The Annals of Statistics
2022, Vol. 50, No. 1, 487–510
https://doi.org/10.1214/21-AOS2117

SEMIPARAMETRIC LATENT-CLASS MODELS FOR MULTIVARIATE LONGITUDINAL AND SURVIVAL DATA

BY KIN YAU WONG[1], DONGLIN ZENG[2,*] AND D. Y. LIN[2,†]

[1]*Department of Applied Mathematics, The Hong Kong Polytechnic University,* kin-yau.wong@polyu.edu.hk
[2]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [*]dzeng@email.unc.edu; [†]lin@bios.unc.edu

In long-term follow-up studies, data are often collected on repeated measures of multivariate response variables as well as on time to the occurrence of a certain event. To jointly analyze such longitudinal data and survival time, we propose a general class of semiparametric latent-class models that accommodates a heterogeneous study population with flexible dependence structures between the longitudinal and survival outcomes. We combine nonparametric maximum likelihood estimation with sieve estimation and devise an efficient EM algorithm to implement the proposed approach. We establish the asymptotic properties of the proposed estimators through novel use of modern empirical process theory, sieve estimation theory and semiparametric efficiency theory. Finally, we demonstrate the advantages of the proposed methods through extensive simulation studies and provide an application to the Atherosclerosis Risk in Communities study.

1. Introduction. Many clinical and epidemiological studies generate data on repeated measures of response variables at multiple time points as well as on time to the occurrence of a clinical event. In cardiovascular cohort studies, for example, data are often recorded for both repeated measures of risk factors (e.g., blood pressures, cholesterol levels) and time to a cardiovascular event (e.g., stroke, heart attack) or death [17]. Shared random-effect models and joint latent-class models have been proposed to investigate the dynamic relationships among such longitudinal and survival data.

In shared random-effect models, a linear mixed model with a set of unobserved random effects is assumed for the longitudinal outcomes, and a proportional hazards model or transformation model with the same random effects as covariates is assumed for the survival time [4, 6, 18, 23, 24]. The shared random effects account for the dependence between the longitudinal and survival outcomes. These models typically assume that, conditional on the random effects, the distribution of the survival time and the effects of covariates on the longitudinal and survival outcomes are the same across subjects.

Joint latent-class models assume that the population consists of subgroups and within each subgroup, subjects have the same distributions of longitudinal and survival outcomes [9, 13]. These models allow the baseline risk of event and the association pattern between the longitudinal and survival outcomes to vary flexibly across subgroups. However, the existing work is mostly confined to fully parametric models. Lin et al. [5] proposed a semiparametric latent-class model with a nonparametric baseline hazard function for the survival time in each latent class but did not investigate the theoretical properties of the proposed nonparametric maximum likelihood estimators (NPMLE). In fact, such NPMLEs are inconsistent [11, 20]; see Section S1 of the Supplementary Material [21].

We propose a general model for the joint analysis of multivariate longitudinal data and survival time. We assume that the population consists of a mixture of latent subgroups such

Received September 2020; revised May 2021.
MSC2020 subject classifications. Primary 62N02; secondary 62G05, 62H30.
Key words and phrases. Censored data, joint analysis, mixture models, nonparametric estimation, sieve estimation.

that within each subgroup, the joint distribution of the longitudinal and survival outcomes is described by a separate random-effect model, in which the survival time is characterized by a separate nonparametric baseline hazard function. This model naturally extends those of Henderson, Diggle and Dobson [4] and Tsiatis and Davidian [18] by allowing the existence of latent subgroups. The model can be used to address important scientific questions:

1. Identification of latent subgroups within a heterogeneous study population;
2. Estimation of the effects of baseline covariates, such as treatment, on longitudinal and survival outcomes within each subgroup;
3. Evaluation of the event risk given baseline covariates and trajectories of longitudinal outcomes; and
4. Estimation of the association between the trajectories of longitudinal outcomes and covariates with proper adjustment of informative dropout due to the occurrence of the event.

The proposed modeling framework also extends existing work by accommodating multivariate longitudinal outcomes measured at multiple time points. This framework is particularly useful in cardiovascular studies, where multiple risk factors, such as blood pressures and cholesterol levels, are repeatedly measured. Including multivariate longitudinal outcomes not only provides a comprehensive depiction of the dynamic relationships among the event of interest and relevant risk factors but also helps identify the latent subgroup structure.

Due to the presence of multiple nonparametric components in the model and the lack of a closed-form expression for the likelihood function, model estimation is highly challenging both theoretically and computationally. To overcome the nonidentifiability of the fully nonparametric likelihood approach, we propose to combine nonparametric likelihood estimation with sieve estimation, such that the cumulative hazard function of a reference latent class is estimated by a step function with jumps at the observed event times, and the ratios of the baseline hazard functions across latent classes are estimated by spline functions. We develop a stable and efficient (accelerated) EM algorithm [3] to compute the proposed estimators.

We prove that the proposed estimators are consistent and the parametric components of the estimators are asymptotically efficient. The derivations involve novel applications of empirical process theory, sieve estimation theory and semiparametric efficiency theory. One major challenge in our theoretical development is to show that the proposed model is identifiable with an invertible information operator. Due to the presence of latent classes, techniques for establishing model identifiability or invertibility of the information operator for semiparametric shared random-effect models are not directly applicable to the current setting. In addition, existing methods for latent-class models are not readily applicable to semiparametric models. To establish model identifiability and the invertibility of the information operator, we note that the likelihood and the score function are the sums of the terms arising from the likelihood of semiparametric shared random-effect models and show that the terms in the summation can be separated by properly varying the observed data values.

The rest of this article is structured as follows. In Section 2, we formulate the model and describe the proposed estimation approach. In Section 3, we discuss the computation of the proposed estimators, and in Section 4, we present the theoretical results. In Section 5, we report the results from our simulation studies. In Section 6, we provide an application to the Atherosclerosis Risk in Communities (ARIC) study [17]. In Section 7, we make some concluding remarks. We relegate technical proofs to the Appendix.

2. Model, likelihood and sieve estimation.

Suppose that there are G latent classes. Let C denote the latent class membership, with $C = g$ if a subject belongs to the gth latent class ($g = 1, \dots, G$). We relate C to a set of time-independent covariates $\boldsymbol{W}$, which generally includes the constant 1, through a multinomial logistic regression model:

$$P(C = g \mid \boldsymbol{W}) = \frac{e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}}}, \tag{1}$$

where $\boldsymbol{\alpha}_g$ is the vector of class-specific regression parameters. For model identifiability, we set $\boldsymbol{\alpha}_G = \mathbf{0}$. Each latent class is characterized by class-specific trajectories of multivariate longitudinal outcomes and a class-specific risk of the event of interest. The longitudinal outcomes and the event time are assumed to be conditionally independent given the latent class membership and a multivariate random effect.

Suppose that there are J types of longitudinal outcomes and the jth type is measured at N_j time points. For $j = 1, \ldots, J$ and $k = 1, \ldots, N_j$, let Y_{jk} denote the kth measurement of the jth longitudinal outcome and $\boldsymbol{X}_{jk}$ and $\widetilde{\boldsymbol{X}}_{jk}$ denote corresponding covariates, which include the constant 1. The covariates $\boldsymbol{X}_{jk}$, $\widetilde{\boldsymbol{X}}_{jk}$ and $\boldsymbol{W}$ may partially or completely overlap. We relate Y_{jk} to $\boldsymbol{X}_{jk}$ and $\widetilde{\boldsymbol{X}}_{jk}$ through the multivariate linear mixed model:

$$Y_{jk}|_{C=g} = \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{jk} + \boldsymbol{b}^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{jk} + \epsilon_{jk} \tag{2}$$

for $g = 1, \ldots, G$, where $\boldsymbol{\beta}_g$ is a vector of class-specific regression parameters, $\boldsymbol{b}$ is a vector of random effects assumed to follow the multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)$, $(\epsilon_{j1}, \ldots, \epsilon_{jN_j})$ are independent zero-mean normal random variables with variance σ_{gj}^2, and $\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)$ is a covariance matrix indexed by a vector of class-specific variance parameters $\boldsymbol{\xi}_g$.

Let T denote the event time of interest. We relate T to a set of potentially time-dependent covariates $\boldsymbol{Z}(\cdot)$ through the proportional hazards model:

$$\lambda(t | \boldsymbol{Z}, \boldsymbol{b}, C = g) = \lambda_g(t) e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}(t) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}}, \tag{3}$$

where $\lambda_g(\cdot)$ is an arbitrary class-specific baseline hazard function, and $\boldsymbol{\gamma}_g$ and $\boldsymbol{\eta}_g$ are class-specific regression parameters. In the presence of censoring, we observe $\widetilde{T} = T \wedge U$ and $\Delta = I(T \le U)$, where U is the censoring time, and $I(\cdot)$ is the indicator function. Let $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1N_1}, \ldots, Y_{J1}, \ldots, Y_{JN_J})^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1N_1}, \ldots, \boldsymbol{X}_{J1}, \ldots, \boldsymbol{X}_{JN_J})^{\mathrm{T}}$, and $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_{11}, \ldots, \widetilde{\boldsymbol{X}}_{1N_1}, \ldots, \widetilde{\boldsymbol{X}}_{J1}, \ldots, \widetilde{\boldsymbol{X}}_{JN_J})^{\mathrm{T}}$. The data consist of n independent observations $\mathcal{O}_i \equiv (N_{i1}, \ldots, N_{iJ}, \boldsymbol{Y}_i, \boldsymbol{X}_i, \widetilde{\boldsymbol{X}}_i, \widetilde{T}_i, \Delta_i, \boldsymbol{W}_i, \{\boldsymbol{Z}_i(t)\}_{t \in [0, \widetilde{T}_i]})$ for $i = 1, \ldots, n$, where τ is the end of study time.

Let $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{G-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G, \sigma_{11}^2, \ldots, \sigma_{1J}^2, \ldots, \sigma_{GJ}^2, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_G, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_G)$ denote the set of all Euclidean parameters and $\Lambda_g(t) = \int_0^t \lambda_g(u)\, \mathrm{d}u$ for $g = 1, \ldots, G$. Under the assumption of noninformative censoring and longitudinal measurement times, rigorously formulated in Section S2 of the Supplementary Material [21], the likelihood function concerning $(\boldsymbol{\theta}, \Lambda_1, \ldots, \Lambda_G)$ is proportional to

$$\begin{aligned} &\prod_{i=1}^{n} \sum_{g=1}^{G} \frac{e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}_i}} \int \prod_{j=1}^{J} \prod_{k=1}^{N_{ij}} \sigma_{gj}^{-1} e^{-\frac{1}{2\sigma_{gj}^2} (Y_{ijk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{ijk} - \boldsymbol{b}^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{ijk})^2} \{\lambda_g(\widetilde{T}_i) e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(\widetilde{T}_i) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}}\}^{\Delta_i} \\ &\quad \times \exp\biggl\{-\int_0^{\widetilde{T}_i} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(t) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}}\, \mathrm{d}\Lambda_g(t)\biggr\} |\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)|^{-1/2} e^{-\frac{1}{2} \boldsymbol{b}^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}}\, \mathrm{d}\boldsymbol{b}. \end{aligned} \tag{4}$$

We reparametrize the model by setting $\Lambda = \Lambda_1$ and $\psi_g = \log(\lambda_g / \lambda_1)$; we then estimate Λ nonparametrically and approximate ψ_g using a sieve of B-spline functions for $g = 2, \ldots, G$. In particular, we treat Λ as a step function that jumps at the observed event times and replace $\lambda_1(\widetilde{T}_i)$ in the likelihood by $\Lambda\{\widetilde{T}_i\}$, where $\Lambda\{t\}$ is the jump size of Λ at t. Let $B_1, \ldots, B_{m_n}$ be B-spline functions on a grid over $[0, \tau]$, where the number of spline functions m_n increases with the sample size. For $g = 2, \dots, G$, we approximate ψ_g by $\sum_{s=1}^{m_n} a_{gs} B_s$, where $\boldsymbol{a} \equiv \{a_{gs}\}_{g=2,\dots,G;s=1,\dots,m_n}$ is a set of regression parameters. Ideally, NPMLE would be adopted for every nonparametric function because it does not require tuning and is more flexible than splines. However, because the NPMLE for $(\Lambda_1, \dots, \Lambda_G)$ is inconsistent, we estimate the cumulative baseline hazard function of a reference group using NPMLE and estimate the remaining nonparametric functions using splines, so as to achieve as much model flexibility as possible while ensuring estimation consistency.

Let $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \widehat{\boldsymbol{a}}_n)$ be the maximizer of

$$\prod_{i=1}^{n} \sum_{g=1}^{G} \frac{e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}_i}} \int \prod_{j=1}^{J} \prod_{k=1}^{N_{ij}} \sigma_{gj}^{-1} e^{-\frac{1}{2\sigma_{gj}^2}(Y_{ijk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{ijk} - \boldsymbol{b}^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{ijk})^2}$$

$$\times \big[\Lambda\{\widetilde{T}_i\} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(\widetilde{T}_i) + \sum_{s=1}^{m_n} a_{gs} B_s(\widetilde{T}_i) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}}\big]^{\Delta_i} \exp\bigg\{-\int_0^{\widetilde{T}_i} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(t) + \sum_{s=1}^{m_n} a_{gs} B_s(t) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}} \, \mathrm{d}\Lambda(t)\bigg\}$$

$$\times \big|\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)\big|^{-1/2} e^{-\frac{1}{2} \boldsymbol{b}^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}} \, \mathrm{d}\boldsymbol{b},$$

and let $\widehat{\psi}_{ng} = \sum_{s=1}^{m_n} \widehat{a}_{ngs} B_s$, where $\widehat{a}_{ngs}$ is the corresponding element of $\widehat{\boldsymbol{a}}_n$. Let $\mathcal{B} = (\psi_2, \dots, \psi_G)$. The sieve NPMLE of $(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ is $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \widehat{\mathcal{B}}_n)$, where $\widehat{\mathcal{B}}_n = (\widehat{\psi}_{n2}, \dots, \widehat{\psi}_{nG})$.

3. Computation of the sieve NPMLE. In this section, we use $\boldsymbol{Z}(\cdot)$ to denote the combination of the original set of time-dependent covariates and the B-spline functions $(B_1, \dots, B_m)$, with $\boldsymbol{\gamma}_g$ being the corresponding vector of regression parameters for the gth latent class. We compute the sieve NPMLE using an accelerated version of the EM algorithm, with C and $\boldsymbol{b}$ treated as missing data. The proposed algorithm iteratively performs the EM steps. Unlike the standard EM algorithm, an E-step may not be performed under the current parameter estimates but under some function of the estimates at the previous steps.

We first introduce the standard EM algorithm. The complete-data log-likelihood function is

$$\sum_{i=1}^{n} \sum_{g=1}^{G} I(C_i = g) \Bigg(\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i - \log\Bigg(\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}_i}\Bigg) - \frac{1}{2} \log\big|\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)\big| - \frac{1}{2} \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}_i$$

$$- \sum_{j=1}^{J} \sum_{k=1}^{N_{ij}} \bigg\{ \frac{1}{2} \log \sigma_{gj}^2 + \frac{(Y_{ijk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{ijk} - \boldsymbol{b}_i^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{ijk})^2}{2\sigma_{gj}^2} \bigg\}$$

$$+ \Delta_i \big[\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(\widetilde{T}_i) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}_i + \log \Lambda\{\widetilde{T}_i\}\big] - \sum_{s \le \widetilde{T}_i} \Lambda\{s\} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(s) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}_i} \Bigg).$$

In the E-step, we compute the expectation of functions of $(\boldsymbol{b}, C)$ involved in the M-step. The conditional density of $\boldsymbol{b}_i$ given $C_i = g$ and the observed data is proportional to

$$f_{ig}(\boldsymbol{b}_i) \equiv \Bigg(\prod_{j=1}^{J} \sigma_{gj}^{-N_{ij}} \Bigg) \prod_{j=1}^{J} \prod_{k=1}^{N_{ij}} \exp\bigg\{ - \frac{(Y_{ijk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{ijk} - \boldsymbol{b}_i^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{ijk})^2}{2\sigma_{gj}^2} \bigg\} \big|\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)\big|^{-1/2}$$

$$\times \exp\bigg\{ -\frac{1}{2} \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}_i \bigg\} e^{\Delta_i \{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(\widetilde{T}_i) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}_i\}} \exp\bigg\{ - \int_0^{\widetilde{T}_i} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(t) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}_i} \, \mathrm{d}\Lambda(t) \bigg\},$$

and the conditional probability of $C_i = g$ given the observed data is proportional to

$$q_{ig} \equiv e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i} \int f_{ig}(\boldsymbol{b}) \, \mathrm{d}\boldsymbol{b}.$$

The conditional expectation of any function h of $(\boldsymbol{b}_i, C_i)$ given the observed data is

$$\mathrm{E}\{h(\boldsymbol{b}_i, C_i) \mid \mathcal{O}_i\} = \sum_{g=1}^{G} \widehat{p}_{ig} \frac{\int h(\boldsymbol{b}, g) f_{ig}(\boldsymbol{b})\,\mathrm{d}\boldsymbol{b}}{\int f_{ig}(\boldsymbol{b})\,\mathrm{d}\boldsymbol{b}},$$

where $\widehat{p}_{ig} = q_{ig} / \sum_{l=1}^{G} q_{il}$. The integrations in the above equation can be approximated with the adaptive Gauss–Hermite quadrature [7].

In the M-step, we update the parameters by maximizing the expected complete-data log-likelihood function given the observed data. In particular, we update $\boldsymbol{\alpha}_g$ $(g = 1, \dots, G-1)$ by maximizing the weighted multinomial log-likelihood

$$\sum_{i=1}^{n} \left\{ \sum_{g=1}^{G} \widehat{p}_{ig} \boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i - \log\left(\sum_{g=1}^{G} e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}_i} \right) \right\}$$

via the Newton–Raphson algorithm. Then we update $\boldsymbol{\beta}_g$ and σ_{gj}^2 $(j = 1, \dots, J; g = 1, \dots, G)$ by maximizing

$$-\frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n} \widehat{p}_{ig} \left[N_{ij} \log \sigma_{gj}^2 + \sum_{k=1}^{N_{ij}} \frac{1}{\sigma_{gj}^2} \widehat{\mathrm{E}}_g\{(Y_{ijk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{ijk} - \boldsymbol{b}_i^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{ijk})^2\} \right]$$

and update $\boldsymbol{\xi}_g$ $(g = 1, \dots, G)$ by maximizing

$$-\frac{1}{2} \sum_{i=1}^{n} \widehat{p}_{ig} [\log|\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)| + \widehat{\mathrm{E}}_g\{\boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}_i\}],$$

where $\widehat{\mathrm{E}}_g$ denotes the conditional expectation with respect to $\boldsymbol{b}_i$ given $C_i = g$ and the observed data. If closed-form solutions for the maximization problems are not available, then we employ the Newton–Raphson algorithm. In addition, we update $(\boldsymbol{\gamma}_g, \boldsymbol{\eta}_g)$ $(g = 1, \dots, G)$ by maximizing the (weighted) log-partial likelihood

$$\sum_{i=1}^{n} \Delta_i \left[\sum_{g=1}^{G} \widehat{p}_{ig} \{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_i(\widetilde{T}_i) + \boldsymbol{\eta}_g^{\mathrm{T}} \widehat{\mathrm{E}}_g(\boldsymbol{b}_i)\} - \log\left\{ \sum_{g=1}^{G} \sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) \widehat{p}_{jg} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}_j(\widetilde{T}_i)} \widehat{\mathrm{E}}_g(e^{\boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}_j}) \right\} \right]$$

via the Newton–Raphson algorithm. Finally, we update the cumulative baseline hazard function Λ by

$$\widehat{\Lambda}\{\widetilde{T}_i\} = \frac{\Delta_i}{\sum_{g=1}^{G} \sum_{j=1}^{n} I(\widetilde{T}_j \geq \widetilde{T}_i) \widehat{p}_{jg} e^{\widehat{\boldsymbol{\gamma}}_g^{\mathrm{T}} \boldsymbol{Z}_j(\widetilde{T}_i)} \widehat{\mathrm{E}}_g(e^{\widehat{\boldsymbol{\eta}}_g^{\mathrm{T}} \boldsymbol{b}_j})}$$

for $i = 1, \dots, n$, where $(\widehat{\boldsymbol{\gamma}}_g, \widehat{\boldsymbol{\eta}}_g)$ are the current estimates of the parameters.

The standard EM algorithm, which iteratively performs the E-step and M-step until convergence, may be slow, especially when the number of parameters is large. To accelerate the convergence, we adopt a modification of the EM algorithm proposed by Varadhan and Roland [19]. Let $\boldsymbol{\vartheta}$ denote the set of all parameters and $\boldsymbol{s}(\boldsymbol{\vartheta})$ be the set of updated parameters after a single EM step if the initial parameter value is $\boldsymbol{\vartheta}$. With $\boldsymbol{\vartheta}^{(k)}$ being the set of current estimates, a step of the accelerated EM algorithm consists of:

1. Calculate $\boldsymbol{\vartheta}_1 = \boldsymbol{s}(\boldsymbol{\vartheta}^{(k)})$.
2. Calculate $\boldsymbol{\vartheta}_2 = \boldsymbol{s}(\boldsymbol{\vartheta}_1)$.
3. Calculate $\boldsymbol{r} = \boldsymbol{\vartheta}_1 - \boldsymbol{\vartheta}^{(k)}$, $\boldsymbol{v} = \boldsymbol{\vartheta}_2 - \boldsymbol{\vartheta}_1 - \boldsymbol{r}$, and $a = -\|\boldsymbol{r}\|_2 / \|\boldsymbol{v}\|_2$.
4. Update the parameter estimates by $\boldsymbol{\vartheta}^{(k+1)} = \boldsymbol{s}(\boldsymbol{\vartheta}^{(k)} - 2a\boldsymbol{r} + a^2\boldsymbol{v})$.

To improve stability, we update the parameters using the standard EM steps at early steps of the algorithm. Once the difference between consecutive parameter estimates becomes smaller than a certain threshold, we perform the accelerated EM steps until convergence. When the assumed number of latent classes is larger than the actual number, the model is nonidentifiable and the parameter estimates may not converge; therefore, we terminate the algorithm when the difference between the log-likelihood values of consecutive iterations is smaller than a certain threshold.

The algorithm may converge to a local maximum of the log-likelihood. To improve the chance of obtaining the global maximum, we can run the algorithm with different initial values and set the estimates to the converged values that yield the largest log-likelihood. One strategy for setting the initial values is to classify subjects into G classes by some clustering method and set the parameter values for each class to be the estimates obtained from subjects assigned to the class.

Upon convergence, we use Louis's formula [10] to compute the observed information matrix, essentially treating the model as parametric, with parameters $\boldsymbol{\theta}$, $\Lambda\{\widetilde{T}_i\}_{i:\Delta_i=1}$, and $\{a_{gs}\}_{g=2,\dots,G;s=1,\dots,m_n}$. The submatrix of the inverse of the observed information matrix corresponding to $\boldsymbol{\theta}$ can be used to estimate the standard errors of $\widehat{\boldsymbol{\theta}}_n$. This submatrix is essentially an estimate of the inverse of the efficient information matrix $\widetilde{\boldsymbol{I}}$ defined in the proof of Theorem 4.2, where the least-favorable directions are estimated by solving the empirical counterparts of the integral equations they satisfy. The consistency of this standard error estimator is established in Theorem 4.3.

We propose to use the Bayesian information criterion (BIC) [15] to select the number of latent classes G. Specifically, for each G, we estimate the model using the sieve NPMLE and compute

$$\mathrm{BIC} = -2\log L_n(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \widehat{\mathcal{B}}_n) + s\log n,$$

where L_n is the likelihood function, and s is the number of free parameters in the model, including the regression parameters for the B-spline functions. We select the G that yields the smallest BIC value.

4. Asymptotic properties of the sieve NPMLE. Assume that the degree of the B-spline functions is fixed at some $p \geq 1$ and that the distance between adjacent knots is within $(K^{-1}m_n^{-1}, Km_n^{-1})$ for some large constant K. Let d be the dimension of the Euclidean parameters and Θ be a known, compact parameter space of $\boldsymbol{\theta}$. Let $(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)$ denote the true parameter values, where $\mathcal{B}_0 = (\psi_{02}, \dots, \psi_{0G})$. Let $\Lambda_g(t) = \int_0^t \lambda_g(u)\,\mathrm{d}u$ and Λ_{0g} be its true value ($g = 1, \dots, G$).

We impose the following conditions.

(C1) The parameter $\boldsymbol{\theta}_0$ lies in the interior of Θ, and the function Λ_{0g} is continuously differentiable up to the third order on $[0, \tau]$ for $g = 1, \dots, G$.

(C2) With probability one, $P\{\widetilde{T} = \tau \mid \boldsymbol{W}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{Z}(\cdot)\} > \delta_0$ for some fixed $\delta_0 > 0$.

(C3) With probability one, $\boldsymbol{Z}(\cdot)$ has left-continuous sample paths on $[0, \tau]$ with right derivatives. In addition, there exists a large constant K such that

$$P\Big\{\max_{j=1,\dots,J} N_j + \|\boldsymbol{W}\|_2 + \|\boldsymbol{X}\|_2 + \|\widetilde{\boldsymbol{X}}\|_2 + \sup_{t\in[0,\tau]} \|\boldsymbol{Z}(t)\|_2 + \sup_{t\in[0,\tau]} \|\boldsymbol{Z}'(t)\|_2 < K\Big\} = 1,$$

where $\boldsymbol{Z}'$ is the (componentwise) left derivative of $\boldsymbol{Z}$.

(C4) The number of knots m_n satisfies $m_n = O(n^q)$ for some $1/12 < q < 1/8$.

The next condition is more technical and ensures model identifiability and invertibility of the information operator. Essentially, it requires that the covariates take enough distinct values such that the class-specific distributions of the longitudinal outcomes can be distinguished and the effect of each covariate on each class-specific distribution can be identified. Let $\mathbf{\Sigma}_{0g} = \operatorname{diag}(\sigma_{0g1}^2 \mathbf{1}_{N_1}, \dots, \sigma_{0gJ}^2 \mathbf{1}_{N_J})$, $\mathbf{\Gamma}_{0g} = \mathbf{\Psi}_{0g}(\boldsymbol{I} + \mathbf{\Psi}_{0g}^{\mathrm{T}} \widetilde{\boldsymbol{X}} \mathbf{\Sigma}_{0g}^{-1} \widetilde{\boldsymbol{X}}^{\mathrm{T}} \mathbf{\Psi}_{0g})^{-1} \mathbf{\Psi}_{0g}^{\mathrm{T}}$, and $\mathbf{\Sigma}_{0Yg} = \widetilde{\boldsymbol{X}} \mathbf{\Psi}_{0g} \mathbf{\Psi}_{0g}^{\mathrm{T}} \widetilde{\boldsymbol{X}}^{\mathrm{T}} + \mathbf{\Sigma}_{0g}$, where $\mathbf{1}_k$ is a k-vector of ones, $\mathbf{\Psi}_{0g}$ is an orthogonal matrix such that $\mathbf{\Sigma}(\boldsymbol{\xi}_{0g}) = \mathbf{\Psi}_{0g} \mathbf{\Psi}_{0g}^{\mathrm{T}}$, and σ_{0gj}^2 and $\boldsymbol{\xi}_{0g}$ are the true values of the corresponding parameters. Note that $\mathbf{\Sigma}_{0Yg}$ is the covariance matrix of $\boldsymbol{Y}$ given $C = g$ and $(N_1, \dots, N_J)$.

(C5) There exist some positive integers $(n_1, \dots, n_J)$ such that $P(N_1 = n_1, \dots, N_J = n_J) > 0$ and that the following holds. Let $\mathcal{X}$ be the set of possible values of $(\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ given $(N_1 = n_1, \dots, N_J = n_J)$ such that $\widetilde{\boldsymbol{X}}^{\mathrm{T}} \widetilde{\boldsymbol{X}}$ is invertible and

$$\begin{aligned} &\widetilde{\boldsymbol{X}} \mathbf{\Sigma}(\boldsymbol{\xi}_{0g}) \widetilde{\boldsymbol{X}}^{\mathrm{T}} + \mathbf{\Sigma}_{0g} \neq \widetilde{\boldsymbol{X}} \mathbf{\Sigma}(\boldsymbol{\xi}_{0l}) \widetilde{\boldsymbol{X}}^{\mathrm{T}} + \mathbf{\Sigma}_{0l} \\ &\quad \text{or} \quad (\boldsymbol{X}\boldsymbol{\beta}_{0g} \neq \boldsymbol{X}\boldsymbol{\beta}_{0l} \text{ and } \mathbf{\Sigma}_{0Yg}^{-1} \boldsymbol{X}\boldsymbol{\beta}_{0g} + \mathbf{\Sigma}_{0g}^{-1} \widetilde{\boldsymbol{X}} \mathbf{\Gamma}_{0g}^{\mathrm{T}} \boldsymbol{\eta}_{0g} \neq \mathbf{\Sigma}_{0Yl}^{-1} \boldsymbol{X}\boldsymbol{\beta}_{0l} + \mathbf{\Sigma}_{0l}^{-1} \widetilde{\boldsymbol{X}} \mathbf{\Gamma}_{0l}^{\mathrm{T}} \boldsymbol{\eta}_{0l}) \end{aligned}$$

whenever $g \neq l$. For $k = 1, \dots, n_j$ and $j = 1, \dots, J$, if $\boldsymbol{W}^{\mathrm{T}} \boldsymbol{h}_W = 0$, $\boldsymbol{X}_{jk}^{\mathrm{T}} \boldsymbol{h}_{Xjk} = 0$, $\widetilde{\boldsymbol{X}}_{jk}^{\mathrm{T}} \boldsymbol{h}_{\widetilde{X}jk} = 0$, and $\boldsymbol{Z}(t)^{\mathrm{T}} \boldsymbol{h}_Z = 0$ almost surely for all $(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) \in \mathcal{X}$ and $t \in [0, \tau]$, then $\boldsymbol{h}_W = \mathbf{0}$, $\boldsymbol{h}_{Xjk} = \mathbf{0}$, $\boldsymbol{h}_{\widetilde{X}jk} = \mathbf{0}$, and $\boldsymbol{h}_Z = \mathbf{0}$, where $\boldsymbol{h}_W$, $\boldsymbol{h}_{Xjk}$, $\boldsymbol{h}_{\widetilde{X}jk}$ and $\boldsymbol{h}_Z$ are fixed vectors of appropriate dimensions.

The final condition ensures that the least-favorable direction for the Euclidean parameters is sufficiently smooth.

(C6) The conditional density of the censoring variable U given the observed covariates is continuously differentiable on the support of U with respect to some dominating measure up to the third order.

REMARK 1. Conditions (C1)–(C3) are common assumptions in the analysis of right-censored data under semiparametric survival models. Condition (C4) pertains to the rate at which the number of B-spline functions increases to infinity. Condition (C5) pertains to the class-specific distributions of the longitudinal outcomes and event time. Instead of directly assuming the identifiability and invertibility of the information operator of the proposed model, we derive these properties under assumptions on individual class-specific distributions. Condition (C5) requires that after removing specific covariate values that yield equality of certain quantities of the class-specific distributions of the observed variables, the set of possible covariate values are linearly independent. For latent-class models in general, linear independence of the covariates and distinctness of parameter values across latent classes are not sufficient for the invertibility of the information operator. To see this, consider a simple model with two latent classes, a known mixture probability of 0.5 for each class, a single binary covariate X, and a single outcome variable Y with $Y \mid (X, C = g) \sim \mathrm{N}(\alpha_g + \beta_g X, 1)$ for $g = 1, 2$, where C denotes the latent class membership. The score statistic along the direction $\alpha_1 = \alpha_{01} + \epsilon$, $\alpha_2 = \alpha_{02} - \epsilon$, $\beta_1 = \beta_{01} - \epsilon$ and $\beta_2 = \beta_{02} + \epsilon$ is zero when $\alpha_{01} = \alpha_{02}$, even if $\beta_{01} \neq \beta_{02}$, where $(\alpha_{01}, \alpha_{02}, \beta_{01}, \beta_{02})$ are the true parameter values. This model does not satisfy (a simplified version of) condition (C5) because the two latent classes are different only at $X \neq 0$, but given $X \neq 0$, $(1, X)$ is no longer linearly independent. A simple sufficient condition for condition (C5) is that all covariates are linearly independent and the class-specific variances of $\boldsymbol{Y}$ are distinct almost surely.

Let $\|\cdot\|_\infty$ be the supremum norm over $[0, \tau]$. We have the following results.

THEOREM 4.1. *Under conditions* (*C*1)–(*C*5), *there exists a local maximum of the nonparametric likelihood in the sieve space*, *denoted by* $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \widehat{\mathcal{B}}_n)$, *such that*

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2 + \|\widehat{\Lambda}_n - \Lambda_0\|_\infty^2 + \sum_{g=2}^{G} \int_0^\tau |\widehat{\psi}_{ng}(t) - \psi_{0g}(t)|^2 \,\mathrm{d}t = o_p(n^{1/2}).$$

This theorem provides a preliminary, combined rate of convergence for the estimators of the Euclidean and infinite-dimensional parameters. Based on this convergence rate, the following theorem establishes that the Euclidean parameter estimators converge at the optimal $n^{1/2}$ rate and attain the semiparametric efficiency bound [1].

THEOREM 4.2. *Under conditions* (*C*1)–(*C*6), $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *converges weakly to the normal distribution with zero mean*, *and its asymptotic variance attains the semiparametric efficiency bound*.

Let $\boldsymbol{I}_n$ be the negative Hessian matrix of the log-likelihood evaluated at the estimated parameters, with the jump sizes of $\widehat{\Lambda}_n$ and the coefficients of the spline functions in $\widehat{\psi}_{n2}, \dots, \widehat{\psi}_{nG}$ treated as Euclidean parameters. Let $\widehat{\boldsymbol{V}}_n$ be the submatrix of $(n^{-1}\boldsymbol{I}_n)^{-1}$ that corresponds to $\boldsymbol{\theta}$.

THEOREM 4.3. *Under conditions* (*C*1)–(*C*6), $\|\widehat{\boldsymbol{V}}_n - \widetilde{\boldsymbol{I}}^{-1}\|_2 = o_p(1)$, *where* $\widetilde{\boldsymbol{I}}$ *is the efficient information matrix of* $\boldsymbol{\theta}$ *defined in the proof of Theorem* 4.2.

The proofs of Theorems 4.1 and 4.2 are given in Appendix A, whereas the proof of Theorem 4.3 is given in Section S3 of the Supplementary Material [21].

5. Simulation studies. We considered a longitudinal study where data were collected on repeated measures of longitudinal outcomes as well as on the time to the occurrence of an event of interest. Each subject was examined periodically until the event of interest occurred or the subject was lost to follow-up. At the initial examination, a set of baseline covariates, which may represent sex, age and other information, were measured, and at each examination, two types of longitudinal outcomes were measured. The latent class for each subject was generated from model (1) with $G = 3$ and $\boldsymbol{W} = (1, X_1, X_2)^{\mathrm{T}}$, where X_1 and X_2 are independent Bernoulli(0.5) and N(0, 1), respectively. We set the examination times at $s_k = 0.15(k-1)$ for $k = 1, \dots, 10$. For $j = 1, 2$ and $k = 1, \dots, 10$, we generated

$$Y_{jk} \,|_{C=g} = \boldsymbol{\beta}_{gj}^{\mathrm{T}} \boldsymbol{X}_k + b_j + b_3 + \epsilon_{jk}, \tag{5}$$

where $\epsilon_{jk} \mid (C = g) \sim \mathrm{N}(0, \sigma_{gj}^2)$, $\boldsymbol{X}_k = (1, s_k, X_1, X_2)^{\mathrm{T}}$, $b_j \mid (C = g) \sim \mathrm{N}(0, \xi_{gj}^2)$, and (b_1, b_2, b_3) are independent of each other and of (X_1, X_2). Note that the random effects b_1 and b_2 account for the dependence among repeated measures of a single type of longitudinal outcome, whereas b_3 accounts for the dependence between the two types of longitudinal outcomes. The event time T was generated from model (3) with a single random effect term b_3 and $\boldsymbol{Z}(t) = (X_1, X_2)^{\mathrm{T}}$ for all t, and the censoring variable U was generated from Uniform$(0, \tau)$ with $\tau = 5$. Note that the number of longitudinal outcome measurements is $\max\{k : k \le 10, s_k \le T \wedge U\}$.

The true values of the Euclidean parameters are given in Table S1 of the Supplementary Material [21]. The class-specific baseline hazard functions are $\lambda_1(t) = 0.5$, $\lambda_2(t) = \exp(0.25t)$ and $\lambda_3(t) = 1$. The proportions of subjects belonging to latent classes 1, 2 and 3 are approximately 35%, 35% and 30%, respectively. The average number of longitudinal outcome measurements per subject is about 5.4. The censoring proportion is about 25%.

We set the degree of the B-spline functions to be 1 and the number of interior knots to be 2; in our experience, the results are largely insensitive to the choice of the number of knots. The locations of the knots were set data-adaptively to be the 33% and 66% empirical quantiles of the observed event times. We considered $G = 2$, 3 and 4 latent classes and used BIC to select G. To set the initial values, we use k-mean clustering based on the event (or censoring) time, the censoring indicator, and the baseline longitudinal outcome values to classify subjects into subgroups with $k = G$. Then we fit the generalized linear models and survival models (without random effects) on each subgroup and set the initial parameter values to be the corresponding estimated values. The initial values for the coefficients of the B-splines and the regression parameters of the random effects are set to 0, the initial values of $\mathrm{Var}(b_j) + \mathrm{Var}(\epsilon_{jk})$ are set to be the estimated variances in the corresponding fitted linear models with $\mathrm{Var}(b_j) = \mathrm{Var}(\epsilon_{jk})$ ($j = 1, 2$; $k = 1, \ldots, 10$), and the variance of b_3 is set to be 0.1. The initial cumulative baseline hazard function is set to be the Breslow estimator. We constrained all Euclidean parameter estimates (including the regression parameters for the B-spline functions and the logarithm of the variance parameters) to be smaller than or equal to 10 in absolute value. This constraint is imposed because in the early iterations of the EM algorithm, the unconstrained estimates may sometimes become too extreme and cause numerical problems. We set the sample size to be $n = 1000$ or 2000 and considered 1000 simulation replicates for each setting.

Under $G = 3$, in no replicates do any parameter estimates (in absolute value) equal the boundary value of 10. Some parameter estimates are equal to the boundary value in about 60% of the replicates for $G = 4$ and in less than 5% of the replicates for $G = 2$. The convergence to the boundary under $G = 4$ is expected, because the model is nonidentifiable. In all but ten replicates under $n = 1000$, BIC selected the correct number of latent classes, and thus we only present the estimation results under $G = 3$. Because the labels of the latent classes are arbitrary, after convergence of the EM algorithm, we redefined the latent classes such that the orders of the estimated values of certain parameters across latent classes match the orders of the corresponding true parameter values. The estimation results for $n = 1000$ and $n = 2000$ are summarized in Tables S1 and S2 in the Supplementary Material [21], respectively. The estimators of all parameters, including the class-specific cumulative baseline hazard functions at particular time points, are virtually unbiased. The standard errors are estimated accurately, and the coverage probabilities of the confidence intervals are close to the nominal level, especially for $n = 2000$. Thus, the proposed estimation method effectively uncovers the latent structure of the population, produces consistent estimators, and yields valid statistical inference.

6. A real study.

The ARIC study is a prospective epidemiological cohort study conducted in the United States. In the study, a total of about 15,000 subjects received a baseline examination in 1987–1989 and potentially six subsequent examinations in 1990–1992, 1993–1995, 1996–1998, 2011–2013, 2016–2017 and 2018–2019. At each examination, medical data, such as body mass index (BMI), blood pressure and cholesterol levels, were collected. The subjects were also followed through reviews of hospital records, and potentially right-censored observations on time to myocardial infarction (MI), stroke and death were also obtained.

We aimed to study the risk of cardiovascular diseases or death among African American subjects and to detect the presence of latent subgroups. The event of interest is MI, stroke or death. The African American subjects were recruited from two centers of study in Forsyth County, NC and Jackson, MS. We set study location, sex and BMI, glucose level, smoking status and age at the first examination as covariates; these are referred to as the baseline covariates in the sequel. We considered systolic blood pressure and total cholesterol level, which were measured at each examination, as longitudinal outcomes. After removing 347 subjects with prior (or unknown status of) stroke or coronary heart disease at baseline and 178 subjects with missing data, the sample size is 3284, and the censoring proportion is 49.2%.

We fit models (1)–(3), where T is the time from the first examination to MI, stroke or death, whichever occurred first, (Y_{1k}, Y_{2k}) are respectively the systolic blood pressure and total cholesterol level at the kth examination, and N_j is the total number of examinations $(k = 1, \dots, N_j; j = 1, 2)$. The set of covariates $\boldsymbol{W}$ consists of the baseline covariates (and the constant 1 for the intercept). For the jth longitudinal outcome at the kth examination, we assumed model (5) with the set of covariates $\boldsymbol{X}_k$ consisting of the baseline covariates and the time of the kth examination. In the survival model, the set of covariates $\boldsymbol{Z}(t)$ is time-independent and consists of the baseline covariates, and the set of random effects consists of a single term b_3. We set the degree of the B-spline functions to be 1 and considered 2–4 interior knots. The locations of the knots were chosen to be empirical quantiles of the observed event times. We ranged the number of latent classes G from 1 to 6.

For any numbers of knots for the B-spline functions, the BIC picked $G = 4$ latent classes. The BIC values at $G = 1, \dots, 6$ under 2 interior knots are plotted in Figure S1 of the Supplementary Material [21]. Since the estimation results across different numbers of knots are similar, we reported the results under 2 interior knots. The point estimates, standard errors and p-values of all Euclidean parameters in the survival model are given in Table 1, and the estimated class-specific cumulative hazard functions are plotted in Figure 1; the estimation results for the remaining Euclidean parameters are given in Tables S3 and S4 of the Supplementary Material [21]. The estimated trajectories of the mean longitudinal outcomes for a typical subject from each latent class are plotted in Figure S2 of the Supplementary Material [21]. We classified a subject to a latent class if the (estimated) posterior probability of the class is larger than 0.7; a subject is unclassified if none of the posterior probabilities is larger

TABLE 1
Estimation results for the Euclidean parameters in the survival model for the ARIC data

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
$\gamma_{1,\text{Center}}$	0.2431	0.3041	4.24E−01	$\gamma_{3,\text{Glucose}}$	0.2304	0.0450	3.15E−07
$\gamma_{1,\text{BMI}}$	−0.0775	0.0949	4.14E−01	$\gamma_{3,\text{Smoke}}$	0.8147	0.1487	4.26E−08
$\gamma_{1,\text{Glucose}}$	0.4086	0.1325	2.04E−03	$\gamma_{3,\text{Sex}}$	0.3840	0.1355	4.61E−03
$\gamma_{1,\text{Smoke}}$	0.7848	0.1505	1.84E−07	$\gamma_{3,\text{Age}}$	0.5433	0.0673	7.13E−16
$\gamma_{1,\text{Sex}}$	0.5965	0.1617	2.25E−04	$\gamma_{4,\text{Center}}$	0.0770	0.3369	8.19E−01
$\gamma_{1,\text{Age}}$	0.6440	0.1303	7.75E−07	$\gamma_{4,\text{BMI}}$	−0.1136	0.1082	2.94E−01
$\gamma_{2,\text{Center}}$	0.1269	0.1887	5.01E−01	$\gamma_{4,\text{Glucose}}$	0.2954	0.0411	7.05E−13
$\gamma_{2,\text{BMI}}$	0.1052	0.0552	5.65E−02	$\gamma_{4,\text{Smoke}}$	0.5983	0.2039	3.34E−03
$\gamma_{2,\text{Glucose}}$	0.0634	0.0403	1.16E−01	$\gamma_{4,\text{Sex}}$	0.4959	0.1986	1.25E−02
$\gamma_{2,\text{Smoke}}$	0.6472	0.1378	2.65E−06	$\gamma_{4,\text{Age}}$	0.2654	0.0980	6.78E−03
$\gamma_{2,\text{Sex}}$	0.3533	0.1298	6.49E−03	η_1	1.8929	2.5689	4.61E−01
$\gamma_{2,\text{Age}}$	0.3426	0.0721	2.00E−06	η_2	1.5561	0.6952	2.52E−02
$\gamma_{3,\text{Center}}$	−0.0954	0.1920	6.19E−01	η_3	0.9861	2.3893	6.80E−01
$\gamma_{3,\text{BMI}}$	0.1853	0.0641	3.86E−03	η_4	1.3614	1.0065	1.76E−01

NOTE: For the parameters labeled γ, the first subscript represents the latent class and the second subscript represents the covariate that corresponds to the parameter. "Center" is the indicator for the Jackson center; "Sex" is the indicator for male; "Smoke" is the indicator for smoker; "Glucose" represents glucose level. All continuous covariates are standardized. The parameter η_g is the regression parameter of b_3 for the gth latent class.

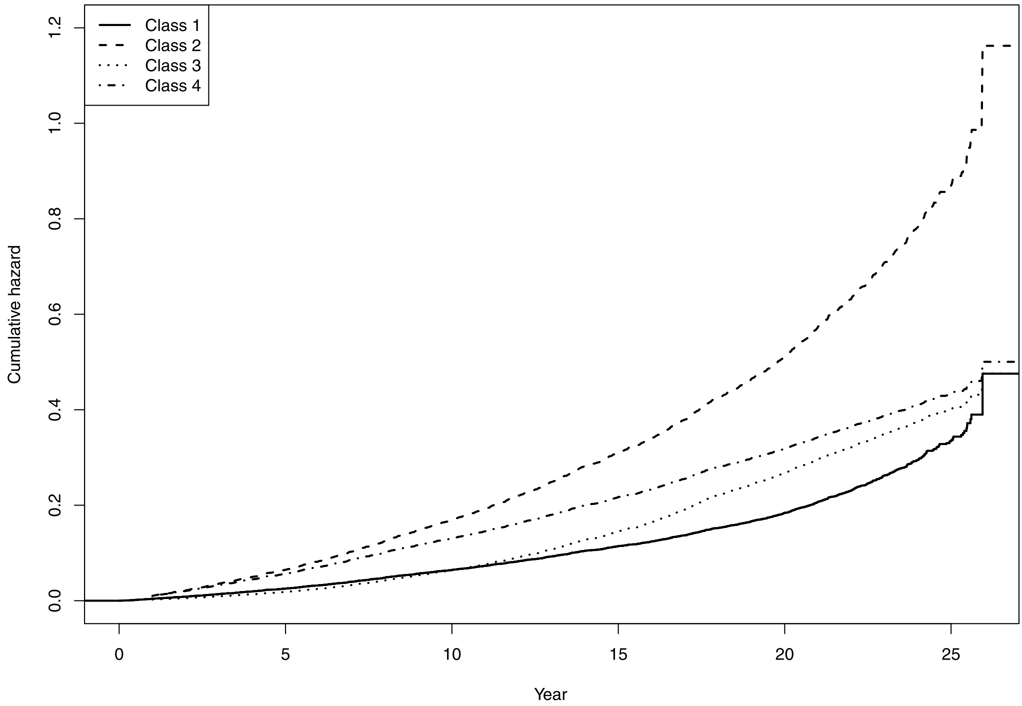

FIG. 1. *Estimated class-specific baseline hazard functions for the ARIC data.*

than 0.7. The Kaplan–Meier curves for the (predicted) latent classes are plotted in Figure S3 of the Supplementary Material [21].

Older subjects, males and smokers have higher risk of MI, stroke or death across all latent classes. Subjects with higher BMI tend to have higher risk of disease or death in the third latent class, but BMI has no significant association with the risk in other latent classes. Glucose level has highly significant positive effect on the risk of disease or death in all but the second latent class. The random effect b_3, which captures the dependence of the systolic blood pressure and the total cholesterol level, is significantly associated with the risk of disease or death only in the second latent class. This suggests that systolic blood pressure and total cholesterol level are associated with the risk of disease or death even conditional on the latent class membership. The estimated class-specific cumulative hazard of the second latent class is substantially higher than those of the other classes, and the empirical survival probabilities of the second latent class are smaller. The mean systolic blood pressure of subjects in the second latent class tends to be higher than those of the other classes. The results suggest that the second latent class is characterized by elevated risk of disease or death. The other groups also exhibit differences in the risk of disease or death, distributions of the longitudinal outcomes, and effects of covariates on the longitudinal and survival outcomes. In the latent-class membership model, the regression parameters for glucose level are significantly negative for the first three latent classes, suggesting that the fourth latent class is characterized by high glucose level. In addition, the second latent class is characterized by older subjects and the third latent class is characterized by males and subjects with higher BMI.

Suppose that we are interested in the conditional survival function for a subject at risk at time s given the trajectories of the longitudinal outcome measurements up to s. For a subject with time-independent covariates in the survival model, this probability function can be estimated by $h(t)/h(s)$ for $t \geq s$, where

$$h(t) = \sum_{g=1}^{G} \frac{e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}}} \int \exp\{-\Lambda_g(t) e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z} + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b}}\} \prod_{j=1}^{J} \prod_{k=1}^{K_j} \sigma_{gj}^{-1} e^{-\frac{1}{2\sigma_{gj}^2}(Y_{jk} - \boldsymbol{\beta}_g^{\mathrm{T}} \boldsymbol{X}_{jk} - \boldsymbol{b}^{\mathrm{T}} \widetilde{\boldsymbol{X}}_{jk})^2}$$
$$\times \left|\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)\right|^{-1/2} e^{-\frac{1}{2}\boldsymbol{b}^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1} \boldsymbol{b}} \, \mathrm{d}\boldsymbol{b},$$

K_j is the number of observations on the jth longitudinal outcome by time s, and the parameters are evaluated at the sieve NPMLE. Figure S4 in the Supplementary Material [21] shows the estimated curves for two hypothetical subjects at $s = 10$.

We use cross-validation to evaluate the robustness of the latent-class structure. We split the data into 20 pairs of training and validation datasets with a ratio of sample sizes of 3 : 2. On each training dataset, we fit the latent-class model with $G = 4$ and 2 interior knots for the B-spline functions, and for each subject in the corresponding validation dataset, we used the estimated model to compute the posterior probabilities of class membership given the subject's covariates and longitudinal outcomes (but not the event time). A subject is predicted to belong to a latent class if the posterior probability of the class is larger than 0.7; a subject is unclassified if none of the posterior probabilities is larger than 0.7. Note that the prediction of latent class does not directly involve the event time of the subjects in the validation dataset.

To evaluate the explanatory power of the (predicted) latent classes, in each validation dataset, we fit the Cox model with covariates, including the baseline systolic blood pressure, the baseline total cholesterol level, and the predicted latent classes; unclassified subjects were discarded. We tested the significance of the latent classes in the model using the likelihood-ratio test. The combined p-value across data splits is 0.0248, where the combined p-value is defined as $\Phi\{0.05 \sum_{s=1}^{20} \Phi^{-1}(p_s)\}$, p_s is the p-value for the sth split and Φ is the standard normal distribution function. In addition, we fit a stratified Cox model, stratifying on the latent classes, with covariates including the baseline covariates, the baseline systolic blood pressure, the baseline total cholesterol level and the interaction between the latent classes and the other covariates. The combined p-value for the likelihood-ratio tests for the interaction terms is 0.0250. These results suggest the existence of heterogeneity in the population that is not captured by the observed covariates. Subjects from different latent classes have not only different baseline hazards but also different association patterns between the covariates and the risk of disease or death.

7. Discussion.

In this article, we consider a semiparametric latent-class model for the joint analysis of longitudinal outcomes and a potentially right-censored event time. We develop a novel estimation approach that combines NPMLE and sieve estimation. We prove that the nonparametric components of the proposed estimators are consistent at a rate of $o(n^{1/4})$. Although sieve estimators generally converge at a rate slower than $n^{1/2}$, the Euclidean components of the estimators are nevertheless $n^{1/2}$-consistent and asymptotically normal.

Under the proposed model, covariates may be associated with the event time through the latent class membership or directly through the class-specific survival models. The regression parameters in the survival models are best interpreted conditional on the latent variables $\boldsymbol{b}$ and C, so that for a subject in a specific latent class, each covariate in the survival model contributes multiplicatively to the baseline hazard. To obtain an "overall" effect of the covariates, we may adopt a Monte-Carlo approach: repeatedly generate data from the estimated model and the observed covariates, and fit the Cox model on the generated event times and covariates. The estimated regression parameters could be interpreted as the overall effects of the covariates, combining the effects on the latent class membership and the class-specific event-time distributions.

We proposed to estimate the standard error of the estimators by the inverse of the observed information matrix. This approach yields satisfactory performance in our extensive numerical studies, but it may be numerically unstable in very large samples or models. If one is interested only in the inference of the Euclidean parameters, then alternative methods based on the profile likelihood can be adopted [22].

The constraints on the number of B-spline functions given by condition (C4) guarantee that $\widehat{\psi}_{ng}$ $(g = 2, \dots, G)$ converges to the true value at a rate faster than $n^{1/4}$, so that the Euclidean parameters can attain the efficiency bound. Because ψ_{0g}'s are continuously differentiable up to the third order, the approximation error of the spline functions is of rate $O(n^{-3q})$ and $q > 1/12$ is necessary for $\|\widehat{\psi}_{ng} - \psi_{0g}\|_2 = o_p(n^{-1/4})$; this bound can be relaxed under stronger assumptions on the smoothness of Λ_{0g}'s. The upper limit $q < 1/8$ arises from the shrinking-neighborhood-based argument for consistency. In the proof, we show that a local maximum of the log-likelihood exists in an $o(n^{-1/4})$-neighborhood of the true parameter values. The upper limit $q < 1/8$ is to guarantee that the second-order term in the linear expansion of the log-likelihood dominates other terms in the expansion.

An intuitively appealing nonparametric estimation approach is to set each class-specific cumulative baseline hazard function to be a step function that jumps at the observed event times. This approach, however, yields inconsistent estimators even in the simple settings considered by Ma and Wang [11] and Wang, Garcia and Ma [20] because the parameter space is overly complex. Each (uncensored) observation belongs to a specific latent class and should only contribute to the jump of the corresponding cumulative baseline hazard function at the observed event time. However, the latent class membership is unknown, and this nonparametric approach incorrectly allows all cumulative baseline hazard functions to jump at the event time. To overcome this difficulty, we only estimate the cumulative baseline hazard function of a reference class nonparametrically and approximate the relative magnitudes of the baseline hazard functions between the reference class and other classes using spline functions. With a properly chosen number of grid points for the spline functions, the complexity of the parameter space is controlled to yield consistent estimators.

During the preparation of this article, independent work of Liu et al. [8] was brought to our attention. Our model is more general than that of Liu et al. [8], which allows only a single type of longitudinal outcome with a random intercept in the longitudinal outcome model, and Liu et al. [8] adopted spline approximation for all nonparametric functions. In addition, we establish the asymptotic properties of the proposed estimators under specific assumptions on the proposed models and the observed data, whereas the assumptions in Liu et al. [8] are expressed in very general terms and are difficult to verify for given models. To demonstrate the extra flexibility of the proposed model over that of Liu et al. [8], we conducted a simulation study, which showed that misspecification of the latent variable structure may yield substantial estimation bias; see Section S4 of the Supplementary Material [21].

Our work can be extended in several directions. First, one may be interested in the joint analysis of multiple event times, such as the times to the occurrence of different diseases. The proposed modeling framework can be readily extended to allow for multivariate event times by assuming a separate regression model for each event time with a set of shared random effects $\boldsymbol{b}$. The sieve NPMLE can be easily extended to the multivariate setting, and its theoretical properties can be established along the lines of the proofs of Theorems 4.1 and 4.2.

Second, one may consider interval-censored event time(s). In ARIC, the onset of asymptomatic diseases, such as diabetes and hypertension, was not directly observed but was known to fall within certain time intervals. To accommodate interval censoring, we can extend the proposed methods and use the NPMLE [28] to estimate the cumulative baseline hazard function of the reference class. However, interval censoring results in a different likelihood function, which poses great challenges to the derivation of the asymptotic properties of the sieve NPMLE.

Finally, it would be of interest to consider high-dimensional longitudinal outcomes or covariates. In current biomedical studies, different types of molecular data, such as DNA alteration and gene expression, are collected along with clinical data. Such molecular data are often high dimensional, with the number of variables much larger than the sample size. These data contain rich genetic information that can be used to classify subjects into biologically distinct disease subtypes [16]. We can set variables for the molecular data as longitudinal outcomes or covariates in models (1)–(3) and adopt a penalized (sieve) likelihood approach for estimation.

APPENDIX A: PROOFS OF THEOREMS

In this Appendix, we prove Theorems 4.1 and 4.2. The proofs make use of the lemmas given in Appendix B. To facilitate the presentation, we introduce the following notation. Let $\mathcal{M}_K = \{\Lambda \in \ell^\infty[0,\tau] : \Lambda \text{ is monotone nondecreasing}, \Lambda(0) = 0, \Lambda(\tau) < K\}$. For some large enough positive constant K, let $\Xi_K \equiv \Theta \times \mathcal{M}_K \times \mathrm{BV}_K[0,\tau]^{G-1}$ be the parameter space of $(\boldsymbol{\theta}, \Lambda, \psi_2, \dots, \psi_G)$, where $\mathrm{BV}_K[0,\tau] = \{\psi \in \ell^\infty[0,\tau] : \|\psi\|_{\mathrm{V}} < K\}$, and $\|\cdot\|_{\mathrm{V}}$ is the total variation over $[0,\tau]$, such that

$$\|f\|_{\mathrm{V}} = \sup_{0=t_0\le t_1<\cdots<t_m=\tau} \sum_{j=1}^{m} |f(t_j) - f(t_{j-1})|.$$

The subscript K for the parameter spaces may be suppressed in the sequel. Let $\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ denote

$$\sum_{g=1}^{G} \frac{e^{\boldsymbol{\alpha}_g^{\mathrm{T}}\boldsymbol{W}}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}}\boldsymbol{W}}} \int \prod_{j=1}^{J}\prod_{k=1}^{N_j} \sigma_{gj}^{-1} e^{-\frac{1}{2\sigma_{gj}^2}(Y_{jk}-\boldsymbol{\beta}_g^{\mathrm{T}}\boldsymbol{X}_{jk}-\boldsymbol{b}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{jk})^2} \{e^{\boldsymbol{\gamma}_g^{\mathrm{T}}\boldsymbol{Z}(\widetilde{T})+\psi_g(\widetilde{T})+\boldsymbol{\eta}_g^{\mathrm{T}}\boldsymbol{b}}\}^{\Delta}$$
$$\times \exp\biggl\{-\int_0^{\widetilde{T}} e^{\boldsymbol{\gamma}_g^{\mathrm{T}}\boldsymbol{Z}(t)+\psi_g(t)+\boldsymbol{\eta}_g^{\mathrm{T}}\boldsymbol{b}}\,\mathrm{d}\Lambda(t)\biggr\} |\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)|^{-1/2} e^{-\frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\xi}_g)^{-1}\boldsymbol{b}}\,\mathrm{d}\boldsymbol{b},$$

so that the likelihood for a generic subject is proportional to $\Lambda\{\widetilde{T}\}^{\Delta}\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$. Let $\dot{\boldsymbol{\Psi}}_\theta(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ denote the derivative of $\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ with respect to $\boldsymbol{\theta}$, $\dot{\Psi}_\Lambda(\boldsymbol{\theta}, \Lambda, \mathcal{B})[H]$ denote the derivative of $\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ with respect to Λ along the direction H, and $\dot{\Psi}_{\psi g}(\boldsymbol{\theta}, \Lambda, \mathcal{B})[h]$ denote the derivative of $\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ with respect to ψ_g along the direction h.

In the sequel, we use $\|\cdot\|$ to denote the Euclidean norm for vectors and the L_2-norm with respect to the Lebesgue measure for functions over $[0,\tau]$. For a set of functions $\mathcal{B} \equiv (\psi_2, \dots, \psi_g)$, let $\|\mathcal{B}\|^2 = \sum_{g=2}^{G} \|\psi_g\|^2$. Let $\mathbb{P}$ and $\mathbb{P}_n$ denote the true and empirical measures, respectively.

Proof of Theorem 4.1. Following Schumaker [14], under condition (C1), there exist functions $(\widetilde{\psi}_{n2}, \dots, \widetilde{\psi}_{nG})$ such that $\|\widetilde{\psi}_{ng} - \psi_{0g}\|_\infty = O(m_n^{-3})$ for $g = 2, \dots, G$, where $\widetilde{\psi}_{ng} = \sum_{s=1}^{m_n} \widetilde{a}_{gs} B_s$ for some regression parameters $\widetilde{a}_{gs}$ $(g = 2, \dots, G; s = 1, \dots, m_n)$. Let

$$\mathcal{N}_{\epsilon_n} = \Biggl\{(\psi_2, \dots, \psi_G) : \psi_g = \sum_{s=1}^{m_n} a_{gs} B_s : \sum_{s=1}^{m_n} |a_{gs} - \widetilde{a}_{gs}|^2 \le \epsilon_n^2, g = 2, \dots, G\Biggr\},$$

where ϵ_n is a positive sequence such that $\epsilon_n = o(m_n^{-3/2})$. For $\mathcal{B}_n \equiv (\psi_{n2}, \dots, \psi_{nG}) \in \mathcal{N}_{\epsilon_n}$,

$$\|\psi_{ng} - \widetilde{\psi}_{0g}\|_{\mathrm{V}} \le \sum_{s=1}^{m_n} |a_{gs} - \widetilde{a}_{gs}| \|B_s'\|_\infty = O(m_n)(\epsilon_n^2 m_n)^{1/2} = o(1).$$

Therefore, each function ψ_{ng} of $\mathcal{N}_{\epsilon_n}$ has bounded total variation and converges uniformly to ψ_{0g}.

The outline of the proof is as follows. For any sequence of $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$, we define

$$\big(\widehat{\boldsymbol{\theta}}_n[\mathcal{B}_n], \widehat{\Lambda}_n[\mathcal{B}_n]\big) = \operatorname*{arg\,max}_{(\boldsymbol{\theta},\Lambda)} \mathbb{P}_n \ell(\boldsymbol{\theta}, \Lambda, \mathcal{B}_n).$$

First, we show that $(\widehat{\boldsymbol{\theta}}_n[\mathcal{B}_n], \widehat{\Lambda}_n[\mathcal{B}_n]) \to_p (\boldsymbol{\theta}_0, \Lambda_0)$ uniformly over $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$. Then we derive the rate of convergence of $(\widehat{\boldsymbol{\theta}}_n[\mathcal{B}_n], \widehat{\Lambda}_n[\mathcal{B}_n])$ in terms of ϵ_n. Finally, we show that the maximum of the profile log-likelihood $\mathbb{P}_n \ell(\widehat{\boldsymbol{\theta}}_n[\mathcal{B}_n], \widehat{\Lambda}_n[\mathcal{B}_n], \mathcal{B}_n)$ over $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$ lies in the interior of $\mathcal{N}_{\epsilon_n}$ for some $\epsilon_n = o(n^{-1/4} m_n^{1/2})$ and for large enough n. For simplicity of presentation, we suppress the argument $\mathcal{B}_n$ in $\widehat{\boldsymbol{\theta}}_n[\mathcal{B}_n]$ and $\widehat{\Lambda}_n[\mathcal{B}_n]$ in the sequel.

Step 1. We prove the existence of the NPMLE, that is, $\widehat{\Lambda}_n(\tau) < \infty$. Let $\pi_g = e^{\boldsymbol{\alpha}_g^{\mathrm{T}} \boldsymbol{W}} / \sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^{\mathrm{T}} \boldsymbol{W}}$ and $f_g(\boldsymbol{Y}, \boldsymbol{b})$ denote the joint density of $(\boldsymbol{Y}, \boldsymbol{b})$ for the gth latent class (given $N_1, \ldots, N_J$); we suppress the parameter or covariate values in the expressions for simplicity of presentation. Note that

$$\begin{aligned}
&\Psi(\mathcal{O}; \boldsymbol{\theta}, \Lambda, \mathcal{B}) \\
&\quad \lesssim \sum_{g=1}^{G} \pi_g \int e^{\{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}(\widetilde{T}) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b} + \psi_g(\widetilde{T})\}\Delta} \bigg\{1 + \int_0^{\widetilde{T}} e^{\boldsymbol{\gamma}_g^{\mathrm{T}} \boldsymbol{Z}(s) + \boldsymbol{\eta}_g^{\mathrm{T}} \boldsymbol{b} + \psi_g(s)} \,\mathrm{d}\Lambda(s)\bigg\}^{-\kappa} f_g(\boldsymbol{Y}, \boldsymbol{b}) \,\mathrm{d}\boldsymbol{b} \\
&\quad \le \sum_{g=1}^{G} \pi_g e^{\psi_g(\widetilde{T})\Delta} \bigg\{1 + \int_0^{\widetilde{T}} e^{\psi_g(s)} \,\mathrm{d}\Lambda(s)\bigg\}^{-\kappa} \int e^{O(1+\|\boldsymbol{b}\|)} f_g(\boldsymbol{Y}, \boldsymbol{b}) \,\mathrm{d}\boldsymbol{b}
\end{aligned}$$

for some constant $\kappa > 1$, where $\lesssim$ denotes "smaller than up to a scaling factor." Therefore, if $\Lambda(\tau) = \infty$, then the right-hand side of the above inequality is zero. We conclude that $\widehat{\Lambda}_n(\tau) < \infty$, so that the NPMLE exists.

Step 2. We show that the NPMLE is uniformly bounded. Note that

$$\begin{aligned}
\frac{1}{n} \log L_n(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) \le \frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \widehat{\Lambda}_n\{\widetilde{T}_i\} + \frac{1}{n} \sum_{i=1}^{n} \log \Bigg[\sum_{g=1}^{G} \pi_{gi} e^{\psi_{ng}(\widetilde{T}_i)\Delta} \\
\times \bigg\{1 + \int_0^{\widetilde{T}_i} e^{\psi_{ng}(s)} \,\mathrm{d}\widehat{\Lambda}_n(s)\bigg\}^{-\kappa} \int e^{O(1+\|\boldsymbol{b}\|)} f_g(\boldsymbol{Y}_i, \boldsymbol{b}) \,\mathrm{d}\boldsymbol{b} \Bigg].
\end{aligned}$$

Let $\widetilde{N}_n = n^{-1} \sum_{i=1}^{n} \Delta_i I(\widetilde{T}_i \le \cdot)$. We have

$$\frac{1}{n} \log L_n(\boldsymbol{\theta}_0, \widetilde{N}_n, \mathcal{B}_n) \ge -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \log n + O_p(1),$$

where the second term on the right-hand side is asymptotically bounded uniformly over $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$. Thus,

$$\begin{aligned}
&\frac{1}{n} \log L_n(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) - \frac{1}{n} \log L_n(\boldsymbol{\theta}_0, \widetilde{N}_n, \mathcal{B}_n) \\
&\quad \le \frac{1}{n} \sum_{i=1}^{n} \Delta_i \log[n \widehat{\Lambda}_n\{\widetilde{T}_i\}] - \frac{\kappa}{n} \sum_{i=1}^{n} \log\{1 + \widehat{\Lambda}_n(\tau)\} + O_p(1).
\end{aligned}$$

Using a partitioning argument similar to that of Murphy [12], we can show that the right-hand side of the above inequality tends to $-\infty$ if $\limsup_n \widehat{\Lambda}_n(\tau) = \infty$. By the definition of $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n)$, the left-hand side of the inequality is nonnegative, so that

$$\limsup_n \sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} \widehat{\Lambda}_n(\tau) < \infty.$$

Step 3. We show that $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n)$ is consistent. Because $\widehat{\Lambda}_n$ belongs to a function space with bounded total variation, by Helly's selection theorem, for every subsequence of $\{n\}_{n=1,2,\dots}$, there exists a further subsequence such that $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ and $\widehat{\Lambda}_n \to \Lambda^*$ for some $(\boldsymbol{\theta}^*, \Lambda^*)$. We show that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda^* = \Lambda_0$ for any subsequence. With an abuse of notation, let $\{n\}_{n=1,2,\dots}$ be the subsequence. Let

$$\widetilde{\Lambda}_n(t) = -\sum_{i=1}^{n} \Delta_i I(\widetilde{T}_i \le t)\left\{\sum_{j=1}^{n} \frac{\dot{\Psi}_\Lambda(\mathcal{O}_j; \boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)[I(\widetilde{T}_i \le \cdot)]}{\Psi(\mathcal{O}_j; \boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)}\right\}^{-1}.$$

Note that $\dot{\Psi}_\Lambda(\boldsymbol{\theta}, \Lambda, \mathcal{B})[I(\cdot \ge t)] = -I(\widetilde{T} \ge t)\sum_{g=1}^{G} \pi_g \int Q_g(\mathcal{O}, \boldsymbol{b}) e^{\boldsymbol{\gamma}_g^{\mathrm{T}}\boldsymbol{Z}(t) + \boldsymbol{\eta}_g^{\mathrm{T}}\boldsymbol{b} + \psi_g(t)}\,\mathrm{d}\boldsymbol{b}$, where

$$Q_g(\mathcal{O}, \boldsymbol{b}) = e^{\{\boldsymbol{\gamma}_g^{\mathrm{T}}\boldsymbol{Z}(\widetilde{T}) + \boldsymbol{\eta}_g^{\mathrm{T}}\boldsymbol{b} + \psi_g(\widetilde{T})\}\Delta} \exp\left\{-\int_0^{\widetilde{T}} e^{\boldsymbol{\gamma}_g^{\mathrm{T}}\boldsymbol{Z}(t) + \boldsymbol{\eta}_g^{\mathrm{T}}\boldsymbol{b} + \psi_g(t)}\,\mathrm{d}\Lambda(t)\right\} f_g(\boldsymbol{Y}, \boldsymbol{b}).$$

By the definition of the NPMLE, $\mathbb{P}_n \ell(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) \ge \mathbb{P}_n \ell(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_n)$, so

$$\mathbb{P}_n \Delta \log \frac{\widehat{\Lambda}_n\{\widetilde{T}\}}{\widetilde{\Lambda}_n\{\widetilde{T}\}} + \mathbb{P}_n \log \frac{\Psi(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n)}{\Psi(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_n)} \ge 0. \tag{6}$$

Note that

$$\begin{aligned}
&\mathbb{P}_n \log \Psi(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) - \mathbb{P} \log \Psi(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0) \\
&\quad = (\mathbb{P}_n - \mathbb{P}) \log \Psi(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) + \mathbb{P}\{\log \Psi(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n) - \log \Psi(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0)\},
\end{aligned}$$

where the first term on the right-hand side goes to zero almost surely because the class of $\log \Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$ is Gilvenko–Cantelli by Lemma B.1, and the second term is $o(1)$ by the dominated convergence theorem; note that both terms converge uniformly over $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$. By a similar argument on $\mathbb{P}_n \log \Psi(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_n)$, the second term on the left-hand side of (6) is

$$\mathbb{P}_n \log \frac{\Psi(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n)}{\Psi(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_n)} = \mathbb{P} \log \frac{\Psi(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0)}{\Psi(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)} + o_p(1),$$

where the $o_p(1)$ term tends to 0 almost surely.

Consider the first term on the left-hand side of (6). Note that

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{\mathbb{P}_n \nu(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0; s)}{\mathbb{P}_n \nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; s)}\,\mathrm{d}\widetilde{\Lambda}_n(s), \tag{7}$$

where $\nu(\boldsymbol{\theta}, \Lambda, \mathcal{B}; t) = \dot{\Psi}_\Lambda(\boldsymbol{\theta}, \Lambda, \mathcal{B})[I(\cdot \ge t)]/\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})$. By Lemma B.1, $\{\nu(\boldsymbol{\theta}, \Lambda, \mathcal{B}; t) : t \in [0, \tau], (\boldsymbol{\theta}, \Lambda, \mathcal{B}) \in \Xi\}$ is Glivenko–Cantelli, so

$$\left|\sup_{t \in [0,\tau]} (\mathbb{P}_n - \mathbb{P}) \nu(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0; t)\right| + \left|\sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} \sup_{t \in [0,\tau]} (\mathbb{P}_n - \mathbb{P}) \nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; t)\right| \to_{\text{a.s.}} 0.$$

By the dominated convergence theorem, $\mathbb{P}\nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; t)$ converges to $\mathbb{P}\nu(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0; t)$ for each t. In addition, it is easy to see that the derivative of $\mathbb{P}\nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; t)$ with respect to t is uniformly bounded, so that $\mathbb{P}\nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; t)$ is equicontinuous with respect to t. Thus, by the Arzela–Ascoli theorem, $\mathbb{P}\nu(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n, \mathcal{B}_n; t) \to \mathbb{P}\nu(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0; t)$ uniformly in $t \in [0, \tau]$. Furthermore, we can follow the argument in Zeng, Lin and Lin [27, p. 374] to show by contradiction that $\min_{t \in [0,\tau]} |\mathbb{P}\nu(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0; t)| > 0$. Taking limit on both sides of (7) yields

$$\Lambda^*(t) = \int_0^t \frac{\mathbb{P}\nu(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0; s)}{\mathbb{P}\nu(\boldsymbol{\theta}^*, \Lambda^*, \mathcal{B}_0; s)}\,\mathrm{d}\Lambda_0(s).$$

We conclude that Λ^* is absolutely continuous with respect to Λ_0, and thus is differentiable. Let λ^* be the derivative of Λ^*. Combining the above results with (6), we have

$$\mathbb{P}\log\frac{\lambda^*(\widetilde{T})^{\Delta}\Psi(\boldsymbol{\theta}^*,\Lambda^*,\mathcal{B}_0)}{\lambda_0(\widetilde{T})^{\Delta}\Psi(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)}\geq 0.$$

By the nonnegativity of the Kullback–Leibler divergence and Lemma B.2, the left-hand side of the above inequality is nonpositive and is equal to zero if and only if $(\boldsymbol{\theta}^*, \Lambda^*) = (\boldsymbol{\theta}_0, \Lambda_0)$. Therefore, $(\widehat{\boldsymbol{\theta}}_n, \widehat{\Lambda}_n)$ is consistent.

Step 4. We derive a bound on $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| + \|\widehat{\Lambda}_n - \Lambda_0\|_\infty$ in terms of $\|\mathcal{B}_n - \mathcal{B}_0\|$. For any $\boldsymbol{h}_\theta \in \mathbb{R}^d$ and $h_\Lambda \in \mathrm{BV}[0,\tau]$, let

$$\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta},\Lambda,\mathcal{B})[\boldsymbol{h}_\theta,h_\Lambda]=\frac{\partial}{\partial\epsilon}\ell\biggl(\boldsymbol{\theta}+\epsilon\boldsymbol{h}_\theta,\Lambda+\epsilon\int h_\Lambda\,\mathrm{d}\Lambda,\mathcal{B}\biggr)\bigg|_{\epsilon=0}.$$

Clearly, $\mathbb{P}_n\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)[\boldsymbol{h}_\theta,h_\Lambda]=0$ and $\mathbb{P}\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)[\boldsymbol{h}_\theta,h_\Lambda]=0$ for any $(\boldsymbol{h}_\theta, h_\Lambda)$. Suppressing the arguments $(\boldsymbol{h}_\theta, h_\Lambda)$, we have

$$\begin{aligned}
&\mathbb{P}\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_0)-\mathbb{P}\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)\\
&\quad=\mathbb{P}\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_0)-\mathbb{P}_n\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)\\
&\quad=-(\mathbb{P}_n-\mathbb{P})\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)-\mathbb{P}\{\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_n)-\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)\}\\
&\qquad-\mathbb{P}[\{\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)-\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_n)\}-\{\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_0)-\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)\}].
\end{aligned}$$

By Lemma B.1, the class $\{\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta},\Lambda,\mathcal{B})[\boldsymbol{h}_\theta,h_\Lambda]:(\boldsymbol{\theta},\Lambda,\mathcal{B})\in\Xi,\|\boldsymbol{h}_\theta\|\leq 1,\|h_\Lambda\|_{\mathrm{V}}\leq 1\}$ is Donsker, so that the first term on the right-hand side above is $O_p(n^{-1/2})$ uniformly over $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$. By repeated applications of the mean-value theorem, we can show that the second term is $O(\|\mathcal{B}_n - \mathcal{B}_0\|)$ and the third term is $o(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| + \|\widehat{\Lambda}_n - \Lambda_0\|_\infty)$. To evaluate the left-hand side of the above display, note that $\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta},\Lambda,\mathcal{B}_0)$ is the score statistic of a survival model with a single nonparametric component; the model falls under the framework of, for example, Zeng and Lin [25]. Using arguments analogous to the proof of Theorem 3.2 of Zeng and Cai [24] and the proof of Theorem 2 of Zeng and Lin [26], we can show that the map $(\boldsymbol{\theta},\Lambda)\mapsto\mathbb{P}\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta},\Lambda,\mathcal{B}_0)$ is Frechet-differentiable with a derivative $\nabla\mathbb{P}\dot{\ell}_{\theta\Lambda}$ that takes the form of a Fredholm operator. By Lemma B.4, $\nabla\mathbb{P}\dot{\ell}_{\theta\Lambda}$ (evaluated at the true parameter values) is one-to-one, so it is continuously invertible. Therefore, there exists some positive constant c_1 such that $\|\nabla\mathbb{P}\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0,\widehat{\Lambda}_n-\Lambda_0)\|\geq c_1(\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|+\|\widehat{\Lambda}_n-\Lambda_0\|_\infty)$, where the norm on the left-hand side of the inequality is the supremum norm over $\{(\boldsymbol{h}_\theta,h_\Lambda):\|\boldsymbol{h}_\theta\|\leq 1,\|h_\Lambda\|_{\mathrm{V}}\leq 1\}$. By the consistency of $(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n)$ and the differentiability of $\mathbb{P}\dot{\ell}_{\theta\Lambda}$,

$$\|\mathbb{P}\dot{\ell}_{\theta\Lambda}(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_0)-\mathbb{P}\dot{\ell}_{\theta\Lambda}(\boldsymbol{\theta}_0,\Lambda_0,\mathcal{B}_0)\|\geq\{c_1+o(1)\}(\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|+\|\widehat{\Lambda}_n-\Lambda_0\|_\infty).$$

Combining the above results, we conclude that

$$\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|+\|\widehat{\Lambda}_n-\Lambda_0\|_\infty\leq A_n(n^{-1/2}+\|\mathcal{B}_n-\mathcal{B}_0\|),$$

where A_n is some random variable that may depend on $\mathcal{B}_n$ and satisfies $\sup_{\mathcal{B}_n\in\mathcal{N}_{\epsilon_n}}|A_n| = O_p(1)$.

Step 5. We show that a local maximum of $\mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)$ with respect to $\mathcal{B}_n$ exists in the interior of $\mathcal{N}_{\epsilon_n}$ for large enough n. It suffices to show that $\sup_{\mathcal{B}_n\in\partial\mathcal{N}_{\epsilon_n}}\mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)<\mathbb{P}_n\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\widetilde{\mathcal{B}}_n)$ with probability going to 1 as $n\to\infty$, where $\widetilde{\mathcal{B}}_n=(\widetilde{\psi}_{n2},\dots,\widetilde{\psi}_{nG})$. Let

$$\begin{aligned}
B_n&=\mathbb{P}_n\ell(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)-\mathbb{P}_n\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\widetilde{\mathcal{B}}_n)\\
&=(\mathbb{P}_n-\mathbb{P})\{\ell(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)-\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\widetilde{\mathcal{B}}_n)\}+\mathbb{P}\{\ell(\widehat{\boldsymbol{\theta}}_n,\widehat{\Lambda}_n,\mathcal{B}_n)-\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\mathcal{B}_0)\}\\
&\quad-\mathbb{P}\{\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\widetilde{\mathcal{B}}_n)-\ell(\boldsymbol{\theta}_0,\widetilde{\Lambda}_n,\mathcal{B}_0)\}.
\end{aligned}\tag{8}$$

By Lemma B.1, the first term on the right-hand side of (8) can be written as $C_n n^{-1/2}$ for some variable C_n such that $\sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} |C_n| = o_p(1)$. To evaluate the second term on the right-hand side above, let

$$\xi(\epsilon; \Lambda) = \mathbb{P}\ell\left\{\boldsymbol{\theta}_0 + \epsilon(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \Lambda + \epsilon \int \widehat{h}_\Lambda \,\mathrm{d}\Lambda, \mathcal{B}_0 + \epsilon(\mathcal{B}_n - \mathcal{B}_0)\right\},$$

where $\widehat{h}_\Lambda$ is a step function that jumps at the observed event times, with $\widehat{h}_\Lambda = \mathrm{d}\widehat{\Lambda}_n/\mathrm{d}\widetilde{\Lambda}_n - 1$ at the jump points. The second term of the right-hand side of (8) is equal to $\xi(1; \widetilde{\Lambda}_n) - \xi(0; \widetilde{\Lambda}_n) = \xi'(0; \widetilde{\Lambda}_n) + \xi''(\epsilon; \widetilde{\Lambda}_n)$ for some $\epsilon \in [0, 1]$. Note that $\xi'(0; \widetilde{\Lambda}_n)$ is equal to

$$\begin{aligned}
&\mathbb{P}\left\{\Delta\widehat{h}_\Lambda(\widetilde{T}) + \dot{\Psi}(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_0)\left[\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \int \widehat{h}_\Lambda \,\mathrm{d}\widetilde{\Lambda}_n, \mathcal{B}_n - \mathcal{B}_0\right]\Big/\Psi(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_0)\right\} \\
&\quad = \mathbb{P}\left\{\Delta\widehat{h}_\Lambda(\widetilde{T}) + \dot{\Psi}(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)\left[\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \int \widehat{h}_\Lambda \,\mathrm{d}\Lambda_0, \mathcal{B}_n - \mathcal{B}_0\right]\Big/\Psi(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)\right\} \\
&\qquad + \mathbb{P}\left\{\dot{\Psi}(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_0)\left[\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \int \widehat{h}_\Lambda \,\mathrm{d}\widetilde{\Lambda}_n, \mathcal{B}_n - \mathcal{B}_0\right]\Big/\Psi(\boldsymbol{\theta}_0, \widetilde{\Lambda}_n, \mathcal{B}_0)\right. \\
&\qquad \left. - \dot{\Psi}(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)\left[\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \int \widehat{h}_\Lambda \,\mathrm{d}\Lambda_0, \mathcal{B}_n - \mathcal{B}_0\right]\Big/\Psi(\boldsymbol{\theta}_0, \Lambda_0, \mathcal{B}_0)\right\} \\
&\quad = O_p\{\|\widetilde{\Lambda}_n - \Lambda_0\|_\infty(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| + \|\widehat{h}_\Lambda\|_{\mathrm{V}} + \|\mathcal{B}_n - \mathcal{B}_0\|_{\mathrm{V}})\},
\end{aligned}$$

where $\dot{\Psi}(\boldsymbol{\theta}, \Lambda, \mathcal{B})[\boldsymbol{h}_\theta, H_\Lambda, h_{\mathcal{B}}] = \dot{\boldsymbol{\Psi}}_\theta(\boldsymbol{\theta}, \Lambda, \mathcal{B})^{\mathrm{T}}\boldsymbol{h}_\theta + \dot{\Psi}_\Lambda(\boldsymbol{\theta}, \Lambda, \mathcal{B})[H_\Lambda] + \sum_{g=2}^{G} \dot{\Psi}_{\psi g}(\boldsymbol{\theta}, \Lambda, \mathcal{B})[h_g]$ for $h_{\mathcal{B}} = (h_2, \ldots, h_G)$. The last equality above follows from the mean-value theorem and that the score statistic is mean zero. By standard arguments for the NPMLE, $\|\widetilde{\Lambda}_n - \Lambda_0\|_\infty = O_p(n^{-1/2})$. Also, $\|\widehat{h}_\Lambda\|_{\mathrm{V}} = o_p(1)$ and $\|\mathcal{B}_n - \mathcal{B}_0\|_{\mathrm{V}} = o(1)$, so the right-hand side of the above equation is $o_p(n^{-1/2})$. To evaluate $\xi''(\epsilon; \widetilde{\Lambda}_n)$, we write

$$\xi''(\epsilon; \widetilde{\Lambda}_n) = \{\xi''(\epsilon; \widetilde{\Lambda}_n) - \xi''(0; \widetilde{\Lambda}_n)\} + \{\xi''(0; \widetilde{\Lambda}_n) - \xi''(0; \Lambda_0)\} + \xi''(0; \Lambda_0).$$

Using the mean-value theorem, we can show that the first term on the right-hand side of the above equation is $O_p(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^3 + \|\widehat{h}_\Lambda\|_\infty^3 + \|\mathcal{B}_n - \mathcal{B}_0\|_3^3) + o_p(\|\widetilde{\Lambda}_n - \Lambda_0\|_\infty)$. Following the arguments for the evaluation of $\xi'(0; \widetilde{\Lambda}_n)$, we can show that the second term is $o_p(n^{-1/2})$. Note that the third term is the negative information of the one-dimensional submodel $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \epsilon\boldsymbol{h}_\theta$, $\mathrm{d}\Lambda = (1 + \epsilon h_\Lambda)\,\mathrm{d}\Lambda_0$, and $\mathcal{B} = \mathcal{B}_0 + \epsilon\boldsymbol{h}_{\mathcal{B}}$, where $\boldsymbol{h}_\theta = \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0$, $h_\Lambda = \widehat{h}_\Lambda$, and $\boldsymbol{h}_{\mathcal{B}} = \mathcal{B}_n - \mathcal{B}_0$. Let $\mathcal{H} = \mathbb{R}^d \times L_2[0, \tau]^G$. For any $h \equiv (\boldsymbol{h}_\theta, h_\Lambda, h_{\psi 2}, \ldots, h_{\psi G}) \in \mathcal{H}$, the score statistic of the submodel along direction h is

$$\begin{aligned}
\dot{\ell}[h] = {}& \sum_{g=1}^{G} \pi_g \int Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b})\left[\left\{1 - \frac{\sum_{l=1}^{G} \pi_l \int Q_l(\widetilde{T}, \Delta, \boldsymbol{Y}, \widetilde{\boldsymbol{b}})\,\mathrm{d}\widetilde{\boldsymbol{b}}}{\int Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \widetilde{\boldsymbol{b}})\,\mathrm{d}\widetilde{\boldsymbol{b}}}\right\}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{h}_{\alpha g}\right. \\
& + \Delta\{\boldsymbol{Z}(\widetilde{T})^{\mathrm{T}}\boldsymbol{h}_{\gamma g} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h}_{\eta g} + h_\Lambda(\widetilde{T}) + h_{\psi g}(\widetilde{T})\} \\
& - \int_0^{\widetilde{T}} e^{\boldsymbol{Z}(s)^{\mathrm{T}}\boldsymbol{\gamma}_{0g} + \boldsymbol{\eta}_{0g}^{\mathrm{T}}\boldsymbol{b} + \psi_{0g}(s)}\{\boldsymbol{Z}(s)^{\mathrm{T}}\boldsymbol{h}_{\gamma g} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h}_{\eta g} + h_\Lambda(s) + h_{\psi g}(s)\}\,\mathrm{d}\Lambda_0(s) \\
& \left. + \frac{\boldsymbol{f}_g^{(1)}(\boldsymbol{Y}, \boldsymbol{b})^{\mathrm{T}}\boldsymbol{h}_{Yg}}{f_g(\boldsymbol{Y}, \boldsymbol{b})}\right]\mathrm{d}\boldsymbol{b}\Big/\sum_{g=1}^{G} \pi_g \int Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b})\,\mathrm{d}\boldsymbol{b} \\
\equiv {}& \mathcal{K}(\widetilde{T}, \Delta, \boldsymbol{Y}; h),
\end{aligned}$$

where $Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b}) = Q_g(\mathcal{O}, \boldsymbol{b})$, $\boldsymbol{f}_g^{(1)}(\boldsymbol{Y}, \boldsymbol{b})$ is the derivative of $f_g(\boldsymbol{Y}, \boldsymbol{b})$ with respect to $(\boldsymbol{\beta}_g, \boldsymbol{\sigma}_g^2, \boldsymbol{\xi}_g)$, $\boldsymbol{h}_{Yg} = (\boldsymbol{h}_{\beta g}^{\mathrm{T}}, h_{\sigma g}, \boldsymbol{h}_{\xi g}^{\mathrm{T}})^{\mathrm{T}}$, $(\boldsymbol{h}_{\alpha g}, \boldsymbol{h}_{\beta g}, \boldsymbol{h}_{\sigma g}, \boldsymbol{h}_{\xi g}, \boldsymbol{h}_{\gamma g}, \boldsymbol{h}_{\eta g})$ are the directions that correspond to the parameters $(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g, \boldsymbol{\sigma}_g^2, \boldsymbol{\xi}_g, \boldsymbol{\gamma}_g, \boldsymbol{\eta}_g)$ for $g = 1, \dots, G$, $\boldsymbol{h}_{\alpha G} = \mathbf{0}$, and $h_{\psi 1}(\cdot) = 0$. For $h^{(1)}, h^{(2)} \in \mathcal{H}$, we can write

$$\mathbb{P}\dot{\ell}[h^{(1)}]\dot{\ell}[h^{(2)}] = \boldsymbol{h}_\theta^{(1)\mathrm{T}} \boldsymbol{G}_1(h^{(2)}) + \sum_{g=1}^{G} \int_0^\tau \{h_\Lambda^{(1)}(t) + h_{\psi g}^{(1)}(t)\} G_{2g}(t; h^{(2)})\, \mathrm{d}t,$$

where $\boldsymbol{G}_1(h)$ is some linear function of h, and $G_{2g}(t; h)$ is equal to

$$\begin{aligned}
&\mathrm{E}\biggl\{\frac{\pi_g \int Q_g(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}}{\sum_{l=1}^G \pi_l \int Q_l(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}} f_T(t \mid \boldsymbol{Y}) S_U(t \mid \boldsymbol{Y}) \mathcal{K}(t, 1, \boldsymbol{Y}; h)\biggr\} \\
&\quad - \mathrm{E}\biggl\{I(t \le \widetilde{T}) \frac{\pi_g \int Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b}) e^{\boldsymbol{Z}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{0g} + \boldsymbol{\eta}_{0g}^{\mathrm{T}}\boldsymbol{b} + \psi_{0g}(t)}\, \mathrm{d}\boldsymbol{b}}{\sum_{l=1}^G \pi_l \int Q_l(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}} \mathcal{K}(\widetilde{T}, \Delta, \boldsymbol{Y}; h)\biggr\} \lambda(t) \\
&\quad = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{h}_\theta + \mathrm{E}\biggl\{\frac{\pi_g \int Q_g(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}}{\sum_{l=1}^G \pi_l \int Q_l(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}} f_T(t \mid \boldsymbol{Y}) S_U(t \mid \boldsymbol{Y})\biggr\} h_\Lambda(t) \\
&\qquad + \sum_{k=2}^{G} \mathrm{E}\biggl[\frac{\pi_g \pi_k \int Q_g(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b} \int Q_k(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}}{\{\sum_{l=1}^G \pi_l \int Q_l(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}\}^2} f_T(t \mid \boldsymbol{Y}) S_U(t \mid \boldsymbol{Y})\biggr] h_{\psi k}(t) \\
&\qquad - \sum_{k=1}^{G} \int_0^\tau \{h_\Lambda(s) + h_{\psi k}(s)\} \biggl(I(s \le t) \mathrm{E}\biggl[\pi_g \pi_k f_T(t \mid \boldsymbol{Y}) S_U(t \mid \boldsymbol{Y}) \\
&\qquad \times \frac{\int Q_g(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b} \int Q_k(t, 1, \boldsymbol{Y}, \boldsymbol{b}) e^{\boldsymbol{Z}(s)^{\mathrm{T}}\boldsymbol{\gamma}_{0k} + \boldsymbol{\eta}_{0k}^{\mathrm{T}}\boldsymbol{b} + \psi_{0k}(s)}\, \mathrm{d}\boldsymbol{b}}{\{\sum_{l=1}^G \pi_l \int Q_l(t, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}\}^2}\biggr] + I(t \le s) \mathrm{E}\biggl[\pi_g \pi_k \\
&\qquad \times \frac{\int Q_g(s, 1, \boldsymbol{Y}, \boldsymbol{b}) e^{\boldsymbol{Z}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{0g} + \boldsymbol{\eta}_{0g}^{\mathrm{T}}\boldsymbol{b} + \psi_{0g}(t)}\, \mathrm{d}\boldsymbol{b} \int Q_k(s, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}}{\{\sum_{l=1}^G \pi_l \int Q_l(s, 1, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}\}^2} f_T(s \mid \boldsymbol{Y}) S_U(s \mid \boldsymbol{Y})\biggr] \\
&\qquad - \mathrm{E}\biggl[I(s \le \widetilde{T}) I(t \le \widetilde{T}) \pi_g \pi_k \\
&\qquad \times \frac{\int Q_g(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b}) e^{\boldsymbol{Z}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{0g} + \boldsymbol{\eta}_{0g}^{\mathrm{T}}\boldsymbol{b} + \psi_{0g}(t)}\, \mathrm{d}\boldsymbol{b} \int Q_k(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b}) e^{\boldsymbol{Z}(s)^{\mathrm{T}}\boldsymbol{\gamma}_{0k} + \boldsymbol{\eta}_{0k}^{\mathrm{T}}\boldsymbol{b} + \psi_{0k}(s)}\, \mathrm{d}\boldsymbol{b}}{\{\sum_{l=1}^G \pi_l \int Q_l(\widetilde{T}, \Delta, \boldsymbol{Y}, \boldsymbol{b})\, \mathrm{d}\boldsymbol{b}\}^2}\biggr]\biggr) \\
&\qquad \times \lambda_0(s)\, \mathrm{d}s,
\end{aligned}$$

where $f_T(\cdot \mid \boldsymbol{Y})$ is the conditional density of the survival time T given $\boldsymbol{Y}$, $S_U(\cdot \mid \boldsymbol{Y})$ is the conditional survival function of the censoring time U given $\boldsymbol{Y}$, and $\boldsymbol{a}$ is a d-dimensional vector. Define an inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{H}$ such that

$$\langle h^{(1)}, h^{(2)} \rangle = \boldsymbol{h}_\theta^{(1)\mathrm{T}} \boldsymbol{h}_\theta^{(2)} + \int_0^\tau \biggl\{h_\Lambda^{(1)}(t) h_\Lambda^{(2)}(t) + \sum_{g=2}^{G} h_{\psi g}^{(1)}(t) h_{\psi g}^{(2)}(t)\biggr\}\, \mathrm{d}t,$$

and let $\dot{\ell}^*$ be the adjoint operator of $\dot{\ell}$. By the definition of $\dot{\ell}^*$, $\mathbb{P}\dot{\ell}[h^{(1)}]\dot{\ell}[h^{(2)}] = \langle h^{(1)}, \dot{\ell}^*\dot{\ell}[h^{(2)}] \rangle$, such that

$$\dot{\ell}^*\dot{\ell}[h] = \biggl(\boldsymbol{G}_1(h), \sum_{g=1}^{G} G_{2g}(\cdot; h), G_{22}(\cdot; h), \dots, G_{2G}(\cdot; h)\biggr).$$

On the space $\mathcal{H}$, we define a seminorm $\|h\|_I = \langle h, \dot{\ell}^*\dot{\ell}[h] \rangle^{1/2}$. By Lemma B.4, $\|h\|_I = 0$ implies that $h = 0$, such that $\|\cdot\|_I$ is a norm in $\mathcal{H}$. Clearly, $\|h\|_I \le c_2 \langle h, h \rangle^{1/2}$ for some constant c_2. By the bounded inverse theorem in Banach spaces, we have $\langle h, h\rangle^{1/2} \le c_3\|h\|_I$ for some constant c_3. We conclude that

$$\begin{aligned}\xi''(0; \Lambda_0) &= -\|(\widehat{\boldsymbol\theta}_n - \boldsymbol\theta_0, \widehat{h}_\Lambda, \mathcal{B}_n - \mathcal{B}_0)\|_I^2 \\ &\le -c_3^{-2}\left(\|\widehat{\boldsymbol\theta}_n - \boldsymbol\theta_0\|^2 + \|\widehat{h}_\Lambda\|^2 + \sum_{g=2}^{G}\|\psi_{ng} - \psi_{0g}\|^2\right).\end{aligned}$$

By Donsker properties of the class of $\nu(\boldsymbol\theta, \Lambda, \mathcal{B}; t)$ and the mean-value theorem,

$$\|\widehat{h}_\Lambda\|_\infty = O_p\big(\|\widehat{\boldsymbol\theta}_n - \boldsymbol\theta_0\| + \|\widehat{\Lambda}_n - \Lambda_0\|_\infty + \|\mathcal{B}_n - \mathcal{B}_0\|_2 + n^{-1/2}\big).$$

In addition, a linear expansion argument shows that the third term of (8) is of order up to $\|\widetilde{\mathcal{B}}_n - \mathcal{B}_0\|_\infty^2$. Combining the above results, we have

$$\begin{aligned}B_n &\le D_n n^{-1/2} + E_n\big(\|\mathcal{B}_n - \widetilde{\mathcal{B}}_n\|_3^3 + \|\widetilde{\mathcal{B}}_n - \mathcal{B}_0\|_\infty^2\big) - c_3^{-2}\sum_{g=2}^{G}\|\psi_{ng} - \widetilde{\psi}_{0g}\|^2 \\ &\le D_n n^{-1/2} + c_4 E_n\big(m_n^{-1}\epsilon_n^3 + m_n^{-6}\big) - c_3^{-2}\sum_{g=2}^{G}\|\psi_{ng} - \widetilde{\psi}_{0g}\|^2\end{aligned}$$

for some sequences of positive variables D_n and E_n such that $\sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} D_n = o_p(1)$ and $\sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} E_n = O_p(1)$ and some positive constant c_4. The second inequality holds because by Theorem 5.2 of de Boor [2],

$$\|\psi_{ng} - \widetilde{\psi}_{ng}\|_3^3 = O\left(m_n^{-1}\sum_{s=1}^{m_n}|a_{gs} - \widetilde{a}_{gs}|^3\right) = O\big(m_n^{-1}\epsilon_n^3\big).$$

Suppose that $\mathcal{B}_n \in \partial\mathcal{N}_{\epsilon_n}$. By the same theorem of de Boor [2], $\|\psi_{ng} - \widetilde{\psi}_{ng}\|^2 \ge c_5 m_n^{-1}\epsilon_n^2$ for some g and $c_5 > 0$. Therefore, by choosing ϵ_n such that $\epsilon_n = o(n^{-1/4}m_n^{1/2})$ and

$$\epsilon_n^2 \gg \sup_{\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}} D_n n^{-1/2} m_n + m_n^{-5},$$

we have $P(B_n < 0) \to 1$; the existence of such an ϵ_n with $\epsilon_n = o(m_n^{-3/2})$ is guaranteed under condition (C4). We conclude that there exists a local maximum of $\mathbb{P}_n\ell(\widehat{\boldsymbol\theta}_n, \widehat{\Lambda}_n, \mathcal{B}_n)$ with respect to $\mathcal{B}_n$ in the interior of $\mathcal{N}_{\epsilon_n}$; let $\widehat{\mathcal{B}}_n$ be the maximizer. Note that by Theorem 5.2 of de Boor [2], $\|\psi_{ng} - \widetilde{\psi}_{ng}\|^2 = O(m_n^{-1}\sum_{s=1}^{m_n}|a_{gs} - \widetilde{a}_{gs}|^2) = O(m_n^{-1}\epsilon_n^2)$ for all $\mathcal{B}_n \in \mathcal{N}_{\epsilon_n}$. We have

$$\begin{aligned}\|\widehat{\boldsymbol\theta}_n - \boldsymbol\theta_0\|^2 + \|\widehat{\Lambda}_n - \Lambda_0\|_\infty^2 + \|\widehat{\mathcal{B}}_n - \mathcal{B}_0\|^2 &= O_p\big(n^{-1} + \|\widehat{\mathcal{B}}_n - \mathcal{B}_0\|^2\big) \\ &= O_p\big(m_n^{-1}\epsilon_n^2 + m_n^{-6}\big) = o_p\big(n^{-1/2}\big). \qquad \square\end{aligned}$$

PROOF OF THEOREM 4.2. Let $\dot{\boldsymbol\ell}_\theta$ be the score statistic for $\boldsymbol\theta$, $\dot\ell_\Lambda[h_\Lambda]$ be the score statistic for Λ along the submodel $\Lambda + \epsilon\int h_\Lambda\, d\Lambda$, and $\dot\ell_{\psi g}[h_{\psi g}]$ be the score statistic for ψ_g along the submodel $\psi_g + \epsilon h_{\psi g}$ $(g = 2, \ldots, G)$. For a set of functions $\boldsymbol h \equiv (h_1, \ldots, h_d)$, let $\dot{\boldsymbol\ell}_\Lambda[\boldsymbol h] = (\dot\ell_\Lambda[h_1], \ldots, \dot\ell_\Lambda[h_d])^{\mathrm T}$ and $\dot{\boldsymbol\ell}_{\psi g}[\boldsymbol h] = (\dot\ell_{\psi g}[h_1], \ldots, \dot\ell_{\psi g}[h_d])^{\mathrm T}$. Let $\widetilde{\boldsymbol h}_\Lambda$ and $\widetilde{\boldsymbol h}_{\psi g}$ be the least favorable directions for the nonparametric functions, such that $(\widetilde{\boldsymbol h}_\Lambda, \widetilde{\boldsymbol h}_{\psi 1}, \ldots, \widetilde{\boldsymbol h}_{\psi G}) = \arg\min_{\boldsymbol h_\Lambda, \boldsymbol h_{\psi 2}, \ldots, \boldsymbol h_{\psi G}} \mathbb{P}\|\dot{\boldsymbol\ell}_\theta - \dot{\boldsymbol\ell}_\Lambda[\int \boldsymbol h_\Lambda\, d\Lambda_0] - \sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}[\boldsymbol h_{\psi g}]\|^2$, where the integration in the second term in the norm is carried out componentwise. The existence of $\widetilde{\boldsymbol h}_\Lambda$ and $\widetilde{\boldsymbol h}_{\psi g}$ follows from the invertibility of the information operator, established in Step 5 of the proof of Theorem 4.1. In addition, from the expressions of $\dot\ell^*\dot\ell$ given in Step 5 of the proof of Theorem

4.1 and condition (C6), each component of $\widetilde{\boldsymbol h}_{\psi g}$ is continuously differentiable up to the third order. Let $\widetilde{\boldsymbol h}_{n,\psi g}$ be the (componentwise) projection of $\widetilde{\boldsymbol h}_{\psi g}$ onto the sieve space, such that $\|\widetilde{\boldsymbol h}_{n,\psi g}-\widetilde{\boldsymbol h}_{\psi g}\|_\infty = O(m_n^{-3})$. By the definition of the sieve NPMLE, $\mathbb{P}_n\dot{\boldsymbol\ell}_\theta(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)=\mathbf{0}$, $\mathbb{P}_n\dot{\boldsymbol\ell}_\Lambda(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\widehat\Lambda_n]=\mathbf{0}$ and $\mathbb{P}_n\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{n,\psi g}]=\mathbf{0}$. Note that

$$
\begin{aligned}
&\mathbb{P}_n\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{\psi g}]\\
&\quad=\mathbb{P}_n\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{n,\psi g}]+\mathbb{P}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}-\widetilde{\boldsymbol h}_{n,\psi g}]\\
&\qquad+(\mathbb{P}_n-\mathbb{P})\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{\psi g}-\widetilde{\boldsymbol h}_{n,\psi g}]\\
&\qquad+\mathbb{P}\{\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{\psi g}-\widetilde{\boldsymbol h}_{n,\psi g}]-\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}-\widetilde{\boldsymbol h}_{n,\psi g}]\}.
\end{aligned}
$$

The first two terms of the right-hand side above are zero. By Lemma B.1, the class of $\dot\ell_{\psi g}(\boldsymbol\theta,\Lambda,\mathcal B)[h]$ is Donsker, so that the third term is $o_p(n^{-1/2})$. By the mean-value theorem, Theorem 4.1, and condition (C4), the fourth term is $o_p(n^{-1/2})$. Obviously, $\mathbb{P}\dot{\boldsymbol\ell}_\theta(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)=\mathbf{0}$, $\mathbb{P}\dot{\boldsymbol\ell}_\Lambda(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\Lambda_0]=\mathbf{0}$, and $\mathbb{P}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}]=\mathbf{0}$. We have

$$
\begin{aligned}
&n^{1/2}(\mathbb{P}_n-\mathbb{P})\biggl\{\dot{\boldsymbol\ell}_\theta(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)-\dot{\boldsymbol\ell}_\Lambda(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)\Bigl[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\widehat\Lambda_n\Bigr]\\
&\qquad-\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{\psi g}]\biggr\}\\
&\quad=-n^{1/2}\mathbb{P}\biggl\{\dot{\boldsymbol\ell}_\theta(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)-\dot{\boldsymbol\ell}_\Lambda(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)\Bigl[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\widehat\Lambda_n\Bigr]\\
&\qquad-\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\widehat{\boldsymbol\theta}_n,\widehat\Lambda_n,\widehat{\mathcal B}_n)[\widetilde{\boldsymbol h}_{\psi g}]\\
&\qquad-\dot{\boldsymbol\ell}_\theta(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)+\dot{\boldsymbol\ell}_\Lambda(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)\Bigl[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\Lambda_0\Bigr]\\
&\qquad+\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}]\biggr\}+o_p(1).
\end{aligned}\tag{9}
$$

By Lemma B.1, the class

$$
\biggl\{\dot{\boldsymbol\ell}_\theta(\boldsymbol\theta,\Lambda,\mathcal B)^{\mathrm T}\boldsymbol v-\dot\ell_\Lambda(\boldsymbol\theta,\Lambda,\mathcal B)[H_\Lambda]-\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta,\Lambda,\mathcal B)[\widetilde{\boldsymbol h}_{\psi g}]:\\
(\boldsymbol\theta,\Lambda,\mathcal B)\in\Xi,\|\boldsymbol v\|\le 1,\|H_\Lambda\|_{\mathrm V}\le 1\biggr\}
$$

is Donsker. Therefore, the left-hand side of (9) is equal to

$$
n^{1/2}(\mathbb{P}_n-\mathbb{P})\biggl\{\dot{\boldsymbol\ell}_\theta(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)-\dot{\boldsymbol\ell}_\Lambda(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)\Bigl[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\Lambda_0\Bigr]-\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}]\biggr\}\\
+o_p(1),
$$

which converges in distribution to $\mathrm{N}(\mathbf{0},\widetilde{\boldsymbol I})$, where

$$
\widetilde{\boldsymbol I}\equiv\mathbb{P}\biggl\{\dot{\boldsymbol\ell}_\theta(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)-\dot{\boldsymbol\ell}_\Lambda(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)\Bigl[\int\widetilde{\boldsymbol h}_\Lambda\,\mathrm{d}\Lambda_0\Bigr]-\sum_{g=2}^{G}\dot{\boldsymbol\ell}_{\psi g}(\boldsymbol\theta_0,\Lambda_0,\mathcal B_0)[\widetilde{\boldsymbol h}_{\psi g}]\biggr\}^{\otimes 2}
$$

is the efficient information matrix for $\boldsymbol{\theta}$. By the Taylor series expansion, Theorem 4.1 and the definition of $\widetilde{\boldsymbol{h}}_\Lambda$ and $\widetilde{\boldsymbol{h}}_{\psi g}$ $(g = 2, \ldots, G)$, the right-hand side of (9) is

$$\begin{aligned} &-n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbb{P}\biggl\{\ddot{\boldsymbol{\ell}}_{\theta\theta} - \ddot{\boldsymbol{\ell}}_{\Lambda\theta}\Bigl[\int \widetilde{\boldsymbol{h}}_\Lambda \,\mathrm{d}\Lambda_0\Bigr] - \sum_{g=2}^{G} \ddot{\boldsymbol{\ell}}_{\psi g\theta}[\widetilde{\boldsymbol{h}}_{\psi g}]\biggr\} + o_p(1) \\ &\quad = n^{1/2}\widetilde{\boldsymbol{I}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(1), \end{aligned}$$

where $\ddot{\boldsymbol{\ell}}_{\theta\theta}$, $\ddot{\boldsymbol{\ell}}_{\Lambda\theta}$ and $\ddot{\boldsymbol{\ell}}_{\psi g\theta}$ are the derivatives of $\dot{\boldsymbol{\ell}}_\theta$, $\dot{\boldsymbol{\ell}}_\Lambda$ and $\dot{\boldsymbol{\ell}}_{\psi g}$ with respect to $\boldsymbol{\theta}$, respectively. As established in Step 5 in the proof of Theorem 4.1, the information operator is invertible, so the efficient information matrix is invertible. We conclude that $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d \mathrm{N}(\boldsymbol{0}, \widetilde{\boldsymbol{I}}^{-1})$. Because $\widehat{\boldsymbol{\theta}}_n$ is an asymptotically linear estimator with the influence function lying in the space spanned by the score functions, $\widehat{\boldsymbol{\theta}}_n$ is asymptotically efficient [1]. □

APPENDIX B: USEFUL LEMMAS

In this Appendix, we present four lemmas that are useful for the proofs of Theorems 4.1 and 4.2. The proofs of the lemmas are given in Section S3 of the Supplementary Material [21].

LEMMA B.1. *For any finite K, the classes of functions*

$$\begin{aligned} \mathcal{G}_1 &= \{\log\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B}) : (\boldsymbol{\theta}, \Lambda, \mathcal{B}) \in \Xi_K\}, \\ \mathcal{G}_2 &= \biggl\{\frac{\dot{\boldsymbol{\Psi}}_\theta(\boldsymbol{\theta}, \Lambda, \mathcal{B})^{\mathrm{T}}\boldsymbol{v}}{\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})} : (\boldsymbol{\theta}, \Lambda, \mathcal{B}) \in \Xi_K, \|\boldsymbol{v}\| < K\biggr\}, \\ \mathcal{G}_3 &= \biggl\{\frac{\dot{\Psi}_\Lambda(\boldsymbol{\theta}, \Lambda, \mathcal{B})[H_\Lambda]}{\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})} : (\boldsymbol{\theta}, \Lambda, \mathcal{B}) \in \Xi_K, \|H_\Lambda\|_{\mathrm{V}} < K\biggr\}, \\ \mathcal{G}_{4g} &= \biggl\{\frac{\dot{\Psi}_{\psi g}(\boldsymbol{\theta}, \Lambda, \mathcal{B})[h_{\psi g}]}{\Psi(\boldsymbol{\theta}, \Lambda, \mathcal{B})} : (\boldsymbol{\theta}, \Lambda, \mathcal{B}) \in \Xi_K, \|h_{\psi g}\|_{\mathrm{V}} < K\biggr\} \end{aligned}$$

are Donsker.

LEMMA B.2. *Under conditions* (C1)–(C3) *and* (C5), *the latent-class model given by* (1)–(3) *is locally identifiable.*

LEMMA B.3. *Consider the following normal mixture model. Let $\boldsymbol{W}$ be a set of covariates and C be a latent class indicator with distribution specified by* (1). *For $g = 1, \ldots, G$, let $\boldsymbol{Y}_g \sim \mathrm{N}(\boldsymbol{\mu}_g, \boldsymbol{\Omega}_g)$, where $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G)$ are vectors of mean parameters, and $(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_G)$ are covariance matrices. The observed outcome variable is $\boldsymbol{Y} = \sum_{g=1}^G I(C = g)\boldsymbol{Y}_g$. Let $(\boldsymbol{\mu}_{0g}, \boldsymbol{\Omega}_{0g})$ be the true values of $(\boldsymbol{\mu}_g, \boldsymbol{\Omega}_g)$. If $(\boldsymbol{\mu}_{01}, \boldsymbol{\Omega}_{01}), \ldots, (\boldsymbol{\mu}_{0G}, \boldsymbol{\Omega}_{0G})$ are distinct and the components of $\boldsymbol{W}$ are linearly independent, then the score statistic along any submodel is nonzero.*

LEMMA B.4. *Under conditions* (C1)–(C3) *and* (C5), *the score statistic along any one-dimensional submodel for the latent-class model given by* (1)–(3) *is nonzero.*

Acknowledgments. The authors thank the staff and participants of the ARIC study for their important contributions.

Funding. This work was supported by a research grant from the Hong Kong Polytechnic University (P0030124), the Hong Kong Research Grants Council grant PolyU 253042/18P and the National Institutes of Health awards R01-HL149683 and R01-HG009974. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, HHSN26820170000 4I).

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric latent-class models for multivariate longitudinal and survival data." (DOI: 10.1214/21-AOS2117SUPP; .pdf). We present additional regularity conditions, the proofs of Theorem 4.3 and Lemmas B.1–B.4, additional simulation results and additional real data analysis results.

REFERENCES

[1] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[2] de Boor, C. (1976). Splines as linear combinations of B-splines. A survey. In *Approximation Theory*, *II* (*Proc. Internat. Sympos.*, *Univ. Texas*, *Austin*, *Tex.*, 1976) 1–47. MR0467092

[3] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–22. MR0501537 https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[4] Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480. https://doi.org/10.1093/biostatistics/1.4.465

[5] Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Amer. Statist. Assoc.* **97** 53–65. MR1947272 https://doi.org/10.1198/016214502753479220

[6] Liu, L., Ma, J. Z. and O'Quigley, J. (2008). Joint analysis of multi-level repeated measures data and survival: An application to the end stage renal disease (ESRD) data. *Stat. Med.* **27** 5679–5691. MR2573776 https://doi.org/10.1002/sim.3392

[7] Liu, Q. and Pierce, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81** 624–629. MR1311107 https://doi.org/10.1093/biomet/81.3.624

[8] Liu, Y., Lin, Y., Zhou, J. and Liu, L. (2020). A semi-parametric joint latent class model with longitudinal and survival data. *Stat. Interface* **13** 411–422. MR4091806 https://doi.org/10.4310/SII.2020.v13.n3.a10

[9] Liu, Y., Liu, L. and Zhou, J. (2015). Joint latent class model of survival and longitudinal data: An application to CPCRA study. *Comput. Statist. Data Anal.* **91** 40–50. MR3368004 https://doi.org/10.1016/j.csda.2015.05.007

[10] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **44** 226–233. MR0676213 https://doi.org/10.1111/j.2517-6161.1982.tb01203.x

[11] Ma, Y. and Wang, Y. (2012). Efficient distribution estimation for data with unobserved sub-population identifiers. *Electron. J. Stat.* **6** 710–737. MR2988426 https://doi.org/10.1214/12-EJS690

[12] Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22** 712–731. MR1292537 https://doi.org/10.1214/aos/1176325492

[13] Proust-Lima, C., Séne, M., Taylor, J. M. G. and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.* **23** 74–90. MR3190688 https://doi.org/10.1177/0962280212445839

[14] Schumaker, L. L. (2007). *Spline Functions*: *Basic Theory*, 3rd ed. Cambridge Univ. Press, Cambridge. MR2348176 https://doi.org/10.1017/CBO9780511618994

[15] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014 https://doi.org/10.1214/aos/1176344136

[16] Shen, R., Olshen, A. B. and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912. https://doi.org/10.1093/bioinformatics/btp543

[17] THE ARIC INVESTIGATORS (1989). The atherosclerosis risk in communities (ARIC) study: Design and objectives. *Am. J. Epidemiol.* **129** 687–702. https://doi.org/10.1093/oxfordjournals.aje.a115184

[18] TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

[19] VARADHAN, R. and ROLAND, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* **35** 335–353. MR2418745 https://doi.org/10.1111/j.1467-9469.2007.00585.x

[20] WANG, Y., GARCIA, T. P. and MA, Y. (2012). Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial. *J. Amer. Statist. Assoc.* **107** 1324–1338. MR3036398 https://doi.org/10.1080/01621459.2012.699353

[21] WONG, K. Y., ZENG, D. and LIN, D. Y. (2022). Supplement to "Semiparametric latent-class models for multivariate longitudinal and survival data." https://doi.org/10.1214/21-AOS2117SUPP

[22] XU, C., BAINES, P. D. and WANG, J.-L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* **15** 731–744. https://doi.org/10.1093/biostatistics/kxu015

[23] XU, J. and ZEGER, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 375–387. MR1856332 https://doi.org/10.1111/1467-9876.00241

[24] ZENG, D. and CAI, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann. Statist.* **33** 2132–2163. MR2211082 https://doi.org/10.1214/009053605000000480

[25] ZENG, D. and LIN, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 507–564. MR2370068 https://doi.org/10.1111/j.1369-7412.2007.00606.x

[26] ZENG, D. and LIN, D. Y. (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statist. Sinica* **20** 871–910. MR2682647

[27] ZENG, D., LIN, D. Y. and LIN, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statist. Sinica* **18** 355–377. MR2384992

[28] ZENG, D., MAO, L. and LIN, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103** 253–271. MR3509885 https://doi.org/10.1093/biomet/asw013