the Supplementary Material. The basic conclusions are the same as those from the first series.

The variance estimation was based on the first-order numerical differentiation with a perturbation constant of $5n^{-1/2}$. The results are quite stable for perturbation constants between $n^{-1/2}$ and $10n^{-1/2}$. We also evaluated variance estimation based on the second-order numerical differentiation and found that the resulting variance estimates may be negative when n is small and the perturbation constant is far away from $5n^{-1/2}$. The two variance estimation methods produced similar estimates in most cases. We recommend using $5n^{-1/2}$ for both the first-order and the second-order numerical differences.

6. An example

The Atherosclerosis Risk in Communities Study recruited a cohort of 14 751 Caucasian and African-American individuals from four U.S. communities: Forsyth County, North Carolina; Jackson, Mississippi; suburbs of Minneapolis, Minnesota; and Washington County, Maryland (The ARIC Investigators, 1989). The participants underwent a baseline examination in 1987–1989, three follow-up examinations at approximately three-year intervals, and a further examination in 2011–2013. One important objective of the study was to investigate risk factors for diabetes and hypertension. The definition of diabetes was a fasting glucose level of 126 mg/dL or above, a nonfasting glucose level of 200 mg/dL or above, self-reported physician diagnosis of diabetes, or use of diabetic medication. The definition of hypertension was systolic blood pressure of 140 mmHg or higher, diastolic blood pressure of 90 mmHg or higher, or use of antihypertensive medication. Both events were determined at the examination times and thus interval-censored.

We related the incidence of diabetes and hypertension to race, gender, communities and five baseline risk factors: age, body mass index, glucose level, systolic blood pressure and diastolic blood pressure. We excluded 5890 individuals with prevalent diabetes or hypertension and 124 individuals with unknown status at baseline. After removing another two individuals with missing values of baseline risk factors, we were left with a total of 8735 individuals. We fitted model (1) with $K = 2, J = 1$ and $b \sim N(0, \sigma^2)$.

The loglikelihood is maximized at $r_1 = 2{\cdot}1$ and $r_2 = 1{\cdot}3$, which is the combination that would be selected by the Akaike information criterion. The loglikelihood values are $-12\,492{\cdot}67$, $-12\,412{\cdot}67$ and $-12\,403{\cdot}46$ at $(r_1, r_2) = (0, 0)$, $(1, 1)$ and $(2{\cdot}1, 1{\cdot}3)$, respectively.

Table 2 shows regression analysis results for the aforementioned three combinations of r_1 and r_2. The p-values are similar. The results indicate that African-Americans are more likely to develop diabetes and hypertension than Caucasians; baseline body mass index is positively associated with the risk of both diabetes and hypertension; and baseline glucose level is positively associated with the risk of diabetes but not hypertension. Not surprisingly, baseline systolic and diastolic blood pressures are positively associated with the risk of hypertension. Because n is large, some of the p-values are extremely small.

The regression parameters have different interpretations under different transformation models. Under the proportional odds model, the regression parameters pertain to the log hazard ratios at baseline, and the hazard ratios decrease over time. Therefore, estimates of the regression parameters tend to have larger magnitudes under the proportional odds model than under the

Table 2. *Regression analysis results for the Atherosclerosis Risk in Communities Study*

		Diabetes			Hypertension		
(r_1, r_2)	Risk factor	Estimate	Std error	p-value	Estimate	Std error	p-value
(0, 0)	Jackson	−0·188	0·194	0·332	−0·251	0·139	0·070
	Minneapolis suburbs	−0·436	0·085	$<10^{-4}$	−0·129	0·054	0·018
	Washington County	0·131	0·081	0·106	0·094	0·055	0·087
	Age	−0·015	0·006	0·011	0·016	0·004	$<10^{-4}$
	Male	−0·082	0·060	0·172	−0·268	0·041	$<10^{-4}$
	Caucasian	−0·563	0·192	0·003	−0·569	0·138	$<10^{-4}$
	Body mass index (kg/m^2)	0·088	0·006	$<10^{-4}$	0·021	0·004	$<10^{-4}$
	Derived glucose value (mg/dl)	0·108	0·003	$<10^{-4}$	0·0003	0·002	0·914
	Systolic blood pressure (mmHg)	0·006	0·003	0·070	0·072	0·003	$<10^{-4}$
	Diastolic blood pressure (mmHg)	0·005	0·005	0·271	0·014	0·003	$<10^{-4}$
(1, 1)	Jackson	−0·189	0·240	0·432	−0·311	0·163	0·056
	Minneapolis suburbs	−0·526	0·101	$<10^{-4}$	−0·164	0·070	0·019
	Washington County	0·149	0·097	0·123	0·113	0·072	0·114
	Age	−0·016	0·007	0·025	0·022	0·005	$<10^{-4}$
	Male	−0·099	0·072	0·170	−0·303	0·053	$<10^{-4}$
	Caucasian	−0·722	0·237	0·002	−0·773	0·163	$<10^{-4}$
	Body mass index (kg/m^2)	0·108	0·008	$<10^{-4}$	0·030	0·006	$<10^{-4}$
	Derived glucose value (mg/dl)	0·130	0·004	$<10^{-4}$	−0·0004	0·003	0·906
	Systolic blood pressure (mmHg)	0·008	0·004	0·053	0·093	0·003	$<10^{-4}$
	Diastolic blood pressure (mmHg)	0·005	0·006	0·351	0·020	0·004	$<10^{-4}$
(2·1, 1·3)	Jackson	−0·201	0·277	0·467	−0·337	0·166	0·043
	Minneapolis suburbs	−0·607	0·116	$<10^{-4}$	−0·174	0·075	0·021
	Washington County	0·161	0·112	0·150	0·119	0·077	0·126
	Age	−0·016	0·008	0·044	0·024	0·005	$<10^{-4}$
	Male	−0·114	0·084	0·178	−0·312	0·057	$<10^{-4}$
	Caucasian	−0·875	0·271	0·001	−0·844	0·168	$<10^{-4}$
	Body mass index (kg/m^2)	0·127	0·010	$<10^{-4}$	0·033	0·006	$<10^{-4}$
	Derived glucose value (mg/dl)	0·150	0·005	$<10^{-4}$	−0·0006	0·003	0·864
	Systolic blood pressure (mmHg)	0·010	0·005	0·036	0·101	0·004	$<10^{-4}$
	Diastolic blood pressure (mmHg)	0·004	0·007	0·496	0·022	0·004	$<10^{-4}$

proportional hazards model. The variance component σ^2 was estimated at 0·591, 0·646 and 0·758 under the proportional hazards, proportional odds and selected models, respectively, and the corresponding standard error estimates were 0·057, 0·087 and 0·111. Thus, there is strong evidence for dependence of diabetes and hypertension.

Figure 2 shows the prediction of development of diabetes and hypertension for a Caucasian female and an African-American female with all other risk factors equal. The risk of both diseases is considerably higher for the African-American individual than the Caucasian individual. The three models yield appreciably different estimates of disease-free probabilities.

7. Remarks

The proposed EM algorithm, which is used for both parameter estimation and variance estimation, performs remarkably well in practical settings, as demonstrated by the simulation studies and real-data example. We have not encountered nonconvergence with any simulated or empirical datasets. The computing time depends on the number of subjects, the number of distinct interval endpoints and the number of covariates, as well as on the convergence criterion. For the results

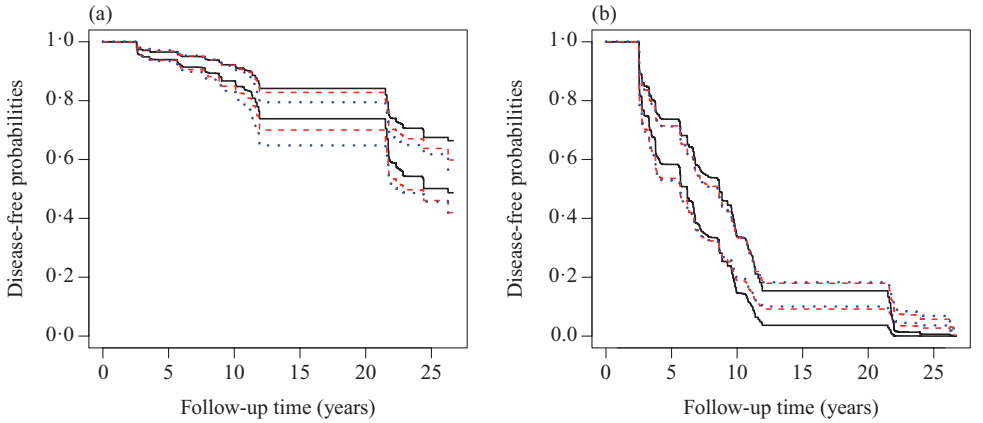

Fig. 2. Estimation of disease-free probabilities for an African-American female and a Caucasian female residing in Forsyth County, North Carolina, of age 53 years, with a body mass index of 30 kg/m^2, glucose level of 97 mg/dl, systolic blood pressure of 125 mmHg and diastolic blood pressure of 70 mmHg: (a) diabetes; (b) hypertension. In each panel the upper solid, dashed and dotted curves represent the Caucasian individual under the proportional hazards, proportional odds and selected models, respectively; the lower solid, dashed and dotted curves pertain to the African-American individual under the proportional hazards, proportional odds and selected models, respectively.

presented in this paper, the convergence criterion was that the maximal relative change in the parameter estimates at two successive iterations should be less than 0·0005. With this criterion, it took less than half a second to analyse one simulated dataset with $n = 200$. It took about 10 hours to analyse the Atherosclerosis Risk in Communities Study data, which involves 8765 subjects with 10 covariates and 2240 or 2303 distinct interval endpoints for diabetes or hypertension, respectively; the computing time was shortened to about one hour when the distinct values were reduced to 133 for diabetes and 138 for hypertension by rounding the examination times to the nearest month. The software implementing the proposed methods is available at http://dlin.web.unc.edu/software.

We have assumed that the support of the examination times for the kth type of event is an interval $[0, \tau_k]$. We can relax this assumption to let the support consist of intervals or a finite number of discrete time-points. The asymptotic results continue to hold, although the consistency for $\hat{\Lambda}_k$ in Theorem 1 should be stated to hold in the support of the examination times. In the proofs, the integration over $[0, \tau_k]$ should be changed to integration over the support.

The framework presented in this paper can be extended to other types of multivariate data. In particular, model (1) can be extended to panel count data (Zhang, 2002) by treating Λ as the intensity function of a counting process rather than the hazard function of a failure time. In addition, model (1) can be combined with a generalized linear mixed model that shares the random effects to jointly model longitudinal and survival data (Henderson et al., 2000; Zeng & Lin, 2007). There are new theoretical and computational challenges in estimating such multivariate models with interval-censored data.

Acknowledgement

This work was supported by the U.S. National Institutes of Health. The authors thank Paul Bunn for programming assistance and the reviewers for helpful comments. The Atherosclerosis

Risk in Communities Study was carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute. The authors thank the staff and participants of the ARIC study for their important contributions.

Supplementary material

Supplementary material available at *Biometrika* online includes three lemmas as well as three figures and six tables presenting additional simulation results.

Appendix

Proofs of the asymptotic results

In this appendix we prove Theorems 1–3. The proofs make use of three lemmas, which are stated and proved in the Supplementary Material. It is convenient to use empirical process notation: $\mathbb{P}_n$ denotes the empirical measure for n independent clusters, $\mathbb{P}$ is the true probability measure, and $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$ is the empirical process. Let $L(\theta, \mathcal{A})$ be the likelihood for a single cluster, such that the loglikelihood is

$$l(\theta, \mathcal{A}) = \log \int \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} D_{jk}(U_{jk}, b; \beta, \Lambda_k) \right\} \phi(b; \Sigma)\, \mathrm{d}b$$

where $D_{jk}(U_{jk}, b; \beta, \Lambda_k) = \sum_{l=0}^{M_{jk}} \Delta_{jkl}\{Q_{jk}(U_{jkl}, b; \beta, \Lambda_k) - Q_{jk}(U_{jk,l+1}, b; \beta, \Lambda_k)\}$, $U_{jk} = (U_{jk1}, \ldots, U_{jk,M_{jk}})$ and $\Delta_{jkl} = I(U_{jkl} \leqslant T_{jk} < U_{jk,l+1})$.

Proof of Theorem 1. We first show that $\limsup_n \hat{\Lambda}_k(\tau_k - \epsilon) < \infty$ with probability 1 for any $\epsilon > 0$ and $k \in \{1, \ldots, K\}$. Write

$$m(\theta, \mathcal{A}) = \log \left\{ \frac{L(\theta, \mathcal{A}) + L(\theta_0, \mathcal{A}_0)}{2} \right\},$$

where $\mathcal{A}_0 = (\Lambda_{01}, \ldots, \Lambda_{0K})$. Since $(\hat{\theta}, \hat{\mathcal{A}})$ maximizes the likelihood,

$$\mathbb{P}_n l(\hat{\theta}, \hat{\mathcal{A}}) \geqslant \mathbb{P}_n l(\theta_0, \mathcal{A}_0) = \mathbb{P}_n m(\theta_0, \mathcal{A}_0).$$

We show in Lemma 1 that $\mathcal{M} = \{m(\theta, \mathcal{A}) : \theta \in \Theta, \mathcal{A} \in \mathcal{L}\}$ is a Glivenko–Cantelli class, where $\mathcal{L}$ is the set of K-dimensional nondecreasing functions $(\Lambda_1, \ldots, \Lambda_K)$ with $\Lambda_k(0) = 0$. Hence, $(\mathbb{P}_n - \mathbb{P})m(\theta_0, \mathcal{A}_0)$ converges to zero almost surely. With probability one,

$$\liminf_n \mathbb{P}_n l(\hat{\theta}, \hat{\mathcal{A}}) \geqslant \liminf_n \mathbb{P}_n m(\theta_0, \mathcal{A}_0) = \mathbb{P} m(\theta_0, \mathcal{A}_0) = O(1).$$

Let $\tilde{M} = \sup_{1\leqslant k\leqslant K} \sup_{t\in[0,\tau_k]}\{\sup_{X_{jk},\beta} |\beta^{\mathrm{T}} X_{jk}(t)| + \sup_{Z_{jk}} |Z_{jk}(t)|\}$, which is finite under Conditions 1–3. For any $\epsilon > 0$,

$$
\begin{aligned}
&\liminf_n \mathbb{P}_n l(\hat{\theta}, \hat{\mathcal{A}}) \\
&\leqslant \limsup_n \mathbb{P}_n \left\{ \log \left(\int \prod_{j=1}^{J} \prod_{k=1}^{K} \left\{ \exp\left(-G_k \left[\int_0^{U_{jkM_{jk}}} \exp\{\hat{\beta}^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\} \, \mathrm{d}\hat{\Lambda}_k(s) \right] \right) \right\}^{\Delta_{jkM_{jk}}} \right. \\
&\qquad\qquad \left. \times \phi\{b; \Sigma(\hat{\gamma})\} \, \mathrm{d}b \right) \Bigg\} \\
&\leqslant \limsup_n \mathbb{P}_n \left[\log \left\{ \int \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\exp\left[-G_k \left\{ \exp(-\tilde{M} - \tilde{M}\|b\|) \hat{\Lambda}_k(U_{jkM_{jk}}) \right\} \right] \right)^{\Delta_{jkM_{jk}}} \right. \right. \\
&\qquad\qquad \left. \left. \times \phi\{b; \Sigma(\hat{\gamma})\} \, \mathrm{d}b \right\} \right] \\
&\leqslant \limsup_n \mathbb{P}_n \left[\log \left\{ \int_{\|b\|\leqslant 1} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\exp\left[-G_k \left\{ \exp(-2\tilde{M}) \hat{\Lambda}_k(U_{jkM_{jk}}) \right\} \right] \right)^{\Delta_{jkM_{jk}}} \phi\{b; \Sigma(\hat{\gamma})\} \, \mathrm{d}b \right\} \right] \\
&\quad + \limsup_n \mathbb{P}_n \left(\log \left[\int_{\|b\|>1} \phi\{b; \Sigma(\hat{\gamma})\} \, \mathrm{d}b \right] \right) \\
&\leqslant -\limsup_n \mathbb{P}_n \left[\sum_{j=1}^{J} \sum_{k=1}^{K} \Delta_{jkM_{jk}} G_k \left\{ \exp(-2\tilde{M}) \hat{\Lambda}_k(U_{jkM_{jk}}) \right\} \right] \\
&\leqslant -\limsup_n \mathbb{P}_n \left[\sum_{j=1}^{J} \sum_{k=1}^{K} \Delta_{jkM_{jk}} I(U_{jkM_{jk}} \geqslant \tau_k - \epsilon) G_k \left\{ \exp(-2\tilde{M}) \hat{\Lambda}_k(\tau_k - \epsilon) \right\} \right].
\end{aligned}
$$

Thus,

$$
\limsup_n \mathbb{P}_n \left[\sum_{j=1}^{J} \sum_{k=1}^{K} \Delta_{jkM_{jk}} I(U_{jkM_{jk}} \geqslant \tau_k - \epsilon) G_k \left\{ \exp(-2\tilde{M}) \hat{\Lambda}_k(\tau_k - \epsilon) \right\} \right] = O(1)
$$

for any $\epsilon > 0$. Since

$$
\mathbb{P}_n \left\{ \sum_{j=1}^{J} \Delta_{jkM_{jk}} I(U_{jkM_{jk}} \geqslant \tau_k - \epsilon) \right\} \to E \left\{ \sum_{j=1}^{J} \Delta_{jkM_{jk}} I(U_{jkM_{jk}} \geqslant \tau_k - \epsilon) \right\},
$$

which is positive under Condition 5, $\limsup_n G_k\{\exp(-2\tilde{M}) \hat{\Lambda}_k(\tau_k - \epsilon)\} < \infty$. If $\limsup_n \hat{\Lambda}_k(\tau_k - \epsilon) = \infty$, then $G_k\{\exp(-2\tilde{M}) \hat{\Lambda}_k(\tau_k - \epsilon)\} = \infty$ under Condition 6. This is a contradiction. Therefore $\limsup_n \hat{\Lambda}_k(\tau_k - \epsilon) < \infty$ with probability 1 for any $\epsilon > 0$ and any $k \in \{1, \ldots, K\}$.

For any $k = 1, \ldots, K$, consider an increasing sequence $\{\tau_{ks}\}$ $(s = 1, 2, \ldots)$ such that $\lim_{s\to\infty} \tau_{ks} = \tau_k$. For any given subsequence of $\hat{\Lambda}_k$, Helly's selection theorem, together with the fact that $\hat{\Lambda}_k(\tau_{ks}) < \infty$, allows us at stage s to choose from the subsequence selected at stage $s - 1$ a further subsequence which converges weakly on $[0, \tau_{ks}]$. We form a final subsequence, still denoted by $\{\hat{\Lambda}_k\}$, whose sth element is the sth element of the sequence selected at stage s. It is clear that $\hat{\Lambda}_k$ converges weakly to some function, say Λ_k^*, in any compact subset of $[0, \tau_k)$. Since the Lebesgue measure for the point τ_k is zero, $\hat{\Lambda}_k \to \Lambda_k^*$ for $t \in [0, \tau_k]$ almost everywhere; that is, the Lebesgue measure of the set $\{t \in [0, \tau_k] : \hat{\Lambda}_k(t)$ does not

converge to $\Lambda_k^*(t)\}$ is zero. Since $\hat{\beta}$ and $\hat{\gamma}$ are bounded, by choosing a further subsequence, which we still denote by $(\hat{\Lambda}_1, \ldots, \hat{\Lambda}_K, \hat{\beta}, \hat{\gamma})$, we can assume that $\hat{\Lambda}_k$ converges to Λ_k^* almost everywhere and that $(\hat{\beta}, \hat{\gamma})$ converges to some constant (β^*, γ^*).

Write $\theta^* = (\beta^*, \gamma^*)$ and $\mathcal{A}^* = (\Lambda_1^*, \ldots, \Lambda_K^*)$. We wish to show that $(\theta^*, \mathcal{A}^*) = (\theta_0, \mathcal{A}_0)$. By the concavity of the log function,

$$\mathbb{P}_n m(\hat{\theta}, \hat{\mathcal{A}}) \geqslant \frac{1}{2}\big\{\mathbb{P}_n \log L(\hat{\theta}, \hat{\mathcal{A}}) + \mathbb{P}_n \log L(\theta_0, \mathcal{A}_0)\big\} \geqslant \mathbb{P}_n m(\theta_0, \mathcal{A}_0).$$

Thus $(\mathbb{P}_n - \mathbb{P})m(\hat{\theta}, \hat{\mathcal{A}}) + \mathbb{P}m(\hat{\theta}, \hat{\mathcal{A}}) \geqslant (\mathbb{P}_n - \mathbb{P})m(\theta_0, \mathcal{A}_0) + \mathbb{P}m(\theta_0, \mathcal{A}_0)$. Since $m(\hat{\theta}, \hat{\mathcal{A}}) \in \mathcal{M}$, $(\mathbb{P}_n - \mathbb{P})m(\hat{\theta}, \hat{\mathcal{A}}) \to 0$ almost surely. Also, since $|\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \beta, \Lambda_k)| < 1$ for any $\beta \in \mathcal{B}$ and $\mathcal{A} \in \mathcal{L}$ with probability 1, we see that with respect to the probability measure for U_{jk} $(j = 1, \ldots, J;\ k = 1, \ldots, K)$,

$$\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \hat{\beta}, \hat{\Lambda}_k) - \prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \beta^*, \Lambda_k^*) \to 0.$$

By the dominated convergence theorem,

$$\begin{aligned}
&\big|\mathbb{P}m(\hat{\theta}, \hat{\mathcal{A}}) - \mathbb{P}m(\theta^*, \mathcal{A}^*)\big| \\
&\quad \leqslant \big|\mathbb{P}m(\hat{\theta}, \hat{\mathcal{A}}) - \mathbb{P}m(\hat{\beta}, \gamma^*, \hat{\Lambda})\big| + \big|\mathbb{P}m(\hat{\beta}, \gamma^*, \hat{\Lambda}) - \mathbb{P}m(\theta^*, \mathcal{A}^*)\big| \\
&\quad = O(1)\|\hat{\gamma} - \gamma^*\| + \mathbb{P}\log \frac{\int \big\{\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \hat{\beta}, \hat{\Lambda}_k)\big\}\phi\{b; \Sigma(\gamma^*)\}\,\mathrm{d}b + L(\theta_0, \mathcal{A}_0)}{\int \big\{\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \beta^*, \Lambda_k^*)\big\}\phi\{b; \Sigma(\gamma^*)\}\,\mathrm{d}b + L(\theta_0, \mathcal{A}_0)}.
\end{aligned}$$

Hence $|\mathbb{P}m(\hat{\theta}, \hat{\mathcal{A}}) - \mathbb{P}m(\theta^*, \mathcal{A}^*)| \to 0$ almost surely, such that

$$\mathbb{P}\log \frac{L(\theta^*, \mathcal{A}^*) + L(\theta_0, \mathcal{A}_0)}{2} \geqslant \mathbb{P}l(\theta_0, \mathcal{A}_0).$$

By the properties of the Kullback–Leibler information, $L(\theta_0, \mathcal{A}_0) = L(\theta^*, \mathcal{A}^*)$ with probability 1. So

$$\int \left\{\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \beta^*, \Lambda_k^*)\right\} \phi\{b; \Sigma(\gamma^*)\}\,\mathrm{d}b = \int \left\{\prod_{j=1}^J \prod_{k=1}^K D_{jk}(U_{jk}, b; \beta_0, \Lambda_{0k})\right\} \phi\{b; \Sigma(\gamma_0)\}\,\mathrm{d}b$$

with probability 1. For any $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$ and $l_{jk} \in \{0, \ldots, M_{jk}\}$, we set $\Delta_{jkl} = 1$ in the above equation for $l = l_{jk}, \ldots, M_{jk}$ and take the sum of the resulting equations to obtain

$$\begin{aligned}
&\int Q_{jk}(U_{jkl_{jk}}, b; \beta^*, \Lambda_k^*) \left\{\prod_{j'=1, j' \neq j}^J \prod_{k'=1, k' \neq k}^K D_{j'k'}(U_{j'k'}, b; \beta^*, \Lambda_{k'}^*)\right\} \phi\{b; \Sigma(\gamma^*)\}\,\mathrm{d}b \\
&\quad = \int Q_{jk}(U_{jkl_{jk}}, b; \beta_0, \Lambda_{0k}) \left\{\prod_{j'=1, j' \neq j}^J \prod_{k'=1, k' \neq k}^K D_{j'k'}(U_{j'k'}, b; \beta_0, \Lambda_{0k'})\right\} \phi\{b; \Sigma(\gamma_0)\}\,\mathrm{d}b.
\end{aligned}$$

This equality holds for arbitrary l_{jk}. Therefore, for any $t_{jk} \in [0, \tau_k]$,

$$\int Q_{jk}(t_{jk}, b; \beta^*, \Lambda_k^*) \left\{ \prod_{j'=1, j' \neq j}^{J} \prod_{k'=1, k' \neq k}^{K} D_{j'k'}(U_{j'k'}, b; \beta^*, \Lambda_{k'}^*) \right\} \phi\{b; \Sigma(\gamma^*)\} \, \mathrm{d}b$$
$$= \int Q_{jk}(t_{jk}, b; \beta_0, \Lambda_{0k}) \left\{ \prod_{j'=1, j' \neq j}^{J} \prod_{k'=1, k' \neq k}^{K} D_{j'k'}(U_{j'k'}, b; \beta_0, \Lambda_{0k'}) \right\} \phi\{b; \Sigma(\gamma_0)\} \, \mathrm{d}b.$$

For some fixed $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$, we repeat this process for $(j', k') \in C_{jk} = \{1, \ldots, j\} \times \{1, \ldots, k\}$ to obtain

$$\int \left[\left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta^*, \Lambda_{k'}^*) \right\} \left\{ \prod_{(j',k') \notin C_{jk}} D_{j'k'}(U_{j'k'}, b; \beta^*, \Lambda_{k'}^*) \right\} \right] \phi\{b; \Sigma(\gamma^*)\} \, \mathrm{d}b$$
$$= \int \left[\left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta_0, \Lambda_{0k'}) \right\} \left\{ \prod_{(j',k') \notin C_{jk}} D_{j'k'}(U_{j'k'}, b; \beta_0, \Lambda_{0k'}) \right\} \right] \phi\{b; \Sigma(\gamma_0)\} \, \mathrm{d}b.$$

Setting $\Delta_{j'k'l} = 1$ in the above equation for $(j', k') \notin C_{jk}$ and $l = 0, \ldots, M_{j'k'}$ and then taking the sum of the resulting equations gives

$$\int \left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta^*, \Lambda_{k'}^*) \right\} \phi\{b; \Sigma(\gamma^*)\} \, \mathrm{d}b = \int \left\{ \prod_{j'=1}^{j} \prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta_0, \Lambda_{0k'}) \right\} \phi\{b; \Sigma(\gamma_0)\} \, \mathrm{d}b.$$

By Condition 7, $\beta^* = \beta_0$, $\gamma^* = \gamma_0$ and $\Lambda_k^*(t) = \Lambda_{0k}(t)$ for $k \in \{1, \ldots, K\}$ and $t \in [0, \tau_k]$. Since $\Lambda_{0k}(t)$ is continuous, $\|\hat{\theta} - \theta_0\| + \sum_{k=1}^{K} \sup_{t \in [0, \tau_k]} |\hat{\Lambda}_k(t) - \Lambda_{0k}(t)| \to 0$ almost surely. □

Proof of Theorem 2. Let $H_{jkl}(t; \theta, \mathcal{A})$ denote

$$\frac{\int B_{jk}(t, U_{jkl}, U_{jk,l+1}, b; \beta, \Lambda_k) \{ \prod_{j'=1, j' \neq j}^{J} \prod_{k'=1, k' \neq k}^{K} D_{j'k'}(U_{j'k'}, b; \beta, \Lambda_{k'}) \} \phi\{b; \Sigma(\gamma)\} \, \mathrm{d}b}{\int \{ \prod_{j'=1}^{J} \prod_{k'=1}^{K} D_{j'k'}(U_{j'k'}, b; \beta, \Lambda_{k'}) \} \phi\{b; \Sigma(\gamma)\} \, \mathrm{d}b},$$

where

$$B_{jk}(t, u, v, b; \beta, \Lambda_k) = \exp\{\beta^{\mathrm{T}} X_{jk}(t) + b^{\mathrm{T}} Z_{jk}(t)\}$$
$$\times \left(Q_{jk}(v, b; \beta, \Lambda_k) G_k' \left[\int_0^v \exp\{\beta^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\} \, \mathrm{d}\Lambda_k(s) \right] I(v \geqslant t) \right.$$
$$\left. - Q_{jk}(u, b; \beta, \Lambda_k) G_k' \left[\int_0^u \exp\{\beta^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\} \, \mathrm{d}\Lambda_k(s) \right] I(u \geqslant t) \right).$$

For a single cluster, the score function for θ is

$$l_\theta(\theta, \mathcal{A}) = \begin{bmatrix} l_\beta(\theta, \mathcal{A}) \\ l_\gamma(\theta, \mathcal{A}) \end{bmatrix},$$

where

$$l_\beta(\theta,\mathcal{A}) = \sum_{j=1}^{J}\sum_{k=0}^{K}\sum_{l=0}^{M_{jk}} \Delta_{jkl}\int_0^{\tau_k} H_{jkl}(t;\theta,\mathcal{A})X_{jk}(t)\,\mathrm{d}\Lambda_k(t),$$

$$l_\gamma(\theta,\mathcal{A}) = \frac{\int\{\prod_{j=1}^{J}\prod_{k=1}^{K} D_{jk}(U_{jk},b;\beta,\Lambda_k)\}\phi'_\gamma\{b;\Sigma(\gamma)\}\,\mathrm{d}b}{\int\{\prod_{j=1}^{J}\prod_{k=1}^{K} D_{jk}(U_{jk},b;\beta,\Lambda_k)\}\phi\{b;\Sigma(\gamma)\}\,\mathrm{d}b}.$$

To obtain the score operator for $\mathcal{A}$, we consider a one-dimensional submodel $\mathcal{A}_\epsilon(h)$ where $h=(h_1,\ldots,h_K)^{\mathrm{T}}$ is a vector of functions in $L_2[0,\tau_k]$. Specifically, the submodel specifies that $\mathrm{d}\Lambda_{k,\epsilon,h_k} = (1+\epsilon h_k)\mathrm{d}\Lambda_k$. The score function for $\mathcal{A}$ along this submodel is

$$l_{\mathcal{A}}(\theta,\mathcal{A})(h) = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=0}^{M_{jk}} \Delta_{jkl}\int_0^{\tau_k} H_{jkl}(t;\theta,\mathcal{A})h_k(t)\,\mathrm{d}\Lambda_k(t).$$

Clearly,

$$\mathbb{G}_n\{l_\theta(\hat\theta,\hat{\mathcal{A}})\} = -n^{1/2}\big[\mathbb{P}\{l_\theta(\hat\theta,\hat{\mathcal{A}})\} - \mathbb{P}\{l_\theta(\theta_0,\mathcal{A}_0)\}\big],$$
$$\mathbb{G}_n\{l_{\mathcal{A}}(\hat\theta,\hat{\mathcal{A}})\}(h) = -n^{1/2}\big[\mathbb{P}\{l_{\mathcal{A}}(\hat\theta,\hat{\mathcal{A}})(h)\} - \mathbb{P}\{l_{\mathcal{A}}(\theta_0,\mathcal{A}_0)(h)\}\big].$$

We apply Taylor series expansion at $(\theta_0,\mathcal{A}_0)$ to the right-hand sides of the above two equations. In light of Lemma 3, the second-order terms are bounded by

$$\begin{aligned} &n^{1/2}E\left(\left[O(1)\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=0}^{M_{jk}}\big\{\hat\Lambda_k(U_{jkl})-\Lambda_{0k}(U_{jkl})\big\}^2\right] + O(1)\|\hat\beta-\beta_0\|^2 + O(1)\|\hat\gamma-\gamma_0\|^2\right)\\ &\quad= n^{1/2}\big\{O_{\mathrm{p}}(n^{-2/3}) + O_{\mathrm{p}}\big(\|\hat\beta-\beta_0\|^2+\|\hat\gamma-\gamma_0\|^2\big)\big\}\\ &\quad= O_{\mathrm{p}}\big(n^{1/2}\|\hat\beta-\beta_0\|^2 + n^{1/2}\|\hat\gamma-\gamma_0\|^2 + n^{-1/6}\big).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{G}_n\{l_\theta(\hat\theta,\hat{\mathcal{A}})\} &= -n^{1/2}\big[\mathbb{P}\{l_{\theta\theta}(\hat\theta-\theta_0) + l_{\theta\mathcal{A}}(\hat{\mathcal{A}}-\mathcal{A}_0)\}\big]\\ &\quad+ O_{\mathrm{p}}\big(n^{1/2}\|\hat\beta-\beta_0\|^2 + n^{1/2}\|\hat\gamma-\gamma_0\|^2 + n^{-1/6}\big),\\ \mathbb{G}_n\{l_{\mathcal{A}}(\hat\theta,\hat{\mathcal{A}})(h)\} &= -n^{1/2}\big[\mathbb{P}\{l_{\mathcal{A}\theta}(h)(\hat\theta-\theta_0) + l_{\mathcal{A}\mathcal{A}}(h,\hat{\mathcal{A}}-\mathcal{A}_0)\}\big]\\ &\quad+ O_{\mathrm{p}}\big(n^{1/2}\|\hat\beta-\beta_0\|^2 + n^{1/2}\|\hat\gamma-\gamma_0\|^2 + n^{-1/6}\big),\end{aligned}$$

where $l_{\theta\theta}$ is the second derivative of $l(\theta,\mathcal{A})$ with respect to θ, $l_{\theta\mathcal{A}}(h)$ is the derivative of l_θ along the submodel $\mathrm{d}\mathcal{A}_{\epsilon,h}$, $l_{\mathcal{A}\theta}(h)$ is the derivative of $l_{\mathcal{A}}(h)$ with respect to θ, and $l_{\mathcal{A}\mathcal{A}}(h,\hat{\mathcal{A}}-\mathcal{A}_0)$ is the derivative of $l_{\mathcal{A}}(h)$ along the submodel $\mathrm{d}\mathcal{A}_0 + \epsilon\,\mathrm{d}(\hat{\mathcal{A}}-\mathcal{A}_0)$. All the derivatives are evaluated at $(\theta_0,\mathcal{A}_0)$.

Let h^* denote the least favourable direction such that

$$l^*_{\mathcal{A}}l_{\mathcal{A}}(h^*) = l^*_{\mathcal{A}}l_\theta, \tag{A1}$$

where $l^*_{\mathcal{A}}$ is the adjoint operator of $l_{\mathcal{A}}$. Note that h^* is a $\{p+d(d+1)/2\}$-dimensional vector of functions in $\mathcal{H} = L_2[0,\tau_1]\times\cdots\times L_2[0,\tau_K]$. We will show later that h^* exists and has bounded variation. It then

follows that

$$E\{l_{\mathcal{A}\mathcal{A}}(h^*, \hat{\mathcal{A}} - \mathcal{A}_0)\} = -E\{l_{\mathcal{A}}(h^*) l_{\mathcal{A}}(\hat{\mathcal{A}} - \mathcal{A}_0)\} = -\int l^*_{\mathcal{A}} l_{\mathcal{A}}(h^*)\, \mathrm{d}(\hat{\mathcal{A}} - \mathcal{A}_0)$$
$$= -\int l^*_{\mathcal{A}} l_{\theta}\, \mathrm{d}(\hat{\mathcal{A}} - \mathcal{A}_0) = E\{l_{\theta\mathcal{A}}(\hat{\mathcal{A}} - \mathcal{A}_0)\},$$

so that

$$\mathbb{G}_n\{l_\theta(\hat{\theta}, \hat{\mathcal{A}}) - l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(h^*)\} = n^{1/2} E\big[\{l_\theta - l_{\mathcal{A}}(h^*)\}^{\otimes 2}\big](\hat{\theta} - \theta_0)$$
$$+ O_{\mathrm{p}}\big(n^{1/2}\|\hat{\beta} - \beta_0\|^2 + n^{1/2}\|\hat{\gamma} - \gamma_0\|^2 + n^{-1/6}\big).$$

In addition, if we can show that $l_\theta(\hat{\theta}, \hat{\mathcal{A}}) - l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(h^*)$ belongs to a Donsker class and that the matrix $E[\{l_\theta - l_{\mathcal{A}}(h^*)\}^{\otimes 2}]$ is invertible, then $n^{1/2}(\hat{\theta} - \theta_0) = O_{\mathrm{p}}(1)$ and

$$n^{1/2}(\hat{\theta} - \theta_0) = (E[\{l_\theta - l_{\mathcal{A}}(h^*)\}^{\otimes 2}])^{-1} \mathbb{G}_n\{l_\theta - l_{\mathcal{A}}(h^*)\} + o_{\mathrm{p}}(1).$$

The influence function for $\hat{\theta}$ is the efficient influence function, such that $n^{1/2}(\hat{\theta} - \theta_0)$ converges weakly to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

It remains to show the existence of h^*, the Donsker property of $l_\theta(\hat{\theta}, \hat{\mathcal{A}}) - l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(h^*)$, and the non-singularity of the matrix $E[\{l_\theta - l_{\mathcal{A}}(h^*)\}^{\otimes 2}]$. To show that there exists a solution h^* to (A1), we equip $\mathcal{H}$ with an inner product defined by

$$\langle h^{(1)}, h^{(2)} \rangle = \sum_{k=1}^{K} \int_0^{\tau_k} h_k^{(1)}(t) h_k^{(2)}(t)\, \mathrm{d}\Lambda_{0k}(t).$$

For any $h^{(1)}, h^{(2)} \in \mathcal{H}$,

$$\mathbb{P}\big\{l_{\mathcal{A}}(h^{(1)}) l_{\mathcal{A}}(h^{(2)})\big\} = \sum_{k=1}^{K} \int_0^{\tau_k} \Gamma_k(h^{(1)})(t) h_k^{(2)}(t)\, \mathrm{d}\Lambda_{0k}(t),$$

where

$$\Gamma_k(h)(t) = \sum_{k'=1}^{K} \int_0^{\tau_{k'}} \sum_{j=1}^{J} E\left[\left\{ \sum_{l=0}^{M_{jk}} \Delta_{jkl} H_{jkl}(s; \theta_0, \mathcal{A}_0) \right\} \right.$$
$$\left. \times \left\{ \sum_{j'=1}^{J} \sum_{l=0}^{M_{j'k'}} \Delta_{j'k'l} H_{j'k'l}(t; \theta_0, \mathcal{A}_0) \right\} \right] h_{k'}(s)\, \mathrm{d}\Lambda_{0k'}(s).$$

We define a seminorm $\|h\|_\Gamma = \langle \Gamma(h), h \rangle^{1/2}$ on the space $\mathcal{H}$. If $\|h\|_\Gamma = 0$ for some $h \in \mathcal{H}$, then $0 = \langle \Gamma(h), h \rangle = \mathbb{P}\{l_{\mathcal{A}}(h)^2\}$. Therefore, with probability 1, $l_{\mathcal{A}}(h) = 0$, i.e.,

$$\int \sum_{j=1}^{J} \sum_{k=1}^{K} \left[\left\{ \prod_{j'=1, j' \neq j}^{J} \prod_{k'=1, k' \neq k}^{K} D_{j'k'}(U_{j'k'}, b; \beta_0, \Lambda_{0k'}) \right\} \right.$$
$$\left. \times \sum_{l=0}^{M_{jk}} \Delta_{jkl} \int_0^{\tau_k} B_{jk}(t, U_{jkl}, U_{jk,l+1}, b; \beta_0, \Lambda_{0k}) h_k(t)\, \mathrm{d}\Lambda_{0k}(t) \right] \phi(b; \Sigma_0)\, \mathrm{d}b = 0.$$

By the arguments in the proof of Lemma 3, $h_k(t) = 0$ for any $k = 1, \ldots, K$ and $t \in [0, \tau_k]$. So $\|\cdot\|_\Gamma$ is a norm in $\mathcal{H}$. Clearly, $\|h\|_\Gamma \leqslant c\|h\|$ for some constant c. By the bounded inverse theorem in Banach space, $\|h\|_\Gamma \geqslant c'\|h\|$ for some constant c'. By the Lax–Milgram theorem (Zeidler, 1995), there exists a solution to (A1). Differentiation of this integral equation with respect to t yields

$$g_{1k}(t)h_k^*(t) + \sum_{k'=1}^{K}\left\{\int_t^{\tau_{k'}} g_{2k'}(s,t)h_{k'}^*(s)\,\mathrm{d}s + \int_0^t g_{3k'}(s,t)h_{k'}^*(s)\,\mathrm{d}s\right\} = g_{4k}(t),$$

where $g_{1k}(t) > 0$ and the g_{jk} $(j = 1, 2, 3, 4)$ are continuously differentiable functions. Thus, $h_k^*(\cdot)$ is continuously differentiable in $[0, \tau_k]$ for $k \in \{1, \ldots, K\}$. By the arguments in the proof of Lemma 2, $l_\theta(\hat\theta, \hat\Lambda) - l_\mathcal{A}(\hat\theta, \hat\Lambda)(h^*)$ belongs to a Donsker class and converges in $L_2(\mathbb{P})$-norm to $l_\theta - l_\mathcal{A}(h^*)$.

Finally, we verify that $E[\{l_\theta - l_\mathcal{A}(h^*)\}^{\otimes 2}]$ is invertible. If the matrix is singular, then there exist vectors $v = (v_1, v_2)$ with $v_1 \in \mathbb{R}^p$ and $v_2 \in \mathbb{R}^{d(d+1)/2}$ such that $v^{\mathrm{T}}E[\{l_\theta - l_\mathcal{A}(h^*)\}^{\otimes 2}]v = 0$. It follows that, with probability 1, the score function along the submodel $\{\theta_0 + \epsilon v, \mathcal{A}_\epsilon(v^{\mathrm{T}}h^*)\}$ is zero. That is,

$$\int \sum_{j=1}^{J}\sum_{k=1}^{K}\left\{\prod_{j'=1,j'\neq j}^{J}\prod_{k'=1,k'\neq k}^{K} D_{j'k'}(U_{j'k'}, b; \beta_0, \Lambda_{0k'})\right\}\sum_{l=0}^{M_{jk}}\Delta_{jkl}\int_0^{\tau_k} B_{jk}(t, U_{jkl}, U_{jk,l+1}, b; \beta_0, \Lambda_{0k})$$
$$\times\{v_1^{\mathrm{T}}X_{jk}(t) - v^{\mathrm{T}}h^*(t)\}\,\mathrm{d}\Lambda_{0k}(t)\phi(b; \Sigma_0)\,\mathrm{d}b - \int\left\{\prod_{j=1}^{J}\prod_{k=1}^{K} D_{jk}(U_{jk}, b; \beta_0, \Lambda_{0k})\right\}v_2^{\mathrm{T}}\phi_\gamma'(b; \Sigma_0)\,\mathrm{d}b = 0$$

with probability 1. For any $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$ and $l_{j'k'} \in \{0, \ldots, M_{j'k'}\}$ $(j' = 1, \ldots, j;\ k' = 1, \ldots, k)$, we evaluate the above equation at all possible values of $\Delta_{j'k'l}$ with $(j', k') \in C_{jk} = \{1, \ldots, j\} \times \{1, \ldots, k\}$ and $l = l_{j'k'}, \ldots, M_{j'k'}$ and take the sum of the resulting equations. We then consider all possible values of $\Delta_{j'k'l}$ with $(j', k') \notin C_{jk}$ and $l = 0, \ldots, M_{j'k'}$ and take the sum of the resulting equations. This yields

$$\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k} Q_{j'k'}(U_{j'k'l_{j'k'}}, b; \beta_0, \Lambda_{0k'})\right\}$$
$$\times\sum_{j'=1}^{j}\sum_{k'=1}^{k} G_k'\left[\int_0^{U_{j'k'l_{j'k'}}}\exp\{\beta_0^{\mathrm{T}}X_{j'k'}(t) + b^{\mathrm{T}}Z_{j'k'}(t)\}\,\mathrm{d}\Lambda_{0k'}(t)\right]$$
$$\times\int_0^{U_{j'k'l_{j'k'}}}\exp\{\beta_0^{\mathrm{T}}X_{j'k'}(t) + b^{\mathrm{T}}Z_{j'k'}(t)\}\{v_1^{\mathrm{T}}X_{j'k'}(t) - v^{\mathrm{T}}h_{k'}^*(t)\}\,\mathrm{d}\Lambda_{0k'}(t)\phi(b; \Sigma_0)\,\mathrm{d}b$$
$$-\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k} Q_{j'k'}(U_{j'k'l_{j'k'}}, b; \beta_0, \Lambda_{0k'})\right\}v_2^{\mathrm{T}}\phi_\gamma'(b; \Sigma_0)\,\mathrm{d}b = 0.$$

This equality holds for any $U_{j'k'l_{j'k'}}$. Hence, for any $t_{j'k'} \in [0, \tau_{k'}]$ $(j' = 1, \ldots, j;\ k' = 1, \ldots, k)$,

$$\int\left\{\prod_{j'=1}^{j}\prod_{k'=1}^{k} Q_{j'k'}(t_{j'k'}, b; \beta_0, \Lambda_{0k'})\right\}\left(\sum_{j'=1}^{j}\sum_{k'=1}^{k} G_k'\left[\int_0^{t_{j'k'}}\exp\{\beta_0^{\mathrm{T}}X_{j'k'}(t) + b^{\mathrm{T}}Z_{j'k'}(t)\}\,\mathrm{d}\Lambda_{0k'}(t)\right]\right.$$
$$\left.\times\int_0^{t_{j'k'}}\exp\{\beta_0^{\mathrm{T}}X_{j'k'}(t) + b^{\mathrm{T}}Z_{j'k'}(t)\}\{v_1^{\mathrm{T}}X_{j'k'}(t) - v^{\mathrm{T}}h_{k'}^*(t)\}\,\mathrm{d}\Lambda_{0k'}(t) - \frac{v_2^{\mathrm{T}}\phi_\gamma'(b; \Sigma_0)}{\phi(b; \Sigma_0)}\right)$$
$$\times\ \phi(b; \Sigma_0)\,\mathrm{d}b = 0.$$

By Condition 8, $v_2 = 0$ and

$$G_k'\left[\int_0^t \exp\{\beta_0^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\}\,\mathrm{d}\Lambda_{0k}(s)\right] \times \left[\int_0^t \exp\{\beta_0^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\}\{v_1^{\mathrm{T}} X_{jk}(s) - v^{\mathrm{T}} h_k^*(s)\}\,\mathrm{d}\Lambda_{0k}(s)\right] = 0$$

for any $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$ and $t \in [0, \tau_k]$. The term $G_k'[\int_0^t \exp\{\beta_0^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\}\,\mathrm{d}\Lambda_{0k}(s)]$ is bounded away from zero. Therefore

$$\int_0^t \exp\{\beta_0^{\mathrm{T}} X_{jk}(s) + b^{\mathrm{T}} Z_{jk}(s)\}\{v_1^{\mathrm{T}} X_{jk}(s) - v^{\mathrm{T}} h_k^*(s)\}\,\mathrm{d}\Lambda_{0k}(s) = 0.$$

Differentiating both sides with respect to t gives $v_1^{\mathrm{T}} X_{jk}(s) - v^{\mathrm{T}} h_k^*(s) = 0$. By Condition 2, $v_1 = 0$. Hence, the matrix $E[\{l_\theta - l_{\mathcal{A}}(h^*)\}^{\otimes 2}]$ is invertible. □

Proof of Theorem 3. By the chain rule,

$$\hat{V}_n = \mathbb{P}_n\left[\left\{\frac{\partial}{\partial\theta} l(\theta, \hat{\mathcal{A}}_\theta)\Big|_{\theta=\hat{\theta}}\right\}^{\otimes 2}\right] = \mathbb{P}_n\left[\left\{l_\theta(\theta, \hat{\mathcal{A}}_\theta)\Big|_{\theta=\hat{\theta}} + l_{\mathcal{A}}(\theta, \hat{\mathcal{A}}_\theta)\left(\frac{\partial \hat{\mathcal{A}}_\theta}{\partial\theta}\right)\Big|_{\theta=\hat{\theta}}\right\}^{\otimes 2}\right]$$
$$= \mathbb{P}_n\left[\left\{l_\theta(\hat{\theta}, \hat{\mathcal{A}}) + l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}})\right\}^{\otimes 2}\right],$$

where $\dot{\mathcal{A}}_{\hat{\theta}} = \partial \hat{\mathcal{A}}_\theta / \partial\theta|_{\theta=\hat{\theta}}$. We first prove that the function $\dot{\mathcal{A}}_{\hat{\theta}}$ has bounded total variation, such that $l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}})$ belongs to a Donsker class by Lemma 2. By the definition of $\hat{\mathcal{A}}_\theta$, $\mathbb{P}_n l_{\mathcal{A}}(\theta, \hat{\mathcal{A}}_\theta)(h) = 0$ for any $\theta \in \Theta$ and $h \in \mathcal{H}$. Differentiation with respect to θ at $\theta = \hat{\theta}$ yields

$$\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}}, h) = -\mathbb{P}_n l_{\mathcal{A}\theta}(\hat{\theta}, \hat{\mathcal{A}})(h)$$

for any $h \in \mathcal{H}$. We consider the linear operator $\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})$ which maps $u \in \mathcal{H}$ to $l^\infty(\mathcal{H})$ by the definition of $\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(u, h)$. According to Theorem 1, $\|\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}}) - \mathbb{P} l_{\mathcal{A}\mathcal{A}}(\theta_0, \mathcal{A}_0)\| \to 0$ in probability. Since $\mathbb{P} l_{\mathcal{A}\mathcal{A}}(\theta_0, \mathcal{A}_0)$ has been shown to be invertible in the proof of Theorem 2, we conclude that $\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})$ is invertible when n is large enough. Hence, there exists a unique solution in $\mathcal{H}$ which solves $\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(u, h) = -\mathbb{P}_n l_{\mathcal{A}\theta}(\hat{\theta}, \hat{\mathcal{A}})(h)$. Therefore $\dot{\mathcal{A}}_{\hat{\theta}}$ is the solution and has bounded total variation.

By the arguments for showing the existence of the least favourable direction h^* in the proof of Theorem 2, $-\mathbb{P} l_{\mathcal{A}\mathcal{A}}(\theta_0, \mathcal{A}_0)(\dot{\mathcal{A}}_{\hat{\theta}} - h^*, \dot{\mathcal{A}}_{\hat{\theta}} - h^*) \geqslant c\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})}^2$ for some positive constant c, so $\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}} - h^*, \dot{\mathcal{A}}_{\hat{\theta}} - h^*) \geqslant c\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})}^2$. Since

$$\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}}, \dot{\mathcal{A}}_{\hat{\theta}} - h^*) = -\mathbb{P}_n l_{\theta\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}} - h^*)$$
$$= o_{\mathrm{p}}(1)\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})} + \mathbb{P} l_{\theta\mathcal{A}}(\theta_0, \mathcal{A}_0)(\dot{\mathcal{A}}_{\hat{\theta}} - h^*)$$

and

$$\mathbb{P}_n l_{\mathcal{A}\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(h^*, \dot{\mathcal{A}}_{\hat{\theta}} - h^*) = o_{\mathrm{p}}(1)\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})} + \mathbb{P} l_{\mathcal{A}\mathcal{A}}(\theta_0, \mathcal{A}_0)(h^*, \dot{\mathcal{A}}_{\hat{\theta}} - h^*)$$
$$= o_{\mathrm{p}}(1)\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})} + \mathbb{P} l_{\theta\mathcal{A}}(\theta_0, \mathcal{A}_0)(\dot{\mathcal{A}}_{\hat{\theta}} - h^*),$$

we obtain $o_{\mathrm{p}}(1)\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})} \geqslant c\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})}^2$. Consequently, $\|\dot{\mathcal{A}}_{\hat{\theta}} - h^*\|_{L_2(\mathbb{P})} = o_{\mathrm{p}}(1)$.

By the consistency of $(\hat{\theta}, \hat{\mathcal{A}})$ and the Donsker property of $l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}})$,

$$\begin{aligned}\hat{V}_n &= \mathbb{P}\Big[\big\{l_\theta(\hat{\theta}, \hat{\mathcal{A}}) + l_{\mathcal{A}}(\hat{\theta}, \hat{\mathcal{A}})(\dot{\mathcal{A}}_{\hat{\theta}})\big\}^{\otimes 2}\Big] + o_{\mathrm{p}}(1)\\ &= \mathbb{P}\Big[\big\{l_\theta(\theta_0, \mathcal{A}_0) + l_{\mathcal{A}}(\theta_0, \mathcal{A}_0)(\dot{\mathcal{A}}_{\hat{\theta}})\big\}^{\otimes 2}\Big] + o_{\mathrm{p}}(1)\\ &= \mathbb{P}\Big[\big\{l_\theta(\theta_0, \mathcal{A}_0) + l_{\mathcal{A}}(\theta_0, \mathcal{A}_0)(h^*)\big\}^{\otimes 2}\Big] + o_{\mathrm{p}}(1),\end{aligned}$$

where the last equality follows from the convergence of $\dot{\mathcal{A}}_{\hat{\theta}}$ to h^*. Hence, the theorem follows from the fact that $\mathbb{P}[\{l_\theta(\theta_0, \mathcal{A}_0) + l_{\mathcal{A}}(\theta_0, \mathcal{A}_0)(h^*)\}^{\otimes 2}]$ is the efficient information for θ_0, which is the inverse covariance matrix of $n^{1/2}(\hat{\theta} - \theta_0)$ by Theorem 2. □

References

Chang, I. S., Wen, C. C. & Wu, Y. J. (2007). A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statist. Sinica* **17**, 1023–46.

Chen, K., Jin, Z. & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659–68.

Chen, M. H., Chen, L. C., Lin, K. H. & Tong, X. (2014). Analysis of multivariate interval censoring by Diabetic Retinopathy Study. *Commun. Statist.* B **43**, 1825–35.

Chen, M. H., Tong, X. & Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statist. Med.* **26**, 5147–61.

Chen, M. H., Tong, X. & Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statist. Med.* **28**, 3424–36.

Chen, M. H., Tong, X. & Zhu, L. (2013). A linear transformation model for multivariate interval-censored failure time data. *Can. J. Statist.* **41**, 275–90.

Cook, R. J. & Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics* **10**, 756–72.

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

Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–80.

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

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, New Jersey: Wiley, 2nd ed.

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

Kor, C. T., Cheng, K. F. & Chen, Y. H. (2013). A method for analyzing clustered interval-censored data based on Cox's model. *Statist. Med.* **32**, 822–32.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statist. Med.* **13**, 2233–47.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

The ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. *Am. J. Epidemiol.* **129**, 687–702.

Tong, X., Chen, M. H. & Sun, J. (2008). Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biomet. J.* **50**, 364–74.

Wang, L., Sun, J. & Tong, X. (2008). Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Anal.* **14**, 134–53.

Wang, N., Wang, L. & Mcmahan, C. S. (2015). Regression analysis of bivariate current status data under the gamma-frailty proportional hazards model using the EM algorithm. *Comp. Statist. Data Anal.* **83**, 140–50.

Wen, C. C. & Chen, Y. H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty Cox model. *Comp. Statist. Data Anal.* **55**, 1053–60.

Wen, C. C. & Chen, Y. H. (2013). A frailty model approach for regression analysis of bivariate interval-censored survival data. *Statist. Sinica* **23**, 383–408.

Zeidler, E. (1995). *Applied Functional Analysis: Applications to Mathematical Physics*. New York: Springer.

Zeng, D. & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with Discussion). *J. R. Statist. Soc.* B **69**, 507–64.

Zeng, D., Mao, L. & Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103**, 253–71.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89**, 39–48.

[*Received on* 6 *June* 2016. *Editorial decision on* 27 *January* 2017]