Maximum likelihood estimation for semiparametric transformation models with interval-censored data

By DONGLIN ZENG, LU MAO AND D. Y. LIN

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

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

SUMMARY

Interval censoring arises frequently in clinical, epidemiological, financial and sociological studies, where the event or failure of interest is known only to occur within an interval induced by periodic monitoring. We formulate the effects of potentially time-dependent covariates on the interval-censored failure time through a broad class of semiparametric transformation models that encompasses proportional hazards and proportional odds models. We consider nonparametric maximum likelihood estimation for this class of models with an arbitrary number of monitoring times for each subject. We devise an EM-type algorithm that converges stably, even in the presence of time-dependent covariates, and show that the estimators for the regression parameters are consistent, asymptotically normal, and asymptotically efficient with an easily estimated covariance matrix. Finally, we demonstrate the performance of our procedures through simulation studies and application to an HIV/AIDS study conducted in Thailand.

Some key words: Current-status data; EM algorithm; Interval censoring; Linear transformation model; Nonparametric likelihood; Proportional hazards; Proportional odds; Semiparametric efficiency; Time-dependent covariate.

1. INTRODUCTION

Interval-censored data arise when the event or failure of interest is known only to occur within a time interval. Such data are commonly encountered in disease research, where the ascertainment of an asymptomatic event is costly or invasive and so can take place only at a small number of monitoring times. For example, in HIV/AIDS studies, blood samples are periodically drawn from at-risk subjects to look for evidence of HIV sero-conversion. Likewise, biopsies are performed on patients at clinic visits to determine the occurrence or recurrence of cancer.

There are several types of interval-censored data. The simplest and most studied type is called case-1 or current-status data, which involves only one monitoring time per subject and is routinely found in cross-sectional studies. When there are two or k monitoring times per subject, the resulting data are referred to as case-2 or case-k interval censoring (Huang & Wellner, 1997). The most general and common type allows for varying numbers of monitoring times among subjects and is termed mixed-case interval censoring (Schick & Yu, 2000).

The fact that the failure time is never observed exactly poses theoretical and computational challenges in semiparametric regression analysis of such data. Huang (1995, 1996) and Huang & Wellner (1997) studied nonparametric maximum likelihood estimation for the proportional hazards and proportional odds models with case-1 and case-2 data. The estimators are obtained by the iterative convex minorant algorithm, which becomes unstable for large datasets. Sieve maximum likelihood estimation for the proportional odds model was considered by

Rossini & Tsiatis (1996) with case-1 data and by Huang & Rossini (1997) and Shen (1998) with case-2 data; however, it is difficult to choose an appropriate sieve parameter space and, especially, to choose the number of knots. For the proportional odds model with case-1 and case-2 data, Rabinowitz et al. (2000) derived an approximate conditional likelihood, which does not perform well in small samples. Gu et al. (2005), Sun & Sun (2005), Zhang et al. (2005) and Zhang & Zhao (2013) constructed rank-based estimators for linear transformation models, but such estimators are computationally demanding and statistically inefficient. None of the existing work accommodates time-dependent covariates or can handle case-k or mixed-case interval censoring.

In this paper we consider interval censoring in the most general form, that is, mixed-case data. We study nonparametric maximum likelihood estimation for a broad class of transformation models that allows time-dependent covariates and includes the proportional hazards and proportional odds models as special cases. We develop an EM-type algorithm, which is demonstrated to perform satisfactorily in a wide variety of settings, even with time-dependent covariates. Using empirical process theory (van der Vaart & Wellner, 1996; van de Geer, 2000) and semiparametric efficiency theory (Bickel et al., 1993), we establish that, under mild conditions, the proposed estimators for the regression parameters are consistent and asymptotically normal and the limiting covariance matrix attains the semiparametric efficiency bound and can be estimated analytically by the profile likelihood method (Murphy & van der Vaart, 2000). The theoretical development requires careful treatment of the time trajectories of covariate processes and the joint distribution for an arbitrary sequence of monitoring times.

2. METHODS

2·1. *Transformation models and likelihood construction*

Let T denote the failure time, and let $Z(\cdot)$ denote a d-vector of potentially time-dependent covariates. Under the semiparametric transformation model, the cumulative hazard function for T conditional on $Z(\cdot)$ takes the form

$$\Lambda(t; Z) = G\left[\int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right], \tag{1}$$

where $G(\cdot)$ is a specific transformation function that is strictly increasing and $\Lambda(\cdot)$ is an unknown increasing function (Zeng & Lin, 2006). The choices of $G(x) = x$ and $G(x) = \log(1 + x)$ yield the proportional hazards and proportional odds models, respectively. It is useful to consider the class of frailty-induced transformations

$$G(x) = -\log \int_0^\infty \exp(-xt) f(t) \, \mathrm{d}t,$$

where $f(t)$ is the density function of a frailty variable with support $[0, \infty)$. The choice of the gamma density with unit mean and variance r for $f(t)$ yields the class of logarithmic transformations, $G(x) = r^{-1} \log(1 + rx)$ $(r \geqslant 0)$, and the choice of the positive stable distribution with parameter $\rho < 1$ yields the class of Box–Cox transformations, $G(x) = \{(1 + x)^\rho - 1\}/\rho$. When all the covariates are time-independent, model (1) can be rewritten as a linear transformation model

$$\log \Lambda(T) = -\beta^{\mathrm{T}} Z + \epsilon,$$

where ϵ is an error term with distribution function $1 - \exp[-G\{\exp(x)\}]$ (Chen et al., 2002). Thus, β can be interpreted as the effects of covariates on a transformation of T.

We formulate the mixed-case interval censoring by assuming that the number of monitoring times, denoted by K, is random and that there exists a random sequence of monitoring times, denoted by $U_1 < \cdots < U_K$. We do not model $(U_1, \ldots, U_K)$. Write $\tilde{U} = (U_0, U_1, \ldots, U_{K+1})$, where $U_0 = 0$ and $U_{K+1} = \infty$. Also, define $\tilde{\Delta} = (\Delta_0, \Delta_1, \ldots, \Delta_K)$, where $\Delta_k = I(U_k < T \leqslant U_{k+1})$ $(k = 0, \ldots, K)$ with $I(\cdot)$ denoting the indicator function. Then the observed data from a random sample of n subjects consist of $(\tilde{U}_i, \tilde{\Delta}_i, Z_i)$ $(i = 1, \ldots, n)$, where $\tilde{U}_i = (U_{i0}, U_{i1}, \ldots, U_{i,K_i+1})$ and $\tilde{\Delta}_i = (\Delta_{i0}, \Delta_{i1}, \ldots, \Delta_{iK_i})$. If $K_i = 1$ or 2 $(i = 1, \ldots, n)$, then the observation scheme becomes case-1 or case-2, respectively.

Suppose that $(\tilde{U}, K)$ is independent of T conditional on $Z(\cdot)$. Then the observed-data likelihood function concerning parameters (β, Λ) takes the form

$$L_n(\beta, \Lambda) = \prod_{i=1}^{n} \prod_{k=0}^{K_i} \left\{ \exp\left(-G\left[\int_0^{U_{ik}} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, d\Lambda(s)\right]\right) \right.$$
$$\left. - \exp\left(-G\left[\int_0^{U_{i,k+1}} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, d\Lambda(s)\right]\right) \right\}^{\Delta_{ik}}.$$

Since only one Δ_{ik} is unity for each subject and the others equal zero,

$$L_n(\beta, \Lambda) = \prod_{i=1}^{n} \left\{ \exp\left(-G\left[\int_0^{L_i} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, d\Lambda(s)\right]\right) \right.$$
$$\left. - \exp\left(-G\left[\int_0^{R_i} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, d\Lambda(s)\right]\right) \right\},$$

where $(L_i, R_i]$ is the smallest interval that brackets T_i, i.e., $L_i = \max\{U_{ik} : U_{ik} < T_i\}$ and $R_i = \min\{U_{ik} : U_{ik} \geqslant T_i\}$. Clearly, $L_i = 0$ indicates that the ith subject is left censored, while $R_i = \infty$ indicates that the subject is right censored.

Remark 1. The sequence of monitoring times may not be completely observed and, in fact, need not be for the purpose of inference. We only need to know the values of L_i and R_i, since the other monitoring times do not contribute to the likelihood. The theoretical development, however, requires consideration of the joint distribution for the entire sequence of monitoring times.

2·2. *Nonparametric maximum likelihood estimation*

To estimate β and Λ, we adopt the nonparametric maximum likelihood approach, under which Λ is regarded as a step function with nonnegative jumps at the endpoints of the smallest intervals that bracket the failure times. Specifically, if $0 = t_0 < t_1 < \cdots < t_m$ denotes the set consisting of 0 and the unique values of $L_i > 0$ and $R_i < \infty$ $(i = 1, \ldots, n)$, then the estimator for Λ is a step function with jump size λ_k at t_k and with $\lambda_0 = 0$. Hence, we maximize the function

$$\prod_{i=1}^{n} \left\{ \exp\left(-G\left[\sum_{t_k \leqslant L_i} \exp\{\beta^{\mathrm{T}} Z_i(t_k)\}\lambda_k\right]\right) - \exp\left(-G\left[\sum_{t_k \leqslant R_i} \exp\{\beta^{\mathrm{T}} Z_i(t_k)\}\lambda_k\right]\right) \right\}. \quad (2)$$

Direct maximization of (2) is difficult due to the lack of an analytical expression for the parameters λ_k $(k = 1, \ldots, m)$. An even more severe challenge is that not all the L_i and R_i are informative about the failure times, so many of the λ_k are zero and hence lie on the boundary of the parameter space. For example, if there are no interval endpoints between some R_i and R_j with $R_i < R_j$,

then the jump size at R_i must be zero in order to maximize (2). The existing iterative convex minorant algorithm works only for the proportional hazards and proportional odds models with time-independent covariates (Huang & Wellner, 1997). In the following, we construct an EM algorithm to maximize (2).

For the class of frailty-induced transformations described in § 2·1, the observed-data likelihood can be written as

$$\prod_{i=1}^{n} \int_{\xi_i} \left(\exp\left[-\xi_i \int_0^{L_i} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, \mathrm{d}\Lambda(s) \right] \right.$$
$$\left. - \exp\left[-\xi_i \int_0^{R_i} \exp\{\beta^{\mathrm{T}} Z_i(s)\}\, \mathrm{d}\Lambda(s) \right] \right) f(\xi_i)\, \mathrm{d}\xi_i,$$

so that the estimation of the transformation model becomes that of the proportional hazards frailty model. With $\Lambda(t)$ as a step function with jumps λ_k at t_k ($k = 1, \ldots, m$), this likelihood becomes

$$\prod_{i=1}^{n} \int_{\xi_i} \exp\left\{ -\xi_i \sum_{t_k \leqslant L_i} \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik}) \right\} \left[1 - \exp\left\{ -\xi_i \sum_{L_i < t_k \leqslant R_i} \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik}) \right\} \right]^{I(R_i < \infty)} f(\xi_i)\, \mathrm{d}\xi_i, \tag{3}$$

where $Z_{ik} = Z_i(t_k)$. We introduce latent variables W_{ik} ($i = 1, \ldots, n;\ k = 1, \ldots, m$) which, conditional on ξ_i, are independent Poisson random variables with means $\xi_i \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik})$. We show below that the nonconcave likelihood function given in (3) is equivalent to a likelihood function for these Poisson variables, so the M-step becomes maximization of a weighted sum of Poisson loglikelihood functions which is strictly concave and has closed-form solutions for λ_k ($k = 1, \ldots, m$). Similar Poisson variables were recently used by Wang et al. (2015) in spline-based estimation of the proportional hazards model with time-independent covariates.

Define $A_i = \sum_{t_k \leqslant L_i} W_{ik}$ and $B_i = I(R_i < \infty) \sum_{L_i < t_k \leqslant R_i} W_{ik}$. Suppose that the observed data consist of $(L_i, R_i, Z_i, A_i = 0, B_i > 0)$ ($i = 1, \ldots, n$), where $A_i = 0$ means that A_i is known to be zero and that $B_i > 0$ means that B_i is known to be positive, such that $W_{ik} = 0$ for $t_k \leqslant L_i$ and at least one $W_{ik} \geqslant 1$ for $L_i < t_k \leqslant R_i$ with $R_i < \infty$. Then the likelihood takes the form

$$\prod_{i=1}^{n} \int_{\xi_i} \left\{ \prod_{t_k \leqslant L_i} \mathrm{pr}(W_{ik} = 0 \mid \xi_i) \right\} \left\{ 1 - \mathrm{pr}\left(\sum_{L_i < t_k \leqslant R_i} W_{ik} = 0 \,\middle|\, \xi_i \right) \right\}^{I(R_i < \infty)} f(\xi_i)\, \mathrm{d}\xi_i, \tag{4}$$

which is the same as (3). Thus, maximization of (3) is equivalent to maximum likelihood estimation based on the data $(L_i, R_i, Z_i, A_i = 0, B_i > 0)$ ($i = 1, \ldots, n$).

We maximize (4) through an EM algorithm by treating ξ_i and W_{ik} as missing data. The complete-data loglikelihood is

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{m} I(t_k \leqslant R_i^*) \left[W_{ik} \log\{\xi_i \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik})\} \right.\right.$$
$$\left.\left. - \xi_i \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik}) - \log W_{ik}! \right] + \log f(\xi_i) \right), \tag{5}$$

where $R_i^* = L_i I(R_i = \infty) + R_i I(R_i < \infty)$. In the M-step, we calculate

$$\lambda_k = \frac{\sum_{i=1}^n I(t_k \leqslant R_i^*)\hat{E}(W_{ik})}{\sum_{i=1}^n I(t_k \leqslant R_i^*)\hat{E}(\xi_i)\exp(\beta^{\mathrm{T}} Z_{ik})} \quad (k = 1, \ldots, m), \tag{6}$$

where $\hat{E}(\cdot)$ denotes the posterior mean given the observed data. After incorporating (6) into the conditional expectation of (5), we update β by solving the following equation using the one-step Newton–Raphson method:

$$\sum_{i=1}^n \sum_{k=1}^m I(t_k \leqslant R_i^*)\hat{E}(W_{ik}) \left\{ Z_{ik} - \frac{\sum_{j=1}^n I(t_k \leqslant R_j^*)\hat{E}(\xi_j)\exp(\beta^{\mathrm{T}} Z_{jk}) Z_{jk}}{\sum_{j=1}^n I(t_k \leqslant R_j^*)\hat{E}(\xi_j)\exp(\beta^{\mathrm{T}} Z_{jk})} \right\} = 0.$$

In the E-step, we evaluate the posterior means $\hat{E}(W_{ik})$ and $\hat{E}(\xi_i)$. The posterior density function of ξ_i given the observed data is proportional to $\{\exp(-\xi_i S_{i1}) - \exp(-\xi_i S_{i2})\}f(\xi_i)$, where $S_{i1} = \sum_{t_k \leqslant L_i} \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik})$ and $S_{i2} = \sum_{t_k \leqslant R_i} \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik})$. Hence, we evaluate the posterior means by noting that for $t_k \leqslant L_i$,

$$\hat{E}(W_{ik}) = E(W_{ik} \mid A_i = 0, B_i > 0, L_i, R_i, Z_i) = 0,$$

and for $L_i < t_k \leqslant R_i$ with $R_i < \infty$,

$$
\begin{aligned}
\hat{E}(W_{ik}) &= E_{\xi_i}\left\{ E\left(W_{ik} \,\Big|\, \xi_i, \sum_{L_i < t_l \leqslant R_i} W_{il} > 0 \right) \,\Bigg|\, A_i = 0, B_i > 0 \right\} \\
&= E_{\xi_i}\left\{ \frac{\xi_i \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik})}{1 - \exp\{-\xi_i(S_{i2} - S_{i1})\}} \,\Bigg|\, A_i = 0, B_i > 0 \right\} \\
&= \lambda_k \exp(\beta^{\mathrm{T}} Z_{ik}) \\
&\quad \times \frac{\int_{\xi_i} \xi_i \{\exp(-\xi_i S_{i1}) - \exp(-\xi_i S_{i2})\}[1 - \exp\{-\xi_i(S_{i2} - S_{i1})\}]^{-1} f(\xi_i)\, \mathrm{d}\xi_i}{\exp\{-G(S_{i1})\} - \exp\{-G(S_{i2})\}},
\end{aligned}
$$

which can be calculated using Gaussian–Laguerre quadrature. In addition,

$$\hat{E}(\xi_i) = \frac{\exp\{-G(S_{i1})\}G'(S_{i1}) - \exp\{-G(S_{i2})\}G'(S_{i2})}{\exp\{-G(S_{i1})\} - \exp\{-G(S_{i2})\}},$$

where $f'(x) = \mathrm{d}f(x)/\mathrm{d}x$ for any function f.

We iterate between the E- and M-steps until the sum of the absolute differences of the estimates at two successive iterations is less than, say, 10^{-3}. This EM algorithm has several desirable features. First, the conditional expectations in the E-step involve at most one-dimensional integration, so they can be evaluated accurately by Gaussian quadrature. Second, in the M-step, the high-dimensional parameters λ_k ($k = 1, \ldots, m$) are calculated explicitly, while the low-dimensional parameter vector β is updated by the Newton–Raphson method. In this way, the algorithm avoids the inversion of any high-dimensional matrices. Finally, the observed-data likelihood is guaranteed to increase after each iteration. To avoid local maxima, we suggest using a range of initial values for β while setting λ_k to $1/m$. We denote the final results by $(\hat{\beta}, \hat{\Lambda})$.

2·3. *Variance estimation*

We use profile likelihood (Murphy & van der Vaart, 2000) to estimate the covariance matrix of $\hat{\beta}$. Specifically, we define the profile loglikelihood

$$\text{pl}_n(\beta) = \max_{\Lambda \in \mathcal{C}} \log L_n(\beta, \Lambda),$$

where $\mathcal{C}$ is the set of step functions with nonnegative jumps at t_k. Then the covariance matrix of $\hat{\beta}$ is estimated by the negative inverse of the matrix whose (j, k)th element is

$$\frac{\text{pl}_n(\hat{\beta}) - \text{pl}_n(\hat{\beta} + h_n e_k) - \text{pl}_n(\hat{\beta} + h_n e_j) + \text{pl}_n(\hat{\beta} + h_n e_k + h_n e_j)}{h_n^2},$$

where e_j is the jth canonical vector in $\mathbb{R}^d$ and h_n is a constant of order $n^{-1/2}$. To calculate $\text{pl}_n(\beta)$ for each β, we reuse the proposed EM algorithm with β held fixed. Thus, the only step in the EM algorithm is to explicitly evaluate $\hat{E}(W_{ik})$ and $\hat{E}(\xi_i)$ so as to update λ_k using (6). The iteration converges quickly with $\hat{\lambda}_k$ as the initial value.

3. Asymptotic theory

We establish the asymptotic properties of $(\hat{\beta}, \hat{\Lambda})$ under the following regularity conditions.

Condition 1.   The true value of β, denoted by β_0, lies in the interior of a known compact set $\mathcal{B}$ in $\mathbb{R}^d$, and the true value of $\Lambda(\cdot)$, denoted by $\Lambda_0(\cdot)$, is continuously differentiable with positive derivatives in $[\zeta, \tau]$, where $[\zeta, \tau]$ is the union of the supports of $(U_1, \ldots, U_K)$.

Condition 2.   The vector $Z(t)$ is uniformly bounded with uniformly bounded total variation over $[\zeta, \tau]$, and its left limit exists for any t. In addition, for any continuously differentiable function $g(\cdot)$, the expectations $E[g\{Z_{(j)}(t)\}]$ $(j = 1, 2)$ are continuously differentiable in $[\zeta, \tau]$, where $Z_{(1)}$ and $Z_{(2)}$ are increasing functions in the decomposition $Z(t) = Z_{(1)}(t) - Z_{(2)}(t)$.

Condition 3.   If $h(t) + \beta^{\mathrm{T}} Z(t) = 0$ for all $t \in [\zeta, \tau]$ with probability 1, then $h(t) = 0$ for $t \in [\zeta, \tau]$ and $\beta = 0$.

Condition 4.   The number of monitoring times, K, is positive, and $E(K) < \infty$. The conditional probability $\text{pr}(U_K = \tau \mid K, Z)$ is greater than some positive constant c. In addition, $\text{pr}(U_{j+1} - U_j \geqslant \eta \mid K, Z) = 1$ $(j = 1, \ldots, K)$ for some positive constant η. Finally, the conditional densities of (U_j, U_{j+1}) given Z and K, denoted by $g_j(u, v \mid Z, K)$ $(j = 1, \ldots, K)$, have continuous second-order partial derivatives with respect to u and v when $v - u > \eta$ and are continuously differentiable with respect to Z.

Condition 5.   The transformation function G is twice continuously differentiable on $[0, \infty)$ with $G(0) = 0$, $G'(x) > 0$ and $G(\infty) = \infty$.

Remark 2.   Condition 1 is standard in survival analysis. Condition 2 allows $Z(t)$ to have discontinuous trajectories, but the expectation of any smooth functional of $Z(t)$ must be differentiable. One example would be that $Z(t)$ is a stochastic process with a finite number of piecewise-smooth trajectories, where the discontinuity points have a continuous joint distribution. This condition excludes taking Brownian motion as a process for $Z(t)$. Condition 3 holds if the matrix $E([1, Z^{\mathrm{T}}(t)]^{\mathrm{T}}[1, Z^{\mathrm{T}}(t)])$ is nonsingular for some t. Condition 4 pertains to the joint

distribution of monitoring times. First, it requires that the monitoring occur anywhere in $[\zeta, \tau]$ and that the largest monitoring time be equal to τ with positive probability. The latter assumption may be removed, at the expense of more complicated proofs. Condition 4 also requires that two adjacent monitoring times be separated by at least η; otherwise, the data may contain exact observations, which would entail a different theoretical treatment. The smoothness condition for the joint density of monitoring times is used to prove the Donsker property of some function classes and the smoothness of the least favourable direction. Finally, Condition 5 pertains to the transformation function and holds for both the logarithmic family $G_r(x) = r^{-1} \log(1 + rx)$ $(r \geqslant 0)$ and the Box–Cox family $G_\rho(x) = \rho^{-1}\{(1 + x)^\rho - 1\}$ $(\rho \geqslant 0)$.

The following theorem establishes the strong consistency of $(\hat{\beta}, \hat{\Lambda})$.

THEOREM 1. *Under Conditions 1–5, $|\hat{\beta} - \beta_0| + \sup_{t \in [\zeta, \tau]} |\hat{\Lambda}(t) - \hat{\Lambda}_0(t)| \to 0$ almost surely as $n \to \infty$, where $|\cdot|$ is the Euclidean norm.*

It is implicitly assumed in Theorem 1 that $\hat{\beta}$ is restricted to $\mathcal{B}$, although in practice $\hat{\beta}$ is allowed to be very large. The proof of Theorem 1 is based on the Kullback–Leibler information and makes use of the strong consistency of empirical processes. Careful arguments are needed to establish a preliminary bound for $\hat{\Lambda}$ and to handle time-dependent covariates. Our next theorem establishes the asymptotic normality and semiparametric efficiency of $\hat{\beta}$.

THEOREM 2. *Under Conditions 1–5, $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution as $n \to \infty$ to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.*

The proof of Theorem 2 relies on the derivation of the least favourable submodel for β_0 and utilizes modern empirical process theory. A key step is to show that $\hat{\Lambda}$ converges to Λ_0 at the $n^{1/3}$ rate. Although the general procedure is similar to that of Huang & Wellner (1997), a major innovation is the derivation of the least favourable submodel for general interval censoring and time-dependent covariates by carefully handling the trajectories of $Z(\cdot)$ and the joint distribution of $(U_1, \ldots, U_K)$. The existence of the least favourable submodel is also used at the end of the Appendix to show consistency of the profile-likelihood covariance estimator given in § 2·3.

4. SIMULATION STUDIES

We conducted simulation studies to assess the operating characteristics of the proposed numerical and inferential procedures. In the first study, we considered two time-independent covariates, $Z_1 \sim \text{Ber}(0·5)$ and $Z_2 \sim \text{Un}(0, 1)$. In the second study, we allowed Z_1 to vary over time by imitating two-stage randomization: $Z_1(t) = B_1 I(t \leqslant V) + B_2 I(t > V)$, where B_1 and B_2 are independent $\text{Ber}(0·5)$ and $V \sim \text{Un}(0, \tau)$ with $\tau = 3$. In both studies, we generated the failure times from the transformation model

$$\Lambda(t; Z_1, Z_2) = G\left[\int_0^t \exp\{\beta_1 Z_1(s) + \beta_2 Z_2\} \, d\Lambda(s)\right],$$

where $G(x) = r^{-1} \log(1 + rx)$ $(r \geqslant 0)$. We set $\beta_1 = 0·5$, $\beta_2 = -0·5$ and $\Lambda(t) = \log(1 + t/2)$. To create interval censoring, we randomly generated two monitoring times, $U_1 \sim \text{Un}(0, 3\tau/4)$ and $U_2 \sim \min\{0·1 + U_1 + \text{Exponential}(1)\tau/2, \tau\}$, so that the time axis $(0, \infty)$ was partitioned into three intervals, $(0, U_1]$, $(U_1, U_2]$ and (U_2, ∞). On average, there were 25–35% left-censored

Table 1. *Summary statistics for the simulation study with time-independent covariates*

		$n = 200$				$n = 400$				$n = 800$			
r		Est	SE	SEE	CP	Est	SE	SEE	CP	Est	SE	SEE	CP
0	$\beta_1 = 0.5$	0.515	0.209	0.216	96	0.506	0.148	0.149	95	0.503	0.103	0.104	95
	$\beta_2 = -0.5$	−0.515	0.366	0.354	94	−0.505	0.254	0.248	95	−0.504	0.176	0.174	95
0.5	$\beta_1 = 0.5$	0.514	0.255	0.259	96	0.507	0.180	0.176	94	0.503	0.125	0.125	95
	$\beta_2 = -0.5$	−0.516	0.451	0.434	94	−0.505	0.311	0.303	94	−0.503	0.215	0.212	94
1	$\beta_1 = 0.5$	0.516	0.294	0.297	95	0.506	0.209	0.207	95	0.504	0.145	0.144	95
	$\beta_2 = -0.5$	−0.517	0.522	0.503	94	−0.505	0.358	0.350	95	−0.502	0.249	0.244	94

Est, empirical average of the parameter estimator; SE, standard error of the parameter estimator; SEE, empirical average of the standard error estimator; CP, empirical coverage percentage of the 95% confidence interval.

Table 2. *Summary statistics for the simulation study with time-dependent covariates*

|		$n = 200$				$n = 400$				$n = 800$			
r		Est	SE	SEE	CP	Est	SE	SEE	CP	Est	SE	SEE	CP
0	$\beta_1 = 0.5$	0.529	0.241	0.239	95	0.518	0.166	0.164	95	0.509	0.114	0.114	95
	$\beta_2 = -0.5$	−0.515	0.363	0.353	95	−0.511	0.253	0.247	94	−0.503	0.175	0.173	95
0.5	$\beta_1 = 0.5$	0.533	0.292	0.280	94	0.522	0.198	0.193	94	0.511	0.138	0.134	94
	$\beta_2 = -0.5$	−0.514	0.441	0.433	95	−0.512	0.307	0.302	95	−0.503	0.214	0.211	95
1	$\beta_1 = 0.5$	0.537	0.336	0.317	94	0.525	0.228	0.219	94	0.514	0.157	0.152	94
	$\beta_2 = -0.5$	−0.518	0.512	0.502	95	−0.513	0.358	0.349	95	−0.505	0.250	0.243	94
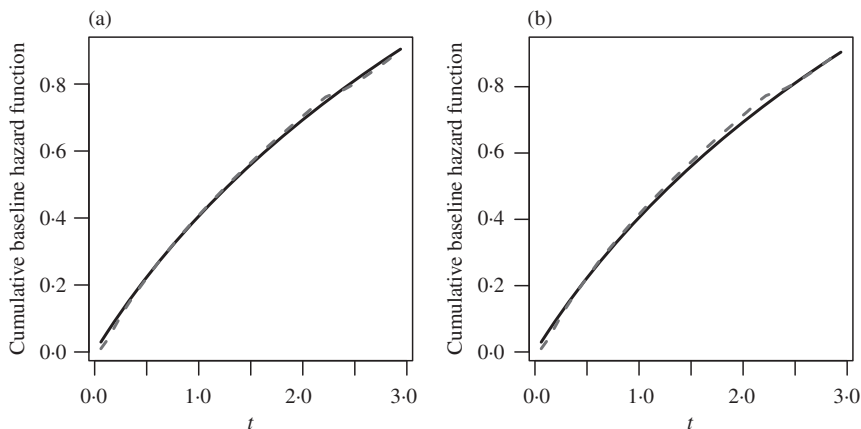
Fig. 1. Estimation of $\Lambda(\cdot)$ with $n = 400$, for (a) $r = 0$ and (b) $r = 1$. The solid and dashed curves represent the true values and mean estimates, respectively.

observations and 50–60% right-censored ones. We set $n = 200$, 400 or 800 and used 10 000 replicates for each sample size.

For each dataset, we applied the proposed EM algorithm by setting the initial value of β to 0 and the initial value of λ_k to $1/m$, and we set the convergence threshold to 10^{-3}. We also tried other initial values for β, but they all led to the same estimates. For the variance estimation, we set $h_n = 5n^{-1/2}$, but the results differed only in the third decimal place if we used $h_n = n^{-1/2}$ or $10n^{-1/2}$. There was no nonconvergence in any of the EM iterations.

Tables 1 and 2 summarize the results of the two simulation studies under $r = 0, 0.5$ or 1. The parameter estimators have small bias, and the bias decreases rapidly as n increases. The
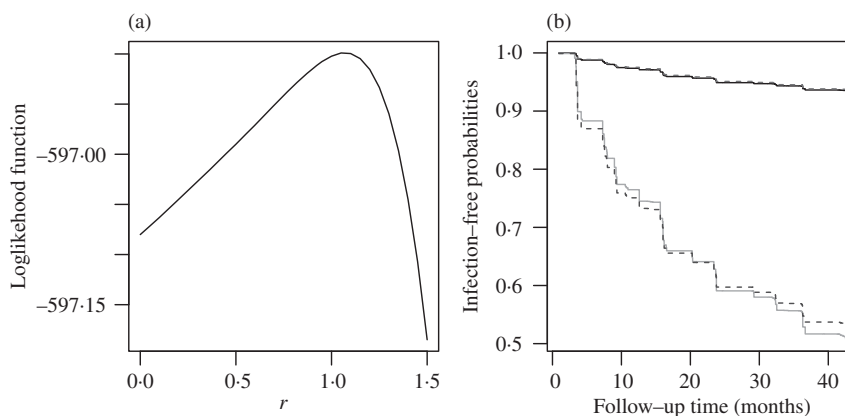
Fig. 2. Analysis of the Bangkok Metropolitan Administration HIV-1 study: (a) the loglike-lihood at the nonparametric maximum likelihood estimates plotted as a function of r in the logarithmic transformations; (b) estimation of infection-free probabilities, where the upper lines correspond to a low-risk subject under the proportional hazards (solid) and proportional odds (dashed) models, and the lower lines correspond to a high-risk subject under the proportional hazards (solid) and proportional odds (dashed) models.

variance estimators accurately reflect the true variabilities, and the confidence intervals have proper coverage probabilities. As shown in Fig. 1, the estimated cumulative hazard functions have negligible bias.

We conducted an additional simulation study with five covariates. We set Z to zero-mean normal with unit variances and pairwise correlations of 0.5 and took $\beta = (0.5, -0.5, 0, 0.3, 0.7)^{\mathrm{T}}$; the other simulation settings were left unchanged. The results are summarized in the Supplementary Material. The proposed methods performed well in this simulation too. Again, there were no cases of nonconvergence.

To evaluate the performance of the EM algorithm in even larger datasets, we set $n = 2000$ and Z to ten standard normal random variables with pairwise correlations of 0.25 and regression coefficients of 0.5. The algorithm converged to values close to 0.5 in all $10\,000$ replicates.

5. APPLICATION

The Bangkok Metropolitan Administration conducted a cohort study of 1209 injecting drug users who were initially sero-negative for the HIV-1 virus. Subjects from 15 drug treatment clinics were followed from 1995 to 1998. At study enrolment and approximately every four months thereafter, subjects were assessed for HIV-1 sero-positivity through blood tests. As of December 1998 there were 133 HIV-1 sero-conversions and roughly 2300 person-years of follow-up.

We aim to identify the factors that influence HIV-1 infection. We fit model (1) with the class of logarithmic transformations $G(x) = r^{-1}\log(1 + rx)$ $(r \geqslant 0)$. The covariates include age at recruitment, gender, history of needle sharing, and drug injection in jail before recruitment; age is measured in years, gender takes value 1 for male and 0 for female, and history of needle sharing and drug injection are binary indicators of yes or no. In addition, we include a time-dependent covariate indicating imprisonment since the last clinic visit.

To select a transformation function, we vary r from 0 to 1.5 in steps of 0.05. For each r, we estimate β and Λ by the EM algorithm and evaluate the loglikelihood at the parameter estimates. Figure 2(a) shows that the loglikelihood changes only very slowly as r varies and is maximized at $r = 1.05$. We choose $r = 1$, which corresponds to the proportional odds model. Table 3 shows the

Table 3. *Regression analysis of the Bangkok Metropolitan Administration HIV-*1 *infection data*

	Proportional hazards			Proportional odds		
Covariates	Estimate	Standard error	*p*-value	Estimate	Standard error	*p*-value
Age	−0·028	0·012	0·021	−0·031	0·013	0·016
Gender	0·424	0·270	0·117	0·539	0·310	0·082
Needle sharing	0·237	0·183	0·196	0·251	0·196	0·200
Drug injection	0·313	0·184	0·089	0·360	0·198	0·069
Imprisonment over time	0·502	0·211	0·017	0·494	0·219	0·024

results under this model. For comparison, we also include the results for $r = 0$, which corresponds to the proportional hazards model.

Under either model, ageing reduces the risk of HIV-1 infection, whereas being male increases it. In addition, drug injection increases the risk. Finally, subjects who have recently been imprisoned have an elevated risk of HIV-1 infection.

Figure 2(b) shows the prediction of HIV-1 infection for a low-risk subject versus a high-risk subject. The low-risk subject is a 50-year-old female with no history of needle sharing, no drug injection in jail before recruitment, and no imprisonment during follow-up; the high-risk subject is a 20-year-old male with a history of needle sharing, drug injection in jail before recruitment, and imprisonment over time. The estimated probabilities of infection for the low-risk subject are similar under the proportional odds and proportional hazards models. For the high-risk subject, however, the proportional odds model yields slightly higher risks of infection than the proportional hazards model during the first part of the follow-up period, with the opposite being true during the later part of the follow-up period.

6. Remarks

The presence of time-dependent covariates poses major computational and theoretical challenges. With time-dependent covariates, the parameters β and λ_k ($k = 1, \ldots, m$) in the likelihood function are entangled. As a result, the diagonal approximation to the Hessian matrix in the iterative convex minorant algorithm (Huang & Wellner, 1997) is inaccurate, and the algorithm becomes unstable. By contrast, each iteration of our EM algorithm only solves a low-dimensional equation for β while calculating the jump sizes of Λ explicitly as weighted sums of Poisson rates. Thus, our algorithm is fast and stable. In extensive numerical studies we have never encountered nonconvergence. Our software is available at http://dlin.web.unc.edu/software.

Our theoretical development requires that the population average of the covariate process be smooth but allows individual covariate trajectories to be discontinuous. We treat $\int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)$ as a bundled process of β and Λ when proving the identifiability of (β_0, Λ_0) in Theorem 1, the convergence rate of $\hat{\Lambda}$ in Lemma A1, and the invertibility of the information operator in Theorem 2. The Donsker property for this class of processes indexed by (β, Λ) is used repeatedly in the proofs. Besides time-dependent covariates, one major theoretical challenge is dealing with general interval censoring, which allows each subject to have a different number of monitoring times. In particular, the derivation of the least favourable direction for β requires careful consideration of the joint distribution for an arbitrary sequence of monitoring times, and the Lax–Milgram theorem is used to prove the existence of a least favourable direction. That theorem greatly simplifies the proof, in contrast to the approach of Huang & Wellner (1997), even for case-2 data.

To apply the transformation model to real data, one must choose a transformation function. In the analysis of the Bangkok Metropolitan Administration HIV-1 data, we used the AIC to select the

transformation function, although the likelihood surface is fairly flat. It would be worthwhile to develop formal diagnostic procedures to check the appropriateness of the transformation function and other model assumptions. One possible strategy is to examine the behaviour of the posterior mean of the martingale residuals (Chen et al., 2012) given the observed intervals.

In many applications, the event of interest may occur repeatedly over time. Recurrent events under interval censoring are called panel count data, which have been studied by Sun & Wei (2000), Zhang (2002) and Wellner & Zhang (2007), among others. There are also studies in which each subject can experience different types of events or where subjects are sampled in clusters such that the failure times with the same cluster are correlated. We are currently developing regression methods to handle such multivariate failure time data.

We are also extending our work to competing risks data. Indeed, the Bangkok Metropolitan Administration HIV-1 study contains information on HIV-1 infection by viral subtypes B and E, which are two competing risks. We propose to formulate the effects of potentially time-dependent covariates on the cumulative incidence functions of competing risks in the form of model (1). We will modify the EM algorithm to deal with multiple subdistribution functions and establish the asymptotic theory under suitable conditions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional simulation results and a zip file containing the computer code and documentation.

APPENDIX

Technical details

We use $\mathbb{P}_n$ to denote the empirical measure from n independent observations and $\mathbb{P}$ to denote the true probability measure. The corresponding empirical process is $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$. Let $l(\beta, \Lambda)$ be the observed-data loglikelihood for a single subject, that is,

$$l(\beta, \Lambda) = \sum_{k=0}^{K} \Delta_k \log\left\{ \exp\left(-G\left[\int_0^{U_k} \exp\{\beta^{\mathrm{T}} Z(s)\} \, d\Lambda(s) \right] \right) \right.$$
$$\left. - \exp\left(-G\left[\int_0^{U_{k+1}} \exp\{\beta^{\mathrm{T}} Z(s)\} \, d\Lambda(s) \right] \right) \right\}.$$

Proof of Theorem 1. We first show that $\limsup_n \hat{\Lambda}(\tau) < \infty$ with probability 1. By Condition 4, the measure generated by the function $E\{\sum_{k=0}^{K} \Delta_k I(U_k \leqslant t)\}$ is dominated by the sum of the Lebesgue measure in $[\zeta, \tau]$ and the counting measure at τ, and its Radon–Nikodym derivative, denoted by $f_1(t)$, is bounded away from zero. We define

$$\tilde{\Lambda}(t) = \int_0^t \frac{\Lambda_0'(s)}{f_1(s)} \, d\left\{ n^{-1} \sum_{i=1}^{n} \sum_{k=0}^{K_i} \Delta_{ik} I(U_{ik} \leqslant s) \right\}.$$

Clearly, $\tilde{\Lambda}$ is a step function with jumps only at $t_1 < \cdots < t_m$. Since

$$n^{-1} \sum_{i=1}^{n} \sum_{k=0}^{K_i} \Delta_{ik} I(U_{ik} \leqslant s) \to E \left\{ \sum_{k=0}^{K} \Delta_k I(U_k \leqslant s) \right\}$$

uniformly in s with probability 1 as $n \to \infty$, we conclude that $\tilde{\Lambda}(t)$ converges uniformly to $\Lambda_0(t)$ with probability 1 for $t \in [\zeta, \tau]$.

By the definition of $(\hat{\beta}, \hat{\Lambda})$, we have $\mathbb{P}_n l(\hat{\beta}, \hat{\Lambda}) \geqslant \mathbb{P}_n l(\beta_0, \tilde{\Lambda})$. Because of its bounded total variation, $Z(t)$ belongs to a Donsker class indexed by t. Hence, the class of functions

$$\mathcal{F}_\omega = \left\{ \int_0^{U_k} \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s) : \beta \in \mathcal{B}, \ \Lambda \in \mathrm{BV}_\omega[0, \tau] \right\},$$

where $\mathrm{BV}_\omega[0, \tau]$ denotes functions which have total variation in $[0, \tau]$ bounded by a given constant ω, is a convex hull of functions $\{I(U_k \geqslant s) \exp\{\beta^{\mathrm{T}} Z(s)\}\}$, so it is a Donsker class. Furthermore,

$$\exp \left(-G \left[\int_0^{U_k} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\tilde{\Lambda}(s) \right] \right) - \exp \left(-G \left[\int_0^{U_{k+1}} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\tilde{\Lambda}(s) \right] \right)$$

is bounded away from zero. Therefore, $l(\beta_0, \tilde{\Lambda})$ belongs to some Donsker class due to the preservation property of the Donsker class under Lipschitz-continuous transformations. We conclude that $|\mathbb{P}_n l(\beta_0, \tilde{\Lambda}) - \mathbb{P} l(\beta_0, \tilde{\Lambda})| \to 0$ almost surely. In addition, by the construction of $\tilde{\Lambda}$, $\mathbb{P} l(\beta_0, \tilde{\Lambda})$ converges almost surely to $\mathbb{P} l(\beta_0, \Lambda_0)$, which is finite. Therefore, with probability 1,

$$\liminf_n \mathbb{P}_n l(\hat{\beta}, \hat{\Lambda}) \geqslant O(1). \tag{A1}$$

Let $c > 0$ be such that $\exp\{\beta^{\mathrm{T}} Z(t)\} \geqslant c$ for $\beta \in \mathcal{B}$. Then the left-hand side of (A1) is less than or equal to

$$-\limsup_n \mathbb{P}_n \left(\Delta_K G \left[\int_0^{U_K} \exp\{\hat{\beta}^{\mathrm{T}} Z(t)\} \, \mathrm{d}\hat{\Lambda}(t) \right] \right) \leqslant -\limsup_n \mathbb{P}_n \left[\Delta_K G\{c\hat{\Lambda}(U_K)\} \right]$$

$$\leqslant -\limsup_n \mathbb{P}_n \left[\Delta_K I(U_K = \tau) G\{c\hat{\Lambda}(\tau)\} \right].$$

Hence $\limsup_n \mathbb{P}_n[\Delta_K I(U_K = \tau) G\{c\hat{\Lambda}(\tau)\}] \leqslant O(1)$. Since as $n \to \infty$, $\mathbb{P}_n\{\Delta_K I(U_K = \tau)\} \to E\{\Delta_K I(U_K = \tau)\}$, which is positive, Condition 5 implies that $\limsup_n \hat{\Lambda}(\tau) < \infty$ with probability 1.

We can now restrict $\hat{\Lambda}$ to a class of functions with uniformly bounded total variation, equipped with the weak topology on $[\zeta, \tau]$. By Helly's selection lemma, for any subsequence of $(\hat{\beta}, \hat{\Lambda})$ we can choose a further subsequence such that $\hat{\Lambda}$ converges weakly to some Λ^* on $[\zeta, \tau]$, $\hat{\Lambda}(\tau)$ converges to $\Lambda^*(\tau)$, and $\hat{\beta}$ converges to β^*. Clearly, $\mathbb{P}_n l(\hat{\beta}, \hat{\Lambda}) \geqslant \mathbb{P}_n l(\beta_0, \tilde{\Lambda})$ implies that $\mathbb{P}_n \log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] \geqslant \mathbb{P}_n \log p(\beta_0, \tilde{\Lambda})$, where $p(\beta, \Lambda) = \exp\{l(\beta, \Lambda)\}$, so that

$$(\mathbb{P}_n - \mathbb{P}) \log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] + \mathbb{P} \log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2]$$

$$\geqslant (\mathbb{P}_n - \mathbb{P}) l(\beta_0, \tilde{\Lambda}) + \mathbb{P} l(\beta_0, \tilde{\Lambda}).$$

By the above arguments for proving the Donsker property of $\mathcal{F}_\omega$, together with the fact that the total variation of $\hat{\Lambda}$ is bounded by a constant, we can show that $\log\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}$ belongs to a Donsker class with a bounded envelope function, so that $(\mathbb{P}_n - \mathbb{P}) \log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] \to 0$. It follows

that $\mathbb{P}l(\beta_0, \Lambda_0) - \mathbb{P}\log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] \leqslant o(1)$. Furthermore,

$$\left| \mathbb{P}\log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] - \mathbb{P}\log[\{p(\beta^*, \Lambda^*) + p(\beta_0, \tilde{\Lambda})\}/2] \right|$$

$$\leqslant O(1) \left(|\hat{\beta} - \beta^*| + E\left[\sum_{k=0}^{K} \left| \int_0^{U_k} \exp\{\beta^{*\mathrm{T}}Z(s)\} \,d\hat{\Lambda}(s) - \int_0^{U_k} \exp\{\beta^{*\mathrm{T}}Z(s)\} \,d\Lambda^*(s) \right| \right] \right)$$

$$\leqslant O(1) \left(|\hat{\beta} - \beta^*| + E\left[\sum_{k=0}^{K} \left\{ \left| \hat{\Lambda}(U_k) - \Lambda^*(U_k) \right| + \int \left| \hat{\Lambda}(s) - \Lambda^*(s) \right| I(U_k \geqslant s) \,|dZ(s)| \right\} \right] \right)$$

$$\leqslant O(1) \left\{ |\hat{\beta} - \beta^*| + \int |\hat{\Lambda}(s) - \Lambda^*(s)| \,d\nu(s) \right\},$$

where $d\nu$ is the measure corresponding to $[dE\{\sum_{k=0}^{K} I(U_k \geqslant s)\} + E\{\sum_{k=0}^{K} I(U_k \geqslant s)|dZ(s)|\}]$. According to Conditions 2 and 4, $d\nu(s)$ is dominated by the Lebesgue measure with bounded derivative in $[\zeta, \tau]$ and has a point mass at τ. Hence $|\mathbb{P}\log[\{p(\hat{\beta}, \hat{\Lambda}) + p(\beta_0, \tilde{\Lambda})\}/2] - \mathbb{P}\log[\{p(\beta^*, \Lambda^*) + p(\beta_0, \tilde{\Lambda})\}/2]| \to 0$ almost surely. We therefore conclude that $\mathbb{P}l(\beta_0, \Lambda_0) - \mathbb{P}\log[\{p(\beta^*, \Lambda^*) + p(\beta_0, \tilde{\Lambda})\}/2] \leqslant 0$.

By the properties of the Kullback–Leibler information, $p(\beta_0, \Lambda_0) = p(\beta^*, \Lambda^*)$ with probability 1. In particular, for any $k \leqslant K$, we choose $\Delta_k = 1$ to obtain

$$\int_0^{U_k} \exp\{\beta_0^{\mathrm{T}}Z(t)\} \,d\Lambda_0(t) = \int_0^{U_k} \exp\{\beta^{*\mathrm{T}}Z(t)\} \,d\Lambda^*(t).$$

Thus, for any $s \in [\zeta, \tau]$,

$$\int_0^s \exp\{\beta_0^{\mathrm{T}}Z(t)\} \,d\Lambda_0(t) = \int_0^s \exp\{\beta^{*\mathrm{T}}Z(t)\} \,d\Lambda^*(t). \tag{A2}$$

Differentiating both sides with respect to s, we have

$$\beta_0^{\mathrm{T}}Z(t) + \log \Lambda_0'(t) = \beta^{*\mathrm{T}}Z(t) + \log \Lambda^{*\prime}(t).$$

By Condition 3, $\beta^* = \beta_0$ and $\Lambda_0'(t) = \Lambda^{*\prime}(t)$ for $t \in [\zeta, \tau]$. We let $Z(t) = 0$ by redefining $Z(t)$ to centre at a deterministic function in the support of $Z(t)$, and we set $s = \zeta$ in (A2) to obtain $\Lambda_0(\zeta) = \Lambda^*(\zeta)$. Hence, $\Lambda_0(t) = \Lambda^*(t)$ for $t \in [\zeta, \tau]$. It follows that $\hat{\beta} \to \beta_0$ and $\hat{\Lambda}(t)$ converges weakly to $\Lambda_0(t)$ almost surely. The latter convergence can be strengthened to uniform convergence since Λ_0 is continuous. Thus, we have proved Theorem 1. □

Since we have established consistency, we may restrict the space of Λ to

$$\mathcal{A} = \{\Lambda : 1/M \leqslant \Lambda(\zeta) \leqslant \Lambda(\tau) \leqslant M\}$$

for some $M > 0$. Thus, when n is large enough, $\hat{\Lambda}$ belongs to $\mathcal{A}$ with probability 1. Before proving Theorem 2, we need to establish the convergence rate for $\hat{\Lambda}$. Specifically, the following lemma holds.

LEMMA A1. *Under Conditions* 1–5,

$$E\left(\sum_{k=1}^{K} \left[\int_0^{U_k} \exp\{\hat{\beta}^{\mathrm{T}}Z(s)\} \,d\hat{\Lambda}(s) - \int_0^{U_k} \exp\{\beta_0^{\mathrm{T}}Z(s)\} \,d\Lambda_0(s) \right]^2 \right)^{1/2} = O_p(n^{-1/3}).$$

Proof. The proof relies on the convergence-rate result in Theorem 3.4.1 of van der Vaart & Wellner (1996). To use that theorem, we define $m_{\beta, \Lambda} = \log\{l(\beta, \Lambda)/l(\beta_0, \Lambda_0)\}$ and let

$$\mathcal{M} = \{m_{\beta, \Lambda} : \beta \in \mathcal{B}, \ \Lambda \in \mathcal{A}\}$$

be a class of functions indexed by β and Λ.

We first calculate the ϵ-bracketing number of $\mathcal{M}$. Because $\mathcal{A}$ consists of increasing and uniformly bounded functions on $[\zeta, \tau]$, Lemma 2.2 of van de Geer (2000) implies that for any $\epsilon > 0$, the bracketing number satisfies

$$N_{[]}(\epsilon, \mathcal{A}, \|\cdot\|_{L_2}) \lesssim \epsilon^{-1},$$

where $\|\cdot\|_{L_2}$ denotes the L_2-norm with respect to the Lebesgue measure on $[0, \tau]$, and $A \lesssim B$ means that $A \leqslant cB$ for a positive constant c. For $\epsilon > 0$, we can find $\exp\{O(1/\epsilon)\}$ number of brackets $\{[\Lambda_j^{\text{L}}, \Lambda_j^{\text{U}}]\}$ with $\|\Lambda_j^{\text{U}} - \Lambda_j^{\text{L}}\|_{L_2} \leqslant \epsilon$ and $|\Lambda_j^{\text{U}}(\tau) - \Lambda_j^{\text{L}}(\tau)| \leqslant \epsilon$ to cover $\mathcal{A}$. In addition, there are $O(\epsilon^{-d})$ number of brackets covering $\mathcal{B}$, such that any two β within the same bracket differ by at most ϵ. Hence, there are a total of $\exp\{O(1/\epsilon)\} \times O(\epsilon^{-d})$ brackets covering $\mathcal{B} \times \mathcal{A}$. For any pair of (β_1, Λ_1) and (β_2, Λ_2), there exist some constants c_1 and c_2 such that

$$
\begin{aligned}
&|m_{\beta_1, \Lambda_1} - m_{\beta_2, \Lambda_2}| \\
&\leqslant |m_{\beta_1, \Lambda_1} - m_{\beta_2, \Lambda_1}| + |m_{\beta_2, \Lambda_1} - m_{\beta_2, \Lambda_2}| \\
&\leqslant c_1 |\beta_1 - \beta_2| + c_1 \sum_{k=0}^{K} \Delta_k \left| \int_0^{U_k} \exp\{\beta_2^{\text{T}} Z(s)\} \, d\Lambda_1(s) - \int_0^{U_k} \exp\{\beta_2^{\text{T}} Z(s)\} \, d\Lambda_2(s) \right| \\
&\leqslant c_2 |\beta_1 - \beta_2| + c_2 \sum_{k=0}^{K} \Delta_k \left\{ |\Lambda_1(U_k) - \Lambda_2(U_k)| + \int |\Lambda_1(s) - \Lambda_2(s)| \, I(U_k \geqslant s) \, |dZ(s)| \right\}.
\end{aligned}
$$

Because the measures $dE\{\sum_{k=0}^{K} I(U_k \geqslant s)\}$ and $E\{\sum_{k=0}^{K} I(U_k \geqslant s) |dZ(s)|\}$ have bounded derivatives with respect to the Lebesgue measure in $[\zeta, \tau)$ and the former has a finite point mass at τ,

$$\|m_{\beta_1, \Lambda_1} - m_{\beta_2, \Lambda_2}\|_{L_2(\mathbb{P})} \leqslant O\{|\beta_1 - \beta_2| + \|\Lambda_1 - \Lambda_2\|_{L_2} + |\Lambda_1(\tau) - \Lambda_2(\tau)|\} = O(\epsilon).$$

Thus, the bracketing number for $\mathcal{M}$ satisfies

$$N_{[]}\{\epsilon, \mathcal{M}, L_2(\mathbb{P})\} \leqslant \exp\{O(1/\epsilon)\} O(\epsilon^{-d}),$$

and so it has a finite entropy integral. Define

$$\varphi(\delta) = \int_0^{\delta} \left[1 + \log N_{[]}\{\epsilon, \mathcal{M}, L_2(\mathbb{P})\}\right]^{1/2} d\epsilon.$$

It is easy to show that $\varphi(\delta) \leqslant O(\delta^{1/2})$. In addition, by Lemma 1.3 of van de Geer (2000),

$$\mathbb{P}(m_{\beta, \Lambda} - m_{\beta_0, \Lambda_0}) \lesssim -H^2\{(\beta, \Lambda), (\beta_0, \Lambda_0)\},$$

where $H\{(\beta, \Lambda), (\beta_0, \Lambda_0)\}$ is the Hellinger distance, defined as

$$H\{(\beta, \Lambda), (\beta_0, \Lambda_0)\} = \left(\int \left[\exp\{l(\beta, \Lambda)/2\} - \exp\{l(\beta_0, \Lambda_0)/2\}\right]^2 d\mu \right)^{1/2}$$

with respect to the dominating measure μ.

The above results, together with the fact that $(\hat{\beta}, \hat{\Lambda})$ maximizes $\mathbb{P}_n m_{\beta, \Lambda}$ and the consistency result in Theorem 1, imply that all the conditions in Theorem 3.4.1 of van der Vaart & Wellner (1996) hold. Thus, we conclude that $H\{(\hat{\beta}, \hat{\Lambda}), (\beta_0, \Lambda_0)\} = O_{\text{p}}(r_n)$, where r_n satisfies $r_n \varphi(r_n^{-1}) \lesssim n^{1/2}$. In particular, we can

choose r_n in the order of $n^{-1/3}$ such that $H\{(\hat{\beta}, \hat{\Lambda}), (\beta_0, \Lambda_0)\} = O_p(n^{-1/3})$. By the mean value theorem,

$$E\left(\sum_{k=0}^{K} \Delta_k \left[\left\{\exp\left(-G\left[\int_0^{U_k} \exp\{\hat{\beta}^{\mathrm{T}} Z(s)\} \, \mathrm{d}\hat{\Lambda}(s)\right]\right) - \exp\left(-G\left[\int_0^{U_{k+1}} \exp\{\hat{\beta}^{\mathrm{T}} Z(s)\} \, \mathrm{d}\hat{\Lambda}(s)\right]\right)\right\}\right.\right.$$

$$\left.\left. - \left\{\exp\left(-G\left[\int_0^{U_k} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda_0(s)\right]\right) - \exp\left(-G\left[\int_0^{U_{k+1}} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda_0(s)\right]\right)\right\}\right]^2\right)$$

$$= O_p(n^{-2/3}).$$

On the left-hand side of the above equation, we consider the event $\Delta_0 = 1$ to find that

$$E\left\{\exp\left(-G\left[\int_0^{U_1} \exp\{\hat{\beta}^{\mathrm{T}} Z(s)\} \, \mathrm{d}\hat{\Lambda}(s)\right]\right) - \exp\left(-G\left[\int_0^{U_1} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda_0(s)\right]\right)\right\}^2 = O_p(n^{-2/3}).$$

Next, we consider $\Delta_1 = 1$ to obtain the same equation as the one above but with U_1 replaced by U_2. By repeating this process, we conclude that, conditional on K and for any $k \leqslant K$,

$$E\left\{\exp\left(-G\left[\int_0^{U_k} \exp\{\hat{\beta}^{\mathrm{T}} Z(s)\} \, \mathrm{d}\hat{\Lambda}(s)\right]\right) - \exp\left(-G\left[\int_0^{U_k} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda_0(s)\right]\right)\right\}^2 = O_p(n^{-2/3}).$$

It then follows from the mean value theorem that

$$E\left(\sum_{k=1}^{K} \left[\int_0^{U_k} \exp\{\hat{\beta}^{\mathrm{T}} Z(s)\} \, \mathrm{d}\hat{\Lambda}(s) - \int_0^{U_k} \exp\{\beta_0^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda_0(s)\right]^2\right) = O_p(n^{-2/3}),$$

and so the lemma is proved. $\qquad \square$

Now we are ready to prove Theorem 2.

Proof of Theorem 2. It is helpful to introduce the following notation: $\pi_k(Z) = \mathrm{pr}(K = k \mid Z)$,

$$Q(u, v; Z, \beta, \Lambda) = \exp\left(-G\left[\int_0^u \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right]\right) - \exp\left(-G\left[\int_0^v \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right]\right)$$

and

$$B(t, u, v; Z, \beta, \Lambda)$$

$$= \frac{-\exp\left(-G\left[\int_0^u \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right]\right) G'\left[\int_0^u \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right] I(u \geqslant t)}{Q(u, v; Z, \beta, \Lambda)} \exp\{\beta^{\mathrm{T}} Z(t)\}$$

$$+ \frac{\exp\left(-G\left[\int_0^v \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right]\right) G'\left[\int_0^v \exp\{\beta^{\mathrm{T}} Z(s)\} \, \mathrm{d}\Lambda(s)\right] I(v \geqslant t)}{Q(u, v; Z, \beta, \Lambda)} \exp\{\beta^{\mathrm{T}} Z(t)\}.$$

The score function for β is

$$l_\beta(\beta, \Lambda) = \int_0^\tau \sum_{k=0}^{K} \Delta_k B(t, U_k, U_{k+1}; Z, \beta, \Lambda) Z(t) \, \mathrm{d}\Lambda(t).$$

To obtain the score operator for Λ, we consider any parametric submodel of Λ defined by $\mathrm{d}\Lambda_{\epsilon,h} = (1 + \epsilon h) \, \mathrm{d}\Lambda$, where $h \in L_2([\zeta, \tau])$. The score function along this submodel is

$$l_\Lambda(\beta, \Lambda)(h) = \frac{\partial}{\partial \epsilon} \log l(\beta, \Lambda_{\epsilon,h})\Big|_{\epsilon=0} = \int_\zeta^\tau \sum_{k=0}^{K} \Delta_k B(t, U_k, U_{k+1}; Z, \beta, \Lambda) h(t) \, \mathrm{d}\Lambda(t).$$

Clearly, $\mathbb{P}_n\{l_\beta(\hat{\beta}, \hat{\Lambda})\} = 0$ and $\mathbb{P}_n\{l_\Lambda(\hat{\beta}, \hat{\Lambda})(h)\} = 0$. Hence

$$\mathbb{G}_n\{l_\beta(\hat{\beta}, \hat{\Lambda})\} = -n^{1/2}\big[\mathbb{P}\{l_\beta(\hat{\beta}, \hat{\Lambda})\} - \mathbb{P}\{l_\beta(\beta_0, \Lambda_0)\}\big],$$

$$\mathbb{G}_n\{l_\Lambda(\hat{\beta}, \hat{\Lambda})(h)\} = -n^{1/2}\big[\mathbb{P}\{l_\Lambda(\hat{\beta}, \hat{\Lambda})(h)\} - \mathbb{P}\{l_\Lambda(\beta_0, \Lambda_0)(h)\}\big].$$

We apply Taylor series expansions about (β_0, Λ_0) to the right-hand sides of the above two equations. By Lemma A1, the second-order terms are bounded by

$$n^{1/2} O\left\{ |\hat{\beta} - \beta_0|^2 + E\left(\sum_{k=1}^{K}\left[\int_0^{U_k}\exp\{\hat{\beta}^{\mathsf{T}}Z(s)\}\,d\hat{\Lambda}(s) - \int_0^{U_k}\exp\{\beta_0^{\mathsf{T}}Z(s)\}\,d\Lambda_0(s)\right]^2\right)\right\}$$

$$= O\big(n^{1/2}|\hat{\beta} - \beta_0|^2 + n^{-1/6}\big).$$

Hence

$$\mathbb{G}_n\{l_\beta(\hat{\beta}, \hat{\Lambda})\} = -n^{1/2}\mathbb{P}l_{\beta\beta}(\hat{\beta} - \beta_0) - n^{1/2}\mathbb{P}l_{\beta\Lambda}(d\hat{\Lambda} - d\Lambda_0) + O(n^{1/2}|\hat{\beta} - \beta_0|^2 + n^{-1/6}), \quad \text{(A3)}$$

$$\mathbb{G}_n\{l_\Lambda(\hat{\beta}, \hat{\Lambda})(h)\} = -n^{1/2}\mathbb{P}l_{\Lambda\beta}(h)(\hat{\beta} - \beta_0) - n^{1/2}\mathbb{P}l_{\Lambda\Lambda}(h, d\hat{\Lambda} - d\Lambda_0)$$
$$+ O(n^{1/2}|\hat{\beta} - \beta_0|^2 + n^{-1/6}), \quad \text{(A4)}$$

where $l_{\beta\beta}$ is the second derivative of $l(\beta, \Lambda)$ with respect to β, $l_{\beta\Lambda}(h)$ is the derivative of l_β along the submodel $d\Lambda_{\epsilon,h}$, $l_{\Lambda\beta}(h)$ is the derivative of $l_\Lambda(h)$ with respect to β, and $l_{\Lambda\Lambda}(h, d\hat{\Lambda} - d\Lambda_0)$ is the derivative of $l_\Lambda(h)$ along the submodel $d\Lambda_0 + \epsilon(d\hat{\Lambda} - d\Lambda_0)$. All the derivatives on the right-hand sides of (A3) and (A4) are evaluated at (β_0, Λ_0).

We choose h to be the least favourable direction h^*, a d-vector with components in $L_2([\zeta, \tau])$ that solves the normal equation

$$l_\Lambda^* l_\Lambda(h^*) = l_\Lambda^* l_\beta, \quad \text{(A5)}$$

where l_Λ^* is the adjoint operator of l_Λ. Then

$$E\{l_{\Lambda\Lambda}(h^*, d\hat{\Lambda} - d\Lambda_0)\} = -E\{l_\Lambda(h^*)l_\Lambda(d\hat{\Lambda} - d\Lambda_0)\}$$

$$= -\int l_\Lambda^* l_\Lambda(h^*)(d\hat{\Lambda} - d\Lambda_0) = -\int l_\Lambda^* l_\beta(d\hat{\Lambda} - d\Lambda_0)$$

$$= -E\{l_\beta l_\Lambda(d\hat{\Lambda} - d\Lambda_0)\} = E\{l_{\beta\Lambda}(d\hat{\Lambda} - d\Lambda_0)\},$$

so that the difference between (A3) and (A4) yields

$$\mathbb{G}_n\{l_\beta(\hat{\beta}, \hat{\Lambda}) - l_\Lambda(\hat{\beta}, \hat{\Lambda})(h^*)\}$$
$$= n^{1/2}E\big[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^{\mathsf{T}}\big](\hat{\beta} - \beta_0) + O\big(n^{1/2}|\hat{\beta} - \beta_0|^2 + n^{-1/6}\big). \quad \text{(A6)}$$

Consequently, Theorem 2 will be established if we can show that:

  (i) equation (A5) has a solution h^*;
 (ii) $l_\beta(\hat{\beta}, \hat{\Lambda}) - l_\Lambda(\hat{\beta}, \hat{\Lambda})(h^*)$ belongs to a Donsker class and converges in the $L_2(\mathbb{P})$-norm to $l_\beta - l_\Lambda(h^*)$;
(iii) the matrix $E[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^{\mathsf{T}}]$ is invertible.

The reason is that when (i)–(iii) hold, (A6) entails $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$ and further yields

$$n^{1/2}(\hat{\beta} - \beta_0) = \big(E\big[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^{\mathsf{T}}\big]\big)^{-1}\mathbb{G}_n\{l_\beta - l_\Lambda(h^*)\} + o_p(1).$$

This implies that the influence function for $\hat{\beta}$ is exactly the efficient influence function, so that $n^{1/2}(\hat{\beta} - \beta_0)$ converges to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound (Bickel et al., 1993, p. 65).

We first verify (i). For any $(h_1, h_2) \in L_2[\zeta, \tau]$,

$$\mathbb{P}\{l_\Lambda(h_1)l_\Lambda(h_2)\} = \int_\zeta^\tau \int_\zeta^\tau E\{g(Z; t, s)\}h_1(t)h_2(s)\,\mathrm{d}\Lambda_0(t)\,\mathrm{d}\Lambda_0(s),$$

where

$$g(Z; t, s) = \sum_{k=1}^\infty \pi_k(Z) \sum_{j=0}^k E\big\{Q(U_j, U_{j+1}; Z, \beta_0, \Lambda_0)B(t, U_j, U_{j+1}; Z, \beta_0, \Lambda_0) \\ \times B(s, U_j, U_{j+1}; Z, \beta_0, \Lambda_0) \mid K = k, Z\big\}.$$

Likewise,

$$\mathbb{P}\{l_\Lambda(h_1)l_\beta\} = \int_\zeta^\tau \int_\zeta^\tau E\{g(Z; t, s)Z(s)\}h_1(t)\,\mathrm{d}\Lambda_0(t)\,\mathrm{d}\Lambda_0(s).$$

Therefore, by the definition of the dual operator l_Λ^*, solving the normal equation (A2) is equivalent to solving the integral equation

$$\int_\zeta^\tau E\{g(Z; t, s)\}h^*(s)\,\mathrm{d}\Lambda_0(s) = \int_\zeta^\tau E\{g(Z; t, s)Z(s)\}\,\mathrm{d}\Lambda_0(s). \tag{A7}$$

We define the left-hand side of (A7) as a linear operator Γ which maps $h \in L_2[\zeta, \tau]$ to itself. In addition, we equip $L_2[\zeta, \tau]$ with an inner product $\langle h_1, h_2 \rangle = \int_\zeta^\tau h_1(t)h_2(t)\,\mathrm{d}\Lambda_0(t)$ so that it becomes a Hilbert space. On the same space, we define $\|h\| = \langle \Gamma(h), h \rangle^{1/2}$. It is easy to show that $\|\cdot\|$ is a seminorm for $L_2[\zeta, \tau]$. Furthermore, if $\|h\| = 0$, then $0 = \langle \Gamma(h), h \rangle = \mathbb{P}\{l_\Lambda(h)^2\}$. Thus, with probability 1, $l_\Lambda(h)$ is zero, i.e., for any $\Delta_k = 0$,

$$- \exp\left(-G\left[\int_0^{U_k} \exp\{\beta_0^\mathsf{T} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]\right) G'\left[\int_0^{U_k} \exp\{\beta_0^\mathsf{T} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]$$

$$\times \int_0^{U_k} \exp\{\beta_0^\mathsf{T} Z(t)\}h(t)\,\mathrm{d}\Lambda_0(t)$$

$$+ \exp\left(-G\left[\int_0^{U_{k+1}} \exp\{\beta_0^\mathsf{T} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]\right) G'\left[\int_0^{U_{k+1}} \exp\{\beta_0^\mathsf{T} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]$$

$$\times \int_0^{U_{k+1}} \exp\{\beta_0^\mathsf{T} Z(t)\}h(t)\,\mathrm{d}\Lambda_0(t) = 0.$$

Setting $k = 0, 1 \ldots,$ we obtain

$$\int_0^{U_1} h(t)\,\mathrm{d}\Lambda_0(t) = 0, \quad \int_0^{U_2} h(t)\,\mathrm{d}\Lambda_0(t) = 0, \quad \ldots.$$

Thus, $h(t) = 0$ for $t \in [\zeta, \tau]$, implying that $\|\cdot\|$ is a norm in $L_2[\zeta, \tau]$. Clearly, $\|h\| \leqslant c\langle h, h \rangle^{1/2}$ for some constant c. According to the bounded inverse theorem in Banach spaces, we have $\langle h, h \rangle^{1/2} \leqslant \tilde{c}\|h\|$ for another constant $\tilde{c}$; that is, we have $\langle \Gamma(h), h \rangle \geqslant \tilde{c}^2\langle h, h \rangle$. By the Lax–Milgram theorem (Zeidler, 1995), the solution to (A7), namely h^*, exists. So we have verified (i).

To verify (ii), we examine $g(Z; t, s)$ by considering $t < s$ and $t \geqslant s$. Along the lines of Huang & Wellner (1997), we differentiate the integral equation (A7) with respect to t to obtain

$$b_1(t)h^*(t) + \int_t^\tau b_2(t, s)h^*(s)\,\mathrm{d}s + \int_\zeta^t b_3(t, s)h^*(s)\,\mathrm{d}s = b_4(t),$$

where $b_1(t) > 0$ and b_j $(j = 1, 2, 3, 4)$ are continuously differentiable with respect to their arguments. Hence $h^*(t)$ is continuously differentiable in $[\zeta, \tau]$. This fact implies that $\int_0^{U_k} h^*(s)\exp\{\hat{\beta}^\mathsf{T} Z(s)\}\,\mathrm{d}\hat\Lambda(s)$ belongs to some Donsker class. It then follows from the Donsker property of the class $\mathcal{F}_\omega$ that (ii) is true.

Finally, we verify (iii). If the matrix is singular, then there exists a nonzero vector v such that

$$v^{\mathsf{T}} E\left[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^{\mathsf{T}}\right] v = 0.$$

It follows that, with probability 1, the score function along the submodel $\{\beta_0 + \epsilon v,\ \mathrm{d}\Lambda_0(1 + \epsilon v^{\mathsf{T}} h^*)\}$ is zero; that is, for any $\Delta_k = 1$,

$$- \exp\left(-G\left[\int_0^{U_k} \exp\{\beta_0^{\mathsf{T}} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]\right) G'\left[\int_0^{U_k} \exp\{\beta_0^{\mathsf{T}} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]$$

$$\times \int_0^{U_k} \exp\{\beta_0^{\mathsf{T}} Z(s)\}\tilde{h}(t)\,\mathrm{d}\Lambda_0(t)$$

$$+ \exp\left(-G\left[\int_0^{U_{k+1}} \exp\{\beta_0^{\mathsf{T}} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]\right) G'\left[\int_0^{U_{k+1}} \exp\{\beta_0^{\mathsf{T}} Z(s)\}\,\mathrm{d}\Lambda_0(s)\right]$$

$$\times \int_0^{U_{k+1}} \exp\{\beta_0^{\mathsf{T}} Z(t)\}\tilde{h}(t)\,\mathrm{d}\Lambda_0(t) = 0,$$

where $\tilde{h}(t) = v^{\mathsf{T}} Z(t) + v^{\mathsf{T}} h^*(t)$. We consider $\Delta_k = 1$ for $k = 0, 1, \ldots$ to obtain

$$\int_0^{U_k} \tilde{h}(t)\,\mathrm{d}\Lambda_0(t) = 0.$$

Therefore, with probability 1, $\int_0^s \tilde{h}(t)\,\mathrm{d}\Lambda_0(t) = 0$ for any $s \in [\zeta, \tau]$. This implies that $v^{\mathsf{T}} Z(t) + v^{\mathsf{T}} h^*(t) = 0$, so $v = 0$ by Condition 3. Thus, we have verified (iii). $\qquad \square$

Remark A1.   For a given β, we define $\hat{\Lambda}_\beta$ as the step function that maximizes $L_n(\beta, \Lambda)$. The arguments in the proof of Theorem 1 can be used to show that $\hat{\Lambda}_\beta$ is bounded asymptotically when β is in a small neighbourhood of β_0, so $\hat{\Lambda}_\beta$ converges to Λ_0 as β converges to β_0. In addition, the arguments in the proof of Lemma A1 yield $\hat{\Lambda}_\beta = \Lambda_0 + O_p(|\beta - \beta_0|) + O_p(n^{-1/3})$ in the $L_2(\mathbb{P})$ space. Finally, in light of the existence of h^* in the proof of Theorem 2, we can define $\mathrm{d}\Lambda_{\theta,\beta}(t) = \{1 + (\theta - \beta)h^*(t)\}\,\mathrm{d}\Lambda(t)$ to obtain the least favourable submodel as $(\theta, \Lambda_{\theta,\beta})$, where θ is in a neighbourhood of β_0. Thus, we can easily verify conditions (8), (9) and (10) of Murphy & van der Vaart (2000) for the likelihood function along this submodel. Along with the Donsker property of the functional classes for the first and second derivatives of $l(\theta, \Lambda_{\theta,\beta})$ with respect to θ and β, we conclude that Theorem 1 of Murphy & van der Vaart (2000) is applicable, and hence the covariance matrix estimator given in §2·3 with $h_n = O(n^{-1/2})$ is consistent for the limiting covariance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

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

Chen, L., Lin, D. Y. & Zeng, D. (2012). Checking semiparametric transformation models with censored data. *Biostatistics* **13**, 18–31.

Gu, M. G., Sun, L. & Zuo, G. (2005). A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Anal.* **11**, 473–88.

Huang, J. (1995). Maximum likelihood estimation for proportional odds regression model with current status data. In *Analysis of Censored Data*, vol. 27 of *IMS Lecture Notes—Monograph Series*, H. L. Koul and J. V. Deshpande, eds. Hayward: Institute of Mathematical Statistics, pp. 129–46.

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

Huang, J. & Rossini, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Statist. Assoc.* **92**, 960–7.

HUANG, J. & WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proc. 1st Seattle Symp. Biostatist.: Survival Anal.*, D. Y. Lin and T. R. Fleming, eds. New York: Springer, pp. 123–69.

MURPHY, S. A. & VAN DER VAART, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

RABINOWITZ, D., BETENSKY, R. A. & TSIATIS, A. A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics* **56**, 511–8.

ROSSINI, A. J. & TSIATIS, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Am. Statist. Assoc.* **91**, 713–21.

SCHICK, A. & YU, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.* **27**, 45–55.

SHEN, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165–77.

SUN, J. & SUN, L. (2005). Semiparametric linear transformation models for current status data. *Can. J. Statist.* **33**, 85–96.

SUN, J. & WEI, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc.* B **62**, 293–302.

VAN DE GEER, S. A. (2000). *Empirical Process Theory and Applications*. Cambridge: Cambridge University Press.

VAN DER VAART, A. W. & WELLNER, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

WANG, L., MCMAHAN, C. S., HUDGENS, M. G. & QURESHI, Z. P. (2015). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72**, 222–31.

WELLNER, J. A. & ZHANG, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35**, 2106–42.

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

ZENG, D. & LIN, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93**, 627–40.

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

ZHANG, Z. & ZHAO, Y. (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *J. Mult. Anal.* **116**, 398–409.

ZHANG, Z., SUN, L., ZHAO, X. & SUN, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Can. J. Statist.* **33**, 61–70.