Maximum likelihood estimation in a partially observed stratified regression model with censored data

Amélie Detais, Jean-François Dupuy

Maximum likelihood estimation in a partially observed stratified regression model with censored data

Detais A. · Dupuy J.-F.

Abstract The stratified proportional intensity model generalizes Cox's proportional intensity model by allowing different groups of the population under study to have distinct baseline intensity functions. In this article, we consider the problem of estimation in this model when the variable indicating the stratum is unobserved for some individuals in the studied sample. In this setting, we construct nonparametric maximum likelihood estimators for the parameters of the stratified model and we establish their consistency and asymptotic normality. Consistent estimators for the limiting variances are also obtained.

Keywords Censored data · Maximum likelihood estimation · Missing data · Stratified intensity model

1 Introduction

This paper considers the problem of estimation in the stratified proportional intensity regression model for survival data, when the stratum information is missing for some sample individuals.

The stratified proportional intensity model (see Andersen *et al.* (1993) or Martinussen and Scheike (2006) for example) generalizes the usual Cox (1972) proportional intensity regression model for survival data, by allowing different groups -the strata- of the population under study to have distinct baseline intensity functions. More precisely, in the stratified model, the strata divide

Detais Amélie
Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 3, 31062 Toulouse, France.
E-mail: detais@math.univ-toulouse.fr

Dupuy Jean-François (corresponding author)
Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 3, 31062 Toulouse, France.
E-mail: dupuy@math.univ-toulouse.fr

the sample individuals into K disjoint groups, each having a distinct baseline intensity function λ_k but a common value for the regression parameter.

The intensity function for the failure time T^0 of an individual in stratum k thus takes the form

$$\lambda_k(t)\exp(\beta' X), \tag{1}$$

where X is a p-vector of covariates, β is a p-vector of unknown regression parameters of interest, and $\{\lambda_k(t) : t \geq 0, k = 1, \ldots, K\}$ are K unknown baseline intensity functions.

A consistent and asymptotically normal estimator of β can be obtained by maximizing the partial likelihood function (Cox 1975). The partial likelihood for the stratified model (1) is the product over strata of the within-stratum partial likelihoods (we refer to Andersen *et al.* (1993) for a detailed treatment of maximum partial likelihood estimation in model (1)). In some applications, it can also be desirable to estimate the cumulative baseline intensity functions $\Lambda_k = \int \lambda_k$. The so-called Breslow (1972) estimators are commonly used for that purpose (see chapter 7 of Andersen *et al.* (1993) for further details on the Breslow estimator and its asymptotic properties).

One major motivation for using the stratified model is that it allows to accomodate in the analysis a predictive categorical covariate whose effect on the intensity is not proportional. To this end, the individuals under study are stratified with respect to the categories of this covariate. In many applications however, this covariate may be missing for some sample individuals (for example, histological stage determination may require biopsy and due to expensiveness, may not be performed on all the study subjects). In this case, the usual statistical inference for model (1), based on the product of within-stratum partial likelihoods, can not be directly applied.

In this work, we consider the problem of estimating β and the $\Lambda_k, k = 1, \ldots, K$ in model (1), when the covariate defining the stratum is missing for some (but not all) individuals. Equivalently said, we consider the problem of estimating model (1) when the stratum information is only partially available.

The problem of estimation in the (unstratified) Cox model $\lambda(t)\exp(\beta' X)$ with missing covariate X has been the subject of intense research over the past decade: see for example Lin and Ying (1993), Paik (1997), Paik and Tsai (1997), Chen and Little (1999), Martinussen (1999), Pons (2002), and the references therein. But to the best of our knowledge and despite its practical relevance, the problem of statistical inference in model (1) with partially available stratum information has not been yet extensively investigated. Recently, Dupuy and Leconte (2009) studied the asymptotic properties of a regression calibration estimator of β in this setting (regression calibration is a general method for handling missing data in regression models, see Carroll *et al.* (1995) for example). The authors proved that this estimator is asymptotically biased, although nevertheless asymptotically normal. No estimators of the cumulative baseline intensity functions were provided.

In this work, we aim at providing an estimator of β that is both consistent and asymptotically normal. Moreover, although the cumulative intensity functions Λ_k are usually not the primary parameters of interest, we also aim at

providing consistent and asymptotically normal estimators of the values $\Lambda_k(t)$, $k = 1, \ldots, K$.

The regression calibration inferential procedure investigated by Dupuy and Leconte (2009) is essentially based on a modified version of the partial likelihood for model (1). In this paper, we propose an alternative method which may be viewed as a fully maximum likelihood approach. Besides assuming that the failure intensity function for an individual in stratum k is given by model (1), we assume that the probability of being in stratum k conditionally on a set of observed covariates W (which may include some components of X) is of the logistic form, depending on some unknown finite-dimensional parameter γ.

A full likelihood for the collected parameter $\theta = (\beta, \gamma, \Lambda_k; k = 1, \ldots, K)$ is constructed from a sample of incompletely observed data. Based on this, we propose to estimate the finite and infinite-dimensional components of θ by using the nonparametric maximum likelihood (NPML) estimation method. We then provide asymptotic results for these estimators, including consistency, asymptotic normality, semiparametric efficiency of the NPML estimator of β, and consistent variance estimation.

Our proofs use some techniques developed by Murphy (1994, 1995) and Parner (1998) to establish the asymptotic theory for the frailty model.

The paper is organized as follows. In Section 2, we describe in greater detail the data structure and the model assumptions. In Section 3, we describe the NPML estimation method for our setting and we establish existence of the NPML estimator of θ. Section 4 establishes the consistency and asymptotic normality of the proposed estimator. Consistent variance estimators are also obtained for both the finite-dimensional parameter estimators and the nonparametric cumulative baseline intensity estimators. We give some concluding remarks in Section 5. Proofs are given in Appendix.

2 Data structure and model assumptions

We describe the notations and model assumptions that will be used throughout the paper.

All the random variables are defined on a probability space $(\Omega, \mathcal{C}, \mathbb{P})$. Let T^0 be a random failure time whose distribution depends on a vector of covariates $X \in \mathbb{R}^p$ and on a stratum indicator $S \in \mathcal{K} = \{1, \ldots, K\}$. We assume that conditionally on X and $S = k$ ($k \in \mathcal{K}$), the intensity function of T^0 is given by model (1). We suppose that T^0 may be right-censored by a positive random variable C and that the analysis is restricted to the time interval $[0, \tau]$, where $\tau < \infty$ denotes the end of the study. Thus we actually observe the potentially censored duration $T = \min\{T^0, \min(C, \tau)\}$ and a censoring indicator $\Delta = 1\{T^0 \leq \min(C, \tau)\}$. If $t \in [0, \tau]$, we denote by $N(t) = 1\{T \leq t\}\Delta$ and $Y(t) = 1\{T \geq t\}$ the failure counting and at-risk processes respectively.

Let $W \in \mathbb{R}^m$ be a vector of surrogate covariates for S (W and X may share some common components). That is, W brings a partial information

about S when S is missing, and it adds no information when S is observed so that the distribution of T^0 conditionally on X, S, and W does not involve the components of W that are not in X. We assume that the conditional probability that an individual belongs to the k-th stratum given his covariate vector W follows a multinomial logistic model:

$$\mathbb{P}(S = k|W) = \frac{\exp(\gamma_k' W)}{\sum_{j=1}^{K} \exp(\gamma_j' W)},$$

where $\gamma_k \in \mathbb{R}^m$ $(k \in \mathcal{K})$. Finally, we let R denote the indicator variable which is 1 if S is observed and 0 otherwise. Then, the data consist of n i.i.d. replicates

$$\mathcal{O}_i = (T_i, \Delta_i, X_i, W_i, R_i, R_i S_i), \quad i = 1, \ldots, n,$$

of $\mathcal{O} = (T, \Delta, X, W, R, RS)$. The data available for the i-th individual are therefore $(T_i, \Delta_i, X_i, W_i, S_i)$ if $R_i = 1$ and $(T_i, \Delta_i, X_i, W_i)$ if $R_i = 0$.

In the sequel, we set $\gamma_K = 0$ for model identifiability purpose and we note $\gamma = (\gamma_1', \ldots, \gamma_{K-1}')' \in (\mathbb{R}^m)^{K-1} \equiv \mathbb{R}^q$. We also note $\pi_{k,\gamma}(W) = \mathbb{P}(S = k|W)$, $k \in \mathcal{K}$. Now, let $\theta = (\beta, \gamma, \Lambda_k; k \in \mathcal{K})$ be the collected parameter and $\theta_0 = (\beta_0, \gamma_0, \Lambda_{k,0}; k \in \mathcal{K})$ denote the true parameter value. Under the true value θ_0, the expectation of random variables will be denoted P_{θ_0}. $\mathbb{P}_n$ will denote the empirical probability measure. In the sequel, the stochastic convergences will be in terms of outer measure.

We now make the following additional assumptions:

(a) The censoring time C is independent of T^0 given (S, X, W), of S given (X, W), and is non-informative. With probability 1, $\mathbb{P}(C \geq T^0 \geq \tau|S, X, W) > c_0$ for some positive constant c_0.

(b) The parameter values β_0 and γ_0 lie in the interior of known compact sets $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ respectively. For every $k \in \mathcal{K}$, the cumulative baseline intensity function $\Lambda_{k,0}$ is a strictly increasing function on $[0, \tau]$ with $\Lambda_{k,0}(0) = 0$ and $\Lambda_{k,0}(\tau) < \infty$. Moreover, for every $k \in \mathcal{K}$, $\Lambda_{k,0}$ is continuously differentiable in $[0, \tau]$, with $\lambda_{k,0}(t) = \partial\Lambda_{k,0}(t)/\partial t$. Let $\mathcal{A}$ denote the set of functions satisfying these properties.

(c) The covariate vectors X and W are bounded (*i.e.* $\|X\| < c_1$ and $\|W\| < c_1$, for some finite positive constant c_1, where $\|\cdot\|$ denotes the Euclidean norm). Moreover, the covariance matrices of X and W are positive definite. Let $c_2 = \min_{\beta \in \mathcal{B}, \|X\| < c_1} e^{\beta' X}$ and $c_3 = \max_{\beta \in \mathcal{B}, \|X\| < c_1} e^{\beta' X}$.

(d) There is a constant $c_4 > 0$ such that for every $k \in \mathcal{K}$, $P_{\theta_0}[1\{S = k\}Y(\tau)R] > c_4$, and the sample size n is large enough to ensure that $\sum_{i=1}^{n} 1\{S_i = k\}Y_i(\tau)R_i > 0$ for every $k \in \mathcal{K}$.

(e) With probability 1, there exists a positive constant c_5 such that for every $k \in \mathcal{K}$, $P_{\theta_0}[\Delta R 1\{S = k\}|T, X, W] > c_5$.

(f) R is independent of S given W, of (T, Δ) given (X, S). The distribution of S conditionally on X and W does not involve the components of X that are not in W. The distributions of R and of the covariate vectors X and W do not depend on the parameter θ.

REMARK. Conditions (b), (c), (d), and (e) are used for identifiability of the parameters and consistency of the proposed estimators. Condition (d) essentially requires that for every stratum k, some subjects are known to belong to k and are still at risk when the study ends. The first assumption in condition (f) states that S is missing at random, which is a fairly general missing data situation (we refer to chapters 6 and 7 in Tsiatis (2006) for a recent exposition of missing data mechanisms).

REMARK. We are now in position to describe our proposed approach to the problem of estimation in model (1) from a sample of incomplete data $\mathcal{O}_i$, $i = 1, \ldots, n$.

Let $\mathcal{S}$ denote the set of subjects with unknown stratum in this sample. The regression calibration method investigated by Dupuy and Leconte (2009) essentially allocates every subject of $\mathcal{S}$ to each of the strata, and estimates β_0 by maximizing a modified version of the partial likelihood for the stratified model, where the contribution of any individual i in $\mathcal{S}$ to the within-k-th-stratum partial likelihood is weighted by an estimate of $\pi_{k,\gamma}(W_i)$ (for every $k \in \mathcal{K}$). The asymptotic bias of the resulting estimator arises from the failure of this method to fully exploit the information carried by $(T_i, \Delta_i, X_i, W_i)$ on the unobserved stratum indicator S_i.

Therefore in this paper, we rather suggest to weight each subject i in $\mathcal{S}$ by an estimate of the conditional probability that subject i belongs to the k-th stratum given the whole observed data $(T_i, \Delta_i, X_i, W_i)$. This suggestion raises two main problems, as is described below.

REMARK. First, we should note that the suggested alternative weights depend on the unknown baseline intensity functions. Therefore, the modified partial likelihood approach considered by Dupuy and Leconte (2009) can not be used to derive an estimator for β_0. Next, the statistics to be involved in the score function for β will depend on the conditional weights and thus, this score will not be expressible as a stochastic integral of some predictable process, as is often the case in models for failure time data. This, in turn, will prevent us from using the counting process martingale theory usually associated with the theoretical developments in failure time models.

To overcome the first problem, we define our estimators from a full likelihood for the whole parameter, that is, for both the finite-dimensional -β (and γ)- and infinite-dimensional -Λ_k, $k \in \mathcal{K}$- components of θ. Empirical process theory (van der Vaart and Wellner 1996) is used to establish asymptotics for the proposed estimators.

3 Maximum likelihood estimation

In the sequel, we assume that there are no ties among the observed death times (this hypothesis is made to simplify notations, but the results below can

be adapted to accomodate ties). The likelihood function for observed data $\mathcal{O}_i$, $i = 1, \ldots, n$ is given by

$$L_n(\theta) = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left\{ \lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i} \Lambda_k(T_i) \right) \pi_{k,\gamma}(W_i) \right\}^{1\{S_i = k\}} \right]^{R_i} \times \left[\sum_{k=1}^{K} \lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i} \Lambda_k(T_i) \right) \pi_{k,\gamma}(W_i) \right]^{1-R_i} \quad (2)$$

It would seem natural to derive a maximum likelihood estimator of θ_0 by maximizing the likelihood (2). However, the maximum of this function over the parameter space $\Theta = \mathcal{B} \times \mathcal{G} \times \mathcal{A}^{\otimes K}$ does not exist. To see this, consider functions Λ_k with fixed values at the T_i, and let $(\partial \Lambda_k(t)/\partial t)|_{t=T_i} = \lambda_k(T_i)$ go to infinity for some T_i with $\Delta_i R_i 1\{S_i = k\} = 1$ or $\Delta_i(1 - R_i) = 1$.

To overcome this problem, we introduce a modified maximization space for (2), by relaxing each $\Lambda_k(\cdot)$ to be an increasing right-continuous step-function on $[0, \tau]$, with jumps at the T_i's such that $\Delta_i R_i 1\{S_i = k\} = 1$ or $\Delta_i(1 - R_i) = 1$. Estimators of $(\beta_0, \gamma_0, \Lambda_{k,0}; k \in \mathcal{K})$ will thus be derived by maximizing a modified version of (2), obtained by replacing $\lambda_k(T_i)$ in (2) with the jump size $\Lambda_k\{T_i\}$ of Λ_k at T_i.

If they exist, these estimators will be referred to as nonparametric maximum likelihood estimators - NPMLEs - (we refer to Zeng and Lin (2007) for a review of the general principle of NPML estimation, with application to various semiparametric regression models for censored data. See also the numerous references therein). In our setting, existence of such estimators is ensured by the following theorem (proof is given in Appendix):

Theorem 1 *Under conditions (a)-(f), the NPMLE* $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\gamma}_n, \widehat{\Lambda}_{k,n}; k \in \mathcal{K})$ *of* θ_0 *exists and is achieved.*

The problem of maximizing L_n over the approximating space described above reduces to a finite dimensional problem, and the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) can be used to calculate the NPMLEs. For $1 \leq i \leq n$ and $k \in \mathcal{K}$, let $w_i(k, \theta)$ be the conditional probability that the i-th individual belongs to the k-th stratum given $(T_i, \Delta_i, X_i, W_i)$ and the parameter value θ, and let $Q(\mathcal{O}_i, k, \theta)$ denote the conditional expectation of $1\{S_i = k\}$ given $\mathcal{O}_i$ and the parameter value θ. Then $Q(\mathcal{O}_i, k, \theta)$ has the form

$$Q(\mathcal{O}_i, k, \theta) = R_i 1\{S_i = k\} + (1 - R_i) w_i(k, \theta).$$

In the M-step of the EM-algorithm, we solve the complete-data score equation conditional on the observed data. In particular, the following expression for the NPMLE of $\Lambda_k(\cdot)$ can be obtained by: (a) taking the derivative with respect to the jump sizes of $\Lambda_k(\cdot)$, of the conditional expectation of the complete-data log-likelihood given the observed data and the NPML estimator, (b) setting this derivative equal to 0:

Lemma 1 *The NPMLE $\widehat{\theta}_n$ satisfies the following equation for every $k \in \mathcal{K}$:*

$$\widehat{\Lambda}_{k,n}(t) = \int_0^t \sum_{i=1}^n \frac{Q(\mathcal{O}_i, k, \widehat{\theta}_n)}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta}_n) \exp(\widehat{\beta}_n' X_j) Y_j(s)} \, dN_i(s), \quad 0 \le t \le \tau.$$

The details of the calculations are omitted (note how the suggested weights $w_i(k, \theta)$ naturally arise from the M-step of the EM algorithm). We refer the interested reader to Zeng and Cai (2005) and Sugimoto and Hamasaki (2006), who recently described EM algorithms for computing NPMLEs in various other semiparametric models with censored data.

In the sequel, we shall denote the conditional expectation of the complete-data log-likelihood given the observed data and the NPML estimator by $E_n[\widetilde{l}_n(\theta)]$.

4 Asymptotic properties

This section states the asymptotic properties of the proposed estimators. We first obtain the following theorem, which states the strong consistency of the proposed NPMLE. The proof is given in Appendix.

Theorem 2 *Under conditions (a)-(f), $\|\widehat{\beta}_n - \beta_0\|$, $\|\widehat{\gamma}_n - \gamma_0\|$, and $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n}(t) - \Lambda_{k,0}(t)|$ (for every $k \in \mathcal{K}$) converge to 0 almost surely as n tends to infinity.*

To derive the asymptotic normality of the proposed estimators, we adapt the function analytic approach developed by Murphy (1995) for the frailty model (see also Chang *et al.* (2005), Kosorok and Song (2007), and Lu (2008), for recent examples of this approach in various other models).

Instead of calculating score equations by differentiating $E_n[\widetilde{l}_n(\theta)]$ with respect to β, γ, and the jump sizes of $\Lambda_k(\cdot)$, we consider one-dimensional submodels $\widehat{\theta}_{n,\eta}$ passing through $\widehat{\theta}_n$ and we differentiate with respect to η. Precisely, we consider submodels of the form

$$\eta \mapsto \widehat{\theta}_{n,\eta} = \left(\widehat{\beta}_n + \eta h_\beta, \widehat{\gamma}_n + \eta h_\gamma, \int_0^{\cdot} \{1 + \eta h_{\Lambda_k}(s)\} d\widehat{\Lambda}_{k,n}(s); k \in \mathcal{K} \right),$$

where h_β and $h_\gamma = (h_{\gamma_1}', \ldots, h_{\gamma_{K-1}}')'$ are p- and q-dimensional vectors respectively ($h_{\gamma_j} \in \mathbb{R}^m$, $j = 1, \ldots, K-1$), and the h_{Λ_k} ($k \in \mathcal{K}$) are functions on $[0, \tau]$. Let $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$. To obtain the score equations, we differentiate $E_n[\widetilde{l}_n(\widehat{\theta}_{n,\eta})]$ with respect to η and we evaluate at $\eta = 0$. $\widehat{\theta}_n$ maximizes $E_n[\widetilde{l}_n(\theta)]$ and therefore satisfies $(\partial E_n[\widetilde{l}_n(\widehat{\theta}_{n,\eta})]/\partial\eta) \mid_{\eta=0} = 0$ for every h, which leads to the score equation $S_n(\widehat{\theta}_n)(h) = 0$ where $S_n(\widehat{\theta}_n)(h)$ takes the form

$$S_n(\widehat{\theta}_n)(h) = \mathbb{P}_n \left[h_\beta' S_\beta(\widehat{\theta}_n) + h_\gamma' S_\gamma(\widehat{\theta}_n) + \sum_{k=1}^K S_{\Lambda_k}(\widehat{\theta}_n)(h_{\Lambda_k}) \right], \qquad (3)$$

where

$$S_\beta(\theta) = \Delta X - \sum_{k=1}^{K} Q(\mathcal{O},k,\theta) X \exp(\beta' X) \Lambda_k(T),$$
$$S_\gamma(\theta) = (S_{\gamma_1}(\theta)', \ldots, S_{\gamma_{K-1}}(\theta)')' \text{ with } S_{\gamma_k}(\theta) = W\left[Q(\mathcal{O},k,\theta) - \pi_{k,\gamma}(W)\right],$$
$$S_{\Lambda_k}(\theta)(h_{\Lambda_k}) = Q(\mathcal{O},k,\theta)\left[h_{\Lambda_k}(T)\Delta - \exp(\beta' X)\int_0^T h_{\Lambda_k}(s)\, d\Lambda_k(s)\right].$$

We take the space of elements $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ to be

$$H = \{(h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K}) : h_\beta \in \mathbb{R}^p, \|h_\beta\| < \infty; h_\gamma \in \mathbb{R}^q, \|h_\gamma\| < \infty;$$
$$h_{\Lambda_k} \text{ is a function defined on } [0,\tau], \|h_{\Lambda_k}\|_v < \infty, k \in \mathcal{K}\},$$

where $\|h_{\Lambda_k}\|_v$ denotes the total variation of h_{Λ_k} on $[0,\tau]$. We further take the functions h_{Λ_k} to be continuous from the right at 0.

Define $\theta(h) = h'_\beta\beta + h'_\gamma\gamma + \sum_{k=1}^K \int_0^\tau h_{\Lambda_k}(s)\, d\Lambda_k(s)$, where $h \in H$. From this, the parameter θ can be considered as a linear functional on H, and the parameter space Θ can be viewed as a subset of the space $l^\infty(H)$ of bounded real-valued functions on H, which we provide with the uniform norm. Moreover, the score operator S_n appears to be a random map from Θ to the space $l^\infty(H)$. Note that appropriate choices of h allow to extract all components of the original parameter θ. For example, letting $h_\gamma = 0$, $h_{\Lambda_k}(\cdot) = 0$ for every $k \in \mathcal{K}$, and h_β be the p-dimensional vector with a one at the i-th location and zeros elsewhere yields the i-th component of β. Letting $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_k}(\cdot) = 0$ for every $k \in \mathcal{K}$ except $h_{\Lambda_j}(s) = 1\{s \le t\}$ (for some $t \in (0,\tau)$) yields $\Lambda_j(t)$.

We need some further notations to state the asymptotic normality of the NPMLE of β_0. Let us first define the linear operator $\sigma = (\sigma_\beta, \sigma_\gamma, \sigma_{\Lambda_k}; k \in \mathcal{K}) : H \to H$ by

$$\begin{aligned}
\sigma_\beta(h) &= P_{\theta_0}[2X\Delta\psi(\mathcal{O},\theta_0)\sum_{k=1}^K Q(\mathcal{O},k,\theta_0)h_{\Lambda_k}(T)] \\
&\quad + P_{\theta_0}\left[\psi(\mathcal{O},\theta_0)X\left\{\psi(\mathcal{O},\theta_0)X'h_\beta + S_\gamma(\theta_0)'h_\gamma\right\}\right], \\
\sigma_\gamma(h) &= P_{\theta_0}[2S_\gamma(\theta_0)\Delta\sum_{k=1}^K Q(\mathcal{O},k,\theta_0)h_{\Lambda_k}(T)] + P_{\theta_0}\left[S_\gamma(\theta_0)S_\gamma(\theta_0)'\right]h_\gamma \\
&\quad + P_{\theta_0}\left[\psi(\mathcal{O},\theta_0)S_\gamma(\theta_0)X'\right]h_\beta, \\
\sigma_{\Lambda_k}(h)(u) &= h_{\Lambda_k}(u)P_{\theta_0}\left[Q(\mathcal{O},k,\theta_0)\phi(u,\mathcal{O},k,\theta_0)\right] \\
&\quad + P_{\theta_0}[2\phi(u,\mathcal{O},k,\theta_0)\sum_{j>k} Q(\mathcal{O},j,\theta_0)\{h_{\Lambda_j}(u) - e^{\beta_0' X}\int_0^u h_{\Lambda_j}\, d\Lambda_{j,0} \\
&\quad - \Delta h_{\Lambda_j}(T) + e^{\beta_0' X}\int_0^T h_{\Lambda_j}\, d\Lambda_{j,0}\}] \\
&\quad - h'_\beta P_{\theta_0}[2X\psi(\mathcal{O},\theta_0)Q(\mathcal{O},k,\theta_0)e^{\beta_0' X}Y(u)] \\
&\quad - h'_\gamma P_{\theta_0}[2S_\gamma(\theta_0)Q(\mathcal{O},k,\theta_0)e^{\beta_0' X}Y(u)],
\end{aligned}$$

where

$$\phi(u, \mathcal{O}, k, \theta_0) = Y(u)Q(\mathcal{O}, k, \theta_0)e^{\beta_0' X}$$

and

$$\psi(\mathcal{O}, \theta_0) = \Delta - \sum_{k=1}^{K} Q(\mathcal{O}, k, \theta_0)e^{\beta_0' X}\Lambda_{k,0}(T).$$

This operator is continuously invertible (Lemma 3 in Appendix). We shall denote its inverse by $\sigma^{-1} = (\sigma_\beta^{-1}, \sigma_\gamma^{-1}, \sigma_{\Lambda_k}^{-1}; k \in \mathcal{K})$.

Next, for every $r \in \mathbb{N}\backslash\{0\}$, the r-dimensional column vector having all its components equal to 0 will be noted by 0_r (or by 0 when no confusion may occur). Let $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K}) \in H$. If $h_\gamma = 0$ and h_{Λ_k} is identically equal to 0 for every $k \in \mathcal{K}$, we note $h = (h_\beta, 0, 0; k \in \mathcal{K})$. Let $\widetilde{\sigma}_\beta^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ be the linear map defined by $\widetilde{\sigma}_\beta^{-1}(u) = \sigma_\beta^{-1}((u, 0, 0; k \in \mathcal{K}))$, for $u \in \mathbb{R}^p$. Let $\{e_1, \ldots, e_p\}$ be the canonical basis of $\mathbb{R}^p$.

Then the following result holds, its proof is given in Appendix.

Theorem 3 *Under conditions (a)-(f), $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ has an asymptotic normal distribution $N(0, \Sigma_\beta)$, where*

$$\Sigma_\beta = (\widetilde{\sigma}_\beta^{-1}(e_1), \ldots, \widetilde{\sigma}_\beta^{-1}(e_p))$$

is the efficient variance in estimating β_0.

REMARK. Although γ_0 and the cumulative baseline intensity functions $\Lambda_{k,0}$ ($k \in \mathcal{K}$) are not the primary parameters of interest, we may also state an asymptotic normality result for their NMPLEs. This requires some further notations.

Define $\widetilde{\sigma}_\gamma^{-1} : \mathbb{R}^q \to \mathbb{R}^q$ by $\widetilde{\sigma}_\gamma^{-1}(u) = \sigma_\gamma^{-1}((0, u, 0; k \in \mathcal{K}))$, let $\{f_1, \ldots, f_q\}$ be the canonical basis of $\mathbb{R}^q$, and define $\Sigma_\gamma = (\widetilde{\sigma}_\gamma^{-1}(f_1), \ldots, \widetilde{\sigma}_\gamma^{-1}(f_q))$. Finally, let $h_{(j,t)} = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ be such that $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_j}(\cdot) = 1\{\cdot \leq t\}$ for some $t \in (0, \tau)$ and $j \in \mathcal{K}$, and $h_{\Lambda_k} = 0$ for every $k \in \mathcal{K}, k \neq j$. Then the following holds (a brief sketch of the proof is given in Appendix):

Theorem 4 *Assume that conditions (a)-(f) hold. Then $\sqrt{n}(\widehat{\gamma}_n - \gamma_0)$ has an asymptotic normal distribution $N(0, \Sigma_\gamma)$. Furthermore, for any $t \in (0, \tau)$ and $j \in \mathcal{K}$, $\sqrt{n}(\widehat{\Lambda}_{j,n}(t) - \Lambda_{j,0}(t))$ is asymptotically distributed as a $N(0, v_j^2(t))$, where*

$$v_j^2(t) = \int_0^t \sigma_{\Lambda_j}^{-1}(h_{(j,t)})(u)\, d\Lambda_{j,0}(u).$$

We now turn to the issue of estimating the asymptotic variances of the estimators $\widehat{\beta}_n$, $\widehat{\gamma}_n$, and $\widehat{\Lambda}_{j,n}(t)$ ($t \in (0, \tau)$, $j \in \mathcal{K}$). It turns out that the asymptotic variances Σ_β, Σ_γ, and $v_j^2(t)$ are not expressible in explicit forms, since

the inverse σ^{-1} has no closed form. However, this is not a problem if we can provide consistent estimators for them. Such estimators are defined below.

For $i = 1, \ldots, n$, let X_{ir} denote the r-th ($r = 1, \ldots, p$) component of X_i, $S_{\gamma,i}(\theta)$ be defined as in (3) with $\mathcal{O}$ and W replaced by $\mathcal{O}_i$ and W_i respectively, and $S_{\gamma,i,s}(\theta)$ be the s-th ($s = 1, \ldots, q$) component of $S_{\gamma,i}(\theta)$. Using these notations, we define the following block matrix

$$\mathbb{A}_n = \begin{pmatrix} A^{\beta\beta} & A^{\beta\gamma} & A^{\beta\Lambda} \\ A^{\gamma\beta} & A^{\gamma\gamma} & A^{\gamma\Lambda} \\ A^{\Lambda\beta} & A^{\Lambda\gamma} & A^{\Lambda\Lambda} \end{pmatrix} \tag{4}$$

where the sub-matrices $A^{\beta\beta}$, $A^{\gamma\gamma}$, $A^{\beta\gamma}$, and $A^{\gamma\beta}$ are defined as follows by their (r, s)-th component:

$$\begin{aligned}
A^{\beta\beta}_{rs} &= \frac{1}{n}\sum_{i=1}^{n}\{\psi(\mathcal{O}_i, \widehat{\theta}_n)\}^2 X_{ir}X_{is}, \quad r, s = 1, \ldots, p, \\
A^{\gamma\gamma}_{rs} &= \frac{1}{n}\sum_{i=1}^{n} S_{\gamma,i,r}(\widehat{\theta}_n)S_{\gamma,i,s}(\widehat{\theta}_n), \quad r, s = 1, \ldots, q, \\
A^{\beta\gamma}_{rs} &= \frac{1}{n}\sum_{i=1}^{n} \psi(\mathcal{O}_i, \widehat{\theta}_n)X_{ir}S_{\gamma,i,s}(\widehat{\theta}_n), \quad r = 1, \ldots, p, \quad s = 1, \ldots, q, \\
A^{\gamma\beta}_{rs} &= A^{\beta\gamma}_{sr}, \quad r = 1, \ldots, q, \quad s = 1, \ldots, p.
\end{aligned}$$

Define the block matrices $A^{\beta\Lambda} = (A^{\beta\Lambda_1}, \ldots, A^{\beta\Lambda_K})$ and $A^{\gamma\Lambda} = (A^{\gamma\Lambda_1}, \ldots, A^{\gamma\Lambda_K})$, where for every $k \in \mathcal{K}$, the sub-matrices $A^{\beta\Lambda_k}$ and $A^{\gamma\Lambda_k}$ are defined by

$$\begin{aligned}
A^{\beta\Lambda_k}_{rs} &= \frac{2}{n}X_{sr}\Delta_s\psi(\mathcal{O}_s, \widehat{\theta}_n)Q(\mathcal{O}_s, k, \widehat{\theta}_n), \quad r = 1, \ldots, p, \quad s = 1, \ldots, n, \\
A^{\gamma\Lambda_k}_{rs} &= \frac{2}{n}S_{\gamma,s,r}(\widehat{\theta}_n)\Delta_s Q(\mathcal{O}_s, k, \widehat{\theta}_n), \quad r = 1, \ldots, q, \quad s = 1, \ldots, n.
\end{aligned}$$

Define also the block matrices

$$A^{\Lambda\beta} = \begin{pmatrix} A^{\Lambda_1\beta} \\ \vdots \\ A^{\Lambda_K\beta} \end{pmatrix} \qquad A^{\Lambda\gamma} = \begin{pmatrix} A^{\Lambda_1\gamma} \\ \vdots \\ A^{\Lambda_K\gamma} \end{pmatrix} \qquad A^{\Lambda\Lambda} = \begin{pmatrix} A^{\Lambda_1\Lambda_1} & \ldots & A^{\Lambda_1\Lambda_K} \\ \vdots & & \vdots \\ A^{\Lambda_K\Lambda_1} & \ldots & A^{\Lambda_K\Lambda_K} \end{pmatrix}$$

where for every $j,k \in \mathcal{K}$,

$$A_{rs}^{\Lambda_k\beta} = -\frac{1}{n}\sum_{i=1}^{n} 2X_{is}\psi(\mathcal{O}_i,\widehat{\theta}_n)Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta}_n' X_i}Y_i(T_r), \quad r=1,\ldots,n, \quad s=1,\ldots,p,$$

$$A_{rs}^{\Lambda_k\gamma} = -\frac{1}{n}\sum_{i=1}^{n} 2S_{\gamma,i,s}(\widehat{\theta}_n)Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta}_n' X_i}Y_i(T_r), \quad r=1,\ldots,n, \quad s=1,\ldots,q,$$

$$\begin{aligned} A_{rs}^{\Lambda_k\Lambda_j} &= 1\{j=k\}1\{r=s\}\frac{1}{n}\sum_{i=1}^{n} Q(\mathcal{O}_i,k,\widehat{\theta}_n)\phi(T_r,\mathcal{O}_i,k,\widehat{\theta}_n) \\ &\quad +1\{j>k\}(1\{r=s\}\frac{1}{n}\sum_{i=1}^{n} 2\phi(T_s,\mathcal{O}_i,k,\widehat{\theta}_n)Q(\mathcal{O}_i,j,\widehat{\theta}_n) \\ &\quad +\frac{2}{n}\sum_{i=1}^{n}\phi(T_r,\mathcal{O}_i,k,\widehat{\theta}_n)Q(\mathcal{O}_i,j,\widehat{\theta}_n)e^{\widehat{\beta}_n' X_i}\widehat{\Delta\Lambda_{j,n}}(T_s)\{1\{T_s\le T_i\}-1\{T_s\le T_r\}\} \\ &\quad -\frac{2}{n}\phi(T_r,\mathcal{O}_s,k,\widehat{\theta}_n)Q(\mathcal{O}_s,j,\widehat{\theta}_n)\Delta_s), \quad r,s=1,\ldots,n, \end{aligned}$$

and $\widehat{\Delta\Lambda_{j,n}}(T_s)$ is the jump size of $\widehat{\Lambda}_{j,n}$ at T_s that is, $\widehat{\Delta\Lambda_{j,n}}(T_s) = \widehat{\Lambda}_{j,n}(T_s) - \widehat{\Lambda}_{j,n}(T_s-)$ ($j\in\mathcal{K}, s=1,\ldots,n$). Note that for notational simplicity, the lower (sample size) indice n has been omitted in the notations for the sub-matrices of $\mathbb{A}_n$.

Now, define

$$\begin{aligned} \widehat{\Sigma}_{\beta,n} &= \big\{A^{\beta\beta} - A^{\beta\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\beta} - (A^{\beta\Lambda} - A^{\beta\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\Lambda}) \\ &\qquad \times(A^{\Lambda\Lambda} - A^{\Lambda\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\Lambda})^{-1}(A^{\Lambda\beta} - A^{\Lambda\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\beta})\big\}^{-1}, \\ \widehat{\Sigma}_{\gamma,n} &= \big\{A^{\gamma\gamma} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma} - (A^{\gamma\Lambda} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda}) \\ &\qquad \times(A^{\Lambda\Lambda} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda})^{-1}(A^{\Lambda\gamma} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma})\big\}^{-1}, \end{aligned}$$

and

$$\begin{aligned} \widehat{\Sigma}_{\Lambda,n} &= \big\{A^{\Lambda\Lambda} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda} - (A^{\Lambda\gamma} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma}) \\ &\qquad \times(A^{\gamma\gamma} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma})^{-1}(A^{\gamma\Lambda} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda})\big\}^{-1}. \end{aligned}$$

Then the following holds:

Theorem 5 *Under conditions (a)-(f), $\widehat{\Sigma}_{\beta,n}$ and $\widehat{\Sigma}_{\gamma,n}$ converge in probability to Σ_β and Σ_γ respectively as n tends to ∞. Moreover, for $t\in(0,\tau)$ and $j\in\mathcal{K}$, let*

$$\widehat{v}_{j,n}^2(t) = \widehat{\Xi}_{(j,t)}^{n'}\widehat{\Sigma}_{\Lambda,n}U_{(j,t)}^n,$$

where

$$\widehat{\Xi}_{(j,t)}^n = \Big(0'_{(j-1)n}, \widehat{\Delta\Lambda_{j,n}}(T_1)1\{T_1\le t\},\ldots,\widehat{\Delta\Lambda_{j,n}}(T_n)1\{T_n\le t\}, 0'_{(K-j)n}\Big)'$$

and

$$U_{(j,t)}^n = (0'_{(j-1)n}, 1\{T_1\le t\},\ldots,1\{T_n\le t\}, 0'_{(K-j)n})'.$$

Then $\widehat{v}_{j,n}^2(t)$ converges in probability to $v_j^2(t)$ as n tends to ∞.

5 Discussion

In this paper, we have constructed consistent and asymptotically normal estimators for the stratified proportional intensity regression model when the sample stratum information is only partially available. The proposed estimator for the regression parameter of interest in this model has been shown to be semiparametrically efficient. Although computationally more challenging, these estimators improve the ones previously investigated in the literature, such as the regression calibration estimators (Dupuy and Leconte 2009).

We have obtained explicit (and computationally fairly simple) formulas for consistent estimators of the asymptotic variances. These formulas may however require the inversion of potentially large matrices. For a large sample, this inversion may be unstable. An alternative solution relies on numerical differentiation of the profile log-likelihood (see Murphy *et al.* (1997) and Chen and Little (1999) for example). Note that in this latter method however, no estimator is available for the asymptotic variance of the cumulative baseline intensity estimator. Some further work is needed to evaluate the numerical performance of the proposed estimators. This is the subject for future research, and requires some extensive simulation work which falls beyond the scope of this paper.

In this paper, a multinomial logistic model (Jobson 1992) is used for modeling the conditional stratum probabilities given covariates. This choice was mainly motivated by the fact that this model is commonly used in medical research for modeling the relationship between a categorical response and covariates. The theoretical results developed here can be extended to the case of other link functions. In addition, the covariate X in model (1) is assumed to be time independent, for convenience. This assumption can be relaxed to accomodate time varying covariates, provided that appropriate regularity conditions are made.

Appendix A. Proofs of Theorems

A.1 Proof of Theorem 1

For every $k \in \mathcal{K}$, define $\mathcal{I}_k^n = \{i \in \{1,\dots,n\} | \Delta_i R_i 1\{S_i = k\} = 1 \text{ or } \Delta_i(1 - R_i) = 1\}$, and let i_k^n denote the cardinality of $\mathcal{I}_k^n$. Let $i_\bullet^n = \sum_{k=1}^K i_k^n$. Consider the set of times $\{T_i, i \in \mathcal{I}_k^n\}$. Let $t_{(k,1)} < \ldots < t_{(k,i_k^n)}$ denote the ordered failure times in this set. For any given sample size n, the NPML estimation method consists in maximizing L_n in (2) over the approximating parameter space

$$\Theta_n = \left\{(\beta, \gamma, \Lambda_k\{t_{(k,j)}\}) : \beta \in \mathcal{B}; \gamma \in \mathcal{G}; \Lambda_k\{t_{(k,j)}\} \in [0,\infty), j = 1,\dots,i_k^n, k \in \mathcal{K}\right\}.$$

Suppose first that $\Lambda_k\{t_{(k,j)}\} \leq M < \infty$, for $j = 1,\dots,i_k^n$ and $k \in \mathcal{K}$. Since L_n is a continuous function of β, γ, and the $\Lambda_k\{t_{(k,j)}\}$'s on the compact set $\mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n}$, L_n achieves its maximum on this set.

To show that a maximum of L_n exists on $\mathcal{B}\times\mathcal{G}\times[0,\infty)^{i_\bullet^n}$, we show that there exists a finite M such that for all $\theta^M=(\beta^M,\gamma^M,(\Lambda_k^M\{t_{(k,j)}\})_{j,k})\in(\mathcal{B}\times\mathcal{G}\times[0,\infty)^{i_\bullet^n})\backslash(\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n})$, there exists a $\theta=(\beta,\gamma,(\Lambda_k\{t_{(k,j)}\})_{j,k})\in\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n}$ such that $L_n(\theta)>L_n(\theta^M)$. A proof by contradiction is adopted for that purpose.

Assume that for all $M<\infty$, there exists $\theta^M\in(\mathcal{B}\times\mathcal{G}\times[0,\infty)^{i_\bullet^n})\backslash(\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n})$ such that for all $\theta\in\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n}$, $L_n(\theta)\leq L_n(\theta^M)$. It can be seen that L_n is bounded above by

$$K^n\prod_{i=1}^n\left[\prod_{k=1}^K\{c_3\Lambda_k\{T_i\}\}^{\Delta_i R_i 1\{S_i=k\}}\exp\left(-c_2R_i1\{S_i=k\}\sum_{j=1}^{i_k^n}\Lambda_k\{t_{(k,j)}\}1\{t_{(k,j)}\leq T_i\}\right)\right].$$

If $\theta^M\in(\mathcal{B}\times\mathcal{G}\times[0,\infty)^{i_\bullet^n})\backslash(\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n})$, then there exists $l\in\mathcal{K}$ and $p\in\{1,\ldots,i_l^n\}$ such that $\Lambda_l^M\{t_{(l,p)}\}>M$. By assumption (d), there exists at least one individual with indice i_M ($i_M\in\{1,\ldots,n\}$) such that $1\{S_{i_M}=l\}=1$, $Y_{i_M}(\tau)=1$ (and therefore $t_{(l,p)}\leq T_{i_M}=\tau$), and $R_{i_M}=1$. Hence

$$R_{i_M}1\{S_{i_M}=l\}\sum_{j=1}^{i_l^n}\Lambda_l^M\{t_{(l,j)}\}1\{t_{(l,j)}\leq T_{i_M}\}\to\infty\text{ as }M\to\infty.$$

It follows that the upper bound of $L_n(\theta^M)$ (and therefore $L_n(\theta^M)$ itself) can be made as close to 0 as desired by increasing M. This is the desired contradiction.

□

A.2 Proof of Theorem 2

We adapt the techniques developed by Murphy (1994), in order to prove consistency of our proposed estimator $\widehat{\theta}_n$. The proof essentially consists of three steps: (i) for every $k\in\mathcal{K}$, we show that the sequence $\widehat{\Lambda}_{k,n}(\tau)$ is almost surely bounded as n goes to infinity, (ii) we show that every subsequence of n contains a further subsequence along which the NPMLE $\widehat{\theta}_n$ converges, (iii) we show that the limit of every convergent subsequence of $\widehat{\theta}_n$ is θ_0.

Proof of (i). Note first that for all $s\in[0,\tau]$ and $k\in\mathcal{K}$, $\frac{1}{n}\sum_{i=1}^n Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta}_n'X_i}Y_i(s)\geq c_2\frac{1}{n}\sum_{i=1}^n R_i1\{S_i=k\}Y_i(\tau)$. Moreover, $Q(\mathcal{O}_i,k,\widehat{\theta}_n)$ is bounded by 1. It follows that for all $k\in\mathcal{K}$,

$$0\leq\widehat{\Lambda}_{k,n}(\tau)\leq\frac{1}{c_2}\int_0^\tau\frac{d\bar{N}_n(s)}{\frac{1}{n}\sum_{i=1}^n R_i1\{S_i=k\}Y_i(\tau)}=\frac{\frac{1}{n}\sum_{i=1}^n\Delta_i}{c_2\frac{1}{n}\sum_{i=1}^n R_i1\{S_i=k\}Y_i(\tau)},$$

where $\bar{N}_n(s)=n^{-1}\sum_{i=1}^n N_i(s)$. Next, $\frac{1}{n}\sum_{i=1}^n R_i1\{S_i=k\}Y_i(\tau)$ converges almost surely to $P_{\theta_0}[R1\{S=k\}Y(\tau)]>c_4>0$ therefore, for each $k\in\mathcal{K}$, as n goes to infinity, $\widehat{\Lambda}_{k,n}(\tau)$ is bounded above almost surely by $\frac{1}{c_2c_4}$.

Proof of (ii). If (i) holds, by Helly's theorem (see Loève (1963), p179), every subsequence of n has a further subsequence along which $\widehat{\Lambda}_{1,n}$ converges weakly to some nondecreasing right-continuous function Λ_1^*, with probability 1. By successive extractions of sub-subsequences, we can further find a subsequence (say n_j) such that $\widehat{\Lambda}_{k,n_j}$ converges weakly to some nondecreasing right-continuous function Λ_k^*, for every $k \in \mathcal{K}$, with probability 1. By the compactness of $\mathcal{B}\times\mathcal{G}$, we can further find a subsequence of n_j (we shall still denote it by n_j for simplicity of notations) such that $\widehat{\Lambda}_{k,n_j}$ converges weakly to Λ_k^* (for every $k \in \mathcal{K}$) and $(\widehat{\beta}_{n_j}, \widehat{\gamma}_{n_j})$ converges to some (β^*, γ^*), with probability 1. We now show that the Λ_k^*'s must be continuous on $[0,\tau]$.

Let ψ be any nonnegative, bounded, continuous function. Then, for any given $k \in \mathcal{K}$,

$$\begin{aligned}\int_0^\tau \psi(s)\, d\Lambda_k^*(s) &= \int_0^\tau \psi(s)\, d\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} \\ &+ \int_0^\tau \psi(s) \left[\frac{1}{n_j}\sum_{l=1}^{n_j} Q(\mathcal{O}_l, k, \widehat{\theta}_{n_j}) e^{\widehat{\beta}'_{n_j} X_l} Y_l(s)\right]^{-1} \frac{1}{n_j}\sum_{i=1}^{n_j} Q(\mathcal{O}_i, k, \widehat{\theta}_{n_j})\, dN_i(s) \\ &\leq \int_0^\tau \psi(s)\, d\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} + \int_0^\tau \psi(s) \left[\frac{c_2}{n_j}\sum_{l=1}^{n_j} R_l 1\{S_l = k\} Y_l(s)\right]^{-1} d\bar{N}_{n_j}(s).\end{aligned}$$

By the Helly-Bray Lemma (see Loève (1963), p180), $\int_0^\tau \psi(s)\, d\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} \to 0$ as $j \to \infty$. Moreover, $\bar{N}_{n_j}(\cdot)$ and $\frac{1}{n_j}\sum_{l=1}^{n_j} R_l 1\{S_l = k\} Y_l(\cdot)$ converge almost surely in supremum norm to

$$\sum_{k=1}^{K} \int_0^{\cdot} P_{\theta_0}\left[1\{S = k\} e^{\beta_0' X} Y(s)\right] d\Lambda_{k,0}(s) \text{ and } P_{\theta_0}\left[R 1\{S = k\} Y(\cdot)\right]$$

respectively, where the latter term is bounded away from 0 on $s \in [0,\tau]$ by assumption (d). Thus, by applying the extended version of the Helly-Bray Lemma (stated by Korsholm (1998) for example) to the second term on the right-hand side of the previous inequality, we get that

$$\begin{aligned}&\int_0^\tau \psi(s)\, d\Lambda_k^*(s) \qquad (5)\\ &\quad\leq c_2 \int_0^\tau \psi(s) \left\{P_{\theta_0}\left[R 1\{S = k\} Y(s)\right]\right\}^{-1} \sum_{k=1}^{K} P_{\theta_0}[1\{S = k\} e^{\beta_0' X} Y(s)] \lambda_{k,0}(s)\, ds. \\ &\quad\leq \frac{c_2 c_3}{c_4} \sum_{k=1}^{K} \int_0^\tau \psi(s) \lambda_{k,0}(s)\, ds.\end{aligned}$$

Suppose that Λ_k^* has discontinuities, and let ψ be close to 0 except at the jump points of Λ_k^*, where it is allowed to have high and thin peaks. While the right-hand side of inequality (5) should be close to 0 ($\lambda_{k,0}$ is continuous by assumption (b)), its left-hand side can be made arbitrarily large, yielding

a contradiction. Thus Λ_k^* must be continuous ($k \in \mathcal{K}$). A second conclusion, arising from Dini's theorem, is that $\widehat{\Lambda}_{k,n_j}$ uniformly converges to Λ_k^* ($k \in \mathcal{K}$), with probability 1. To summarize: for any given subsequence of n, we have found a further subsequence n_j and an element $(\beta^*, \gamma^*, \Lambda_k^*, k \in \mathcal{K})$ such that $\|\widehat{\beta}_{n_j} - \beta^*\|$, $\|\widehat{\gamma}_{n_j} - \gamma^*\|$, and $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n_j}(t) - \Lambda_k^*(t)|$ (for every $k \in \mathcal{K}$) converge to 0 almost surely.

Proof of (iii). To prove (iii), we first define random step functions

$$\overline{\Lambda}_{k,n}(t) = \int_0^t \sum_{i=1}^n \frac{Q(\mathcal{O}_i, k, \theta_0)}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \theta_0) \exp(\beta_0' X_j) Y_j(s)}\, dN_i(s),\ 0 \le t \le \tau, k \in \mathcal{K},$$

and we show that for every $k \in \mathcal{K}$, $\overline{\Lambda}_{k,n}$ almost surely uniformly converges to $\Lambda_{k,0}$ on $[0,\tau]$. First, note that

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \left| \overline{\Lambda}_{k,n}(t) - P_{\theta_0}\left[\frac{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \,|_{s=T}} \right] \right| \\
&\quad \le \sup_{t\in[0,\tau]} \left| \frac{1}{n} \sum_{i=1}^n \Delta_i 1\{T_i \le t\} Q(\mathcal{O}_i, k, \theta_0) \right. \\
&\qquad \left. \times \left\{ \frac{1}{\mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)\right]} - \frac{1}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right]} \right\}_{|_{s=T_i}} \right| \\
&\qquad + \sup_{t\in[0,\tau]} \left| (\mathbb{P}_n - P_{\theta_0}) \left[\frac{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \,|_{s=T}} \right] \right| \\
&\quad \le \sup_{s\in[0,\tau]} \left| \frac{1}{\mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)\right]} - \frac{1}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right]} \right| \\
&\qquad + \sup_{t\in[0,\tau]} \left| (\mathbb{P}_n - P_{\theta_0}) \left[\frac{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \,|_{s=T}} \right] \right| \qquad (6)
\end{aligned}$$

The class $\{Y(s) : s \in [0,\tau]\}$ is Donsker and $Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X}$ is a bounded measurable function, hence $\{Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s) : s \in [0,\tau]\}$ is Donsker (Corollary 9.31, Kosorok (2008)), and therefore Glivenko-Cantelli. Moreover, $P_{\theta_0}[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)] = P_{\theta_0}[P_{\theta_0}[1\{S=k\}|\mathcal{O}] e^{\beta_0' X} Y(s)] = P_{\theta_0}[1\{S = k\} e^{\beta_0' X} Y(s)]$. Thus

$$\sup_{s\in[0,\tau]} \left| \mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s) \right] - P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s) \right] \right|$$

converges to 0 a.e. Next, $P_{\theta_0}[1\{S=k\} e^{\beta_0' X} Y(s)]$ is larger than $c_2 . P_{\theta_0}[1\{S = k\} Y(\tau)]$ and thus, by assumption (d), $P_{\theta_0}[1\{S=k\} e^{\beta_0' X} Y(s)] > 0$. It follows that the first term on the right-hand side of inequality (6) converges to 0 a.e.. Similar aguments show that the class $\{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0) / P_{\theta_0}[1\{S =$

$k\}e^{\beta_0' X} Y(s)] \mid_{s=T} : t \in [0,\tau]\}$ is also a Glivenko-Cantelli class, and therefore $\overline{\Lambda}_{k,n}$ almost surely uniformly converges to

$$P_{\theta_0}\left[\frac{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \mid_{s=T}}\right].$$

Now, note that $\Lambda_{k,0}(t) = \int_0^t \frac{P_{\theta_0}[1\{S=k\}\, dN(s)]}{P_{\theta_0}[1\{S=k\} e^{\beta_0' X} Y(s)]}$, which can be reexpressed as

$$\Lambda_{k,0}(t) = \frac{P_{\theta_0}\left[1\{S=k\}\Delta 1\{T \le t\}\right]}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \mid_{s=T}} = P_{\theta_0}\left[\frac{\Delta 1\{T \le t\} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[1\{S=k\} e^{\beta_0' X} Y(s)\right] \mid_{s=T}}\right].$$

Thus $\overline{\Lambda}_{k,n}$ almost surely uniformly converges to $\Lambda_{k,0}$ on $[0,\tau]$.

Next, using somewhat standard arguments (see Parner (1998) for example), we can show that $0 \le n_j^{-1}\{\log L_{n_j}(\widehat{\theta}_{n_j}) - \log L_{n_j}(\overline{\theta}_{n_j})\}$ converges to the negative Kullback-Leibler information $P_{\theta_0}[\log(L_1(\theta^*)/L_1(\theta_0))]$. Thus, the Kullback-Leibler information must be zero, and it follows that with probability 1, $L_1(\theta^*) = L_1(\theta_0)$. The proof of consistency is completed if we show that this equality implies $\theta^* = \theta_0$. For that purpose, consider $L_1(\theta^*) = L_1(\theta_0)$ under $\Delta = 1$, $R = 1$, and $1\{S=k\} = 1$ (for each $k \in \mathcal{K}$ in turn). Note that this is possible by assumption (e). This yields the following equation for almost all $t \in [0,\tau], \|x\| < c_1, \|w\| < c_1$:

$$\log \frac{\lambda_k^*(t)}{\lambda_{k,0}(t)} + (\beta^* - \beta_0)' x - \Lambda_k^*(t) e^{\beta^{*\prime} x} + \Lambda_{k,0}(t) e^{\beta_0' x} + \log \frac{\pi_k^*(w)}{\pi_{k,0}(w)} = 0.$$

This equation is analogous to equation (A.2) in Chen and Little (1999). The rest of the proof of identifiability thus proceeds along the same lines as the proof of Lemma A.1.1 in Chen and Little (1999), and is omitted.

Hence, for any given subsequence of n, we have found a further subsequence n_j such that $\|\widehat{\beta}_{n_j} - \beta_0\|$, $\|\widehat{\gamma}_{n_j} - \gamma_0\|$, and $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n_j}(t) - \Lambda_{k,0}(t)|$ (for every $k \in \mathcal{K}$) converge to 0 almost surely, which implies that the sequence of NPMLE $\widehat{\theta}_n$ converges almost surely to θ_0.

□

A.3 Proof of Theorem 3

The proof of Theorem 3 uses similar arguments as the proof of Theorem 3 of Fang *et al.* (2005), so we only highlight the parts that are different. We need a few lemmas before presenting the proof.

Lemma 2 *Let $h \in H$. Then the following holds: $P_{\theta_0}[S_1(\theta_0)(h)] = P_{\theta_0}[h_\beta' S_\beta(\theta_0) + h_\gamma' S_\gamma(\theta_0) + \sum_{k=1}^K S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})] = 0$.*

Proof. From the properties of the conditional expectation, we first note that

$$\begin{aligned} P_{\theta_0}\left[S_\beta(\theta_0)\right] &= P_{\theta_0}\left[\Delta X - \sum_{k=1}^{K} Q(\mathcal{O},k,\theta_0) X \exp(\beta_0' X)\Lambda_{k,0}(T)\right] \\ &= P_{\theta_0}\left[\Delta X - \sum_{k=1}^{K} 1\{S=k\} X \exp(\beta_0' X)\Lambda_{k,0}(T)\right] \\ &= P_{\theta_0}\left[XM(\tau)\right], \end{aligned}$$

where $M(t) = N(t) - \int_0^t \sum_{k=1}^{K} 1\{S = k\} e^{\beta_0' X} Y(u)\, d\Lambda_{k,0}(u)$ is the counting process martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N(u), 1\{C \leq u\}, X, S, W : 0 \leq u \leq t\}$. X is bounded and $\mathcal{F}_t$-measurable, hence it follows that $P_{\theta_0}[S_\beta(\theta_0)] = 0$. Using similar arguments, we can verify that $P_{\theta_0}[S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})] = 0$, $k \in \mathcal{K}$. Finally, for $k = 1, \ldots, K-1$,

$$\begin{aligned} P_{\theta_0}\left[S_{\gamma_k}(\theta_0)\right] &= P_{\theta_0}\left[W\left[Q(\mathcal{O},k,\theta_0) - \pi_{k,\gamma_0}(W)\right]\right] \\ &= P_{\theta_0}\left[W P_{\theta_0}\left[1\{S=k\} - \pi_{k,\gamma_0}(W)|W\right]\right] \\ &= 0. \end{aligned}$$

Combining these results yields that $P_{\theta_0}\left[S_1(\theta_0)(h)\right] = 0$.

□

We now come to the continuous invertibility of the continuous linear operator σ defined in Section 4.

Lemma 3 *The operator σ is continuously invertible.*

Proof. Since H is a Banach space, to prove that σ is continuously invertible, it is sufficient to prove that σ is one-to-one and that it can be written as the sum of a bounded linear operator with a bounded inverse and a compact operator (Lemma 25.93 of van der Vaart (1998)).

Define the linear operator $A(h) = \left(h_\beta, h_\gamma, P_{\theta_0}\left[1\{S=k\}\phi(\cdot,\mathcal{O},k,\theta_0)\right] h_{\Lambda_k}(\cdot); k \in \mathcal{K}\right)$, this is a bounded operator due to the boundedness of X. Moreover, for all $u \in [0,\tau]$ and $k \in \mathcal{K}$, $P_{\theta_0}\left[1\{S=k\}\phi(u,\mathcal{O},k,\theta_0)\right] \geq c_2 c_4 > 0$ by assumptions (c) and (d). This implies that A is invertible with bounded inverse $A^{-1}(h) = \left(h_\beta, h_\gamma, P_{\theta_0}\left[1\{S=k\}\phi(\cdot,\mathcal{O},k,\theta_0)\right]^{-1} h_{\Lambda_k}(\cdot); k \in \mathcal{K}\right)$. The operator $\sigma - A$ can be shown to be compact by using the same techniques as in Lu (2008) for example.

To prove that σ is one-to-one, let $h \in H$ such that $\sigma(h) = 0$. If $\sigma(h) = 0$, $P_{\theta_0}\left[S_1(\theta_0)(h)^2\right] = 0$, and therefore $S_1(\theta_0)(h) = 0$ almost surely. Let $j \in \mathcal{K}$. By assumption (e), for almost every $t \in [0,\tau]$, $\|x\| \leq c_1$, and $\|w\| \leq c_1$, there is a non-negligible set $\Omega_{t,x,w} \subseteq \Omega$ such that $\Delta(\omega) = 1$, $R(\omega) = 1$, and $1\{S(\omega) = j\} = 1$ when $\omega \in \Omega_{t,x,w}$. If $S_1(\theta_0)(h) = 0$ almost surely, then in particular, for almost every $t \in [0,\tau]$, $\|x\| \leq c_1$, and $\|w\| \leq c_1$, $S_1(\theta_0)(h) = 0$ when

$\omega \in \Omega_{t,x,w}$, which yields the following equation:

$$h_{\Lambda_j}(t) + h'_\beta x + w' h_{\gamma_j} - \sum_{k=1}^{K-1} w' h_{\gamma_k} \pi_{k,\gamma_0}(w) \\ -e^{\beta'_0 x} \left[\int_0^t h_{\Lambda_j}(s)\, d\Lambda_{j,0}(s) + h'_\beta x \Lambda_{j,0}(t) \right] = 0, \qquad (7)$$

with $h_{\gamma_j} = 0$ when $j = K$. Then, by choosing t arbitrarily close to 0, and since $\Lambda_{j,0}$ is continuous, $\Lambda_{j,0}(0) = 0$, and h_{Λ_j} is continuous from the right at 0, we get that

$$h_{\Lambda_j}(0) + h'_\beta x + w' h_{\gamma_j} - \sum_{k=1}^{K-1} w' h_{\gamma_k} \pi_{k,\gamma_0}(w) = 0. \qquad (8)$$

Taking the difference (7)-(8) yields that

$$h_{\Lambda_j}(t) - h_{\Lambda_j}(0) = e^{\beta'_0 x} \left[\int_0^t h_{\Lambda_j}(s)\, d\Lambda_{j,0}(s) + h'_\beta x \Lambda_{j,0}(t) \right] \qquad (9)$$

for almost every $t \in [0, \tau]$ and $\|x\| \leq c_1$. Since $\Lambda_{j,0}$ is increasing (by assumption (b)), for every $t > 0$, $\Lambda_{j,0}(t) > \Lambda_{j,0}(0) = 0$ and therefore (9) can be rewritten as

$$\frac{h_{\Lambda_j}(t) - h_{\Lambda_j}(0)}{\Lambda_{j,0}(t)} = e^{\beta'_0 x} \left[r(t) + h'_\beta x \right], \qquad (10)$$

where $r(t) = \int_0^t h_{\Lambda_j}(s)\, d\Lambda_{j,0}(s)/\Lambda_{j,0}(t)$. Consider first the case where $\beta_0 = 0$. Since the left-hand side of (10) does not depend on x, h_β must equal 0. Next, consider the case where $\beta_0 \neq 0$. Let $t_1, t_2 > 0$. Then $e^{\beta'_0 x}[r(t_1) - r(t_2)]$ does not depend on x. Since the covariance matrix of X is positive definite, we can find two distinct values x_1 and x_2 of X such that $e^{\beta'_0 x_1}[r(t_1) - r(t_2)] = e^{\beta'_0 x_2}[r(t_1) - r(t_2)]$. This implies that $r(t_1) = r(t_2)$, from which we deduce that $h_{\Lambda_j}(t)$ has to be constant (say, equal to α) for almost every $t \in (0, \tau]$. From (10), we then deduce that $h_{\Lambda_j}(0) = \alpha$, which further implies that $h_\beta = 0$, $\alpha = 0$, and thus $h_{\Lambda_j}(t) = 0$ for almost every $t \in [0, \tau]$ ($j \in \mathcal{K}$). This, together with (8) implies that $h_{\gamma_j} = 0$, $j \in \mathcal{K}$.

Let $k = K$. Then $\sigma_{\Lambda_K}(h)(u) = P_{\theta_0}\left[1\{S = K\}\phi(u, \mathcal{O}, K, \theta_0)\right] h_{\Lambda_K}(u) = 0$ for all $u \in [0, \tau]$ since $h_\beta = 0$ and $h_\gamma = 0$. By assumptions (c) and (d), for every $u \in [0, \tau]$ and $k \in \mathcal{K}$,

$$P_{\theta_0}\left[1\{S = k\}\phi(u, \mathcal{O}, k, \theta_0)\right] = P_{\theta_0}\left[1\{S = k\}Y(u)Q(\mathcal{O}, k, \theta_0)e^{\beta'_0 X}\right] \\ \geq P_{\theta_0}\left[1\{S = k\}Y(\tau)Re^{\beta'_0 X}\right] > 0,$$

hence we conclude that h_{Λ_K} is identically equal to 0 on $[0, \tau]$. Next, considering $\sigma_{\Lambda_{K-1}}(h)(u) = 0$ with $h_\beta = 0$, $h_\gamma = 0$ and $h_{\Lambda_K} = 0$, we conclude similarly that $h_{\Lambda_{K-1}}(u) = 0$ for every $u \in [0, \tau]$. It follows that h_{Λ_j} is identically equal

to 0 on $[0,\tau]$ for every $j \in \mathcal{K}$. Therefore, σ is one-to-one.

□

We now turn to the proof of Theorem 3 itself. Similar to Fang *et al.* (2005), we get that

$$\sqrt{n}\left(h'_\beta(\widehat{\beta}_n-\beta_0)+h'_\gamma(\widehat{\gamma}_n-\gamma_0)+\sum_{k=1}^{K}\int_0^\tau h_{\Lambda_k}(s)\,d(\widehat{\Lambda}_{k,n}-\Lambda_{k,0})(s)\right)$$
$$=\sqrt{n}\left(S_n(\theta_0)(\sigma^{-1}(h))-P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(h))\right]\right)+o_p(1),$$

where S_n is given by (3). Consider the subset $\{(h_\beta,0,0;k\in\mathcal{K})|h_\beta\in\mathbb{R}^p\}\subset H$ and let $\widetilde{h}$ be an element of this subset. Setting $h=\widetilde{h}$ in the above equation yields

$$\sqrt{n}h'_\beta(\widehat{\beta}_n-\beta_0)=$$
$$\sqrt{n}\left(S_n(\theta_0)(\sigma^{-1}(\widetilde{h}))-P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))\right]\right)+o_p(1). \quad (11)$$

By Lemma 2, the central limit theorem, and Slutsky's theorem, $\sqrt{n}h'_\beta(\widehat{\beta}_n-\beta_0)$ is asymptotically normal with mean 0 and variance $P_{\theta_0}[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2]$. If $h\in H$, direct calculation yields

$$\begin{aligned}S_1(\theta_0)(h)^2 &= h'_\beta S_\beta(\theta_0)S_\beta(\theta_0)'h_\beta+h'_\gamma S_\gamma(\theta_0)S_\gamma(\theta_0)'h_\gamma+2h'_\beta S_\beta(\theta_0)S_\gamma(\theta_0)'h_\gamma\\ &+2h'_\beta S_\beta(\theta_0)\left(\sum_{k=1}^{K}S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})\right)+2h'_\gamma S_\gamma(\theta_0)\left(\sum_{k=1}^{K}S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})\right)\\ &+\sum_{k=1}^{K}\left(Q(\mathcal{O},k,\theta_0)\left[h_{\Lambda_k}(T)\Delta-\exp(\beta_0'X)\int_0^T h_{\Lambda_k}(s)\,d\Lambda_{k,0}(s)\right]\right)^2\\ &+2\sum_{k=1}^{K}\sum_{j>k}Q(\mathcal{O},k,\theta_0)\left[h_{\Lambda_k}(T)\Delta-\exp(\beta_0'X)\int_0^T h_{\Lambda_k}(s)\,d\Lambda_{k,0}(s)\right]\\ &\times Q(\mathcal{O},j,\theta_0)\left[h_{\Lambda_j}(T)\Delta-\exp(\beta_0'X)\int_0^T h_{\Lambda_j}(s)\,d\Lambda_{j,0}(s)\right].\end{aligned}$$

Taking expectation followed by some tedious algebraic manipulations and rearrangement of terms yield that

$$P_{\theta_0}\left[S_1(\theta_0)(h)^2\right]=$$
$$h'_\beta\sigma_\beta(h)+h'_\gamma\sigma_\gamma(h)+\sum_{k=1}^{K}\int_0^\tau\sigma_{\Lambda_k}(h)(u)h_{\Lambda_k}(u)\,d\Lambda_{k,0}(u). \quad (12)$$

Therefore

$$P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2\right] = \sigma_\beta^{-1}(\widetilde{h})'\sigma_\beta(\sigma^{-1}(\widetilde{h})) + \sigma_\gamma^{-1}(\widetilde{h})'\sigma_\gamma(\sigma^{-1}(\widetilde{h})) + \sum_{k=1}^{K}\int_0^\tau \sigma_{\Lambda_k}^{-1}(\widetilde{h})(u)\sigma_{\Lambda_k}(\sigma^{-1}(\widetilde{h}))(u)\, d\Lambda_{k,0}(u)$$
$$= h'_\beta\sigma_\beta^{-1}(\widetilde{h}),$$

where the last equality comes from the fact that

$$\sigma(\sigma^{-1}(\widetilde{h})) = (\sigma_\beta(\sigma^{-1}(\widetilde{h})), \sigma_\gamma(\sigma^{-1}(\widetilde{h})), \sigma_{\Lambda_k}(\sigma^{-1}(\widetilde{h})); k \in \mathcal{K}) = \widetilde{h}.$$

Now, recall that the linear map $\widetilde{\sigma}_\beta^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ was defined in Section 4 as a restricted version of σ_β^{-1}, by setting $\widetilde{\sigma}_\beta^{-1}(h_\beta) = \sigma_\beta^{-1}(\widetilde{h})$ for any $\widetilde{h}$ of the form $(h_\beta, 0, 0; k \in \mathcal{K})$. Let $\{e_1, \ldots, e_p\}$ be the canonical basis of $\mathbb{R}^p$ and $\Sigma_\beta = (\widetilde{\sigma}_\beta^{-1}(e_1), \ldots, \widetilde{\sigma}_\beta^{-1}(e_p))$. Then for any $h_\beta \in \mathbb{R}^p$, we have $\widetilde{\sigma}_\beta^{-1}(h_\beta) = \Sigma_\beta h_\beta$ and thus $P_{\theta_0}[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2] = h'_\beta\Sigma_\beta h_\beta$. Hence, for every $h_\beta \in \mathbb{R}^p$, $\sqrt{n}h'_\beta(\widehat{\beta}_n - \beta_0)$ converges in distribution to $\mathcal{N}(0, h'_\beta\Sigma_\beta h_\beta)$. By the Cramér-Wold device, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to $\mathcal{N}(0, \Sigma_\beta)$.

We now turn to semiparametric efficiency of $\widehat{\beta}_n$. Our proof relies on some geometric arguments. A detailed exposition of such arguments can be found in Bickel *et al.* (1993) and Tsiatis (2006) for example. Note that similar arguments have recently been used to derive efficiency in various other semiparametric hazard regression models (see Chang *et al.* (2005) and Zeng *et al.* (2005) for example). For $j = 1 \ldots, p$, denote $\widetilde{h}_j = (e_j, 0, 0; k \in \mathcal{K})$. Letting $h = \widetilde{h}_j$ for each $j = 1 \ldots, p$ in turn in (11) yields

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_\beta(\mathcal{O}_i, \theta_0) + o_p(1),$$

where

$$l_\beta(\mathcal{O}, \theta_0) = \Sigma_\beta S_\beta(\theta_0) + \Xi S_\gamma(\theta_0) + \sum_{k=1}^{K} S_{\Lambda_k}(\theta_0)(\Xi^*),$$

Ξ and Ξ^* are $(p \times q)$ and $(p \times 1)$ matrices respectively defined by

$$\Xi = \begin{pmatrix} \sigma_\gamma^{-1}(\widetilde{h}_1)' \\ \vdots \\ \sigma_\gamma^{-1}(\widetilde{h}_p)' \end{pmatrix} \quad \text{and} \quad \Xi^* = \begin{pmatrix} \sigma_{\Lambda_k}^{-1}(\widetilde{h}_1) \\ \vdots \\ \sigma_{\Lambda_k}^{-1}(\widetilde{h}_p) \end{pmatrix},$$

and $S_{\Lambda_k}(\theta_0)$ is applied componentwise to Ξ^*. Thus $\widehat{\beta}_n$ is an asymptotically linear estimator for β_0, and its influence function $l_\beta(\mathcal{O}, \theta_0)$ belongs to the tangent space spanned by the score functions. It follows that $l_\beta(\mathcal{O}, \theta_0)$ is the efficient influence function for β_0, and that $\widehat{\beta}_n$ is semiparametrically efficient

(see Bickel *et al.* (1993) or Tsiatis (2006)).

REMARK. Note that $l_\beta(\mathcal{O},\theta_0)$ is orthogonal to the space generated by the score functions for γ and the Λ_k, $k\in\mathcal{K}$. This can be ascertained as follows. Let $i\in\{1,\dots,p\}$. The following hold: $e_i'\Sigma_\beta=[\sigma_\beta^{-1}((e_i,0,0;k\in\mathcal{K}))]'$, $e_i'\Xi=[\sigma_\gamma^{-1}((e_i,0,0;k\in\mathcal{K}))]'$, and $e_i'S_{\Lambda_k}(\theta_0)(\Xi^*)=S_{\Lambda_k}(\theta_0)(\sigma_{\Lambda_k}^{-1}((e_i,0,0;k\in\mathcal{K})))$. Now,

$$
\begin{aligned}
&P_{\theta_0}\left[e_i'l_\beta(\mathcal{O},\theta_0)\sum_{j=1}^K S_{\Lambda_j}(\theta_0)(h_{\Lambda_j})\right]\\
&\quad=P_{\theta_0}\left[e_i'\left(\Sigma_\beta S_\beta(\theta_0)+\Xi S_\gamma(\theta_0)+\sum_{k=1}^K S_{\Lambda_k}(\theta_0)(\Xi^*)\right)\left(\sum_{j=1}^K S_{\Lambda_j}(\theta_0)(h_{\Lambda_j})\right)\right]\\
&\quad=P_{\theta_0}[\{[\sigma_\beta^{-1}((e_i,0,0;k\in\mathcal{K}))]'S_\beta(\theta_0)+[\sigma_\gamma^{-1}((e_i,0,0;k\in\mathcal{K}))]'S_\gamma(\theta_0)\\
&\qquad\qquad+\sum_{k=1}^K S_{\Lambda_k}(\theta_0)(\sigma_{\Lambda_k}^{-1}((e_i,0,0;k\in\mathcal{K})))\}(\sum_{j=1}^K S_{\Lambda_j}(\theta_0)(h_{\Lambda_j}))]\\
&\quad=P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}((e_i,0,0;k\in\mathcal{K})))\cdot S_1(\theta_0)((0,0,h_{\Lambda_j};j\in\mathcal{K}))\right],
\end{aligned}
$$

where the last equality follows from equation (3). Now, similar calculations as the ones leading to (12) yield that

$$
\begin{aligned}
&P_{\theta_0}\left[e_i'l_\beta(\mathcal{O},\theta_0)\sum_{j=1}^K S_{\Lambda_j}(\theta_0)(h_{\Lambda_j})\right]=\\
&\qquad\sum_{k=1}^K\int_0^\tau h_{\Lambda_k}(u)\sigma_{\Lambda_k}(\sigma^{-1}((e_i,0,0;k\in\mathcal{K})))(u)\,d\Lambda_{k,0}(u)=0.
\end{aligned}
$$

Thus, $l_\beta(\mathcal{O},\theta_0)$ is orthogonal to the score $\sum_{j=1}^K S_{\Lambda_j}(\theta_0)(h_{\Lambda_j})$. Similarly, $l_\beta(\mathcal{O},\theta_0)$ is orthogonal to $S_\gamma(\theta_0)$.

A.4 Proof of Theorem 4

The proof of asymptotic normality of $\sqrt{n}(\widehat{\gamma}_n-\gamma_0)$ proceeds along the same line as for $\sqrt{n}(\widehat{\beta}_n-\beta_0)$, and is therefore omitted.

Next, for any $t\in(0,\tau)$ and $j\in\mathcal{K}$, the asymptotic normality of $\sqrt{n}(\widehat{\Lambda}_{j,n}(t)-\Lambda_{j,0}(t))$ can be proved by using a similar argument with $\widetilde{h}$ replaced by $h_{(j,t)}=(h_\beta,h_\gamma,h_{\Lambda_k};k\in\mathcal{K})$, where $h_\beta=0$, $h_\gamma=0$, $h_{\Lambda_j}(\cdot)=1\{\cdot\le t\}$ $(t\in(0,\tau)$ and $j\in\mathcal{K})$, and $h_{\Lambda_k}=0$ for every $k\in\mathcal{K},k\neq j$. Details are omitted.

□

A.5 Proof of Theorem 5

The proof of Theorem 5 parallels the proof of Theorem 3 in Parner (1998) and thus, will be kept brief. Let $\widehat{\sigma}_n = (\widehat{\sigma}_{\beta,n}, \widehat{\sigma}_{\gamma,n}, \widehat{\sigma}_{\Lambda_k,n}; k \in \mathcal{K})$ be defined as σ with all of the θ_0 and P_{θ_0} replaced by $\widehat{\theta}_n$ and $\mathbb{P}_n$ respectively. Similar to the proof of Theorem 3 in Parner (1998), it can be shown that $\widehat{\sigma}_n$ converges in probability to σ uniformly over H and that its inverse $\widehat{\sigma}_n^{-1} = (\widehat{\sigma}_{\beta,n}^{-1}, \widehat{\sigma}_{\gamma,n}^{-1}, \widehat{\sigma}_{\Lambda_k,n}^{-1}; k \in \mathcal{K})$ is such that $\widehat{\sigma}_n^{-1}(h)$ converges to $\sigma^{-1}(h)$ in probability.

For every h_β, the asymptotic variance of $\sqrt{n}h'_\beta(\widehat{\beta}_n - \beta_0)$ is $h'_\beta \sigma_\beta^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$, which is consistently estimated by $h'_\beta \widehat{\sigma}_{\beta,n}^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$. Let $h_n = (h_{\beta,n}, h_{\gamma,n}, h_{\Lambda_k,n}; k \in \mathcal{K}) = \widehat{\sigma}_n^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$. Then $\widehat{\sigma}_n(h_n) = (h_\beta, 0, 0; k \in \mathcal{K})$, or

$$\begin{cases} \widehat{\sigma}_{\beta,n}(h_n) = h_\beta \\ \widehat{\sigma}_{\gamma,n}(h_n) = 0 \\ \widehat{\sigma}_{\Lambda_k,n}(h_n)(u) = 0, \quad k \in \mathcal{K}, \quad u \in [0,\tau]. \end{cases} \tag{13}$$

In particular, letting $u = T_1, \ldots, T_n$ in (13) yields the following system of equations:

$$\mathbb{A}_n \begin{pmatrix} h_{\beta,n} \\ h_{\gamma,n} \\ h_{\Lambda,n} \end{pmatrix} = \begin{pmatrix} h_\beta \\ 0_q \\ 0_{Kn} \end{pmatrix}, \tag{14}$$

where $h_{\Lambda,n} = (h_{\Lambda_1,n}(T_1), \ldots, h_{\Lambda_1,n}(T_n), \ldots, h_{\Lambda_K,n}(T_1), \ldots, h_{\Lambda_K,n}(T_n))'$, and $\mathbb{A}_n$ is defined by (4). Some simple algebra on (14) yields that $h_{\beta,n} = \widehat{\Sigma}_{\beta,n} h_\beta$ where $\widehat{\Sigma}_{\beta,n}$ is defined in Section 4, and therefore $h'_\beta \widehat{\Sigma}_{\beta,n} h_\beta$ is a consistent estimator of the asymptotic variance of $\sqrt{n}h'_\beta(\widehat{\beta}_n - \beta_0)$ for every h_β. We conclude that $\widehat{\Sigma}_{\beta,n}$ converges in probability to Σ_β. The consistency of $\widehat{\Sigma}_{\gamma,n}$ proceeds along the same lines and is therefore omitted.

We now turn to the estimation of the asymptotic variance of $\widehat{\Lambda}_{j,n}(t)$, for $t \in (0,\tau)$ and $j \in \mathcal{K}$. By the dominated convergence theorem and the consistency of $\widehat{\sigma}_n^{-1}$,

$$\int_0^t \widehat{\sigma}_{\Lambda_j,n}^{-1}(h_{(j,t)})(u)\, d\widehat{\Lambda}_{j,n}(u)$$

converges to $v_j^2(t) = \int_0^t \sigma_{\Lambda_j}^{-1}(h_{(j,t)})(u)\, d\Lambda_{j,0}(u)$, where we recall that $h_{(j,t)}$ is the element $(h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ such that $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_j}(\cdot) = 1\{\cdot \le t\}$ for some $t \in (0,\tau)$ and $j \in \mathcal{K}$, and $h_{\Lambda_k} = 0$ for every $k \in \mathcal{K}, k \neq j$. Letting $\widetilde{h}_n = (\widetilde{h}_{\beta,n}, \widetilde{h}_{\gamma,n}, \widetilde{h}_{\Lambda_k,n}; k \in \mathcal{K}) = \widehat{\sigma}_n^{-1}(h_{(j,t)})$, we get that $\widehat{\sigma}_n(\widetilde{h}_n) = h_{(j,t)}$ or:

$$\begin{cases} \widehat{\sigma}_{\beta,n}(\widetilde{h}_n) = 0 \\ \widehat{\sigma}_{\gamma,n}(\widetilde{h}_n) = 0 \\ \widehat{\sigma}_{\Lambda_j,n}(\widetilde{h}_n)(u) = 1\{u \le t\}, \quad u \in [0,\tau] \\ \widehat{\sigma}_{\Lambda_k,n}(\widetilde{h}_n)(u) = 0, \quad k \in \mathcal{K}, \quad k \neq j, \quad u \in [0,\tau]. \end{cases} \tag{15}$$

In particular, letting $u = T_1, \ldots, T_n$ in (15) yields the system of equations

$$\mathbb{A}_n \begin{pmatrix} \widetilde{h}_{\beta,n} \\ \widetilde{h}_{\gamma,n} \\ \widetilde{h}_{\Lambda,n} \end{pmatrix} = \begin{pmatrix} 0_p \\ 0_q \\ U^n_{(j,t)} \end{pmatrix},$$

with the notations $\widetilde{h}_{\Lambda,n} = (\widetilde{h}_{\Lambda_1,n}(T_1), \ldots, \widetilde{h}_{\Lambda_1,n}(T_n), \ldots, \widetilde{h}_{\Lambda_K,n}(T_1), \ldots, \widetilde{h}_{\Lambda_K,n}(T_n))'$ and $U^n_{(j,t)} = (0'_{(j-1)n}, 1\{T_1 \leq t\}, \ldots, 1\{T_n \leq t\}, 0'_{(K-j)n})'$. Similar algebra as above yields

$$\widetilde{h}_{\Lambda,n} = \widehat{\Sigma}_{\Lambda,n} U^n_{(j,t)},$$

where $\widehat{\Sigma}_{\Lambda,n}$ is defined in Section 4. Now, $\int_0^t \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(u)\, d\widehat{\Lambda}_{j,n}(u)$ verifies

$$\begin{aligned} \int_0^t \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(u)\, d\widehat{\Lambda}_{j,n}(u) &= \sum_{i=1}^n \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(T_i)\widehat{\Delta\Lambda_{j,n}}(T_i)1\{T_i \leq t\} \\ &= \widehat{\Xi}^{n'}_{(j,t)} \widetilde{h}_{\Lambda,n}, \end{aligned}$$

where $\widehat{\Xi}^n_{(j,t)} = \left(0'_{(j-1)n}, \widehat{\Delta\Lambda_{j,n}}(T_1)1\{T_1 \leq t\}, \ldots, \widehat{\Delta\Lambda_{j,n}}(T_n)1\{T_n \leq t\}, 0'_{(K-j)n}\right)'$. It follows that $\widehat{\Xi}^{n'}_{(j,t)} \widehat{\Sigma}_{\Lambda,n} U^n_{(j,t)}$ is a consistent estimator of $v_j^2(t)$, which concludes the proof.

□

Acknowledgements

The authors gratefully acknowledge the helpful comments on an earlier version from the Associate Editor and a Referee.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N. (1993). Statistical models based on counting processes. New York: Springer.

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.

Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society. Series B, 34*, 216–217.

Carroll, R. J., Ruppert, D., Stefanski, L. A. (1995). Measurement error in nonlinear models. London: Chapman & Hall.

Chang, I-S., Hsuing, C. A., Wang, M.-C., Wen, C.-C. (2005). An asymptotic theory for the nonparametric maximum likelihood estimator in the Cox gene model. *Bernoulli, 11*, 863–892.

Chen, H. Y., Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association, 94*, 896–908.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B, 34*, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*, 269–276.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, 39*, 1–38.

Dupuy, J.-F., Leconte, E. (2009). A study of regression calibration in a partially observed stratified Cox model. *Journal of Statistical Planning and Inference, 139*, 317–328.

Fang, H.-B., Li, G., Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics, 32*, 59–75.

Jobson, J. D. (1992). Applied multivariate data analysis. Volume II: Categorical and multivariate methods. New York: Springer.

Korsholm, L. (1998). Likelihood Ratio Test in the Correlated Gamma-Frailty Model. *Technical Report 98-11. Centre for Labour Market and Social Research - University of Aarhus School of Business.*

Kosorok, M. R. (2008). Introduction to empirical processes and semiparametric inference. New York: Springer.

Kosorok, M. R., Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics, 35*, 957–989.

Lin, D. Y., Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association, 88*, 1341–1349.

Loève, M. (1963). Probability theory. Toronto: D. Van Nostrand Co., Inc..

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics, 60*, 545–574.

Martinussen, T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics, 26*, 479–491.

Martinussen, T., Scheike, T. H. (2006). Dynamic regression models for survival data. New York: Springer.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics, 22*, 712–731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics, 23*, 182–198.

Murphy, S. A., Rossini, A. J., van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association, 92*, 968–976.

Paik, M. C. (1997). Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis, 3*, 289-298.

Paik, M. C., Tsai, W.-Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika, 84*, 579–593.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics, 26*, 183–214.

Pons, O. (2002). Estimation in the Cox model with missing covariate data. *Journal of Nonparametric Statistics, 14*, 223–247.

Sugimoto, T., Hamasaki, T. (2006). Properties of estimators of baseline hazard functions in a semiparametric cure model. *Annals of the Institute of Statistical Mathematics, 58*, 647–674.

Tsiatis, A. A. (2006). Semiparametric theory and missing data. New York: Springer.

van der Vaart, A. W. (1998). Asymptotic statistics. Cambridge: Cambridge University Press.

van der Vaart, A. W., Wellner, J. A. (1996). Weak convergence and empirical processes. New York: Springer.

Zeng, D., Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics, 33*, 2132–2163.

Zeng, D., Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society. Series B, 69*, 507–564.

Zeng, D., Lin, D. Y., Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association, 100*, 470–483.