Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Teng Fei Date

Latent Class Methods for Complex Chronic Disease Data

By

Teng Fei

Doctor of Philosophy

Biostatistics

Limin Peng, Ph.D.
Advisor

John Hanfelt, Ph.D.
Committee Member

Yijian (Eugene) Huang, Ph.D.
Committee Member

James J. Lah, MD, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Latent Class Methods for Complex Chronic Disease Data

By

Teng Fei
B.Sc. City University of Hong Kong, 2016

Advisor: Limin Peng, Ph.D.

An Abstract of
a dissertation submitted to the Faculty of the
James T. Laney Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2021

Abstract

Latent Class Methods for Complex Chronic Disease Data

By Teng Fei

Latent class analysis (LCA) is a powerful but intuitive data-driven tool to characterize the heterogeneity of chronic disease phenotypes. Motivated by the different research questions on neurodegenerative disease, we develop novel latent class methods in this dissertation, aiming to overcome various limitations of existing methods, such as estimation bias, restrictive parametric model assumptions, and expensive computation. We apply our methods to analyze the Uniform Data Set (UDS) for a cohort with mild cognitive impairment (MCI).

In the first topic, we propose a novel structural time-dependent competing risks model, which is sensibly formulated to assess the association between latent classes of baseline cognitive performance in MCI patients and their subsequent neuropathological features. We develop a two-step estimation procedure which circumvents latent class membership assignment and is rigorously justified in terms of accounting for the uncertainty in classifying latent classes. The new method also properly addresses the complications for competing risks outcomes, such as censoring and missing failure types. Our application on UDS uncovers a detailed picture of the neuropathological relevance of the baseline MCI subgroups.

Next, we develop a semi-parametric LCA framework with proportional hazards submodel to investigate the heterogeneity of baseline patient characteristics and its implications for survival. We novelly utilize non-parametric maximum likelihood estimator (NPMLE) to derive estimation procedure and asymptotic theories, which addresses considerable complications due to the presence of infinite-dimensional baseline hazard component in the finite mixture framework. The framework also flexibly considers class-specific covariate effects on both class membership and hazard. We apply the method on the UDS data, which reveals MCI subgroups with distinctive baseline factors on class-specific survival, and further helps to improve the prediction of survival based on baseline covariates.

In the third topic, we study a finite mixture framework for joint longitudinal and survival data, which effectively incorporates semi-parametric generalized estimating equation (GEE) and proportional hazards submodels. Critically, we account for the within-class correlation between longitudinal trajectories and time-to-event by treating longitudinal outcomes as time-dependent internal covariates for the survival submodel. We derive unbiased estimator which properly addresses challenging data characteristics, including time-dependent internal covariates and informative censoring of longitudinal observations due to a terminal event. Our application on the UDS data recognizes multiple latent MCI subgroups with distinguishable neurodegeneration trajectories and survival probability curves.

Latent Class Methods for Complex Chronic Disease Data

By

Teng Fei
B.Sc. City University of Hong Kong, 2016

Adviser: Limin Peng, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2021

Acknowledgements

I would like to express my sincere thankfulness to my advisor, Dr. Limin Peng, for her marvelous guidance at each stage of my PhD career. When I was hesitating on which PhD program I should attend, it was Dr. Peng who convinced me to start this wonderful 60-month experience at Emory. When I was exploring different research areas, it was Dr. Peng who opened the door of latent class analysis and survival analysis for me. During my job hunting, it was again Dr. Peng who provided most considerate support in preparing my application package and job talk. I enjoyed each weekly meeting with Dr. Peng, which always extracted the full intellectual potential of me to address the challenges in our research, giving me more confidence but also keeping me humble in the ocean of knowledge. The experience working with Dr. Peng is certainly one of my biggest fortunes in my career development.

It is my privilege to have Dr. John Hanfelt, Dr. Eugene Huang and Dr. James Lah on my committee. I am very grateful that Dr. Hanfelt establishes the latent class research group and starts a number of exciting research projects for us. Particularly, Dr. Hanfelt generously offered me an opportunity to finalize a research project with Dr. Kari Hart, which helped me to gain crucial theoretical and computational experiences for my own dissertation research. Dr. James Lah is the very important person in our research group, who always provides useful insights on the application side of statistical methods. I cannot thank Dr. Lah enough for his ideas in finding interesting research problems and sensible ways of application. I am also very lucky to have Dr. Eugene Huang on my committee, who is extremely knowledgeable in advanced survival analysis and provided concise and in-depth discussions when I encountered difficulties.

I am extremely lucky to be able to collaborate with Dr. Tianwei Yu and Dr. Reneé Moore on a number of projects aside my dissertation research. Special thanks to Dr. Yu for leading my earliest methodological research projects at Emory, growing my sense of conducting

research and progressing under challenges, and providing invaluable career advice to me. Enormous thanks to Dr. Moore for nurturing me as a collaborative biostatistician, creating leadership and mentoring opportunities for me, teaching me about diversity, equity and inclusion in academia, and supporting me through the job hunting.

I would like to thank the Department of Biostatistics and Bioinformatics at Emory University for creating such a wonderful PhD experience for me during the past 60 months. Such a great experience has been made real by the magnificent efforts from department chairs Dr. Lance Waller, Dr. John Hanfelt, and Dr. Robert Krafty, PhD program directors Dr. Limin Peng and Dr. Steve Qin, and a lovely team including Mary Abosi, Angela Guinyard, Joy Hearn, Melissa Sherrer, and Bob Waggoner. I also want to thank my fellow PhD students, especially my office mates in Grace Crum Rollins 323 and 351, my mentees and mentors, for the great discussions we have had. All of you have made this journey so special. All the best to Dr. Robert Krafty and the Department for the bright future.

Emory created so many great memories in my life. The most important one is that I met my wife and my best friend, Xin, in the very same PhD program. Together with Xin, we stepped over one milestone after another. We will keep doing the research we like and keep enjoying the promising life we live.

Last but not least, I want to make a wish. I wish we can have world peace, and I wish our planet will be filled with laughter and happiness.

- Teng on June 16 2021, in the middle of global COVID-19 pandemic

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Latent class analysis (LCA) is a powerful statistical method to reveal the structure of heterogeneous disease syndromes. Typically, LCA fits likelihood-based models for the observed clinically-relevant variables, which are believed to be a manifestation of underlying classes, to provide inferences that guide the clustering of patients into latent classes or subgroups.

Latent class analysis facilitates better understanding of disease heterogeneity in multiple perspectives. First, the flexibility of LCA framework allows utilizing various formats of observed variables from complex chronic disease data, including cross-sectional covariates, longitudinal biomarkers, and time-to-event data. Consequently, researchers are able to compare data-driven latent classes by class-specific distributions or trajectories, which further helps interpret clinical relevance. In addition, the clinical interpretation contributes to justifying established or revealing new disease subpopulations. The disease subgroups determined by an established LCA model, moreover, can be further featured in structural downstream analysis that investigates whether and how the subgroups are related to other phenotypes that are not used for the classification. Studying the association of established subgroups and phenotypes as new evidences can provide justification or new insight with regard to the previously established LCA model.

1.2 Motivating example

The understanding of disease heterogeneity is evolving as the disease data collection becomes more extensive and complicated. For instance, when the term mild cognitive impairment (MCI) was introduced in the 1990s, it was defined by homogeneous memory-related criteria including memory complaint, objective evidence of abnormal memory for age, normal general cognitive function, and normal activities of daily living (Petersen et al., 1999). As more data were collected, high heterogeneity was discovered in the clinical presentations of MCI. Accordingly, a new four-subtype classification system for MCI became well recognized: Amnestic MCI, Multidomain MCI-Amnestic, Multidomain MCI-Non-Amnestic, or Single Non-Memory MCI (Winblad et al., 2004). This popular four-class system was established on the number of affected cognitive domains and whether the affected cognitive domains are memory-related, thus the classification mainly relied on patients' cognitive characteristics. Later on, with other clinically-relevant characteristics being collected, such as neuropsychiatric features and functional assessments on activities of daily living, together with cognitive features, it became unstraightforward to naively rely on one or two variables for classification. Instead, LCA was conducted to define data-driven subgroups of MCI (Hanfelt et al., 2011). In more recent investigation on MCI heterogeneity, longitudinal biomarkers and competing risks were considered to assist justifying the characteristics of different data-driven MCI subgroups (Hanfelt et al., 2018).

The Uniform Data Set (UDS), collected by 39 past and present NIH-funded Alzheimer's disease centers (ADCs) which are coordinated by National Alzheimer's Coordinating Center (NACC), provides an ideal platform for the investigation. The dataset consists of information for thousands of participants with a baseline diagnosis of MCI, including longitudinal cognitive, functional and neuropsychiatric characteristics measured at each visit, time to event outcomes such as time to the diagnosis of dementia or time to death. In addition, the associated neuropathology data set further provides brain autopsy data for a subset of UDS participants.

Specifically, cognitive tests were conducted based on the UDS Neuropsychological Battery

Categories	Manifest Variables	Interpretations	Data Characteristics
Demographics	age, sex, race		baseline categorical
	decades of smoking	risk factor of cognitive decline	baseline continuous
	MMSE at first visit	overall cognitive status	baseline continuous
	elevated Hachinski score	cerebrovascular disease indicator	baseline binary
Cognitive tests	MMSE	overall cognitive status	longitudinal continuous
	trail making test	executive functioning	
	Boston naming test	language	
	category fluency	language	
	digit span	attention	
	digit symbol	attention	
	logical memory	episodic memory	
	story A	episodic memory	
Functional	IADLs	functional abilities on daily tasks	longitudinal count
Neuropsychiatric	GDS	depression indicator	longitudinal binary
	NPI-Q	bahavioral changes	longitudinal count
Time-to-event	time-to-death	time to death since first visit	time-to-event subject to censoring
	time-to-dementia	time to diagnosis of dementia since first visit	time-to-event subject to censoring
Neuropathological	CERAD score	density of neocortical neuritic plaques	competing risks

Table 1.1: Manifest variables available from Uniform Data Set.

version 2.0, including mini-mental state exam (Folstein et al., 1975, MMSE), trail-making test (Reitan, 1958), Boston naming test (Williams et al., 1989), category fluency (Cooper et al., 2004), digit span (Heinly et al., 2005), digit symbol, logical memory and story A (Wechsler, 1945); functional abilities were evaluated by quantifying the performance on instrumental activities of daily living (Pfeffer et al., 1982, IADLs) over the past four weeks; neuropsychiatric information was assessed by the Geriatric Depression Scale (Sheikh and Yesavage, 1986, GDS); neuropathological phenotypes determined by brain autopsies included the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) scores (Welsh et al., 1991), which measures the density of neocortical neuritic plaques. A more detailed summary can be found in Table 1.1 for the manifest variables available from UDS. Thanks to its richness and complexity of data types for numerous clinically-relevant characteristics of MCI patients, the UDS data enables in-depth investigations on MCI heterogeneity.

Challenges from the perspective of data, however, also prevail when analyzing UDS. For example, the neuropathological features, such as the density of neuritic plaques, were progressive but only observable from brain autopsies after death. This indicates that the neuropathological features are inseparable from the survival outcomes of MCI patients, thus requires careful coding to maintain both autopsy and time-to-event information. As another example, the longitudinal observations are observed at the baseline visit and follow-up visits. The frequency of visits, however, might be correlated with the progression of dementia. Moreover, the longitudinal observations are also censored by either drop-out or death, where the failure process is also likely to be correlated with disease progression. Therefore, it is important to account for informative visit, drop-out, or censoring by terminal event when conducting estimation using the longitudinal data. Data challenges will be discussed in detail for each topic of this dissertation.

1.3 Outline

In this dissertation work, we propose novel latent class methods with following aims. First, the methods should better address the specific research questions motivated by the UDS data. In addition, complex data characteristics, as described in Section 1.2, need to be accounted for. Furthermore, the proposed methods should be carefully designed to address the computational challenges.

In Chapter 2, we are interested in investigating how the latent classes defined by baseline patient characteristics are related to the phenotypes extracted from brain autopsy. To address this research interest, we propose and study a time-dependent structural model to evaluate the association between latent classes and competing risk outcomes that are subject to missing failure types. We develop a two-step estimation procedure which circumvents latent class membership assignment and is rigorously justified in terms of accounting for the uncertainty in classifying latent classes. The new method also properly addresses the realistic complications for competing risks outcomes, including random censoring and missing failure types. The asymptotic properties of the resulting estimator are established. Given that the standard bootstrapping inference is not feasible in the current problem setting, we develop analytical inference procedures, which are easy to implement. Our simulation studies demonstrate the advantages of the proposed method over benchmark approaches. We present an application to the MCI data from UDS, which uncovers a detailed picture of the neuropathological relevance of the baseline MCI subgroups.

In Chapter 3, we aim to investigate the heterogeneity of baseline patient characteristics and its implications for disease progression reflected by time to diagnosis of dementia. Correspondingly, we develop a semi-parametric LCA framework for time-to-event data. In the proposed framework, we adapt infinite-dimensional baseline hazard function, and class-specific covariate effects for both class membership and survival, to enable higher flexibility in capturing heterogeneous data patterns. We novelly utilize non-parametric maximum likelihood estimator (NPMLE) technique to address the challenges caused by the entanglement of finite and infinite dimensional parameters in our model, and derive a stable

expectation-maximization (EM) algorithm that is robust to different initialization schemes. We also establish rigorous asymptotic theories for the proposed estimator. We apply the method on the UDS data, which reveals MCI subgroups with distinctive risk factors.

Chapter 4 is motivated by the research plan of a joint analysis of longitudinal and survival information from the UDS. Accordingly, we extend existing semi-parametric LCA frameworks to jointly handle longitudinal and survival data by generalized estimating equation (GEE) and proportional hazard model. The proposed approach regards longitudinal outcomes as time-dependent covariates of the class-specific survival model, which naturally accounts for the within-class correlation of longitudinal and survival outcomes. We also address the informative censoring of longitudinal observation caused by a terminal event such as death, by inverse probability weighting technique. We derive unbiased estimating equations and the corresponding iterative algorithms. Our numerical experience indicates reliable performance of the algorithm under non-informative initialization. Our application on UDS recognizes four MCI latent subgroups with clinically interpretable differences in cognitive trajectories and time-to-death distributions.

Chapter 2

A Time-Dependent Structural Model Between Latent Classes and Competing Risks Outcomes

2.1 Introduction

In chronic disease studies, data on multiple phenotypes are often collected to provide a comprehensive view of disease manifestation and progression. Investigation across different disease phenotypes can shed valuable insight of disease heterogeneity, which further improves disease diagnosis and management. For example, the Uniform Data Set (UDS) has collected various baseline cognitive characteristics for a cohort of mild cognitive impairment (MCI) patients. Upon occurrences of death, neuropathological findings, such as the density of neuritic plaques, were recorded when brain autopsies were available. It is of interest how the heterogeneity of MCI presented at baseline is associated with the progression of neuropathological features that reflect the brain etiology of the Alzheimer disease.

To address such an interest, we require a workflow which first explains the heterogeneous structure of baseline MCI phenotypes, then captures the association between baseline heterogeneity and autopsy data. Latent class analysis (LCA) is a powerful tool, for the first

step of the workflow, to reveal the structure of heterogeneous disease syndromes. Assuming the observed phenotypes are a manifestation of latent classes or subgroups, LCA can provide inferences to guide the clustering of subjects into subgroups. Traditionally, the MCI subtypes were defined based on the number and type of affected cognitive domains (Winblad et al., 2004). More recently, data-driven subgroups of MCI by LCA were developed based on the multi-phenotype data collected from MCI patients in UDS (Hanfelt et al., 2011).

For the second step of workflow, an intuitive and classic approach is to use the results from LCA to assign a latent class membership to each subject, and then regress the neuropathological phenotype of interest over the assigned latent class membership. This type of approach was referred to as "three-step" methods (Clogg, 1995; Bolck et al., 2004). An apparent issue with the three-step methods is that the assigned latent class memberships may not be the true ones defined by LCA and this can lead to biased estimation in the subsequent regression analysis. Various techniques have been proposed to mitigate the bias from the three-step methods (Bandeen-Roche et al., 1997; Bolck et al., 2004; Wang et al., 2005; Vermunt, 2010; Bakk and Vermunt, 2016; Dias and Vermunt, 2008; Bakk et al., 2013, for example), but most maintained the assignment step and few considered directly incorporating the variability of the LCA parameter estimates into downstream analysis. Alternatively, one may consider joint modeling, which includes both models used to define latent classes and to regress over the latent classes, leading to the so-called "one-step" methods (Proust-Lima et al., 2009; Rowley et al., 2017; Elliott et al., 2020; Hart et al., 2020, for example). Due to the nature of joint modeling, however, the resulting latent classes depend on both baseline and distal information, thus cannot be interpreted as revealing *baseline* heterogeneity. To ensure that the obtained latent classes reflect *baseline* heterogeneity but also to account for the misclassification issue, researchers have investigated "two-step" approaches. Instead of estimating all parameters in the joint models simultaneously, the two-step methods fit a latent class model as step 1, then estimate the regression model that evaluates the effect of the unobservable latent classes on the phenotype of interest as step 2, while accounting for the variability of the step 1 results. As commented in Bakk and Kuha (2018), two-step methods, though less efficient than one-step methods, can greatly relieve the computational burden

and avoid the ambiguous latent class interpretation involved in the one-step methods.

In practice, investigating the association of latent classes with a separate phenotype faces further complexities originated from data collection. In the UDS example, the progressive neuropathological features were measured from brain autopsies that were only available in deceased subjects, meaning they are inseparable from the survival outcomes of MCI patients. Each feature is presented as one of mutually exclusive forms, such as different levels of neuritic plaque density, while sharing the same survival component. Therefore, one natural way to utilize all useful information is to formulate observed neuropathological features as competing risk outcomes, for example, time to death with none or sparse, moderate, or frequent neuritic plaque. Moreover, only a proportion of study participants agreed brain donation after death. As a result, the autopsy data are missing in the deceased subjects without donation agreement. Under the competing risks formulation, this causes the so-called missing failure type problem. Strategies have been proposed to address the missing failure type issue, such as inverse probability weighting (IPW) technique (Ma et al., 2018, for example) and imputation of missing event type (Schaubel and Cai, 2006, for example).

Despite the availability of a wide range of methods, to the best of our knowledge, very limited attention was paid to the real data issues exemplified above. To more effectively investigate the association between the *baseline* cognitive heterogeneity of MCI patients and time-to-death with types of progressive neuropathological features (levels of neuritic plaques, for example) in the presence of missing failure type issue, we study a structural model between latent classes and competing risks outcomes. This model mimics the semi-parametric direct binomial regression model for competing risks (Scheike et al., 2008). By allowing for time-dependent latent class effects, it provides a flexible platform to explore the interested association. Our estimation procedure shares some similarity with two-step approaches. Specifically, our method skips the membership assignment and accounts for the variability of LCA parameter estimates in the estimation of the effects of latent classes. Furthermore, the proposed method addresses random right censoring to the competing risks outcome and properly handles the missing failure type issue with the technique of IPW.

2.2 Data, notation and models

Let T and $\epsilon \in \{1, 2, ..., K\}$ respectively denote time to failure (e.g. death) and the associated failure type (e.g. form of a neuropathological feature). Let C denote time to independent censoring of T. Let R indicate the potential availability of the failure type ϵ. In the UDS example, $R = 1$ if the study participant is willing to donate his/her brain after death, and 0 otherwise. Note that ϵ is observed only if $T \leq C$ and $R = 1$. Let $\boldsymbol{Y}$ denote a $p \times 1$ vector of baseline covariates. Define $X = T \wedge C$, $D = \epsilon R \cdot I(T \leq C)$, and $D_c = I(T \leq C)$, where $\wedge$ is the minimum operator and $I(\cdot)$ is the indicator function. The observed data consist of n iid replicates of $(X, D, D_c, \boldsymbol{Y})$, denoted by $\{(X_i, D_i, D_{c,i}, \boldsymbol{Y}_i), i = 1, ..., n\}$.

The type-d cumulative incidence function conditioning on a random vector $\boldsymbol{V}$ is defined as $F_d(t|\boldsymbol{V}) = \Pr(T \leq t, \epsilon = d|\boldsymbol{V}), d = 1, ..., K$, which represents, for example, the probability of death with a neuropathological feature presented in type d form, given $\boldsymbol{V}$. The latent classes of interest (e.g. MCI subgroups defined by baseline cognitive performance) are represented by a set of binary indicators, $\{\delta_l, \ \ l = 1, ..., L\}$, where L is the number of latent classes, and $\delta_l = 1$ if belonging to class l, and 0 otherwise. Define $\boldsymbol{\Delta} = (I(\delta_1 = 1), I(\delta_2 = 1), ..., I(\delta_L = 1))^T$.

We assume (A1) C is independent of (T, ϵ) given $\boldsymbol{Y}$; and (A2) R is independent of (ϵ, C) given $\boldsymbol{Y}, T$. The assumption (A1) assumes conditional independent censoring given the baseline covariates. The assumption (A2) implies that the failure type ϵ is missing at random; the missing mechanism is specified later by formula (2.3.2).

2.2.1 Latent class model

For the problem considered in this work, the latent classes of interest are defined based on the observed baseline covariates $\boldsymbol{Y}$. Specifically, we assume $\boldsymbol{\Delta}$ follows a multinomial distribution, $Multinomial\{1, (p_1, p_2, ..., p_L)^T\}$, where $p_l > 0$ and $\sum_{l=1}^{L} p_l = 1$. Given being in class l, $\boldsymbol{Y}$ is assumed to follow a distribution with the density, $f(\boldsymbol{Y}; \boldsymbol{\xi}_l)$, where $\boldsymbol{\xi}_l$ is an

unknown vector of parameters ($l = 1, \ldots, L$). The density of $\boldsymbol{Y}$ then takes the form,

$$f(\boldsymbol{Y}; \boldsymbol{\theta}_0) = \sum_{l=1}^{L} p_l f(\boldsymbol{Y}; \boldsymbol{\xi}_l) \tag{2.2.1}$$

with the $h \times 1$ vector $\boldsymbol{\theta}_0 = (p_1, \ldots, p_L, \boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_L^T)^T$. In a real application, one may determine the value of L by employing domain knowledge alone, or in combination with an evaluation of the model fitting based on information criteria, statistical tests, entropy, replicability, or other criteria. We illustrate such a procedure via the UDS example presented in Section 2.5.

2.2.2 Structural competing risks model

To capture how the latent classes defined on the baseline covariates are associated with the progression of the competing risks outcome, we adopt a structural competing risks model, which formulates the effects of latent classes on the cumulative incidence function of the competing risks. Specifically, we assume that

$$F_d(t|\boldsymbol{\Delta}, \boldsymbol{Y}) = g\{\boldsymbol{\lambda}_{0,d}(t)^T \boldsymbol{\Delta}^* + \boldsymbol{\beta}_{0,d}(t)^T \bar{\boldsymbol{Y}}\}, \tag{2.2.2}$$

where $g(\cdot)$ is a known monotone and differentiable link function, $\boldsymbol{\lambda}_{0,d}(t)$ and $\boldsymbol{\beta}_{0,d}(t)$ are vectors of unknown functional coefficients without parametric forms of lengths L and q, respectively, $\boldsymbol{\Delta}^* = \{1, I(\delta_2 = 1), \ldots, I(\delta_L = 1)\}^T$, and $\bar{\boldsymbol{Y}}$ is a $q \times 1$ subvector of $\boldsymbol{Y}$ or $\boldsymbol{Y}$ itself ($q \leq p$). Note that $\boldsymbol{\Delta}^*$ has a one-to-one correspondence with $\boldsymbol{\Delta}$. The key interest under model (2.2.2) is $\boldsymbol{\lambda}_{0,d}(t)$. The lth component of $\boldsymbol{\lambda}_{0,d}(t)$ represents the difference in the g^{-1} transformed type-d cumulative incidence rate at time t between latent class l and the reference latent class 1 ($l = 2, 3, ..., L$). Including the term, $\boldsymbol{\beta}_{0,d}(t)^T \bar{\boldsymbol{Y}}$, allows us to capture any remaining effects of baseline covariates on the type-d cumulative incidence function after adjusting for the effects of latent classes. It is important to note that the structural competing risks model (2.2.2) is defined and interpreted conditionally on $\boldsymbol{\Delta}$, which assumes a pre-specified number of classes L. This model set-up serves to evaluate the effects of

pre-formulated baseline latent classes on a competing risks outcome of interest, which is the main motivation of this work.

Model (2.2.2) takes the same form as the direct binomial regression proposed by Scheike et al. (2008) but includes the unobservable latent class label as a covariate. When $g(x) = 1-\exp\{-\exp(x)\}$ and all non-intercept components of $\boldsymbol{\lambda}_{0,d}(t)$ and $\boldsymbol{\beta}_{0,d}(t)$ are constant over t, model (2.2.2) has the same form as Fine and Gray (1999)'s proportional subdistribution hazards model. Compared to Fine and Gray's model, model (2.2.2) provides extra flexibility by allowing time-dependent latent class effects on the progression of competing risks outcomes, which, in the UDS example, capture the neuropathological development in MCI patients. For notation simplicity, in the sequel, we shall omit the subscript "$_d$" in $\boldsymbol{\lambda}_{0,d}(t)$ and $\boldsymbol{\beta}_{0,d}(t)$ whenever a confusion does not arise.

2.2.3 Missing failure type model

To account for missing failure type, we utilize a logistic regression model to characterize the distribution of the indicator R given $\tilde{\boldsymbol{Y}} = (1, \boldsymbol{Y}, T)^T$. Specifically, the model assumes

$$\pi(\tilde{\boldsymbol{Y}}; \boldsymbol{\gamma}_0) \doteq \Pr(R = 1|\tilde{\boldsymbol{Y}}) = \frac{e^{\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}}}{1 + e^{\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}}}, \tag{2.2.3}$$

where $\boldsymbol{\gamma_0} = (\gamma_{00}, \gamma_{10}, ..., \gamma_{p0}, \gamma_{(p+1)0})^T$ is a $(p+2) \times 1$ vector of unknown regression coefficients. The odds of observing ϵ (i.e. $R = 1$) is increased by $e^{\gamma_{j0}}$ with a unit increase in the j-th component of $\boldsymbol{Y}$ $(j = 1, 2, \ldots, p)$ and by $e^{\gamma_{(p+1)0}}$ with a unit increase in T.

2.3 Estimation and inference

2.3.1 Estimation for the latent class model

Under the assumptions for the latent class model (2.2.1), the likelihood function for $\boldsymbol{\theta}_0$ based on observations $\{\boldsymbol{Y}_i, i = 1, ..., n\}$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{Y_i}; \boldsymbol{\theta}).$$

In practice, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, which is the solution to $\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = 0$, often does not have a closed form. Alternatively, $\hat{\boldsymbol{\theta}}$ may be numerically solved by standard EM algorithm. Consider a common example where $\boldsymbol{Y}$ given $\delta_l = 1$ follows the p-variate Normal distribution, i.e. $f(\boldsymbol{Y}; \boldsymbol{\xi}_l) = (2\pi)^{-p/2}|\mathbf{V}_l|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu_l})'\mathbf{V}_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_l})\}$, with $\boldsymbol{\xi}_l = \{\boldsymbol{\mu}_l, \boldsymbol{V}_l\}$. In this special case, $\hat{\boldsymbol{\theta}}$ can be obtained by the EM algorithm developed for the finite Gaussian mixture model (McLachlan and Peel, 2000, p.48), which is implemented by the R package, `mclust` (Fraley and Raftery, 2006). The corresponding asymptotic results (Boldea and Magnus, 2009) can be used to assess the variability of $\hat{\boldsymbol{\theta}}$.

By the Bayes Rule, the posterior membership probability of subject i belonging to class l is given by

$$\tilde{\delta}_l(\boldsymbol{Y_i}; \boldsymbol{\theta}_0) \doteq \Pr(\delta_l = 1|\boldsymbol{Y_i}) = \frac{p_l f(\boldsymbol{Y_i}; \boldsymbol{\mu_l}, \boldsymbol{V}_l)}{\sum_j p_j f(\boldsymbol{Y_i}; \boldsymbol{\mu_j}, \boldsymbol{V}_j)} = \frac{p_l f(\boldsymbol{Y_i}; \boldsymbol{\mu_l}, \boldsymbol{V}_l)}{f(\boldsymbol{Y_i}; \boldsymbol{\theta}_0)}. \tag{2.3.1}$$

This posterior membership probability can be estimated by $\tilde{\delta}_l(\boldsymbol{Y}_i; \hat{\boldsymbol{\theta}})$, which is obtained by plugging in the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ into (2.3.1).

An entropy index is defined as $1 - \sum_{i=1}^{n}\sum_{l=1}^{L}\tilde{\delta}_l(\boldsymbol{Y_i}; \hat{\boldsymbol{\theta}})[-\log\{\tilde{\delta}_l(\boldsymbol{Y_i}; \hat{\boldsymbol{\theta}})\}]/n \log L$, which is often calculated to evaluate the extent of separation of latent classes established by a latent class model (Muthén et al., 2002). The underlying rationale is that if the latent classes are well separated, the posterior membership probabilities are close to either zero or one and consequently the entropy index is expected to be close to one.

2.3.2 Estimation for the model for missing failure types

To estimate model (2.2.3), it is important to note that T and thus $\tilde{\boldsymbol{Y}}$ is not always observable. However, as shown in Appendix section 2.7.2, model (2.2.3), coupled with assumption (A2), implies

$$\Pr(R=1|\tilde{\boldsymbol{Y}}) = \Pr(R=1|D_c=1,\boldsymbol{Y}^*) = \frac{e^{\boldsymbol{\gamma}_0^T\boldsymbol{Y}^*}}{1+e^{\boldsymbol{\gamma}_0^T\boldsymbol{Y}^*}}, \tag{2.3.2}$$

where $\boldsymbol{Y}^* = (1, \boldsymbol{Y}, X)^T$. By this result, one can obtain a valid estimate for $\boldsymbol{\gamma}_0$ by performing standard logistic regression of responses R_i's over the covariates $(1, \boldsymbol{Y}_i, X_i)$'s in subjects with uncensored T_i (i.e. $D_{c,i}=1$). Denote the resulting maximum likelihood estimate for $\boldsymbol{\gamma}_0$ by $\hat{\boldsymbol{\gamma}}$. The asymptotic properties of $\hat{\boldsymbol{\gamma}}$ follows the theory for the logistic regression (Agresti, 2003).

2.3.3 Estimation for the structural competing risks model

Estimating the $\boldsymbol{\lambda}_0(t)$ in the structural competing risks model (2.2.2) is the key interest of this work. It is important to note that since $\boldsymbol{\Delta}$ (or $\boldsymbol{\Delta}^*$) is not observable, the existing methods for the direct binomial regression (Scheike et al., 2008) cannot be directly applied.

A critical step to tackle this difficulty is to note that

$$F_d(t|\boldsymbol{Y}) = E\{F_d(t|\Delta,\boldsymbol{Y})|\boldsymbol{Y}\} = \sum_{l=1}^{L} g\{\boldsymbol{\lambda}_0(t)^T\boldsymbol{\kappa}_l + \boldsymbol{\beta}_0(t)^T\bar{\boldsymbol{Y}}\}\Pr(\delta_l=1|\boldsymbol{Y}), \tag{2.3.3}$$

where $\boldsymbol{\kappa}_l$ is a $L\times 1$ vector with the first and the lth elements equal to 1 and the rest of elements equal to 0 ($l=1,...,L$). To deal with the random censoring to T and missing failure types, by employing the technique of IPW and inverse probability censoring weighting (IPCW), we can show that under assumptions (A1) and (A2),

$$E\left\{\frac{I(X\le t, D=d)}{G(X|\boldsymbol{Y})\pi(\tilde{\boldsymbol{Y}};\boldsymbol{\gamma_0})}\middle|\boldsymbol{Y}\right\} = F_d(t|\boldsymbol{Y}), \tag{2.3.4}$$

where $G(x|\boldsymbol{Y}) = \Pr(C\ge x|\boldsymbol{Y})$; please see Appendix section 2.7.3 for more detailed justifications. In practice, we require a model and an estimator, $\hat{G}(x|\boldsymbol{Y})$, for $G(x|\boldsymbol{Y})$. For illustra-

tive purpose, we consider Kaplan-Meier estimator in subsequent derivations and numerical analysis, which is consistent under unconditionally independent censoring assumption.

Motivated by the facts (2.3.3) and (2.3.4), writing $\boldsymbol{\alpha}(t) = (\boldsymbol{\lambda}(t)^T,\ \boldsymbol{\beta}(t)^T)^T$, we propose the following estimating equation for $\boldsymbol{\alpha}_0(t) \doteq (\boldsymbol{\lambda}_0(t)^T,\ \boldsymbol{\beta}_0(t)^T)^T$:

$$\boldsymbol{U}_n\{\boldsymbol{\alpha}(t), \hat{\boldsymbol{\theta}}, t\} = 0, \tag{2.3.5}$$

where $\boldsymbol{U}_n\{\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t\}$ equals

$$n^{-1/2}\sum_{i=1}^{n}\left(\sum_{l=1}^{L} g'\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\boldsymbol{W}_{l,i}\tilde{\delta}_{il}(\boldsymbol{\theta})\right)\left[\frac{I(X_i \le t, D_i = d)}{\hat{G}(X_i)\hat{\pi}_i} - \sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\boldsymbol{\theta})\right].$$

Here $g'(\cdot)$ denotes the derivative function of $g(\cdot)$, $\boldsymbol{W}_{l,i} = (\boldsymbol{\kappa}_l^T, \bar{\boldsymbol{Y}}_i^T)^T$, $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of the survival function of C, $\hat{\pi}_i = \pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \hat{\boldsymbol{\gamma}})$, and $\tilde{\delta}_{il}(\boldsymbol{\theta}) = \tilde{\delta}_l(\boldsymbol{Y}_{\boldsymbol{i}}; \boldsymbol{\theta})$.

To solve the estimating equation (2.3.5), we find the solution by minimizing the following objective function $S_n\{\boldsymbol{\alpha}(t), t\}$, the derivative of which with respect to $\boldsymbol{\alpha}(t)$ is $\boldsymbol{U}_n\{\boldsymbol{\alpha}(t), \hat{\boldsymbol{\theta}}, t\}$:

$$S_n\{\boldsymbol{\alpha}(t), t\} = \sum_{i=1}^{n}\left(\sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\hat{\boldsymbol{\theta}})\right)\left[\frac{I(X_i \le t, D_i = d)}{\hat{G}(X_i)\hat{\pi}_i} - \frac{1}{2}\sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\hat{\boldsymbol{\theta}})\right].$$

Note that $S_n\{\boldsymbol{\alpha}(t), t\}$ is a nonconvex function of $\boldsymbol{\alpha}(t)$ given t. The standard Newton-Raphson algorithm, therefore, may require multiple initializations to find the global minimum. To meet this need, we solve equation (2.3.5) by utilizing the differential evolution algorithm (Ardia et al., 2011; Mullen et al., 2011) implemented by R package `DEoptim`, which searches global optimum in a prespecified range of $\boldsymbol{\alpha}(t)$ with automatically generated multiple initial values.

In practice, it is often of interest to predict the cumulative incidence of each failure type for a new patient. Given the new patient's baseline covariates $\boldsymbol{Y}$, we can first use for-

mula (2.3.1) to calculate the posterior latent class membership probabilities, $\tilde{\delta}_l(\boldsymbol{Y};\hat{\boldsymbol{\theta}}) = \frac{\hat{p}_l f(\boldsymbol{Y};\hat{\boldsymbol{\mu}}_{\boldsymbol{l}},\hat{\boldsymbol{V}}_l)}{\sum_j \hat{p}_j f(\boldsymbol{Y};\hat{\boldsymbol{\mu}}_{\boldsymbol{j}},\hat{\boldsymbol{V}}_j)}, l = 1,\ldots,L$. Then, by formula (2.3.3), we can obtain the predicted cumulative incidence functions for the new patient as

$$\hat{F}_d(t|\boldsymbol{Y}) = \sum_{l=1}^{L} g\{\hat{\boldsymbol{\lambda}}(t)^T\boldsymbol{\kappa}_l + \hat{\boldsymbol{\beta}}(t)^T\bar{\boldsymbol{Y}}\}\tilde{\delta}_l(\boldsymbol{Y};\hat{\boldsymbol{\theta}}), \quad d = 1,\ldots,K.$$

2.3.4 Asymptotic properties of the proposed estimator

In this section, we establish the uniform consistency and weak convergence of the proposed estimator $\hat{\boldsymbol{\alpha}}(t)$ for $t \in [l,u]$. Define $\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t) = \sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\boldsymbol{\theta})$ and $\boldsymbol{J}(t) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\boldsymbol{\alpha}(t)}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\right\}^{\otimes 2}$, where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$. Let $\boldsymbol{\iota}_i(\cdot)$ denote the influence function of $\boldsymbol{\gamma}_0$. Further define the following quantities:

$$\begin{aligned}
\boldsymbol{H}(t) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\alpha}(t)}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)^T\right],\\
N_i^G(t) &= I(X_i \le t, D_i = 0), \quad \Upsilon_i(t) = I(X_i \ge t), \quad \upsilon(t) = \Pr(X \ge t),\\
\boldsymbol{w}_G(\boldsymbol{\alpha},\boldsymbol{\theta},t) &= E\big[\frac{\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)I(X_i \le t, D_i = d)\Upsilon(s)}{G(X)\pi_i}\big],\\
\lambda^G(t) &= \lim_{\Delta\to 0}\Pr(X \in (t,t+\Delta)|X \ge t)/\Delta,\\
\Lambda^G(t) &= \int_0^t \lambda^G(s)ds, \quad M_i^G(t) = N_i^G(t) - \int_0^\infty \Upsilon_i(s)d\Lambda^G(s),\\
\boldsymbol{\Xi}_{1i}(t) &= \int_0^\infty \boldsymbol{w}_G(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\upsilon(s)^{-1}dM_i^G(s),\\
\boldsymbol{w}_\pi(\boldsymbol{\alpha},\theta,t) &= E\big[\frac{\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)I(X_i \le t, D_i = d, R_i = 1)\{\frac{\partial}{\partial\boldsymbol{\gamma}}\pi_i\}^T}{G(X)\pi_i^2}\big],\\
\boldsymbol{\Xi}_{2i}(t) &= \boldsymbol{w}_\pi(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0),
\end{aligned}$$

$$\boldsymbol{A}_i^{G\pi}(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t) = \left(\sum_{l=1}^{L} g'\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\boldsymbol{W}_{l,i}\tilde{\delta}_{il}(\boldsymbol{\theta})\right)\left[\frac{I(X_i \le t, D_i = d)}{G(X_i)\pi_i} - \sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\boldsymbol{\theta})\right].$$

We assume the following regularity conditions:

(C1) There exists $\nu > 0$ such that $\Pr(C = \nu) > 0$ and $\Pr(C > \nu) = 0$.

(C2) $\boldsymbol{\alpha}_0(t)$ is Lipschitz continuous for $t \in [l, u]$, $\sup_{t\in[l,u]}||\boldsymbol{\alpha}_0(t)|| < \infty, ||\boldsymbol{\theta}_0|| < \infty$. In addition, $\boldsymbol{Y}_i$ is bounded with probability one for all i.

(C3) $\Psi_i(\boldsymbol{u}(t), \boldsymbol{v}, t)$ and all components of $\frac{\partial\Psi_i(\boldsymbol{u}(t),\boldsymbol{v},t)}{\partial(\boldsymbol{u}(t),\boldsymbol{v})}$ are Lipschitz continuous.

(C4) $\inf_{t\in[l,u]}\text{eigmin}J(t) > 0$, where eigmin$(\cdot)$ denotes the smallest eigenvalue of a matrix.

(C5) All components of $\frac{\partial^2\Psi_i(\boldsymbol{u}(t),\boldsymbol{v},t)}{\partial(\boldsymbol{u}(t),\boldsymbol{v})\partial(\boldsymbol{u}(t),\boldsymbol{v})^T}$ are Lipschitz continuous.

The regularity conditions are reasonable in practical applications. Condition (C1) is often met in scenarios with administrative censoring, and it can facilitate proving the uniform consistency of $\hat{G}(t), 0 < t < \nu$. Condition (C2) assumes the smoothness of coefficient process $\boldsymbol{\alpha}_0(t)$ as well as the boundedness of $\boldsymbol{\alpha}_0(t)$, $\boldsymbol{\theta}_0$, and baseline covariates $\boldsymbol{Y}_i$'s. Conditions (C3) and (C5) impose mild smoothness assumptions for $\Psi_i(\boldsymbol{u}(t), \boldsymbol{v}, t)$. Condition (C4) is a technical assumption that plays a role to ensure the identifiability of $\boldsymbol{\alpha}_0(t)$. In addition, it is also assumed that $\hat{\boldsymbol{\theta}}$ is uniformly consistent and $\sqrt{n}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\}$ has an asymptotic i.i.d. sum representation. These assumptions are often satisfied for MLE estimators under mild regularity conditions.

The uniform consistency and weak convergence results are stated in the following theorems. Proofs are correspondingly provided in Appendix sections 2.7.4 and 2.7.5.

Theorem 2.3.1. *Suppose conditions C1 - 4 hold and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$. Then for n large enough, there exists a uniformly bounded solution of $\boldsymbol{U}_n\{\boldsymbol{\alpha}(t), \hat{\boldsymbol{\theta}}, t\} = 0$, $\hat{\boldsymbol{\alpha}}(t)$, such that $sup_{t\in[l,u]}||\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)|| \to 0$ in probability.*

Theorem 2.3.2. *Suppose conditions of Theorem 1 and condition C5 hold, and there exist iid random functions $\{\boldsymbol{\phi}_i(\boldsymbol{\theta}_0)\}_{i=1}^{\infty}$ such that*

$$||\sqrt{n}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\} - n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\phi}_i(\boldsymbol{\theta}_0)|| \xrightarrow{p} 0$$

and $sup_i|\boldsymbol{\phi}_i(\boldsymbol{\theta}_0)| < \infty$. Then $\sqrt{n}[\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)]$ converges weakly to a zero-mean Gaussian

process for $t \in [l, u]$ *with covariance function* $\boldsymbol{\Sigma}(s,t) = E\{\boldsymbol{Q}_1(s)\boldsymbol{Q}_1(t)^T\}$, *where*

$$\boldsymbol{Q}_i(t) = \boldsymbol{J}(t)^{-1}[\boldsymbol{A}_i^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) - \boldsymbol{\Xi}_{1i}(t) - \boldsymbol{\Xi}_{2i}(t) - \boldsymbol{H}(t)\boldsymbol{\phi}_i(\boldsymbol{\theta}_0)], i = 1, ..., n.$$

2.3.5 Inference procedures

Inferences about $\boldsymbol{\alpha}_0(t)$ are important for assessing the association between latent classes and the competing risks outcome. It is worth noting that conducting bootstrapping in LCA is often subject to the label switching issue. More specifically, the latent classes identified in different runs of resampling may not have clear correspondences. To circumvent this challenge, we develop an analytical approach to deriving the estimator of $\boldsymbol{\Sigma}(s,t)$, the asymptotic covariance of the consistent estimate $\hat{\boldsymbol{\alpha}}(t)$, which is established in Theorem 2.3.2. Define

$$\boldsymbol{A}_i(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t) = \left(\sum_{l=1}^{L} g'\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\boldsymbol{W}_{l,i}\tilde{\delta}_{il}(\boldsymbol{\theta})\right)\left[\frac{I(X_i \le t, D_i = d)}{\hat{G}(X_i)\hat{\pi}_i} - \sum_{l=1}^{L} g\{\boldsymbol{\alpha}(t)^T\boldsymbol{W}_{l,i}\}\tilde{\delta}_{il}(\boldsymbol{\theta})\right].$$

Suppose there exists statistics $\hat{\boldsymbol{\phi}}_i(\hat{\boldsymbol{\theta}})$ satisfying $\sup_i||\hat{\boldsymbol{\phi}}_i(\hat{\boldsymbol{\theta}}) - \boldsymbol{\phi}_i(\boldsymbol{\theta}_0)|| \xrightarrow{p} 0$. Let

$$\hat{\boldsymbol{Q}}_i(t) = \hat{\boldsymbol{J}}(t)^{-1}[\boldsymbol{A}_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \hat{\boldsymbol{\Xi}}_{1i}(t) - \hat{\boldsymbol{\Xi}}_{2i}(t) - \hat{\boldsymbol{H}}(t)\hat{\boldsymbol{\phi}}_i(\hat{\boldsymbol{\theta}})],$$

$$\hat{\boldsymbol{\Sigma}}(s,t) = n^{-1}\sum_{i}^{n} \hat{\boldsymbol{Q}}_i(s)\hat{\boldsymbol{Q}}_i(t)^T,$$

where

$$\hat{\boldsymbol{J}}(t) = \frac{1}{n}\sum_{i=1}^{n} \partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)^{\otimes 2},$$

$$\hat{\boldsymbol{H}}(t) = \frac{1}{n}\sum_{i=1}^{n} \partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)\partial_{\boldsymbol{\theta}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)^T,$$

$$\hat{\boldsymbol{\Xi}}_{1i}(t) = I(D_i=0)\frac{\sum_{j=1}^n I(X_j \geq X_i)\partial_{\boldsymbol{\alpha}}\Psi_j(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_j \leq t, D_j=d)\{\hat{G}(X_j)\}^{-1}\hat{\pi}_i^{-1}}{\sum_{j=1}^n I(X_j \geq X_i)},$$

$$\hat{\boldsymbol{\Xi}}_{2i}(t) = \frac{1}{n}\sum_{j=1}^n \frac{\partial_{\boldsymbol{\alpha}}\Psi_j(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_j \leq t, D_j=d)\{\frac{\partial}{\partial\boldsymbol{\gamma}}\hat{\pi}_i\}^T}{\hat{G}(X_j)\hat{\pi}_i^2}\boldsymbol{\iota}_i(\hat{\boldsymbol{\gamma}}).$$

We show in Appendix section 2.7.5 that $\sup_{s,t\in[l,u]}||\hat{\boldsymbol{\Sigma}}(s,t)-\boldsymbol{\Sigma}(s,t)|| \xrightarrow{p} 0$. The time-specific 95% confidence intervals can be constructed based on $\hat{\boldsymbol{\Sigma}}(s,t)$ with normal approximations. To conduct inference about $\hat{\boldsymbol{\alpha}}(\cdot)$ on a time range $[t_l,t_u]$, we further construct simultaneous confidence bands using similar resampling strategy as Yin and Cai (2004) did. First, we generate n_B samples, indexed by k, of $\sqrt{n}\check{\boldsymbol{\alpha}}_k(t) \equiv n^{-1/2}\sum_{i=1}^n \hat{\boldsymbol{Q}}_i(t)Z_{k,i}$ for all $t \in [t_l,t_u]$. Second, for the lth element of $\hat{\boldsymbol{\alpha}}(\cdot)$, $\hat{\alpha}_l(\cdot)$, we find constant $q_{l,0.05}$ which satisfies

$$\Pr\left\{\sup_{k,t\in[t_l,t_u]}\left|\frac{\hat{\boldsymbol{\alpha}}_l(t)\sqrt{n}\check{\alpha}_{l,k}(t)}{\hat{\Sigma}_{l,l}(t,t)}\right| > q_{l,0.05}\right\} = 0.05,$$

where $\check{\alpha}_{l,k}(\cdot)$ is the lth element of $\check{\boldsymbol{\alpha}}_k(\cdot)$ and $\hat{\Sigma}_{l,l}(\cdot,\cdot)$ is the (l,l) entry of $\hat{\boldsymbol{\Sigma}}(\cdot,\cdot)$. Then the confidence band for $\hat{\alpha}_l(t)$ is constructed as $\hat{\alpha}_l(t) \mp q_{l,0.05}\hat{\sigma}_{l,l}(t,t)/\sqrt{n}\hat{\alpha}_l(t)$.

2.4 Simulations

2.4.1 Data generation and analysis procedures

Simulation studies were conducted to evaluate the finite-sample performance of the proposed method. For sample size $N = 2000$, 1000 simulated datasets were generated and analyzed following the same procedures for scenarios with different choices of model parameters. With $L = 3$, we first generated latent class membership vector $\boldsymbol{\Delta} = (\delta_1,\delta_2,\delta_3)^T$ from a multinomial distribution with relative frequency $(p_1,p_2,p_3)^T$. Given $\delta_l = 1$, baseline covariates $\boldsymbol{Y} \in \mathbb{R}^2$ was generated from a bivariate normal distribution specified for class l.

Given the true latent class membership $\boldsymbol{\Delta}$, we further generated competing risks outcome with two failure types, following a simulation scheme similar to that in Scheike et al. (2008). Specifically, we specified $F_1(t|\boldsymbol{\Delta},\boldsymbol{Y}) = 1-\{1-0.66(1-e^{-t})\}^{\exp(\zeta(t)^T\boldsymbol{\Delta}^*)}$ and $F_2(t|\boldsymbol{\Delta},\boldsymbol{Y}) = 1-0.34^{\exp(\zeta(t)^T\boldsymbol{\Delta}^*)}\{1-e^{-t\exp(\zeta(t)^T\boldsymbol{\Delta}^*)}\}$, where $\zeta(t) = \{\zeta_1,\zeta_2,\zeta_3(t)\}^T$. Here $\zeta_1 = 0.5, \zeta_2 = 0.5$

and $\zeta_3(t) = \log\{\frac{\log\{(1-p_c)^{e^{-1}-0.38}[1-p_c(1-e^{-t})]^{0.38}\}}{\log\{1-p_c(1-e^{-t})\}}\}$. Then the proposed competing risks model (2.2.2) was satisfied with $F_1(t|\boldsymbol{\Delta}, \boldsymbol{Y}) = g\{\boldsymbol{\lambda}_0(t)^T\boldsymbol{\Delta}^*\}$, where $g(x) = 1-\exp\{-\exp(x)\}$ and $\boldsymbol{\lambda}_0(t) = \{\zeta_1 + \log[-\log\{1-p_c(1-e^{-t})\}], \zeta_2, \zeta_3(t)\}^T$. Based on the specified $F_d(t|\boldsymbol{\Delta}, \boldsymbol{Y})$, we first generated failure type ϵ based on the facts that ϵ follows $Multinomial(1, \{\Pr(\epsilon = 1|\boldsymbol{\Delta}), 1-\Pr(\epsilon = 1|\boldsymbol{\Delta})\}^T)$, where $\Pr(\epsilon = 1|\boldsymbol{\Delta}) = \lim_{t\to\infty} F_1(t|\boldsymbol{\Delta}, \boldsymbol{Y})$. Given failure type ϵ, T was generated from $\Pr(T \le t|\epsilon = d) = \frac{F_d(t|\boldsymbol{\Delta},\boldsymbol{Y})}{\Pr(\epsilon=d|\boldsymbol{\Delta})}$. Furthermore, censoring time C was independently generated from $Uniform(0.19, 1.09)$. The resulting proportion of censoring was around 50%. To simulate missing failure types, we generated the indicator of observing failure types, R, from $Bernoulli\{\Pr(R = 1|\boldsymbol{Y}, T)\}$, where $\Pr(R = 1|\boldsymbol{Y}, T) = \frac{\exp(\boldsymbol{\gamma}^T\boldsymbol{Y}-T)}{1+\exp(\boldsymbol{\gamma}^T\boldsymbol{Y}-T)}$.

Three strategies were applied to analyze the simulated datasets: (i) the proposed method; (ii) a modal assignment method replacing $\tilde{\delta}_{il}(\hat{\boldsymbol{\theta}})$ by $I\{l = \arg\max_{1\le j\le L}\tilde{\delta}_{ij}(\hat{\boldsymbol{\theta}})\}$ in the estimating equation (2.3.5); (iii) Scheike et al. (2008)'s direct binomial regression implemented by R package `timereg` (Scheike and Zhang, 2011), without adjustment for missing failure types but using true latent class labels as covariates. Compared to the proposed method, the modal assignment method ignored the uncertainty of the estimated class membership but accounted for missing failure types, while the direct binomial regression overlooked the missing failure types but used the correct class memberships as covariates. By comparing the three methods, it is straightforward to examine the sources of biases. When applying methods to fit the competing risks data, we set the competing outcome of interest as $\epsilon = 1$ and assumed that the number of latent classes L was already known as 3. Further discussion and simulation results about selecting L can be found in Appendix section 2.7.6.

2.4.2 Simulation scenarios

As shown in Table 2.1, we designed several simulation settings resulting in different data characteristics. By adjusting the relative frequency, $\boldsymbol{p} = (p_1, p_2, p_3)$, of latent classes, the data were simulated with balanced or imbalanced latent class proportions. To create different extents of overlapping of $\boldsymbol{Y}$ among latent classes, we also specified several choices of class-specific mean, $\boldsymbol{\mu}_l, l = 1, 2, 3$, of $\boldsymbol{Y}$. In addition, different $\boldsymbol{\gamma}$ parameters were used to

simulate moderate (around 25%) to heavy (around 50%) rate of missing failure types given failure.

Table 2.1: Different simulation settings and the corresponding interpretations.

Settings	$\boldsymbol{p} = (p_1, p_2, p_3)$	Class proportions
(1a)	(0.3,0.35,0.35)	balanced
(1b)	(0.2,0.35,0.45)	moderately imbalanced
(1c)	(0.15,0.35,0.5)	severely imbalanced

Settings*	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	Overlapping (entropy)
(2a)	(3,3)	(5,5)	mild ($\sim^\dagger$ 0.9)
(2b)	(2.5,2.5)	(4,4)	moderate ($\sim$ 0.8)
(2c)	(2.25,2.25)	(3.5,3.5)	moderately severe ($\sim$ 0.7)
(2d)	(2,2)	(3,3)	severe ($\sim$ 0.6)

Settings	$\boldsymbol{\gamma}$	Missing failure type (missing rate)
(3a)	(0.25,0.5)	moderate ($\sim^\dagger$ 25%)
(3b)	(-0.35,0.5)	heavy ($\sim$ 50%)

* $\boldsymbol{\mu}_1$ is fixed as (1,1).
† Approximated levels observed from simulations.

Based on the settings in Table 2.1, we evaluated the performance of the three methods comprehensively, as detailed in Table 2.2. Controlling other factors as fixed, we conducted simulations under (A) different overlapping levels of $\boldsymbol{Y}$; (B) different class proportions; (C) moderate or heavy rates of missing failure types. In addition, we investigated (D) how nuisance or noisy baseline covariates affected the analysis and (E) how our method performed under a setting similar to the UDS data. Performance under different choices of sample size was also assessed in (F). For (E), $\boldsymbol{Y} \in \mathbb{R}^{10}$ was generated from a multivariate normal distribution with the mean vector and covariance matrix equal to the estimates from fitting a 3-class finite Gaussian mixture model for the UDS data.

2.4.3 Convergence of algorithm

Under challenging simulation scenarios, we typically observed a small portion of outlying estimates which may not converge to the global optimum. We empirically defined the

Table 2.2: Sets of investigations and the corresponding simulation scenarios.

Investigations	Scenarios simulated†
(A) overlapping	(1b) + (2a) + (3a) (1b) + (2b) + (3a) (1b) + (2c) + (3a) (1b) + (2d) + (3a)
(B) class proportion	(1a) + (2b) + (3a) (1b) + (2b) + (3a) (1c) + (2b) + (3a)
(C) missing failure type	(1b) + (2b) + (3a) (1b) + (2b) + (3b)
(D) nuisance covariates	(1a) + (2b) + (3a) + Y_3^* + Y_4^*
(E) real data scenario	Same distribution of $\boldsymbol{Y}$ as in UDS
(F) Sample size	(1b) + (2c) + (3a) $N \in \{500, 1000, 2000\}$

† Refer to Table 2.1
* Adding two independent nuisance covariates in $\boldsymbol{Y}$.

outlying estimates, whose distances to the median of the total 1000 estimates were larger than four times median absolute deviation (MAD), as non-convergent results. As shown in Table 2.3, non-convergence became more frequent for more overlapped or imbalanced scenarios, and smaller sample size. In addition, the non-convergent results tended to be associated with large standard error estimates throughout the time period (Figure 2.1). In practice, therefore, we should be careful about interpreting results with large standard error estimates, which may be caused by the non-convergence of the algorithm.

2.4.4 Simulation results

Figure 2.2 displays the simulation results for the comparisons (A)-(F), including the empirical biases (based on mean) and the empirical coverage rates of 95% confidence intervals for $t \in [0.1, 1.0]$, after excluding outlying estimates. The empirical coverage rates of 95% confidence bands for $t \in [0.1, 1.0]$ and detailed standard error estimation results can be found in Tables 2.6-2.11 in Appendix section 2.7.7. Overall, the proposed approach not only achieved the smallest empirical biases, but also obtained reasonable empirical coverage

Table 2.3: The percentage of excluded estimates, defined as the outlying estimates whose distances to the median of the total 1000 estimates were larger than four times median absolute deviation, for different simulation scenarios.

Comparisons	Scenarios	Excluded estimates (%)
(A)	(1b)+(2a)+(3a)	0
	(1b)+(2b)+(3a)	0.7
	(1b)+(2c)+(3a)	2.5
	(1b)+(2d)+(3a)	6.8
(B)	(1a)+(2b)+(3a)	0.8
	(1b)+(2b)+(3a)	0.7
	(1c)+(2b)+(3a)	8
(C)	(1b)+(2c)+(3a)	2.5
	(1b)+(2c)+(3b)	2.9
(D)	Nuisance parameters	2.1
(E)	UDS scenario	2.3
(F)	N=500	6.1
	N=1000	4.1
(1b)+(2c)+(3a)	N=2000	2.5

rates for both confidence interval and confidence band. Compared to the proposed method, the modal assignment method suffered from moderate biases caused by misclassification of latent classes, and the `timereg` method obtained more severe biases by ignoring missing failure types in uncensored subjects.

Figure 2.2(A) and 2.2(B) show the results under the trend of increasingly fuzzier latent class pattern, either due to more severe overlapping of $\boldsymbol{Y}$ or more imbalanced latent class proportion. The proposed method maintained low biases and standard coverage rates even under the severe overlapping (entropy lower than 0.6) and the severely imbalanced scenarios. In contrast, the modal assignment method obtained increasingly larger biases as the latent class pattern became fuzzier, which was caused by the misclassification of modally assigned class labels. The `timereg` method, on the other hand, used true class membership as covariates, thus showing constant biases in (A) and (B), where the missing failure type probability was unchanged and the biases were solely caused by ignoring the missing failure type problem. The larger bias of the `timereg` method also suggested that the missing failure type was a stronger source of bias than latent class misclassification in our simulation settings.

Comparison (C) assesses the performance under moderate (25%) or heavy (50%) missing failure type proportion among the uncensored events. Displayed in Figure 2.2(C), the proposed method achieved similar results under both moderate and heavy missingness of failure types with the smallest biases among the three methods. The modified method accounted for missing failure type issue, thus also showing similar performances under the two scenarios where the latent class pattern was unchanged. The `timereg` approach, on the other hand, ignored missing causes of failure and obtained much larger bias for $\hat{\lambda}_1(t)$ as the missingness became more severe. From Figure 2.2(D) and 2.2(E), we observed that the proposed method had good performance even in the presence of nuisance variables and under a scenario based on real world data. This endorses our application of the proposed method to the MCI data from UDS. Moreover, Figure 2.2(F) shows promising finite sample performances at sample size 500 and 1000 under a challenging scenario where the entropy index was around 0.7. As sample size grew, the proposed method achieved smaller biases, better coverage rates and a higher convergence rate.

2.5 An application to the MCI data from UDS

We applied the proposed approach to analyze the Uniform Data Set (Beekly et al., 2007; Weintraub et al., 2009, UDS). As of June 2015, ten different cognitive test scores at the first visit and the follow-up survival information were available for 6034 MCI patients. There were 818 deaths during the follow-up, including 411 individuals with autopsy data.

In our analysis, we investigated the association between baseline cognitive MCI subtypes and the levels of density of neocortical neuritic plaques. The levels of plaque density, namely none or sparse, moderate, and frequent, were recorded as the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) scores (Welsh et al., 1991). Existing knowledge has linked amnestic MCI population, with damaged brain domain related to memory functions, to frequent density of neuritic plaques (Dugger et al., 2015).

To enable our analysis, we formulated the neuropathological endpoint of interest as time from first visit to death (T) with none or sparse ($\epsilon = 1$), moderate ($\epsilon = 2$), and frequent ($\epsilon =$

3) levels of neocortical neuritic plaques, determined by CERAD scores. Since one subject can die only with one given level of CERAD score, such defined endpoint forms a competing risk outcome. Moreover, for individuals who died but did not agree to donate brain, their competing risks failure types, which correspond to the CERAD features at death, were missing. For subjects who survived at the end of the follow-up, their survival times were right-censored at the last follow-up visit. Here, We assumed independent censoring and used Kaplan-Meier estimator for IPCW. We discuss our choice of censoring assumption in detail in Appendix section 2.7.9.

2.5.1 Latent class model and missing failure type model

As the first step, we conducted a LCA based on ten standardized baseline cognitive test scores adjusted for age, education and gender. The ten scores evaluated different perspectives of patients' cognitive performance, which were described in details in the Section 2.2.1 of Hanfelt et al. (2018). Figure 2.3 shows the histograms and kernel density estimates for the ten cognitive test scores.

The latent classes of baseline cognitive performance were defined based on a 3-class finite Gaussian mixture model with flexible covariance structure that captured the dependency between cognitive test scores. The number of latent classes, $L = 3$, was determined by comparing the integrated classification likelihood-BIC (ICL-BIC) for LCA models with 1 to 9 latent classes. We chose ICL-BIC over simpler BIC because ICL-BIC penalizes models with excessive fuzziness; thus the selected model would have a better balance between model fitting and class membership certainty. While the results, shown in Figure 2.4, preferred a 2-class model, existing knowledge about MCI (as confirmed by our collaborators in neurology) suggested at least three subgroups to capture the clinical heterogeneity of MCI patients. Considering that a 3-class model corresponded to the second highest ICL-BIC, we adopted $L = 3$ to account for both latent class model fitting and clinical plausibility.

Table 2.4 summarizes the cognitive characteristics of the three subgroups derived by the 3-class LCA model and applying modal class assignment rule. Class 1 was interpreted

as mildly impaired, with test scores showing the best overall cognition; class 2 was non-amnestic, with impairments in attention and executive function domains; class 3 was amnestic, identified by the heavier impairment in delayed memory. The overall entropy index for the fitted 3-class model was 0.70, corresponding to the *moderately severe overlapping* scenario in our simulation.

One can further examine the fuzziness of latent class classification with respect to a given class c, by the class-specific entropy index for class c,

$$1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{L} I(i \in \hat{M}_c)\tilde{\delta}_l(\boldsymbol{Y_i};\hat{\boldsymbol{\theta}})[-\log\{\tilde{\delta}_l(\boldsymbol{Y_i};\hat{\boldsymbol{\theta}})\}]}{\sum_{i=1}^{n} I(i \in \hat{M}_c)\log L}$$

, where $\hat{M}_c$ denotes the set of subjects that are classified as belonging to class c by modal assignment. Based on the modally assigned class memberships, the class-specific entropy indices were 0.62, 0.87, and 0.62 respectively for classes 1-3. These suggested that the class 2 was more easily separable from class 1 and 3, while the patterns of the cognitive test scores might be more overlapped between class 1 and 3.

In order to account for missing neuropathological phenotypes in deceased subjects, we fit a logistic regression model of R using the ten baseline test scores and survival times. After variable selection based on Akaike information criterion (Akaike, 1974), we obtained the final model $\text{Logit}\{\Pr(R=1)\} = -0.39 + 0.06 \cdot \text{MMSE} - 0.17 \cdot \text{Category Fluency} - 0.07 \cdot \text{Trails A} + 0.11 \cdot \text{Boston Naming} + 0.13 \cdot T$. Based on this model, we calculated the estimated inverse probability weight $\hat{\pi}_i$ for each deceased subject. Hosmer–Lemeshow's goodness-of-fit test (p = 0.90) indicated a good fitting of the logistic regression model (Hosmer Jr et al., 2013).

2.5.2 Structural competing risks model

Next, we fit the proposed structural competing risks model (2.2.2). For each CERAD level $\epsilon = d$ $(d = 1, 2, 3)$, we used link function $g(x) = 1 - \exp\{-\exp(x)\}$ and obtained the point estimates for the true coefficient functions $\lambda_{l,d}(t)$, denoted by $\hat{\lambda}_{l,d}(t)$, for $t \in [0.25, 8]$, where t denotes years since first visit, and $l = 1, 2, 3$ denotes the three baseline MCI latent

Table 2.4: Summary of the ten baseline standardized cognitive test scores for the three latent classes.

Standardized scores	Mean Values Class 1 mildly impaired	Class 2 non-amnestic	Class 3 amnestic
Mini-Mental State Examination (MMSE)	-0.45	-2.24	-2.12
Logical Memory: Immediate	-0.94	-1.34	-1.30
Logical Memory: Delayed	-0.94	-1.38	-1.55
Semantic Memory: Category Fluency	-0.64	-1.12	-0.96
Attention: Trails A	0.16	1.72	0.01
Attention: Digit Span Forward	-0.21	-0.49	-0.17
Language: Boston Naming	0.00	-1.95	-1.15
Executive: Trails B	0.40	3.08	0.20
Executive: Digit Span Backward	-0.33	-0.72	-0.36
Visuo-Motor: Digit Symbol	-0.58	-1.28	-0.51

Note: The class memberships were assigned by modal assignment scheme.
For the trails A and trails B scores, a higher score indicates greater impairment.
For the other scores, a lower score indicates greater impairment.

subtypes. Here, $\hat{\lambda}_{1,d}(t)$ represents the estimated g^{-1} transformed cumulative incidence at time t for latent class 1, and $\hat{\lambda}_{l,d}(t)$ $(l = 2, 3)$ represents the estimated difference in the g^{-1} transformed cumulative incidence at time t between latent class l and class 1. The corresponding 95% confidence intervals were also computed based on estimated covariance function $\hat{\Sigma}_{l,d}(t, t)$ for class l at time t. To compare the proposed method with alternative strategies, the point estimates were also obtained by the modal assignment method used in the simulation studies and a naive method using modal assignment and excluding subjects with missing failure type.

Figure 2.5 displays $\hat{\lambda}_{l,d}(t)$ and the corresponding 95% confidence intervals versus time for the three CERAD competing phenotypes $(d = 1, 2, 3)$. Similarly, Figure 2.6 shows the estimated g^{-1} cumulative incidences. As observed in Figure 2.5, both the modal assignment and naive approaches may tend to underestimate the cumulative incidence differences between latent class 2 versus latent class 1 and those between latent class 3 versus latent class 1, captured by $\lambda_{2,d}(t)$ and $\lambda_{3,d}(t)$, $d = 1, 2, 3$. Compared to the modal assignment method that ignored latent class uncertainty, the naive approach further overlooked missing CERAD phenotypes thus yielding more biases.

To better display the separation in neuropathological features by the baseline MCI subgroups, we present in Figure 2.7 the predicted cumulative incidence functions based on the proposed method and the empirical cumulative incidence functions. According to the results from the proposed method, the amnestic group (class 3, red solid line) were more likely to develop frequent neuritic plaques, as compared to non-amnestic (class 2, blue solid line) and mildly impaired (class 1, green solid line) groups in the 8-year follow-up period. In contrast, non-amnestic and mildly impaired MCI patients were more likely to develop sparse or moderate neuritic plaques than the amnestic patients. These findings are consistent with existing knowledge of amnestic MCI's high susceptibility of Alzheimer's Disease (Adler et al., 2010; Guillozet et al., 2003). Moreover, the cumulative incidence estimates by the proposed method demonstrated different trends from the naive empirical cumulative incidence estimates that overlooked missing phenotypes and latent class uncertainty.

Combining Figures 2.5 and 2.7, we also noticed time-varying differences between latent classes in developing different phenotypes of neuritic plaques. For deaths with none or sparse neuritic plaques, the cumulative incidence in the non-amnestic MCI group was significantly higher than mildly impaired group between 3 and 6 years after first visit, while mildly impaired group's cumulative incidence rapidly caught up between 6 and 8 years after enrollment. For deaths with frequent neuritic plaques, the cumulative incidence in the amnestic MCI group was lower than that in the non-amnestic MCI group until 6 years after first visit and rapidly increased between 6 and 8 years. These time-varying differences in cumulative incidences may provide a more detailed picture about the progression of the neuropathological phenotypes in terms of neuritic plaques.

We present in Figure 2.8 the predicted cumulative incidence functions for a "new" patient whose baseline cognitive test scores are equal to the median test scores observed in the MCI data. It is observed that the predicted cumulative incidence functions for this new patient are similar to the cumulative incidence functions estimated for the mildly impaired latent class. This is reasonable because the estimated posterior membership probabilities for this new patient are 0.79, 0.20, 0.01 respectively for the mildly impaired, amnestic, and non-amnestic latent classes, suggesting a high likelihood of belonging to the mildly impaired

MCI latent class. Comparing the predicted cumulative incidence functions across different CERAD pheotypes in Figure 2.8, this new patient is more likely to show none or sparse neocortical neuritic plague than the other two CERAD phenotypes at death.

In summary, the application of the proposed method provides useful insight about the neuropathological relevance of the baseline MCI subgroups. Our results successfully link the amnestic MCI group to higher risk of frequent neuritic plaques, which is consistent with existing knowledge about Alzheimer's Disease. In addition, the flexibility in accommodating time-varying latent class effects entails more robust and more in-depth investigations. Similar analysis can also be easily conducted on other neuropathological features of interests.

2.6 Discussion

In this article, we propose a flexible approach to investigating the association between latent classes and competing risks outcomes. Through involving separate steps for LCA and fitting the structural competing risks model, the proposed method properly integrates the results from these two steps without being plagued by the estimation bias from directly plugging in latent class membership assignment. Our method also properly handles realistic complications for competing risks outcome such as missing failure types. Compared to the popular one-step methods, our method has more proper interpretability of *baseline* heterogeneity and is computationally more economic. We also derive analytical forms for consistent variance and covariance estimates. Such inferences otherwise are not straightforward and computationally intensive with bootstrapping.

As guided by simulation results, for sample size 2000, our method still works well when the entropy index is around 0.6. In addition, the method is robust under imbalanced latent class proportion, severe missing failure type problems, and signal interference of nuisance covariates. When latent class pattern is overly fuzzy, such as severe overlapping plus severely imbalanced class proportion, the proposed method will have relatively high non-convergence rate and unstable estimation with sample size less than 2000 (Figure 2.9 in Appendix section 2.7.8). However, our method can still perform well with sufficiently large sample size, which

is also the case of the UDS data. To check non-convergence in practice, we recommend checking the standard error estimates and be cautious when encountering unusually large standard error estimates across the time range.

Theoretically, one may choose any latent class framework to define disease heterogeneity given baseline covariates. Once the posterior membership probabilities are computed, the point estimation of the proposed structural competing risks model can be obtained by solving the estimating equation (2.3.5). The influence function of the LCA model parameters $\boldsymbol{\theta}_0$, however, is required for variance estimation and inference. Therefore, we recommend using LCA methods with developed asymptotic results, such that the variability of the LCA parameter estimates can be accounted for in the estimating procedure of the structural competing risks model.

In our real data application, the IPW method accounting for missing failure type issue may be less efficient if the missing failure type problem is severe among the uncensored observations. Advanced strategies, such as augmented methods (Gao and Tsiatis, 2005), pseudo-likelihood (Bakoyannis et al., 2020), or multiple imputation (Lu and Tsiatis, 2001) can be considered to improve the potential efficiency loss.

Worth further mention, when there is only one cause of failure ($\epsilon \in \{1\}$), the proposed model (2.2.2) is equivalent to the proportional hazard model, if specified with constant non-intercept terms in $\{\boldsymbol{\lambda}(t)_{0,1}^T, \boldsymbol{\beta}(t)_{0,1}^T\}^T$, and a proper choice of link function g, such as $g(x) = 1 - \exp\{-\exp(x)\}$. Thus, our approach is also applicable to analyzing data with single time-to-event outcome. Since the missing failure type issue is not present for the single time-to-event situation, the missing failure type model (2.2.3), and the corresponding components will be removed from the estimation (2.3.5) and inference procedures.

2.7 Appendices

2.7.1 Notations

Define following quantities:

$$\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\frac{\partial\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)}{\partial\boldsymbol{\alpha}(t)}=\sum_{l=1}^{L}g'\{\boldsymbol{\alpha}(t)^T\kappa_l\}\kappa_l\tilde{\delta}_{il}(\boldsymbol{\theta}),$$

$$\partial_{\theta}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\frac{\partial\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)}{\partial\boldsymbol{\theta}}=\sum_{l=1}^{C}g'\{\boldsymbol{\alpha}(t)^T\kappa_l\}\frac{\partial\tilde{\delta}_{il}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

$$\partial^2_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\frac{\partial\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)}{\partial\{\boldsymbol{\alpha}(t)\ \boldsymbol{\theta}\}^T},$$

$$\boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\left[\frac{I(X_i\le t,D_i=d)}{G(X_i)\pi_i}-\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\right],$$

$$\boldsymbol{K}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},t)\left[\frac{I(X_i\le t,D_i=d)}{\hat{G}(X_i)\hat{\pi}_i}-\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\right],$$

$$\boldsymbol{K}_i^{G\pi}(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},t)\left[\frac{I(X_i\le t,D_i=d)}{G(X_i)\pi_i}-\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\right],$$

$$\boldsymbol{L}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\left[\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},t)\right]\left[\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\right]^T,$$

$$\boldsymbol{L}(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=E[\boldsymbol{L}_1(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)],$$

$$\boldsymbol{B}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)=\left[\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},t)\right]\left[\partial_{\boldsymbol{\theta}}\Psi_i(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)\right]^T.$$

2.7.2 Proof of Equation (2.3.2)

Assumptions (A1) and (A2) jointly imply that C, R and ε have mutually conditional independence given $\boldsymbol{Y}, T$. Thus,

$$\begin{aligned}\Pr(R=1|\tilde{\boldsymbol{Y}})&=\Pr(R=1|\boldsymbol{Y},T)\\&=\frac{\Pr(R=1|\boldsymbol{Y},T)\Pr(C\ge T|\boldsymbol{Y},T)}{\Pr(C\ge T|\boldsymbol{Y},T)}=\frac{\Pr(R=1,C\ge T|\boldsymbol{Y},T)}{\Pr(C\ge T|\boldsymbol{Y},T)}\\&=\Pr(R=1|C\ge T,\boldsymbol{Y},T)=\Pr(R=1|D_c=1,\boldsymbol{Y},X)\\&=\Pr(R=1|D_c=1,\boldsymbol{Y}^*)\end{aligned}$$

which completes the proof.

2.7.3 Proof of Equation (2.3.4)

Assumption (A1) implies

$$f(c|t,\varepsilon,\boldsymbol{y}) = f(c|\boldsymbol{y}). \tag{2.7.1}$$

Assumptions (A1) and (A2) jointly imply that C, R and ε have mutually conditional independence given $\boldsymbol{Y}, T$, which further implies

$$f(r|t,\varepsilon,\boldsymbol{y}) = f(r|t,\boldsymbol{y}). \tag{2.7.2}$$

In addition by (A1) and Bayes Theorem

$$f(c,\varepsilon|t,\boldsymbol{y}) = \frac{f(c,t,\varepsilon,\boldsymbol{y})}{f(t,\boldsymbol{y})} = \frac{f(c|\boldsymbol{y})f(t,\varepsilon|\boldsymbol{y})f(\boldsymbol{y})}{f(t,\boldsymbol{y})} = f(c|\boldsymbol{y})f(\varepsilon|t,\boldsymbol{y}). \tag{2.7.3}$$

Thus we can show the conditional independence of C and R given $(T,\boldsymbol{\epsilon},\boldsymbol{Y})$ by

$$\begin{aligned} f(r,c|t,\varepsilon,\boldsymbol{y}) = \frac{f(r,c,\varepsilon|t,\boldsymbol{y})}{f(\varepsilon|t,\boldsymbol{y})} &\overset{(A2)}{=} \frac{f(r|t,\boldsymbol{y})f(c,\varepsilon|t,\boldsymbol{y})}{f(\varepsilon|t,\boldsymbol{y})} \overset{(2.7.3)}{=} f(r|t,\boldsymbol{y})\frac{f(c|\boldsymbol{y})f(\varepsilon|t,\boldsymbol{y})}{f(\varepsilon|t,\boldsymbol{y})} \\ &= f(r|t,\boldsymbol{y})f(c|\boldsymbol{y}) \overset{(2.7.1)(2.7.2)}{=} f(r|t,\varepsilon,\boldsymbol{y})f(c|t,\varepsilon,\boldsymbol{y}). \end{aligned}$$

Therefore the weighted response can be justified by

$$\begin{aligned} E\left\{\frac{I(X\le t, D=d)}{G(X|\boldsymbol{Y})\pi(\tilde{\boldsymbol{Y}};\boldsymbol{\gamma_0})}\middle|\boldsymbol{Y}\right\} &= E\left[E\left\{\frac{I(T\le t,\boldsymbol{\epsilon}=d,C\ge T,R=1)}{G(T|\boldsymbol{Y})\pi(\tilde{\boldsymbol{Y}};\boldsymbol{\gamma_0})}\middle|T,\boldsymbol{\epsilon},\boldsymbol{Y}\right\}\middle|\boldsymbol{Y}\right] \\ &= E\left[\frac{I(T\le t,\boldsymbol{\epsilon}=d)G(T|\boldsymbol{Y})\pi(\tilde{\boldsymbol{Y}};\boldsymbol{\gamma_0})}{G(T|\boldsymbol{Y})\pi(\tilde{\boldsymbol{Y}};\boldsymbol{\gamma_0})}\middle|\boldsymbol{Y}\right] \\ &= \Pr(T\ge t,\boldsymbol{\epsilon}=d|\boldsymbol{Y}) = F_d(t|\boldsymbol{Y}). \end{aligned}$$

2.7.4 Proof of Theorem 2.3.1

Lemma 2.7.1. *The inverse probability weight $\hat{\pi}_i = \pi(\boldsymbol{Y_i};\hat{\boldsymbol{\gamma}})$ satisfies $sup_i|\hat{\pi}_i - \pi_i| = o_p(1)$.*

Proof. Before getting into direct proof of the lemma, we first establish the asymptotic properties for $\hat{\boldsymbol{\gamma}}$, which is also useful in later proofs. As introduced in Section 2.2, a logistic regression model is proposed for the completeness R that satisfies equation 2.3.2.

Specifically, the log-likelihood of the model is

$$l(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} D_{ci}[R_i(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}}) - \log\{1 + \exp(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}})],$$

with the corresponding score function

$$\boldsymbol{S}(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} D_{ci}\boldsymbol{S}_i(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} D_{ci}\left\{R_i - \frac{\exp(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}})}{1 + \exp(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}})}\right\}\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}$$

and information matrix

$$\boldsymbol{I}(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} D_{ci}\boldsymbol{I}_i(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} D_{ci}\frac{\exp(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}})}{\{1 + \exp(\boldsymbol{\gamma}_0^T \tilde{\boldsymbol{Y}}_{\boldsymbol{i}})\}^2}\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}^T.$$

The maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$ can be obtained by Newton-Raphson method (Agresti, 2003, p.194). By Wald (1943), $\hat{\boldsymbol{\gamma}}$ satisfies $|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0| = o_p(1)$ and $\sqrt{n_D}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$, where $n_D = \sum_{i=1}^{n} D_{ci}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = E[\{E[\boldsymbol{I}_1(\boldsymbol{\gamma}_0)]^{-1}\boldsymbol{S}_1(\boldsymbol{\gamma}_0)\}^{\otimes 2}]$. Assuming that $\frac{n}{n_D} \to q$. Then we also have the influence function

$$\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0) = qD_{ci}E[\boldsymbol{I}_1(\boldsymbol{\gamma}_0)]^{-1}\boldsymbol{S}_i(\boldsymbol{\gamma}_0)$$

satisfying $||\sqrt{n}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)|| \xrightarrow{p} 0$ and $\sup_i|\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)| < \infty$. $\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)$ can be estimated by $\boldsymbol{\iota}_i(\hat{\boldsymbol{\gamma}})$.

In order to prove the lemma, we notice that $\pi(\tilde{\boldsymbol{Y}}; \boldsymbol{\gamma})$ defined in equation (2) is a continuous function of $\boldsymbol{\gamma}$. By Taylor's expansion, we have for any i

$$\pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \hat{\boldsymbol{\gamma}}) - \pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma}_{\boldsymbol{0}}) = \frac{\partial}{\partial \boldsymbol{\gamma}}\pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma}_{\boldsymbol{0}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + o_p(1).$$

Since $\frac{\partial}{\partial \boldsymbol{\gamma}}\pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma}) = \pi_i(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma})\{1 - \pi_i(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma})\}\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}$ is bounded under regularity conditions (C1) and (C2), we have $|\frac{\partial}{\partial \boldsymbol{\gamma}}\pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma})| \leq M$, where M is constant and $M > 0$. Thus for any i

$$|\hat{\pi}_i - \pi_i| = |\pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \hat{\boldsymbol{\gamma}}) - \pi(\tilde{\boldsymbol{Y}}_{\boldsymbol{i}}; \boldsymbol{\gamma}_{\boldsymbol{0}})| \leq M \cdot o_p(1) + o_p(1) = o_p(1),$$

and $\sup_i|\hat{\pi}_i - \pi_i| = o_p(1)$.

□

Proof of Theorem 2.3.1

Note that the estimating function $\boldsymbol{U}_n(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t) = n^{-1/2}\sum_{i=1}^n \boldsymbol{K}_i(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, \boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)$. First, by condition (C1), for every $r > 0$, we have $\sup_{t<\boldsymbol{\nu}}|\hat{G}(t) - G(t)| = o_p(n^{-1/2+r})$. Moreover, by Lemma 1, $\sup_i|\hat{\pi}_i - \pi_i| = o_p(1)$. These, combined with (C2) and (C3), implies that

$$\sup_{i,t,\boldsymbol{\alpha}}||\boldsymbol{K}_i(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}(t), \boldsymbol{\theta}, t) - \boldsymbol{K}_i^{G\pi}(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)|| = o_p(1). \tag{2.7.4}$$

Define $\mathcal{G} = \left\{\boldsymbol{K}_i^{G\pi}(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}(t), \boldsymbol{\theta}, t) : \boldsymbol{\tau}(t), \boldsymbol{\alpha}(t) \in \{\ell_c^\infty([l,u])\}^{L+q}, \boldsymbol{\eta}, \boldsymbol{\theta} \in \mathbb{R}^h, t \in [l,u]\right\}$. The class $\mathcal{G}$ is Donsker since indicator function is Donsker; since Lipschitz transformation maintains Donsker class' property. Consequently $\mathcal{G}$ is Glivenko-Cantelli since a Donsker class is a Glivenko-Cantelli class in probability (van der Vaart and Wellner, 1996). It then follows from Glivenko-Cantelli Theorem (van der Vaart and Wellner, 1996) that for any $\boldsymbol{\tau}(t)$ and $\boldsymbol{\eta}$,

$$\begin{aligned}\sup_{t\in[l,u]}||n^{-1}\sum_{i=1}^n \boldsymbol{K}_i^{G\pi}(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}_0(t),&\boldsymbol{\theta}_0, t)|| \\ &\xrightarrow{p} E\{\boldsymbol{K}_1^{G\pi}(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)\} = 0.\end{aligned} \tag{2.7.5}$$

Then (2.7.4) and (2.7.5) implies for any $\boldsymbol{\tau}(t)$ and $\boldsymbol{\eta}$.

$$\sup_{t\in[l,u]}||n^{-1}\sum_{i=1}^n \boldsymbol{K}_i(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)|| \xrightarrow{p} 0. \tag{2.7.6}$$

For any $\tilde{\boldsymbol{\alpha}}(t) \in \{\ell_c^\infty([l,u])\}^{L+q}$, by some algebra

$$
\begin{aligned}
n^{-1/2}\boldsymbol{U}_n(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{K}_i(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t) \\
&\qquad - \partial_{\boldsymbol{\alpha}}\Psi_i(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)\big[\Psi_i(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) - \Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{K}_i(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t) \\
&\quad - \partial_{\boldsymbol{\alpha}}\Psi_i(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)\Big[\partial_{\boldsymbol{\alpha}}\Psi_i(\check{\boldsymbol{\alpha}}(t),\check{\boldsymbol{\theta}},t)^T[\tilde{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)] + \partial_{\theta}\Psi_i(\check{\boldsymbol{\alpha}}(t),\check{\boldsymbol{\theta}},t)^T[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0]\Big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{K}_i(\tilde{\boldsymbol{\alpha}}(t),\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t) \\
&\qquad - \partial_{\boldsymbol{\alpha}}\Psi_i(\tilde{\boldsymbol{\alpha}}(t),\boldsymbol{\theta}_0,t)\Big[\partial_{\boldsymbol{\alpha}}\Psi_i(\check{\boldsymbol{\alpha}}(t),\boldsymbol{\theta}_0,t)^T[\tilde{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)]\Big]\Big) + \boldsymbol{\epsilon}_n(t),
\end{aligned}
\tag{2.7.7}
$$

where $\{\check{\boldsymbol{\alpha}}(t)^T, \check{\boldsymbol{\theta}}^T\}^T$ is between $\{\tilde{\boldsymbol{\alpha}}(t)^T, \hat{\boldsymbol{\theta}}^T\}^T$ and $\{\boldsymbol{\alpha}_0(t)^T, {\boldsymbol{\theta}_0}^T\}^T$. Since $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, $\sup_{t\in[l,u]}||\boldsymbol{\epsilon}_n(t)|| \xrightarrow{p} 0$ follows from boundedness of $\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0$ by (C2) and the Lipschitz continuity of $\partial_{\boldsymbol{\alpha}}\Psi_i$ and $\partial_{\boldsymbol{\theta}}\Psi_i$ indicated by (C3).

By similar arguments with respect to $\mathcal{G}$, we can establish the Glivenko-Cantelli and Donsker properties for

$$
\mathcal{G}^* = \big\{\boldsymbol{L}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t), \boldsymbol{\tau}(t),\boldsymbol{\alpha}(t) \in \{\ell_c^\infty([l,u])\}^{L+q}, \boldsymbol{\eta},\boldsymbol{\theta}\in\mathbb{R}^h, t\in[l,u]\big\}
$$

and

$$
\mathcal{G}^{**} = \big\{\boldsymbol{B}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t), \boldsymbol{\tau}(t),\boldsymbol{\alpha}(t) \in \{\ell_c^\infty([l,u])\}^{L+q}, \boldsymbol{\eta},\boldsymbol{\theta}\in\mathbb{R}^h, t\in[l,u]\big\}.
$$

Then for $\boldsymbol{\tau}(t),\boldsymbol{\alpha}(t) \in \{\ell_c^\infty([l,u])\}^{L+q}, \boldsymbol{\eta},\boldsymbol{\theta}\in\mathbb{R}^h$, $n^{-1}\sum_{i=1}^n \boldsymbol{L}_i(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)$ converges to $\boldsymbol{L}(\boldsymbol{\tau}(t),\boldsymbol{\eta},\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)$ uniformly in $t\in[l,u]$. Together with (2.7.6) and (2.7.7), we have

$$
n^{-1/2}\boldsymbol{U}_n(\tilde{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) = -\boldsymbol{L}(\tilde{\boldsymbol{\alpha}}(t),\boldsymbol{\theta}_0,\check{\boldsymbol{\alpha}}(t),\boldsymbol{\theta}_0,t)[\tilde{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)] + \boldsymbol{\epsilon}_n^*(t)
$$

with $\sup_{t\in[l,u]}||\boldsymbol{\epsilon}_n^*(t)|| \xrightarrow{p} 0$.

By condition (C4) and continuity of $\partial_{\boldsymbol{\alpha}}\Psi_i\partial_{\boldsymbol{\alpha}}\Psi_i^T$, there exists a small neighborhood of $(\boldsymbol{\alpha}_0, \boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_0)$ in $[\{\ell_c^\infty([l,u])\}^{L+q} \times \mathbb{R}^h]^2$, where

$$\inf_{t\in[l,u]}\text{eigmin}\boldsymbol{L}(\boldsymbol{\tau}(t), \boldsymbol{\eta}, \boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)$$

is bounded below by a positive constant k. Therefore, there exists a uniformly bounded root $\hat{\boldsymbol{\alpha}}(t)$ of $\boldsymbol{U}_n$ for n large enough, and the uniform consistency follows from

$$o_p(1) = ||n^{-1/2}\boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)|| \geq k||\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)|| + \boldsymbol{\epsilon}_n^*(t).$$

2.7.5 Proof of Theorem 2.3.2

First, the estimating equation can be decomposed as following:

$$\begin{aligned} 0 =& \boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) \\ =& \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) \\ &+ [\boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)] \\ &+ [\boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)] \end{aligned} \tag{2.7.8}$$

Then we can analyze each component in Equation (2.7.8). First, decompose $\boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)$ $-\boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)$ as following:

$$\begin{aligned} &\boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) \\ &= n^{-1/2}\sum_{i=1}^n \left(\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)\right)\left[\frac{I(X_i \leq t, D_i = d)}{G(X_i)\pi_i}\right. \\ &\qquad\qquad \left. - \Psi_i(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)\right] \\ &\quad - n^{-1/2}\sum_{i=1}^n \partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)[\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \Psi_i(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)]. \end{aligned} \tag{2.7.9}$$

Let

$$\boldsymbol{U}_n^{[1]}(t) = n^{-1/2}\sum_{i=1}^{n}\left(\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) - \partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\right)\Bigg[\frac{I(X_i\le t, D_i=d)}{G(X_i)\pi_i} - \Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\Bigg].$$

Then by Taylor expansion of $\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)$ at $(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0)^T$

$$\boldsymbol{U}_n^{[1]}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\partial^2_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\right)\Bigg[\frac{I(X_i\le t, D_i=d)}{G(X_i)\pi_i} - \Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\Bigg]\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix} + \boldsymbol{\nu}_n(t),$$

where $\sup_{t\in[l,u]}\boldsymbol{\nu}_n(t)\xrightarrow{p}0$ by uniform consistency of $\hat{\boldsymbol{\alpha}}(t)$ and consistency of $\hat{\boldsymbol{\theta}}$. Since conditions (C2) and (C5) imply that $\partial^2_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)$ is uniformly bounded and $E[\frac{I(X_1\le t,D_1=d)}{G(X_1)\pi_1} - \Psi_1(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)|\boldsymbol{Y}_1] = 0$, $\boldsymbol{U}_n^{[1]}(t)$ can be written as

$$\boldsymbol{U}_n^{[1]}(t) = \tilde{\boldsymbol{\nu}}_n(t)\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix}, \tag{2.7.10}$$

where $\sup_{t\in[l,u]}\tilde{\boldsymbol{\nu}}_n(t)\xrightarrow{p}0$.

Also let $\boldsymbol{U}_n^{[2]}(t) = -n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)[\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) - \Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)]$. By Taylor expansion of $\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)$ at $(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0)^T$,

$$\boldsymbol{U}_n^{[2]}(t) = \frac{1}{n}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)\Big[\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)^T\sqrt{n}(\hat{\boldsymbol{\alpha}}(t)-\boldsymbol{\alpha}_0(t)) + \boldsymbol{\nu}_n^*(t) + \partial_{\boldsymbol{\theta}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)^T\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \boldsymbol{\nu}_n^{**}(t)\Big],$$

where $\boldsymbol{\nu}_n^*(t)$ and $\boldsymbol{\nu}_n^{**}(t)$ uniformly converge to 0 in probability for $t\in[l,u]$, implied by uniform consistency of $\hat{\boldsymbol{\alpha}}(t)$ and consistency of $\hat{\boldsymbol{\theta}}$. Again by consistency properties of $\hat{\boldsymbol{\alpha}}(t)$ and $\hat{\boldsymbol{\theta}}$, we have $\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t) = \partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t) + \boldsymbol{\nu}_n^{***}(t)$ where $\sup_{t\in[l,u]}\boldsymbol{\nu}_n^{***}(t)\xrightarrow{p}0$. Then since $\mathcal{G}^*$ and $\mathcal{G}^{**}$ are Glivenko-Cantelli, by Glivenko-Cantelli Theorem (van der Vaart

and Wellner, 1996) and the existence of influence function $\boldsymbol{\phi}_i$ for $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned}\boldsymbol{U}_n^{[2]}(t) &= \{\boldsymbol{J}(t) + \boldsymbol{\varepsilon}_n^*(t)\}\sqrt{n}(\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)) + \{\boldsymbol{H}(t) + \boldsymbol{\varepsilon}_n^{**}(t)\}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \\ &= \boldsymbol{J}(t)\sqrt{n}(\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)) + \boldsymbol{H}(t)n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\phi}_i(\boldsymbol{\theta}_0) + \tilde{\boldsymbol{\varepsilon}}(t)\end{aligned} \tag{2.7.11}$$

where $\boldsymbol{\varepsilon}_n^*(t)$, $\boldsymbol{\varepsilon}_n^{**}(t)$ and $\tilde{\boldsymbol{\varepsilon}}(t)$ uniformly converge to 0 in probability for $t \in [l, u]$.

Next, we assess another component in Equation (2.7.8), $\boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)$. As observed in Equation (2.7.12), this component also has two parts, one related to $\hat{G}(X_i) - G(X_i)$ (denoted by $\boldsymbol{U}_n^{[3]}(t)$) and the other related to $\hat{\pi}_i - \pi_i$ (denoted by $\boldsymbol{U}_n^{[4]}(t)$).

$$\begin{aligned}&\boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) \\ &= -n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)\frac{I(X_i \le t, D_i = d)[\hat{G}(X_i)\hat{\pi}_i - G(X_i)\pi_i]}{\hat{G}(X_i)\hat{\pi}_i G(X_i)\pi_i} \\ &= -n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)I(X_i \le t, D_i = d)\left[\frac{\hat{G}(X_i) - G(X_i)}{\hat{G}(X_i)G(X_i)\pi_i} + \frac{\hat{\pi}_i - \pi_i}{\hat{G}(X_i)\hat{\pi}_i\pi_i}\right]. \\ &= \boldsymbol{U}_n^{[3]}(t) + \boldsymbol{U}_n^{[4]}(t)\end{aligned} \tag{2.7.12}$$

From (C1) and Pepe (1991),

$$\sup_{t\in[0,\boldsymbol{\nu}]}||n^{1/2}\{\hat{G}(t) - G(t)\} - n^{-1/2}\sum_{i}^{n}G(t)\int_0^t \upsilon(s)^{-1}dM_i^G(s)|| \to 0.$$

Also by standard empirical process argument, we can show that

$$n^{-1}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)I(X_i \le t, D_i = d)\Upsilon(s)G(X)^{-1}\pi_i^{-1}$$

converges to $\boldsymbol{w}_G(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)$ uniformly. Together with the consistency properties of $\hat{\boldsymbol{\alpha}}(t)$ and

$\hat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}
\boldsymbol{U}_n^{[3]}(t) &= -n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_i\le t,D_i=d)\frac{\hat{G}(X_i)-G(X_i)}{\hat{G}(X_i)G(X_i)\pi_i}\\
&\approx -n^{-1}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)\frac{n^{-1/2}\sum_{j=1}^{n}I(X_i\le t,D_i=d)\int_0^\infty\Upsilon_i(s)\upsilon^{-1}(s)dM_j^G(s)}{G(X_i)\pi_i}\\
&= -n^{-1/2}\sum_{j=1}^{n}\int_0^\infty\frac{1}{n}\Big(\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)\frac{I(X_i\le t,D_i=d)\Upsilon_i(s)}{G(X_i)\pi_i}\Big)\frac{dM_j^G(s)}{\upsilon(s)}\\
&\approx -n^{-1/2}\sum_{j=1}^{n}\int_0^\infty\boldsymbol{w}_G(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\frac{dM_j^G(s)}{\upsilon(s)}\\
&= -n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Xi}_{1i}(t),
\end{aligned}\tag{2.7.13}$$

where $\approx$ represents asymptotic equivalence uniformly in $t\in[l,u]$. For $\boldsymbol{U}_n^{[4]}(t)$, in the proof of Lemma 1 we have derived the influence function $\boldsymbol{\iota}_i$ for the logistic regression estimates $\hat{\boldsymbol{\gamma}}$. Moreover, by empirical process arguments we can show

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\partial_{\boldsymbol{\alpha}}\Psi_j(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)I(X_j\le t,D_j=d)\{\frac{\partial}{\partial\boldsymbol{\gamma}}\pi_i\}^T}{G(X_j)\pi_i^2}$$

converges to $\boldsymbol{w}_\pi(\boldsymbol{\alpha}(t),\boldsymbol{\theta},t)$ uniformly. Taking Taylor expansion of $\hat{\pi}_i$ at $\boldsymbol{\gamma_0}$ and by the consistency properties of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}$ we have

$$\begin{aligned}
\boldsymbol{U}_n^{[4]}(t) &= -n^{-1/2}\sum_{i=1}^{n}\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_i\le t,D_i=d)\frac{\hat{\pi}_i-\pi_i}{\hat{G}(X_i)\hat{\pi}_i\pi_i}\\
&\approx -n^{-1}\sum_{i=1}^{n}\frac{\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_i\le t,D_i=d)}{G(X_i)\pi_i^2}\big(\frac{\partial}{\partial\boldsymbol{\gamma}}\pi_i\big)^T\sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\\
&\approx -n^{-1}\sum_{i=1}^{n}\frac{\partial_{\boldsymbol{\alpha}}\Psi_i(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_i\le t,D_i=d)\{\frac{\partial}{\partial\boldsymbol{\gamma}}\pi_i\}^T}{G(X_i)\pi_i^2}n^{-1/2}\sum_{j=1}^{n}\boldsymbol{\iota}_j(\boldsymbol{\gamma}_0)\\
&= -n^{-1/2}\sum_{i=1}^{n}\Big[\frac{1}{n}\sum_{j=1}^{n}\frac{\partial_{\boldsymbol{\alpha}}\Psi_j(\hat{\boldsymbol{\alpha}}(t),\hat{\boldsymbol{\theta}},t)I(X_j\le t,D_j=d)\{\frac{\partial}{\partial\boldsymbol{\gamma}}\pi_j\}^T}{G(X_j)\pi_j^2}\Big]\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)\\
&\approx -n^{-1/2}\sum_{i=1}^{n}\boldsymbol{w}_\pi(\boldsymbol{\alpha}_0(t),\boldsymbol{\theta}_0,t)\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)\\
&= -n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Xi}_{2i}(t).
\end{aligned}\tag{2.7.14}$$

We claim that $\mathcal{G}^{***} = \{\boldsymbol{\Xi}_{1i}(t), t \in [l,u]\}$ and $\mathcal{G}^{****} = \{\boldsymbol{\Xi}_{2i}(t), t \in [l,u]\}$ are Donsker. It can be shown that $\int_0^\infty \boldsymbol{w}_G(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)\frac{dM_j^G(s)}{y(s)}$ and $\boldsymbol{w}_\pi(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t)\boldsymbol{\iota}_i(\boldsymbol{\gamma}_0)$ are Lipschitz in t, thus Donsker property follows under Lipschitz transformations.

Therefore, combining formulas (2.7.8), (2.7.10), (2.7.11), (2.7.13) and (2.7.14)

$$\begin{aligned}
0 =& \boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) \\
=& \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) + [\boldsymbol{U}_n(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t)] \\
&+ [\boldsymbol{U}_n^{G\pi}(\hat{\boldsymbol{\alpha}}(t), \hat{\boldsymbol{\theta}}, t) - \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta_0}, t)] \\
=& \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) + \boldsymbol{U}_n^{[1]}(t) - \boldsymbol{U}_n^{[2]}(t) + \boldsymbol{U}_n^{[3]}(t) + \boldsymbol{U}_n^{[4]}(t). \\
\approx& \boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) - \boldsymbol{J}(t)\sqrt{n}\{\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)\} - \boldsymbol{H}(t)n^{-1/2}\sum_{i=1}^n \phi_i(\boldsymbol{\theta}_0) \\
&- n^{-1/2}\sum_{i=1}^n \boldsymbol{\Xi}_{1i}(t) - n^{-1/2}\sum_{i=1}^n \boldsymbol{\Xi}_{2i}(t).
\end{aligned} \tag{2.7.15}$$

Note that $\boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t) = n^{-1/2}\sum_{i=1}^n \boldsymbol{A}_i^{G\pi}(\boldsymbol{\alpha}(t), \boldsymbol{\theta}, t)$. Thus,

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0(t)) \\
&= \boldsymbol{J}(t)^{-1}\Bigg\{\boldsymbol{U}_n^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) \\
&\qquad + n^{-1/2}\sum_{i=1}^n \big[- \boldsymbol{H}(t)\phi_i(\boldsymbol{\theta}_0) - \boldsymbol{\Xi}_{1i}(t) - \boldsymbol{\Xi}_{2i}(t)\big]\Bigg\} + \boldsymbol{\tau}_n(t) \\
&= \boldsymbol{J}(t)^{-1}\Bigg\{n^{-1/2}\sum_{i=1}^n \big[\boldsymbol{A}_i^{G\pi}(\boldsymbol{\alpha}_0(t), \boldsymbol{\theta}_0, t) \\
&\qquad - \boldsymbol{H}(t)\phi_i(\boldsymbol{\theta}_0) - \boldsymbol{\Xi}_{1i}(t) - \boldsymbol{\Xi}_{2i}(t)\big]\Bigg\} + \boldsymbol{\tau}_n(t),
\end{aligned} \tag{2.7.16}$$

where $\sup_{t\in[l,u]}\boldsymbol{\tau}_n(t) \xrightarrow{p} 0$. Weak convergence follows since $\boldsymbol{A}_i, \boldsymbol{\Xi}_{1i}$ and $\boldsymbol{\Xi}_{2i}$ has been shown as Donsker and $\boldsymbol{\phi}_i$ is assumed to be Donsker (van der Vaart and Wellner, 1996).

Uniform consistency of $\hat{\boldsymbol{J}}(t)$ and $\hat{\boldsymbol{H}}(t)$ follows from Glivenko-Cantelli property of $\mathcal{G}^*$ and $\mathcal{G}^{**}$ and uniform consistency of $\hat{\boldsymbol{\alpha}}(t)$ and $\hat{\boldsymbol{\theta}}$. Since

$$\{\boldsymbol{A}_i^{G\pi}(\boldsymbol{\alpha}(t), \theta, t), t \in [l,u]\}, \{\boldsymbol{\Xi}_{1i}(t), t \in [l,u]\}, \{\boldsymbol{\Xi}_{2i}(t), t \in [l,u]\}$$

are Glivenko-Cantelli and $\{\boldsymbol{\phi}_i\}$, $\boldsymbol{J}(t)^{-1}$, $\boldsymbol{H}(t)$ is bounded for $t \in [l, u]$, $\{\boldsymbol{Q}_i(t), t \in [l, u]\}$ is also Glivenko-Cantelli. By Slutsky's theorem and uniform law of large numbers, $\hat{\boldsymbol{\Sigma}}(s, t)$ converges to $\boldsymbol{\Sigma}(s, t)$ uniformly.

2.7.6 Further simulation about selecting the number of latent classes

We conduct further simulation studies to investigate how well the ICL-BIC information criterion selects the number of latent classes. Specifically, we generated 5000 datasets for each configuration used in the comparison (A) in our main simulation study for sample size $N \in \{500, 1000, 2000\}$. Then we fitted finite Gaussian mixture models with number of latent classes $L \in \{1, \ldots, 9\}$. Then we used integrated classification likelihood-BIC (ICL-BIC) to select the best L for each simulated dataset.

As shown in Table 2.5, the ICL-BIC criterion worked best when the sample size was large and the baseline covariates were mildly or moderately overlapped. When the latent class pattern was severely overlapped (entropy index smaller than 0.7), however, the information criterion may not be able to detect the true number of latent classes. This is consistent with our real application results, where ICL-BIC preferred a two-class model while our domain knowledge perferred at least three classes. Therefore, we recommend using information criteria together with domain knowledge in choosing the number of latent classes.

Table 2.5: Percentage of selected number L of latent classes, by ICL-BIC, under different choices of simulation scenario and sample size N from 5000 simulations.

		Percentage selected (%)			
Scenarios	N	$L = 1$	$L = 2$	$L = 3$	$L = 4$
(1b)+(2a)+(3a)	500	0	0.88	99.08	0.04
mild overlapping	1000	0	0.06	99.94	0
entropy around 0.9	2000	0	0	100	0
(1b)+(2b)+(3a)	500	2.64	53.54	43.74	0.08
moderate overlapping	1000	0.06	39.84	60.08	0.02
entropy around 0.8	2000	0	21.72	78.28	0
(1b)+(2c)+(3a)	500	30.68	60.82	8.50	0
moderately severe ovarlapping	1000	12.36	78.58	9.06	0
entropy around 0.7	2000	2.72	90.14	7.14	0
(1b)+(2d)+(3a)	500	76.76	22.78	0.46	0
severe ovarlapping	1000	73.94	25.82	0.24	0
entropy around 0.6	2000	70.12	29.88	0	0

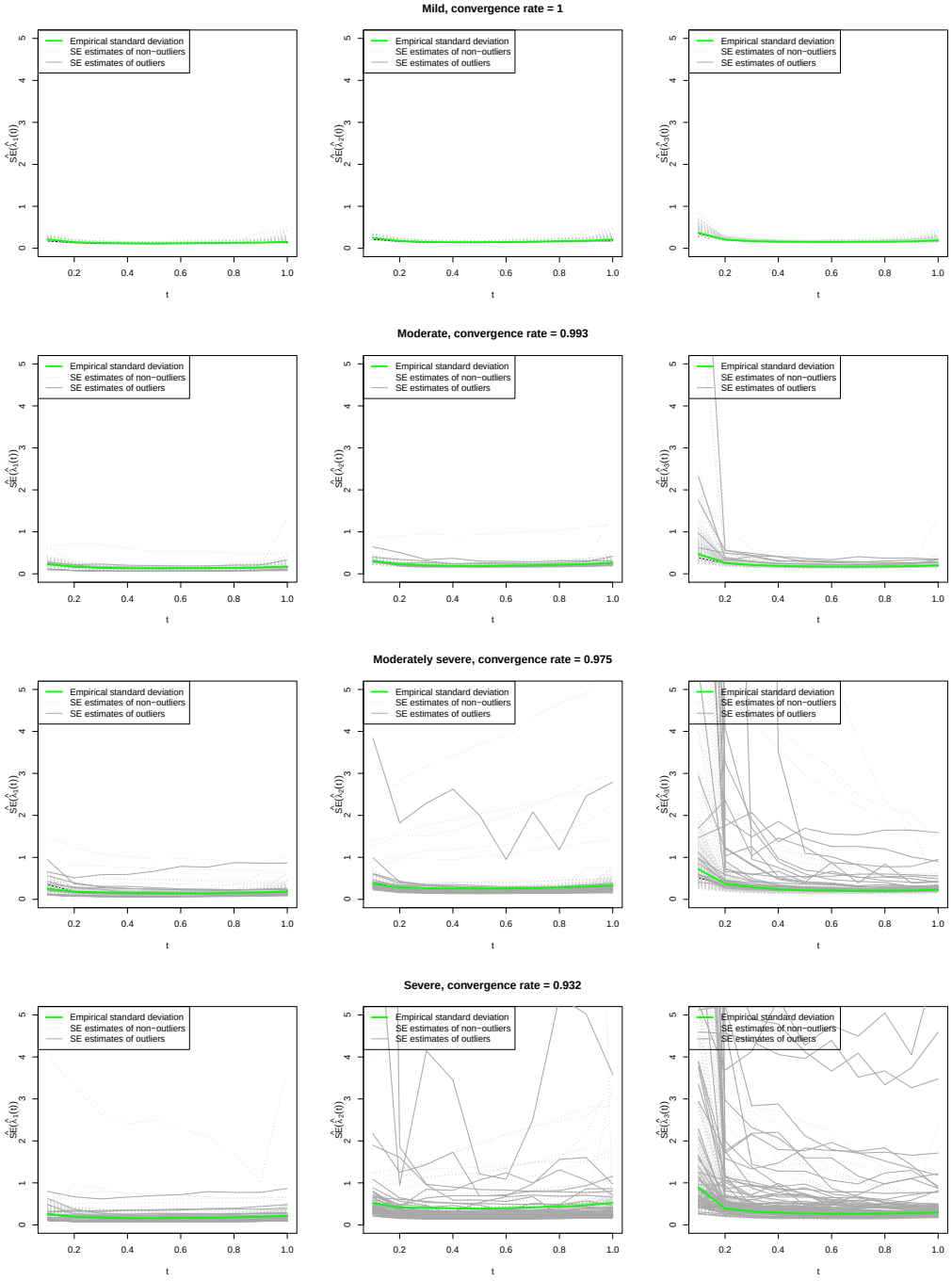

Figure 2.1: Standard error (SE) estimation of $\hat{\boldsymbol{\lambda}}(t) = \{\hat{\lambda}_1(t), \hat{\lambda}_2(t), \hat{\lambda}_3(t)\}$ under the four scenarios of overlapping in comparison (A), namely mild, moderate, moderately severe, and severe. Green line denotes the empirical standard deviation of the estimates. Light gray dotted lines represent SE estimates of non-outliers. Dark gray lines display SE estimates of outliers.

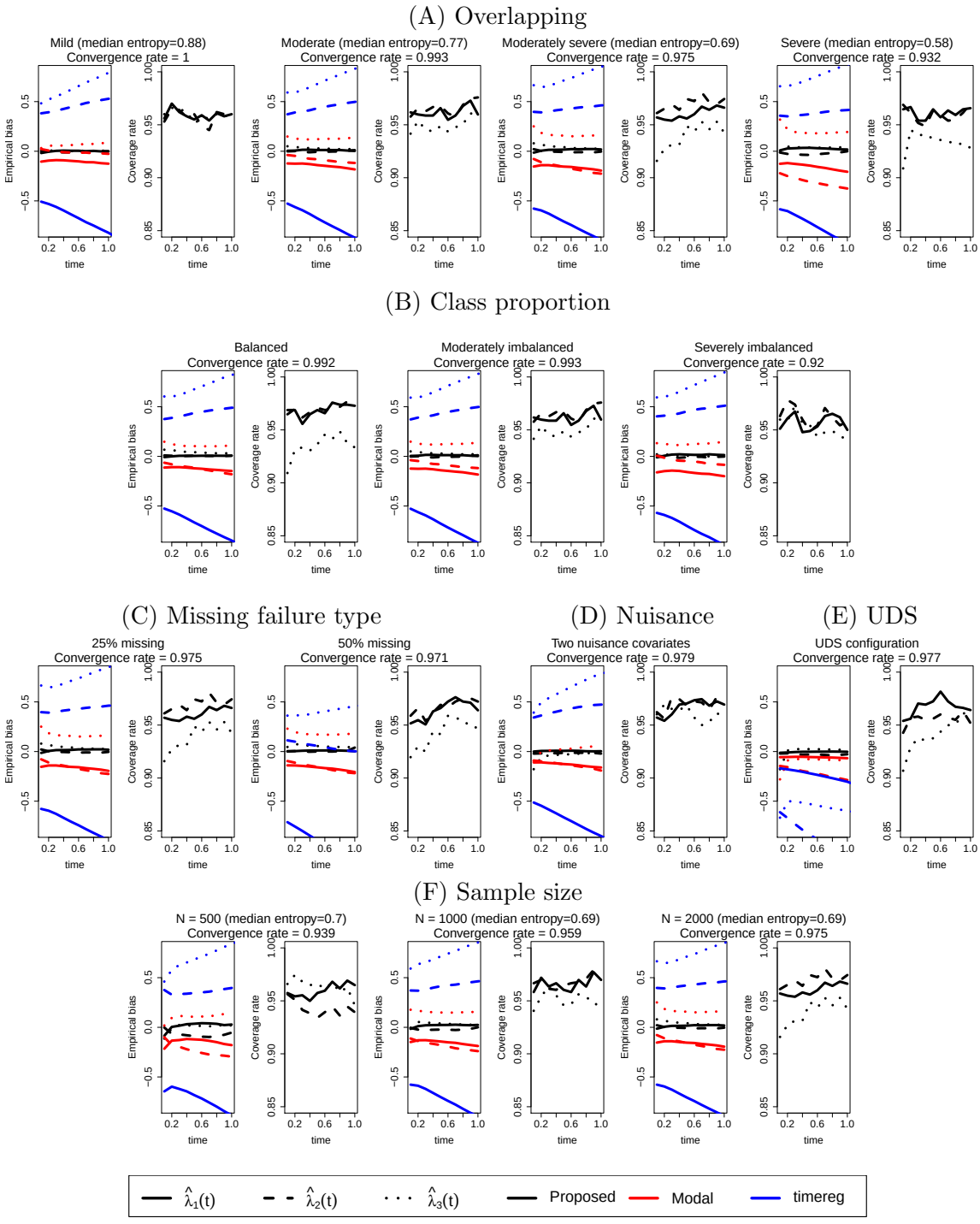

Figure 2.2: Simulation results for investigation purposes (A)-(F). Quantities associated with the three regression coefficients in $\boldsymbol{\lambda}_0(t)$ are represented by solid, dashed and dotted lines. The proposed, modal assignment, and `timereg` strategies are respectively shown in black, red and blue lines.

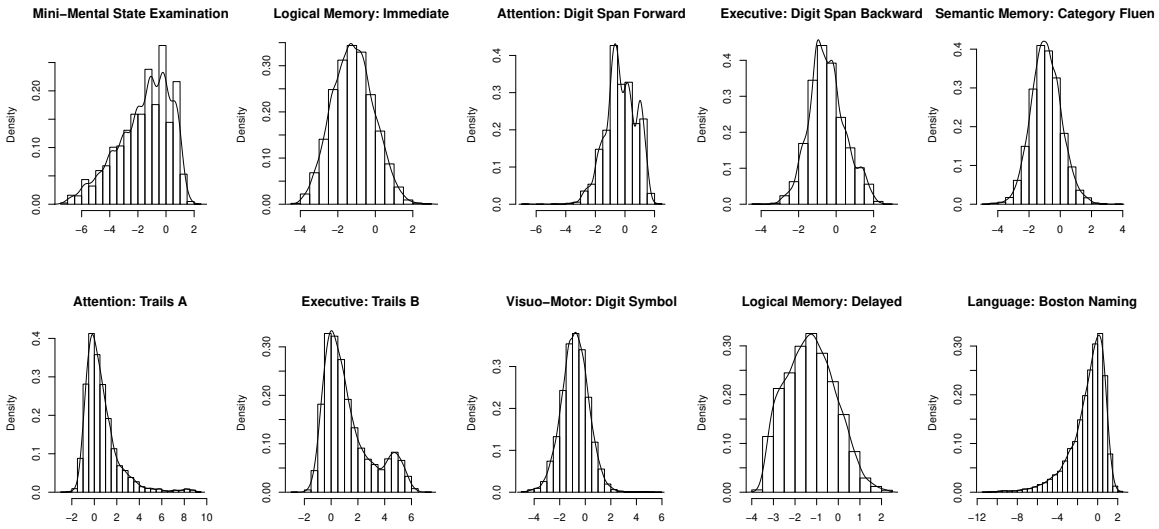

Figure 2.3: Histogram and kernel density estimation for the ten cognitive test scores.

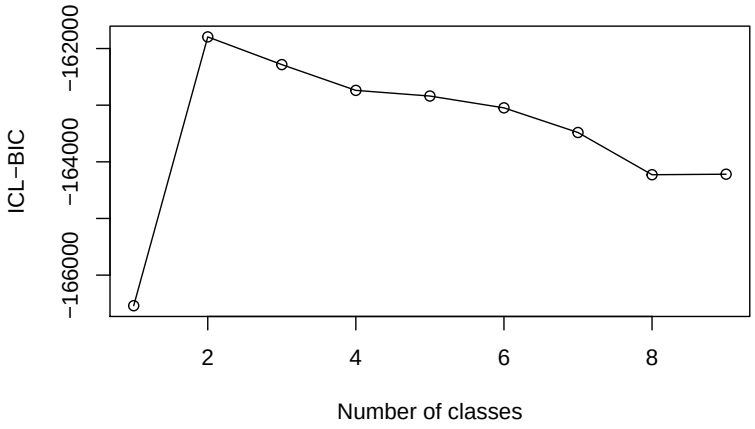

Figure 2.4: ICL-BIC for the finite Gaussian mixture models with different number of classes of the ten baseline cognitive test scores.

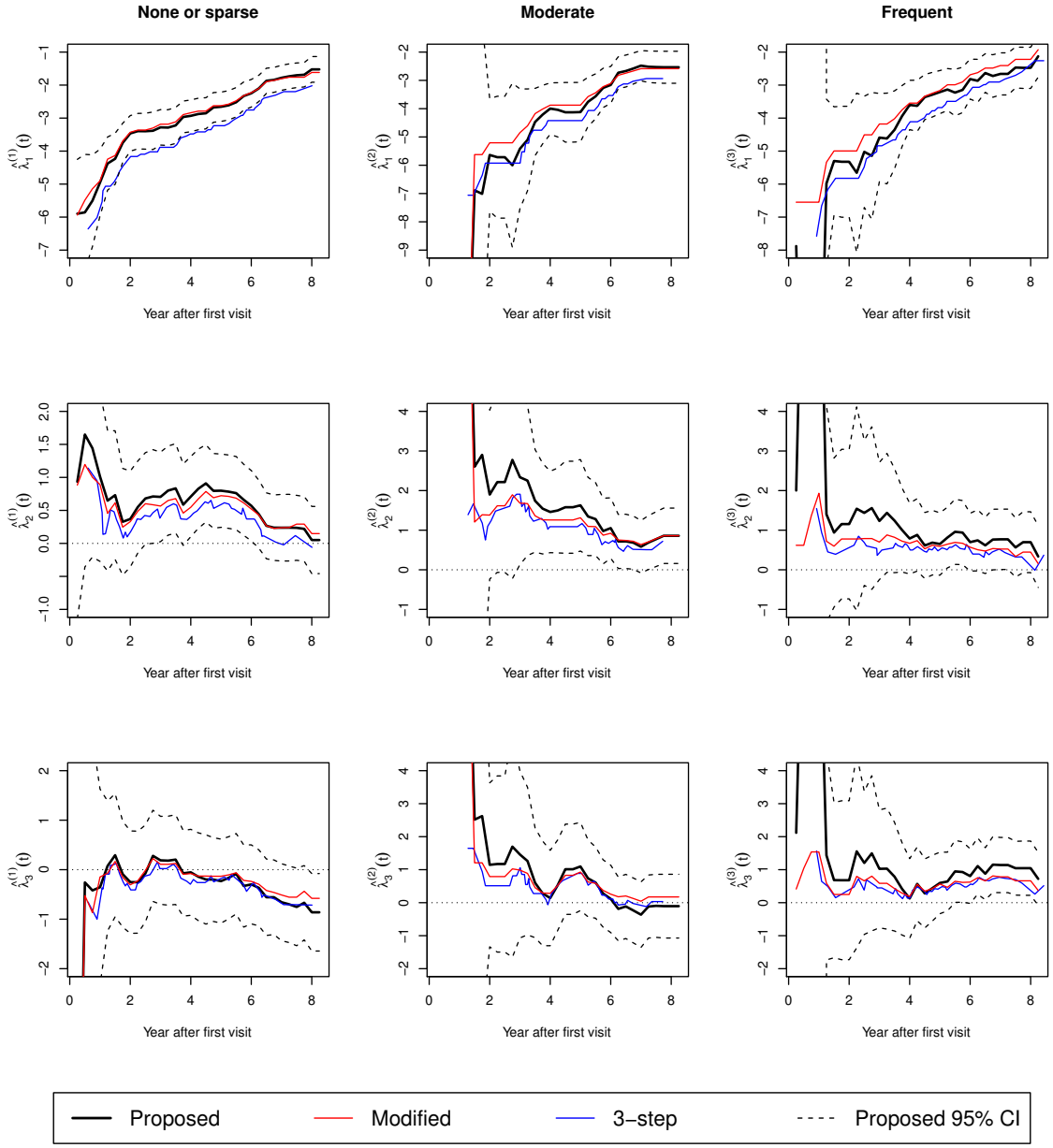

Figure 2.5: Point estimates $\hat{\lambda}_l^{(i)}(t), l = 1, 2, 3, i = 1, 2, 3, t \in [0.25, 8]$, for the l-th regression coefficient for the i-th competing risk outcome. Each column shows the three regression coefficients for the corresponding competing risk. Point estimates obtained by the proposed, modal and naive approaches are respectively represented by black, red and blue solid lines. Black dashed lines represent the 95% confidence intervals for the proposed estimator.

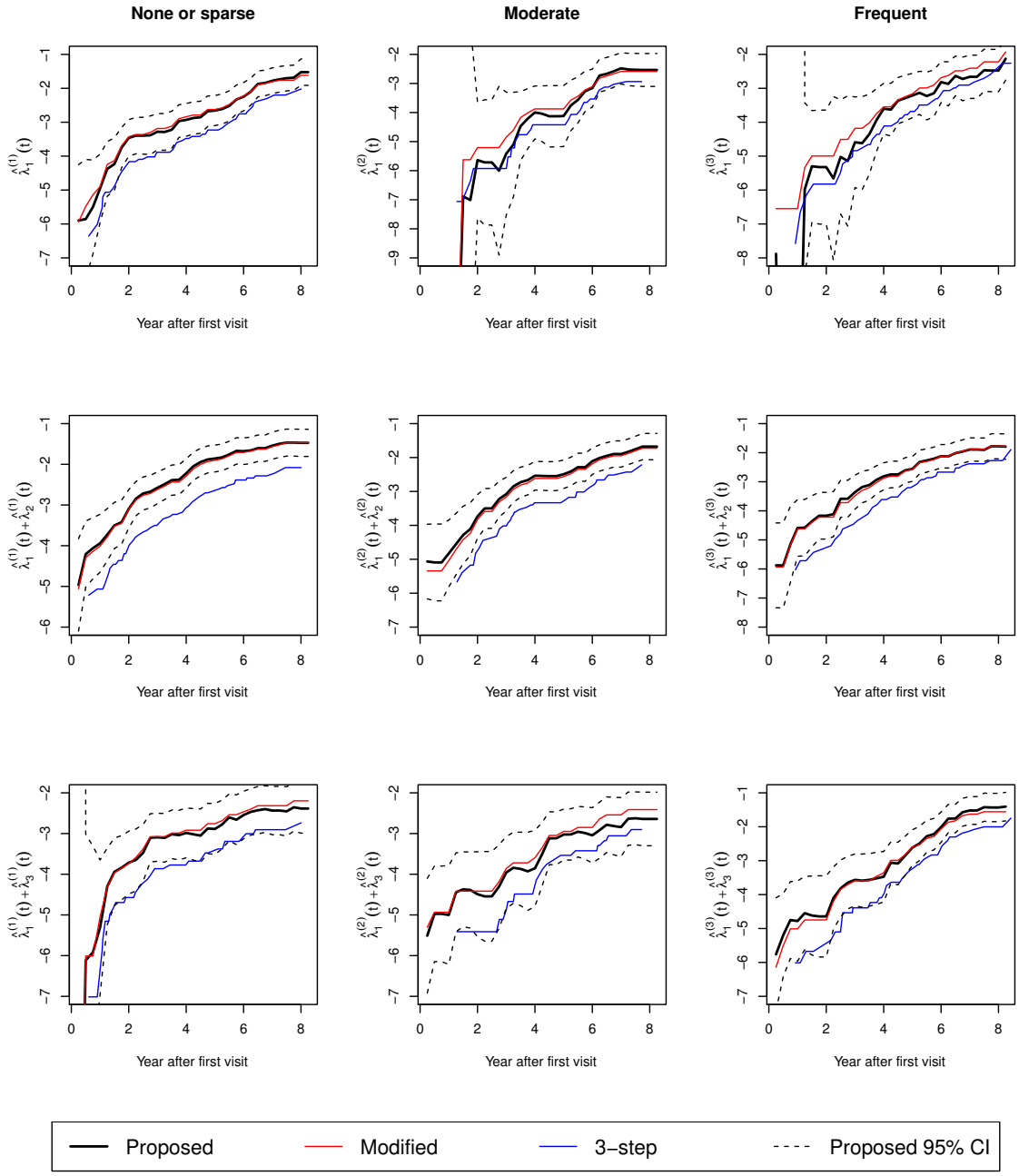

Figure 2.6: Estimated g^{-1} transformed cumulative incidence for class $l, l = 1, 2, 3$ for competing risk $i, i = 1, 2, 3$ at $t \in [0.25, 8]$. Each column shows the three quantities for the corresponding competing risk. Estimates obtained by the proposed, modal and naive approaches are respectively represented by black, red and blue solid lines. Black dashed lines represent the 95% confidence intervals for the proposed estimator.

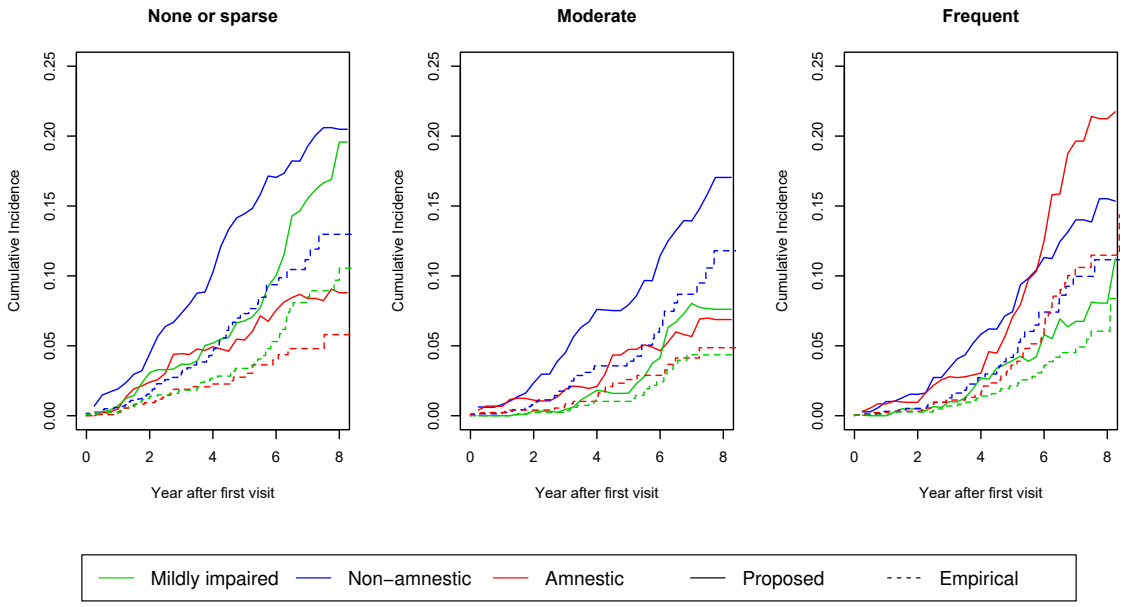

Figure 2.7: Cumulative incidence curves of death with the three CERAD phenotypes. Solid lines represent the predicted cumulative incidence by the proposed method. Dashed lines represent the empirical cumulative incidence curves. Mildly impaired, non-amnestic MCI and amnestic MCI groups are represented in green, blue and red, respectively.

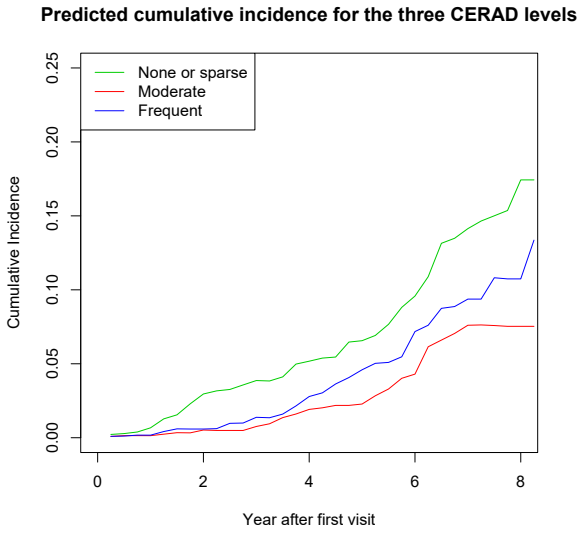

Figure 2.8: Predicted cumulative incidence functions corresponding to different CERAD phenotypes for a new patient with baseline cognitive test scores equal to median values in the observed MCI data.

2.7.7 Additional tables for simulation results

Table 2.6: The empirical coverage rates of the 95% confidence band on $t \in [0.1, 1.0]$ for $\boldsymbol{\lambda}(t)$, for different simulation scenarios, after excluding outlying estimates.

		Coverage rate*		
Comparisons	Scenarios	$\hat{\lambda}_1(t)$	$\hat{\lambda}_2(t)$	$\hat{\lambda}_3(t)$
(A)	(1b)+(2a)+(3a)	0.968	0.966	0.975
	(1b)+(2b)+(3a)	0.968	0.966	0.975
	(1b)+(2c)+(3a)	0.961	0.961	0.977
	(1b)+(2d)+(3a)	0.951	0.968	0.979
(B)	(1a)+(2b)+(3a)	0.970	0.966	0.964
	(1b)+(2b)+(3a)	0.961	0.961	0.977
	(1c)+(2b)+(3a)	0.961	0.960	0.961
(C)	(1b)+(2c)+(3a)	0.951	0.968	0.979
	(1b)+(2c)+(3b)	0.962	0.958	0.984
(D)	Nuisance parameters	0.960	0.978	0.970
(E)	UDS scenario	0.961	0.975	0.975
(F)	N=500	0.953	0.902	0.989
	N=1000	0.972	0.957	0.982
(1b)+(2c)+(3a)	N=2000	0.951	0.968	0.979

* Confidence band was constructed on $t \in [0.1, 1.0]$.

Table 2.7: Standard deviation (SE), average standard error estimate (SEE), and coverage probability (CP) of $\hat{\boldsymbol{\lambda}}(\cdot)$ at different choices of t, for the four scenarios of overlapping, after excluding outlying estimates.

	(1b)+(2a)+(3a) Mild			(1b)+(2b)+(3a) Moderate			(1b)+(2c)+(3a) Moderately severe			(1b)+(2d)+(3a) Severe		
	SE	SEE	CP	SE	SEE	CP	SE	SEE	CP	SE	SEE	CP
$\hat{\lambda}_1(0.1)$	0.212	0.213	0.956	0.225	0.229	0.962	0.251	0.266	0.957	0.253	0.279	0.965
$\hat{\lambda}_1(0.2)$	0.145	0.155	0.970	0.166	0.169	0.960	0.175	0.194	0.955	0.185	0.210	0.967
$\hat{\lambda}_1(0.3)$	0.126	0.133	0.962	0.140	0.146	0.959	0.152	0.168	0.954	0.166	0.184	0.954
$\hat{\lambda}_1(0.4)$	0.120	0.125	0.958	0.128	0.136	0.959	0.143	0.158	0.958	0.152	0.172	0.954
$\hat{\lambda}_1(0.5)$	0.117	0.122	0.955	0.125	0.133	0.966	0.135	0.153	0.956	0.148	0.168	0.965
$\hat{\lambda}_1(0.6)$	0.119	0.123	0.959	0.128	0.134	0.955	0.135	0.154	0.960	0.152	0.168	0.957
$\hat{\lambda}_1(0.7)$	0.123	0.127	0.954	0.130	0.138	0.959	0.135	0.158	0.967	0.157	0.172	0.964
$\hat{\lambda}_1(0.8)$	0.129	0.134	0.961	0.139	0.145	0.968	0.147	0.166	0.964	0.168	0.179	0.959
$\hat{\lambda}_1(0.9)$	0.136	0.146	0.958	0.150	0.157	0.973	0.158	0.179	0.968	0.183	0.192	0.965
$\hat{\lambda}_1(1)$	0.158	0.168	0.960	0.172	0.180	0.960	0.184	0.203	0.966	0.211	0.220	0.966
$\hat{\lambda}_2(0.1)$	0.248	0.249	0.953	0.279	0.281	0.958	0.313	0.337	0.961	0.346	0.375	0.969
$\hat{\lambda}_2(0.2)$	0.175	0.184	0.965	0.206	0.212	0.966	0.226	0.259	0.965	0.271	0.293	0.961
$\hat{\lambda}_2(0.3)$	0.151	0.160	0.964	0.175	0.187	0.963	0.201	0.234	0.969	0.248	0.268	0.952
$\hat{\lambda}_2(0.4)$	0.144	0.151	0.957	0.164	0.178	0.967	0.190	0.228	0.964	0.236	0.262	0.948
$\hat{\lambda}_2(0.5)$	0.144	0.149	0.960	0.161	0.178	0.968	0.184	0.229	0.974	0.232	0.265	0.964
$\hat{\lambda}_2(0.6)$	0.150	0.151	0.949	0.169	0.182	0.961	0.188	0.237	0.972	0.247	0.275	0.959
$\hat{\lambda}_2(0.7)$	0.157	0.158	0.945	0.177	0.191	0.958	0.195	0.251	0.981	0.262	0.292	0.954
$\hat{\lambda}_2(0.8)$	0.167	0.170	0.962	0.192	0.206	0.969	0.219	0.271	0.971	0.282	0.318	0.955
$\hat{\lambda}_2(0.9)$	0.179	0.187	0.960	0.213	0.227	0.975	0.243	0.298	0.969	0.313	0.363	0.961
$\hat{\lambda}_2(1)$	0.211	0.218	0.958	0.248	0.263	0.976	0.282	0.344	0.974	0.395	0.398	0.970
$\hat{\lambda}_3(0.1)$	0.371	0.363	0.960	0.459	0.468	0.942	0.623	0.733	0.916	0.845	1.335	0.909
$\hat{\lambda}_3(0.2)$	0.206	0.219	0.966	0.252	0.250	0.954	0.298	0.318	0.927	0.352	0.381	0.945
$\hat{\lambda}_3(0.3)$	0.170	0.180	0.965	0.203	0.202	0.946	0.236	0.256	0.931	0.278	0.297	0.940
$\hat{\lambda}_3(0.4)$	0.157	0.163	0.962	0.182	0.181	0.944	0.211	0.223	0.931	0.241	0.256	0.941
$\hat{\lambda}_3(0.5)$	0.152	0.155	0.952	0.170	0.171	0.949	0.190	0.205	0.949	0.227	0.234	0.937
$\hat{\lambda}_3(0.6)$	0.150	0.153	0.955	0.167	0.167	0.944	0.185	0.196	0.944	0.218	0.220	0.936
$\hat{\lambda}_3(0.7)$	0.152	0.156	0.957	0.166	0.167	0.948	0.176	0.191	0.953	0.215	0.214	0.935
$\hat{\lambda}_3(0.8)$	0.155	0.162	0.960	0.168	0.173	0.951	0.184	0.194	0.946	0.222	0.215	0.931
$\hat{\lambda}_3(0.9)$	0.164	0.174	0.960	0.176	0.185	0.961	0.192	0.204	0.953	0.231	0.223	0.932
$\hat{\lambda}_3(1)$	0.186	0.198	0.954	0.197	0.208	0.960	0.219	0.225	0.943	0.256	0.249	0.928

Table 2.8: Standard deviation (SE), average standard error estimate (SEE), and coverage probability (CP) of $\hat{\boldsymbol{\lambda}}(\cdot)$ at different choices of t, for the three scenarios for latent class proportion, after excluding outlying estimates.

	(1a)+(2b)+(3a) Balanced			(1b)+(2b)+(3a) Moderately imbalanced			(1c)+(2b)+(3a) Severely imbalanced		
	SE	SEE	CP	SE	SEE	CP	SE	SEE	CP
$\hat{\lambda}_1(0.1)$	0.173	0.183	0.969	0.225	0.229	0.962	0.295	0.278	0.951
$\hat{\lambda}_1(0.2)$	0.129	0.134	0.969	0.166	0.169	0.960	0.196	0.202	0.961
$\hat{\lambda}_1(0.3)$	0.111	0.115	0.956	0.140	0.146	0.959	0.162	0.175	0.967
$\hat{\lambda}_1(0.4)$	0.099	0.108	0.965	0.128	0.136	0.959	0.157	0.163	0.948
$\hat{\lambda}_1(0.5)$	0.097	0.105	0.969	0.125	0.133	0.966	0.152	0.159	0.949
$\hat{\lambda}_1(0.6)$	0.096	0.105	0.966	0.128	0.134	0.955	0.153	0.160	0.953
$\hat{\lambda}_1(0.7)$	0.102	0.108	0.976	0.130	0.138	0.959	0.156	0.165	0.963
$\hat{\lambda}_1(0.8)$	0.104	0.114	0.974	0.139	0.145	0.968	0.169	0.175	0.965
$\hat{\lambda}_1(0.9)$	0.111	0.124	0.974	0.150	0.157	0.973	0.186	0.189	0.962
$\hat{\lambda}_1(1)$	0.131	0.143	0.973	0.172	0.180	0.960	0.215	0.215	0.950
$\hat{\lambda}_2(0.1)$	0.237	0.241	0.965	0.279	0.281	0.958	0.329	0.326	0.963
$\hat{\lambda}_2(0.2)$	0.174	0.182	0.970	0.206	0.212	0.966	0.230	0.243	0.978
$\hat{\lambda}_2(0.3)$	0.153	0.161	0.962	0.175	0.187	0.963	0.197	0.214	0.974
$\hat{\lambda}_2(0.4)$	0.141	0.154	0.967	0.164	0.178	0.967	0.193	0.204	0.960
$\hat{\lambda}_2(0.5)$	0.138	0.154	0.971	0.161	0.178	0.968	0.189	0.202	0.949
$\hat{\lambda}_2(0.6)$	0.143	0.158	0.969	0.169	0.182	0.961	0.194	0.207	0.954
$\hat{\lambda}_2(0.7)$	0.154	0.167	0.970	0.177	0.191	0.958	0.200	0.217	0.971
$\hat{\lambda}_2(0.8)$	0.163	0.180	0.972	0.192	0.206	0.969	0.219	0.233	0.962
$\hat{\lambda}_2(0.9)$	0.180	0.199	0.977	0.213	0.227	0.975	0.248	0.256	0.954
$\hat{\lambda}_2(1)$	0.211	0.233	0.981	0.248	0.263	0.976	0.289	0.295	0.958
$\hat{\lambda}_3(0.1)$	0.529	0.557	0.909	0.459	0.468	0.942	0.536	0.523	0.960
$\hat{\lambda}_3(0.2)$	0.256	0.254	0.931	0.252	0.250	0.954	0.269	0.274	0.959
$\hat{\lambda}_3(0.3)$	0.200	0.201	0.933	0.203	0.202	0.946	0.210	0.223	0.973
$\hat{\lambda}_3(0.4)$	0.180	0.178	0.930	0.182	0.181	0.944	0.195	0.201	0.957
$\hat{\lambda}_3(0.5)$	0.167	0.165	0.936	0.170	0.171	0.949	0.185	0.190	0.954
$\hat{\lambda}_3(0.6)$	0.160	0.159	0.946	0.167	0.167	0.944	0.186	0.186	0.943
$\hat{\lambda}_3(0.7)$	0.160	0.157	0.942	0.166	0.167	0.948	0.189	0.188	0.948
$\hat{\lambda}_3(0.8)$	0.161	0.161	0.949	0.168	0.173	0.951	0.200	0.196	0.949
$\hat{\lambda}_3(0.9)$	0.173	0.170	0.938	0.176	0.185	0.961	0.215	0.209	0.945
$\hat{\lambda}_3(1)$	0.190	0.190	0.933	0.197	0.208	0.960	0.242	0.236	0.940

Table 2.9: Standard deviation (SE), average standard error estimate (SEE), and coverage probability (CP) of $\hat{\boldsymbol{\lambda}}(\cdot)$ at different choices of t, for the two scenarios for missing failure type, after excluding outlying estimates.

	(1b)+(2c)+(3a) 25% missing			(1b)+(2c)+(3b) 50% missing		
	SE	SEE	CP	SE	SEE	CP
$\hat{\lambda}_1(0.1)$	0.251	0.266	0.957	0.275	0.275	0.952
$\hat{\lambda}_1(0.2)$	0.175	0.194	0.955	0.200	0.204	0.955
$\hat{\lambda}_1(0.3)$	0.152	0.168	0.954	0.174	0.178	0.951
$\hat{\lambda}_1(0.4)$	0.143	0.158	0.958	0.156	0.167	0.963
$\hat{\lambda}_1(0.5)$	0.135	0.153	0.956	0.151	0.164	0.966
$\hat{\lambda}_1(0.6)$	0.135	0.154	0.960	0.148	0.165	0.972
$\hat{\lambda}_1(0.7)$	0.135	0.158	0.967	0.151	0.171	0.976
$\hat{\lambda}_1(0.8)$	0.147	0.166	0.964	0.166	0.181	0.972
$\hat{\lambda}_1(0.9)$	0.158	0.179	0.968	0.182	0.197	0.971
$\hat{\lambda}_1(1)$	0.184	0.203	0.966	0.218	0.226	0.964
$\hat{\lambda}_2(0.1)$	0.313	0.337	0.961	0.360	0.362	0.959
$\hat{\lambda}_2(0.2)$	0.226	0.259	0.965	0.270	0.279	0.966
$\hat{\lambda}_2(0.3)$	0.201	0.234	0.969	0.244	0.252	0.954
$\hat{\lambda}_2(0.4)$	0.190	0.228	0.964	0.232	0.246	0.962
$\hat{\lambda}_2(0.5)$	0.184	0.229	0.974	0.233	0.249	0.969
$\hat{\lambda}_2(0.6)$	0.188	0.237	0.972	0.240	0.259	0.967
$\hat{\lambda}_2(0.7)$	0.195	0.251	0.981	0.255	0.276	0.972
$\hat{\lambda}_2(0.8)$	0.219	0.271	0.971	0.283	0.301	0.975
$\hat{\lambda}_2(0.9)$	0.243	0.298	0.969	0.316	0.337	0.975
$\hat{\lambda}_2(1)$	0.282	0.344	0.974	0.494	0.397	0.972
$\hat{\lambda}_3(0.1)$	0.623	0.733	0.916	0.720	0.859	0.920
$\hat{\lambda}_3(0.2)$	0.298	0.318	0.927	0.332	0.329	0.930
$\hat{\lambda}_3(0.3)$	0.236	0.256	0.931	0.261	0.262	0.925
$\hat{\lambda}_3(0.4)$	0.211	0.223	0.931	0.237	0.232	0.944
$\hat{\lambda}_3(0.5)$	0.190	0.205	0.949	0.218	0.217	0.941
$\hat{\lambda}_3(0.6)$	0.185	0.196	0.944	0.202	0.210	0.959
$\hat{\lambda}_3(0.7)$	0.176	0.191	0.953	0.197	0.210	0.957
$\hat{\lambda}_3(0.8)$	0.184	0.194	0.946	0.209	0.218	0.955
$\hat{\lambda}_3(0.9)$	0.192	0.204	0.953	0.225	0.234	0.949
$\hat{\lambda}_3(1)$	0.219	0.225	0.943	0.259	0.267	0.947

Table 2.10: Standard deviation (SE), average standard error estimate (SEE), and coverage probability (CP) of $\hat{\boldsymbol{\lambda}}(\cdot)$ at different choices of t, for simulation (D) and (E), after excluding outlying estimates.

	(D) Nuisance parameters N=2000			(E) UDS scenario N=2000		
	SE	SEE	CP	SE	SEE	CP
$\hat{\lambda}_1(0.1)$	0.181	0.197	0.957	0.149	0.146	0.954
$\hat{\lambda}_1(0.2)$	0.135	0.149	0.954	0.106	0.110	0.956
$\hat{\lambda}_1(0.3)$	0.113	0.130	0.960	0.090	0.097	0.970
$\hat{\lambda}_1(0.4)$	0.101	0.121	0.969	0.084	0.092	0.969
$\hat{\lambda}_1(0.5)$	0.099	0.118	0.970	0.080	0.090	0.972
$\hat{\lambda}_1(0.6)$	0.098	0.118	0.973	0.081	0.091	0.982
$\hat{\lambda}_1(0.7)$	0.102	0.121	0.974	0.084	0.094	0.972
$\hat{\lambda}_1(0.8)$	0.107	0.126	0.968	0.089	0.100	0.967
$\hat{\lambda}_1(0.9)$	0.118	0.136	0.972	0.097	0.107	0.966
$\hat{\lambda}_1(1)$	0.134	0.154	0.969	0.110	0.120	0.964
$\hat{\lambda}_2(0.1)$	0.240	0.265	0.962	0.237	0.230	0.943
$\hat{\lambda}_2(0.2)$	0.183	0.208	0.956	0.176	0.177	0.957
$\hat{\lambda}_2(0.3)$	0.156	0.189	0.977	0.154	0.160	0.958
$\hat{\lambda}_2(0.4)$	0.142	0.184	0.974	0.149	0.157	0.955
$\hat{\lambda}_2(0.5)$	0.145	0.186	0.968	0.148	0.160	0.956
$\hat{\lambda}_2(0.6)$	0.146	0.193	0.972	0.154	0.168	0.960
$\hat{\lambda}_2(0.7)$	0.156	0.203	0.971	0.167	0.180	0.957
$\hat{\lambda}_2(0.8)$	0.168	0.218	0.964	0.184	0.197	0.955
$\hat{\lambda}_2(0.9)$	0.186	0.239	0.977	0.206	0.224	0.963
$\hat{\lambda}_2(1)$	0.214	0.274	0.972	0.301	0.269	0.952
$\hat{\lambda}_3(0.1)$	0.689	1.043	0.960	1.027	1.778	0.907
$\hat{\lambda}_3(0.2)$	0.294	0.388	0.969	0.434	0.417	0.929
$\hat{\lambda}_3(0.3)$	0.227	0.298	0.965	0.299	0.303	0.938
$\hat{\lambda}_3(0.4)$	0.197	0.250	0.964	0.274	0.266	0.936
$\hat{\lambda}_3(0.5)$	0.178	0.224	0.968	0.251	0.246	0.938
$\hat{\lambda}_3(0.6)$	0.171	0.207	0.957	0.243	0.237	0.944
$\hat{\lambda}_3(0.7)$	0.171	0.198	0.950	0.235	0.235	0.950
$\hat{\lambda}_3(0.8)$	0.175	0.194	0.953	0.235	0.239	0.951
$\hat{\lambda}_3(0.9)$	0.182	0.198	0.961	0.245	0.250	0.960
$\hat{\lambda}_3(1)$	0.202	0.214	0.958	0.273	0.272	0.967

Table 2.11: Standard deviation (SE), average standard error estimate (SEE), and coverage probability (CP) of $\hat{\boldsymbol{\lambda}}(\cdot)$ at different choices of t, for the three scenarios with different sample size, after excluding outlying estimates.

	(1b)+(2c)+(3a) N=500			(1b)+(2c)+(3a) N=1000			(1b)+(2c)+(3a) N=2000		
	SE	SEE	CP	SE	SEE	CP	SE	SEE	CP
$\hat{\lambda}_1(0.1)$	0.596	0.718	0.957	0.379	0.373	0.958	0.251	0.266	0.957
$\hat{\lambda}_1(0.2)$	0.365	0.532	0.954	0.244	0.269	0.972	0.175	0.194	0.955
$\hat{\lambda}_1(0.3)$	0.300	0.453	0.955	0.203	0.231	0.964	0.152	0.168	0.954
$\hat{\lambda}_1(0.4)$	0.276	0.414	0.950	0.192	0.216	0.967	0.143	0.158	0.958
$\hat{\lambda}_1(0.5)$	0.270	0.409	0.957	0.191	0.210	0.960	0.135	0.153	0.956
$\hat{\lambda}_1(0.6)$	0.275	0.404	0.960	0.191	0.211	0.958	0.135	0.154	0.960
$\hat{\lambda}_1(0.7)$	0.283	0.393	0.968	0.192	0.216	0.970	0.135	0.158	0.967
$\hat{\lambda}_1(0.8)$	0.301	0.401	0.962	0.206	0.228	0.964	0.147	0.166	0.964
$\hat{\lambda}_1(0.9)$	0.330	0.419	0.969	0.224	0.247	0.977	0.158	0.179	0.968
$\hat{\lambda}_1(1)$	0.449	0.465	0.965	0.294	0.283	0.970	0.184	0.203	0.966
$\hat{\lambda}_2(0.1)$	0.762	0.971	0.956	0.469	0.472	0.967	0.313	0.337	0.961
$\hat{\lambda}_2(0.2)$	0.516	0.777	0.951	0.327	0.355	0.970	0.226	0.259	0.965
$\hat{\lambda}_2(0.3)$	0.453	0.731	0.941	0.283	0.317	0.961	0.201	0.234	0.969
$\hat{\lambda}_2(0.4)$	0.425	0.708	0.938	0.269	0.306	0.960	0.190	0.228	0.964
$\hat{\lambda}_2(0.5)$	0.430	0.732	0.934	0.270	0.307	0.965	0.184	0.229	0.974
$\hat{\lambda}_2(0.6)$	0.445	0.753	0.939	0.277	0.317	0.967	0.188	0.237	0.972
$\hat{\lambda}_2(0.7)$	0.467	0.791	0.945	0.284	0.334	0.968	0.195	0.251	0.981
$\hat{\lambda}_2(0.8)$	0.502	0.891	0.936	0.313	0.361	0.971	0.219	0.271	0.971
$\hat{\lambda}_2(0.9)$	0.547	0.968	0.945	0.347	0.400	0.979	0.243	0.298	0.969
$\hat{\lambda}_2(1)$	0.745	1.073	0.939	0.492	0.488	0.984	0.282	0.344	0.974
$\hat{\lambda}_3(0.1)$	1.297	7.304	0.966	0.910	1.502	0.941	0.623	0.733	0.916
$\hat{\lambda}_3(0.2)$	0.708	1.384	0.974	0.435	0.449	0.954	0.298	0.318	0.927
$\hat{\lambda}_3(0.3)$	0.483	0.839	0.968	0.319	0.341	0.964	0.236	0.256	0.931
$\hat{\lambda}_3(0.4)$	0.409	0.744	0.965	0.280	0.300	0.955	0.211	0.223	0.931
$\hat{\lambda}_3(0.5)$	0.385	0.660	0.966	0.268	0.279	0.946	0.190	0.205	0.949
$\hat{\lambda}_3(0.6)$	0.368	0.597	0.962	0.258	0.268	0.947	0.185	0.196	0.944
$\hat{\lambda}_3(0.7)$	0.371	0.562	0.967	0.255	0.266	0.956	0.176	0.191	0.953
$\hat{\lambda}_3(0.8)$	0.381	0.517	0.958	0.266	0.272	0.953	0.184	0.194	0.946
$\hat{\lambda}_3(0.9)$	0.395	0.522	0.963	0.282	0.289	0.950	0.192	0.204	0.953
$\hat{\lambda}_3(1)$	0.509	0.564	0.946	0.338	0.326	0.941	0.219	0.225	0.943

2.7.8 Simulation under severe overlapping plus severely imbalanced class proportion

To evaluate the performance of the proposed method under extremely challenging conditions, we conducted simulation under scenario (1c)+(2d)+(3a) as defined in Table 2.1, with sample size $N = 500, 1000, 2000$ and 4000. We also excluded the outlying estimates using the same procedure as described in the main article.

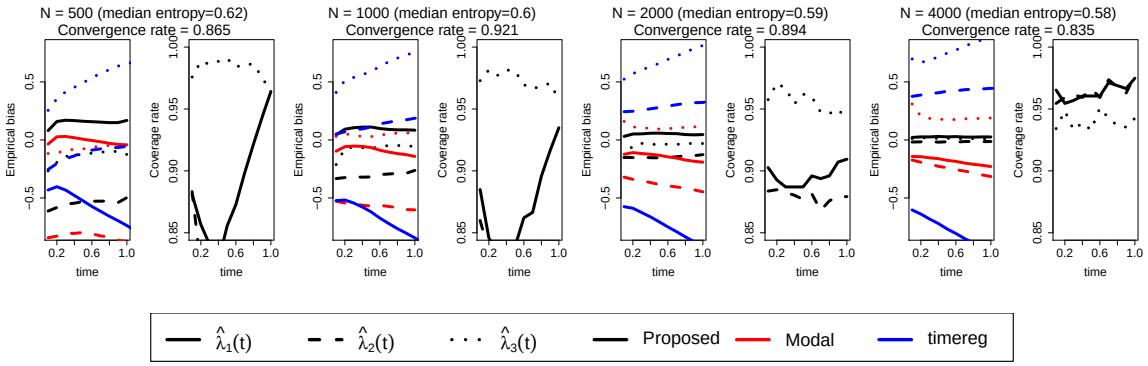

Figure 2.9: Simulation results for the scenario described by Section 2.7.8. Quantities associated with the three regression coefficients in $\boldsymbol{\lambda}_0(t)$ are represented by solid, dashed and dotted lines. The proposed, modal assignment, and `timereg` strategies are respectively shown in black, red and blue lines.

As shown in Figure 2.9, the estimation was unstable when the sample size is small ($N = 500$ and 1000), which was expected. Specifically, the estimation for the coefficient associated with the third class (dotted line) was the most stable with smaller sample size, because the third class has the largest population. In contrast, the estimation for the coefficient associated with class one (solid line) and class two (dashed line) were relatively more biased under smaller sample size. The bias associated with the parameter for class two was larger because the class two has overlapping with both classes one and three, while classes one and three do not overlap with each other, which means heavier misclassification for subjects in class two.

As sample size increased to 2000 and 4000, the proposed method achieved much smaller biases, together with more standard coverage probabilities. This shows that our method can handle data with highly overlapped and highly imbalanced latent class pattern when

the sample size is sufficiently large.

2.7.9 Discussions about the independent censoring assumption

As discussed in the section 2.3, our model framework is flexible to be generalized to incorporate the conditional independence assumption for censoring and the corresponding model-based inverse probability censoring weighting (IPCW) terms. In terms of point estimation, one can directly use the model-based $\hat{G}(X_i|\boldsymbol{Y}_i) = \Pr(C_i \geq X_i|\boldsymbol{Y}_i)$ in place of $\hat{G}(X_i)$ in the estimating equation (8). For variance estimation, we also provided a flexible template for inference with model-based IPCW. That is, instead of considering the variability of Kaplan-Meier estimator, as denoted by $\hat{\boldsymbol{\Xi}}_{1i}$ in the manuscript, one can replace $\hat{\boldsymbol{\Xi}}_{1i}$ by its counterpart from Cox regression to conduct inference with model-based IPCW.

In our real data application, we assumed unconditionally independence of censoring in the UDS data. This assumption creates easier implementation of the inverse probability censoring weighting and the inference procedure for the structural competing risks model, by directly incorporating results from a Kaplan-Meier estimator. As pointed out by a referee, the above assumption may not satisfy in the UDS data and in more general cases. Instead, a conditional independence assumption of censoring given covariates may be more appropriate.

We investigated the association between censoring time and four covariates (baseline overall cognition MMSE, age at baseline, gender, and race) by fitting a Cox proportional hazard model. Although the covariate effects were statistically significant (Table 2.12) for all covariates, the effect sizes were fairly small. We further compared the estimation results of the structural competing risks model with IPCW based on (i) independent censoring assumption (Kaplan-Meier estimator of censoring) and (ii) conditionally independent censoring assumption (Cox model of censoring). As Figure 2.10 shows, the point estimates based on the two assumptions were close to each other for the UDS data, indicating that the assumption about censoring mechanism does not dramatically affect the point estimation.

Table 2.12: Hazard ratios and p-values for the covariates of the Cox proportional hazard model for the censoring time.

Covariates	Hazard ratio	P value
Baseline MMSE	0.97	< 0.001
Baseline age	0.88	< 0.001
gender (female vs male)	1.11	< 0.001
race (white vs others)	0.83	< 0.001

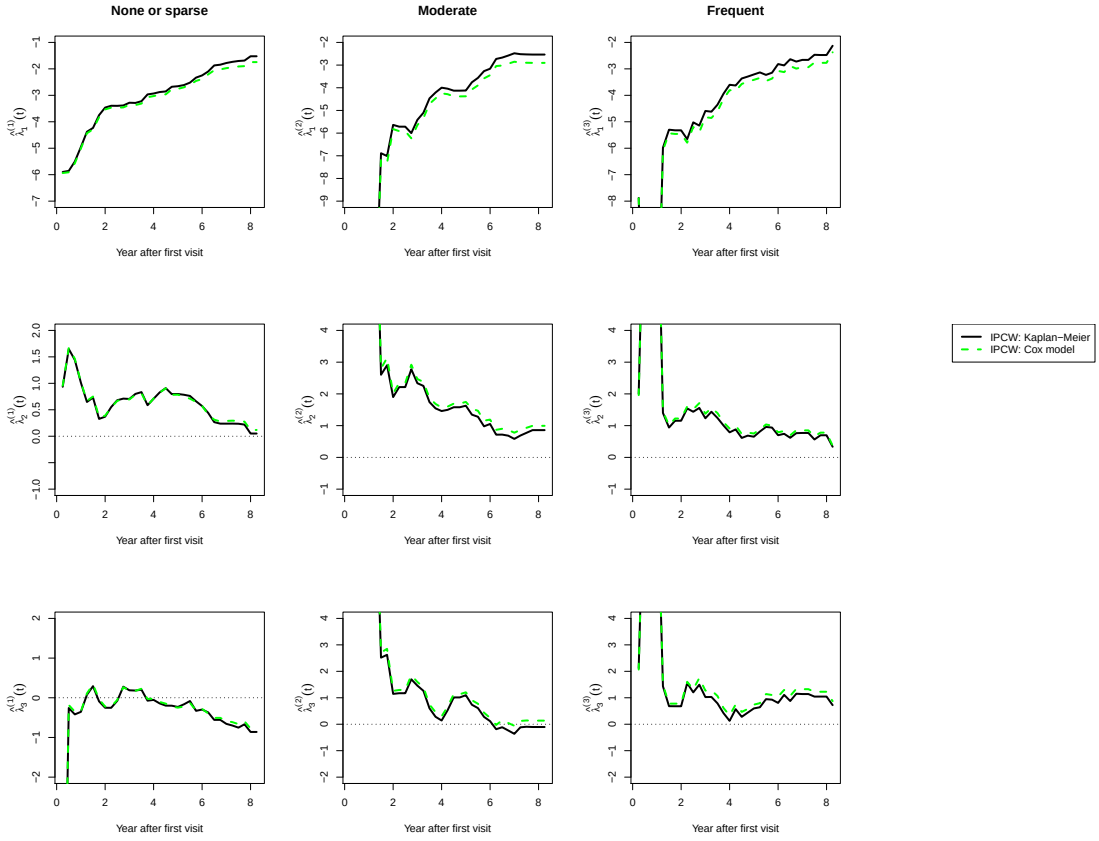

Figure 2.10: Point estimates of the structural competing risks model with inverse probability censoring weighting based on Kaplan-Meier estimator (solid black) and Cox regression (dashed green).

Chapter 3

Latent Class Analysis for Time-to-event Data Based on Semi-parametric Proportional Hazards Submodel

3.1 Introduction

The nature of heterogeneity has been recognized in a number of diseases, such as mild cognitive impairment (MCI) and prostate cancer. Usually, the clinicians classify the patients into certain disease *subtypes*, such as the amnestic and the non-amnestic subtypes of MCI (Winblad et al., 2004), where each subtype represents a particular etiology with potentially unique pattern of disease progression. It is of critical clinical interest to understand the heterogeneity of disease population and its implications for the onset of clinical events, which will contribute to better prediction of time-to-event, such as predicting the time to dementia for MCI patients based on baseline patient characteristics.

Common practice to address this interest is regression analysis, such as fitting a Cox propor-

tion hazard model (Cox, 1972). Typically, traditional survival models attempt to explain the survival distribution for the whole population by a single model with unified covariate effects, such that different values or levels of covariates indicate earlier or later onset of disease. For a heterogeneous population, however, traditional survival models are oversimplified. This is because the heterogeneity of disease population indicates different underlying etiologies, which means the disease progression and the importance of associated risk factors can vary among disease subtypes. In statistical modeling, this implies varied baseline hazard functions and covariate effects for different disease subtypes.

Latent class analysis (LCA) is a useful tool to address the above challenges in analyzing survival data of heterogeneous populations. Extended from finite mixture models (McLachlan and Peel, 2000), the LCA framework is able to incorporate class-specific survival submodels to capture heterogeneous patterns in disease progression, and a class membership probability submodel which addresses the uncertainty of belonging to certain latent subtypes given patient characteristics. In addition, the latent classes (or subtypes) defined by LCA are jointly determined by the membership probability submodel and the class-specific survival submodels, which means the obtained latent classes are data-driven with high relevance to the survival outcome of interest.

Various mixture models have been proposed in the past few decades for the clustering analysis of survival data. The two-component mixture cure models (Kuk and Chen, 1992; McLachlan and McGiffin, 1994; Lambert et al., 2010, for example) were studied to investigate differences in survival between cured and uncured population. However, mixture cure models assume only two classes in the population, which is not applicable if there are more than two classes. In addition, mixture Weibull models (Bučar et al., 2004; Mair and Hudec, 2009, for example) and mixture exponential models (Hilton et al., 2018, for example) were proposed to investigate heterogeneous lifetime distribution for two or more underlying classes. Nevertheless, these methods are less flexible due to the imposed parametric assumption of survival distribution.

In recent development of mixture modeling for biomedical data, large efforts were made in the development of joint latent class models of survival data and other phenotypes, such

as longitudinal data (Lin et al., 2002; Proust-Lima et al., 2009, 2017, for example) and responses to a questionnaire (Larsen, 2004). Typically, these models assume conditional independence between phenotypes and survival data given the class membership, which ignores within-class correlation between survival time and other phenotypes. Under this assumption, these methods may obtain redundant latent classes which are largely attributed to the heterogeneity of phenotypes, instead of the time-to-event of interest. Hypothetically, some phenotypes can have high heterogeneity in disease population but little correlation with time-to-event. Under such hypothetical setting, the joint models will still recognize the latent classes with respect to the phenotypes, which can hardly contribute to our research question of interest in understanding the heterogeneity in survival. In terms of survival submodels, the majority of existing joint latent class models use class-specific Cox proportional hazard model. Proust-Lima et al. (2017) utilized Weibull distribution, piecewise constant with limited number of jumps, and cubic M-splines to formulate baseline hazard functions, which creates challenges in model specification. In contrast, Lin et al. (2002) and Larsen (2004) incorporated unspecified baseline hazard functions which imposes weaker assumptions. However, both methods assumed common covariate effects on survival for different latent classes, which is less flexible in capturing the heterogeneity of covariate effects. In addition, little attention was paid to deriving asymptotic theories of the estimators for the semi-parametric latent class models.

Motivated by the limitations of the existing methods, we provide a semi-parametric framework for the latent class analysis of survival data, to investigate the heterogeneity of disease population and its implications for disease progression. We impose weaker assumptions of unspecified baseline cumulative hazard function, which improves flexibility compared to the existing methods with parametric assumptions. We also enables class-specific covariate effects in both latent class probability submodel and survival submodel, such that the heterogeneity can be better detected and interpreted. In addition, our framework focuses on the heterogeneity in time-to-event distribution, such that the resulting latent class patterns are not interfered by other phenotypes with limited contribution to understanding survival heterogeneity.

Technically, we utilize non-parametric maximum likelihood estimator (Zeng and Lin, 2007, NPMLE) approach to incorporate the the infinite-dimensional component of baseline cumulative hazard function. Due to the finite mixture structure of latent class problems, the finite-dimensional and infinite-dimensional components are entangled in the likelihood function, which creates further challenges of establishing asymptotic properties and conducting variance estimation. We address this difficulty using an approach similar to that used by Mao and Lin (2017). To handle unobservable latent class labels, we derive a stable expectation-maximization (EM) algorithm which can be easily assembled by existing software or algorithm. According to our numerical experience, the algorithm is robust to initialization and achieves impressive results with non-informative initial values. Moreover, asymptotic theories are rigorously established by empirical process arguments (van der Vaart and Wellner, 1996) and semi-parametric efficiency results (Bickel et al., 1993). We also provide alternative strategies for inference, based on either information matrix, or profile likelihood (Murphy and van der Vaart, 2000). Furthermore, we give recommendations in model selection criteria in selecting the most appropriate number of latent classes.

3.2 Data, notation and models

3.2.1 Data and notations

Let T and C respectively denote time to event of interest and time to independent censoring of T. Let $\boldsymbol{x}$ denote a $p \times 1$ vector of baseline covariates. Define $\tilde{T} = T \wedge C$ and $\Delta = I(T \leq C)$. The observed data consist of n independent and identically distributed replicates of $\boldsymbol{O} = (\tilde{T}, \Delta, \boldsymbol{x})$, denoted by $\{\boldsymbol{O}_i = (\tilde{T}_i, \Delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. The latent classes are denoted by an unobservable $L \times 1$ vector of binary indicators, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, where L is the number of latent classes, and $\xi_l = 1$ if belonging to the lth class and 0 otherwise.

3.2.2 The assumed models

In this work we assume that the marginal density of time-to-event observation $(\tilde{T}, \Delta)$ can be captured by a finite mixture model (McLachlan and Peel, 2000) with L components

$$f(\tilde{T}, \Delta) = \sum_{l=1}^{L} p_l f_l(\tilde{T}, \Delta),$$

where p_l is the probability of belonging to class l and $f_l(\tilde{T}, \Delta)$ is the class-specific density of $(\tilde{T}, \Delta)$ for class l. Our modeling strategy involves a class membership probability submodel for p_l and a class-specific survival submodel for $f_l(\tilde{T}, \Delta)$.

For the class membership probability submodel, we utilize a standard latent polytomous logistic regression model (Bandeen-Roche et al., 1997) to account for the effect of baseline covariates on the relative frequency of latent classes:

$$\Pr(\xi_l = 1|\boldsymbol{x}) = p_l(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\exp(\tilde{\boldsymbol{x}}^T \boldsymbol{\alpha}_l)}{\sum_{d=1}^{L} \exp(\tilde{\boldsymbol{x}}^T \boldsymbol{\alpha}_d)}, l = 1, \ldots, L, \tag{3.2.1}$$

where $\tilde{\boldsymbol{x}} = (1, \boldsymbol{x}^T)^T$, $\boldsymbol{\alpha}_1 = \mathbf{0}$ for identifiability consideration, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L)^T$ is the vector of unknown parameters with length $(p+1) \times (L-1)$.

For the class-specific survival submodel, we propose a semi-parametric class-specific proportional hazards model. Without loss of generality, we let the first class be the reference class with hazard function $\lambda(t|\xi_1 = 1) = \lambda_0(t) \exp(\bar{\boldsymbol{x}}^T \boldsymbol{\zeta}_1)$, where $\lambda_0(t)$ is the unspecified baseline hazard function for the reference class, $\bar{\boldsymbol{x}}$ is a $q \times 1$ subvector of $\boldsymbol{x}$ with $q \leq p$, and $\boldsymbol{\zeta}_1$ is the $q \times 1$ unknown covariate effect in the reference class model. For other classes $l = 2, \ldots, L$ that is not the first class, we assume $\lambda(t|\xi_l = 1) = \lambda_0(t) \exp\{a_l + \bar{\boldsymbol{x}}^T(\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_l)\}$, where $\exp(a_l)$ is a constant ratio between the baseline hazard functions of class l and class 1, and $\boldsymbol{\zeta}_l$ is the $q \times 1$ difference of covariate effects between class l and class 1. Let $\boldsymbol{z}_l = (\bar{\boldsymbol{x}}^T, \mathbf{0}^T_{(q+1)\times(L-1)})^T \cdot I(l = 1) + (\bar{\boldsymbol{x}}^T, (\boldsymbol{e}_{l-1} \otimes \tilde{\boldsymbol{x}})^T)^T \cdot I(l > 1)$ and $\boldsymbol{\gamma} = (\boldsymbol{\zeta}_1^T, a_2, \boldsymbol{\zeta}_2^T, a_3, \boldsymbol{\zeta}_3^T, \ldots, a_L, \boldsymbol{\zeta}_L^T)^T$, where $\mathbf{0}_d$ represents a $d-$vector of zeros, $\boldsymbol{e}_{l-1}$ represents a $(L-1)-$vector whose $(l-1)$th element is 1 and other elements are zero, $\tilde{\boldsymbol{x}} = (1, \bar{\boldsymbol{x}}^T)^T$, and $\otimes$ denotes Kronecker product operator. Then it follows a universal expression of the class-specific hazard functions for class 1 to

class L

$$\lambda(t|\xi_l = 1) = \lambda_0(t)\exp(\boldsymbol{z}_l^T\boldsymbol{\gamma}),\ l = 1, \ldots, L, \tag{3.2.2}$$

where $\boldsymbol{\gamma}$ is the vector of unknown parameters with length $q \times L + (L-1)$. The according class-specific density of $(\tilde{T}, \Delta)$ satisfies

$$f_l(\tilde{T}, \Delta|\boldsymbol{x}; \boldsymbol{\theta}) = \{\lambda_0(\tilde{T})\exp(\boldsymbol{z}_l^T\boldsymbol{\gamma})\}^{\Delta}\exp\{-\Lambda_0(\tilde{T})\exp(\boldsymbol{z}_l^T\boldsymbol{\gamma})\},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ and $\boldsymbol{\theta} = \{\boldsymbol{\gamma}^T, \Lambda(\cdot)\}^T$. Then the finite mixture framework implies that the conditional density of $(\tilde{T}, \Delta)$ given $\boldsymbol{x}$ satisfies

$$f(\tilde{T}, \Delta|\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{l=1}^{L} p_l(\boldsymbol{x}, \boldsymbol{\alpha}) f_l(\tilde{T}, \Delta|\boldsymbol{x}; \boldsymbol{\theta}). \tag{3.2.3}$$

3.3 Estimation and inference

Based on the assumed models in Section 3.2, in this section we derive the likelihood function and the associated estimation and inference procedures. Due to the complications caused by the missingness of latent class memberships $\boldsymbol{\xi}$, and the non-parametric assumption of the baseline cumulative hazard function $\Lambda_0(\cdot)$, it is not straightforward to maximize the likelihood function for the observed data. As a natural solution, we utilize non-parametric maximum likelihood estimators (NPMLE) technique to account for the unobservable $\boldsymbol{\xi}$ by an Expectation-Maximization (EM) algorithm, and to facilitate inference for the non-parametric estimator $\hat{\Lambda}(\cdot)$ of $\Lambda_0(\cdot)$.

3.3.1 Observed data likelihood

Under the assumed finite mixture model (3.2.3), and submodels (3.2.1) and (3.2.2) we obtain the observed data likelihood

$$L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{O}) = \prod_{i=1}^{n}\left\{\sum_{l=1}^{L} p_l(\boldsymbol{x}_i; \boldsymbol{\alpha})\{\lambda(\tilde{T}_i)\exp(\boldsymbol{z}_{il}^T\boldsymbol{\gamma})\}^{\Delta_i}\exp\{-\Lambda(\tilde{T}_i)\exp(\boldsymbol{z}_{il}^T\boldsymbol{\gamma})\}\right\} f_{\boldsymbol{X}}(\boldsymbol{x}_i), \tag{3.3.1}$$

where $f_{\boldsymbol{X}}(\cdot)$ is the density function of $\boldsymbol{x}$. Note that $f_{\boldsymbol{X}}(\boldsymbol{x}_i)$ is a constant with respect to unknown parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ and Λ, thus is omitted in further derivations.

3.3.2 EM algorithm for point estimation

Assuming $\boldsymbol{\xi}$ are observed, the complete data likelihood corresponding to the observed data likelihood (3.3.1) satisfies

$$L_c(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{\xi}, \boldsymbol{O}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \left\{ p_l(\boldsymbol{x}_i; \boldsymbol{\alpha}) \{\lambda(\tilde{T}_i) \exp(\boldsymbol{z}_{il}^T \boldsymbol{\gamma})\}^{\Delta_i} \exp\{-\Lambda(\tilde{T}_i) \exp(\boldsymbol{z}_{il}^T \boldsymbol{\gamma})\} \right\}^{I(\xi_{il}=1)}.$$

We further treat $\Lambda(\cdot)$ as piecewise constant between observed event times. That is, $\Lambda(t) = \sum_{j:t_j \le t} \Lambda\{t_j\}$ with $\Lambda\{t_j\} = d_j$, where $t_1 < t_2 < \ldots < t_m$ are distinct uncensored event times. Denote the cumulative hazard function $\Lambda(t_j)$, at $t_j, j = 1, \ldots, m$, as Λ_j. Then the corresponding log complete data likelihood satisfies

$$\begin{aligned} \ell_c(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{\xi}, \boldsymbol{O}) = & \sum_{j=1}^{m} \sum_{l=1}^{L} \xi_{(j)l} \left\{ \log \Lambda\{t_j\} + \boldsymbol{z}_{(j)l}^T \boldsymbol{\gamma} - e^{\boldsymbol{z}_{(j)l}^T \boldsymbol{\gamma}} \Lambda_j \right\} \\ & - \sum_{j=1}^{m} \sum_{k: t_j \le \tilde{T}_k < t_{j+1}} I(\Delta_k = 0) \sum_{l=1}^{L} \xi_{kl} e^{\boldsymbol{z}_{kl}^T \boldsymbol{\gamma}} \Lambda_j \\ & + \sum_{i=1}^{n} \sum_{l=1}^{L} \xi_{il} \log p_l(\boldsymbol{x}_i; \boldsymbol{\alpha}), \end{aligned} \tag{3.3.2}$$

where $\xi_{(j)l}$ and $\boldsymbol{z}_{(j)l}$ represents the membership indicator ξ_l and covariate vector $\boldsymbol{z}_l$ for the observation with uncensored failure time $t_j, j = 1, \ldots, m$.

In the E-step, we calculate the expectation, $E\{\ell_c(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{\xi}, \boldsymbol{O}) | \boldsymbol{O}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\gamma}^{(j)}, \Lambda^{(j)}\}$, of the log complete data likelihood (3.3.2), conditioned on observable data $\boldsymbol{O}$ and the current estimates of unknown parameters $\boldsymbol{\alpha}^{(j)}, \boldsymbol{\gamma}^{(j)}, \Lambda^{(j)}$ at the arbitrary jth iteration. Because of the simplicity of (3.3.2) with respect to $\boldsymbol{\xi}$, it is straightforward to see

$$E\{\ell_c(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{\xi}, \boldsymbol{O}) | \boldsymbol{O}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\gamma}^{(j)}, \Lambda^{(j)}\} = \ell_c\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \hat{E}(\boldsymbol{\xi}), \boldsymbol{O}\},$$

where $\hat{E}(\xi_{il}) \equiv E(\xi_{il} | \boldsymbol{O}_i; \boldsymbol{\alpha}^{(j)}, \boldsymbol{\gamma}^{(j)}, \Lambda^{(j)})$. Note $E(\xi_{il} | \boldsymbol{O}_i; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) = \Pr(\xi_{il} = 1 | \boldsymbol{O}_i; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)$

is the posterior membership probability which can be derived by Bayes' Rule $\Pr(\xi_{il} = 1|\boldsymbol{O}_i;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) = \Pr(\xi_{il}=1,\tilde{T}_i,\Delta_i|\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)/\Pr(\tilde{T}_i,\Delta_i|\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)$. That is,

$$\hat{E}(\xi_{il}) = \Pr(\xi_{il}=1|\boldsymbol{O}_i;\boldsymbol{\alpha}^{(h)},\boldsymbol{\gamma}^{(h)},\Lambda^{(h)}) = \frac{p_l(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(h)})f_l(\tilde{T}_i,\Delta_i|\boldsymbol{x}_i;\boldsymbol{\gamma}^{(h)},\Lambda^{(h)})}{\sum_{d=1}^{L} p_d(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(h)})f_d(\tilde{T}_i,\Delta_i|\boldsymbol{x}_i;\boldsymbol{\gamma}^{(h)},\Lambda^{(h)})}. \tag{3.3.3}$$

The resulting conditional expectation $\ell_c\{\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda;\hat{E}(\boldsymbol{\xi}),\boldsymbol{O}\}$, denoted by $Q(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)$, serves as the target function to be maximized in the subsequent M-step.

In the M-step, we adopt a profile likelihood strategy to maximize $Q(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)$ by profiling out Λ, where Λ is treated as an m-dimensional unknown parameter $\Lambda\{t_k\} = d_k, k = 1,\ldots,m$. First, with fixed $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, we find $\hat{\Lambda}(t;\boldsymbol{\gamma}) = \arg\max_{\Lambda} Q(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)$ by solving

$$\frac{\partial}{\partial d_k}Q(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) = \frac{1}{d_k} - \sum_{i:\tilde{T}_i\geq t_k}\sum_{l=1}^{L}\hat{E}(\xi_{il})e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}} = 0, \quad k = 1,\ldots,m.$$

That is, $\hat{d}_k(\boldsymbol{\gamma}) = \{\sum_{i:\tilde{T}_i\geq t_k}\sum_{l=1}^{L}\hat{E}(\xi_{il})e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}}\}^{-1}, k = 1,\ldots,m$ and

$$\hat{\Lambda}(t;\boldsymbol{\gamma}) = \sum_{k:t_k\leq t}\hat{d}_k(\boldsymbol{\gamma}) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(s)}{\sum_{i=1}^{n}\sum_{l=1}^{L}\hat{E}(\xi_{il})Y_i(s)e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}}}, \tag{3.3.4}$$

where $N(t) = I(\tilde{T}\leq t, \Delta = 1)$ and $Y(t) = I(\tilde{T}\geq t)$. Then by plugging in $\hat{\Lambda}(t;\boldsymbol{\gamma})$, we obtain the profile complete data log likelihood $Q_p(\boldsymbol{\alpha},\boldsymbol{\gamma}) \equiv Q\{\boldsymbol{\alpha},\boldsymbol{\gamma},\hat{\Lambda}(t;\boldsymbol{\gamma})\}$:

$$\begin{aligned} Q_p(\boldsymbol{\alpha},\boldsymbol{\gamma}) = &\sum_{i=1}^{n}\sum_{l=1}^{L}\int_0^{t^*}\hat{E}(\xi_{il})\Big\{\log\frac{1}{\sum_{i=1}^{n}\sum_{l=1}^{L}\hat{E}(\xi_{il})Y_i(s)e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}}} + \boldsymbol{z}_{il}^T\boldsymbol{\gamma}\Big\}dN_i(s) \\ &+\sum_{i=1}^{n}\sum_{l=1}^{L}\hat{E}(\xi_{il})\log p_l(\boldsymbol{x}_i;\boldsymbol{\alpha}), \end{aligned} \tag{3.3.5}$$

where t^* is a finite constant satisfying $t^* > t_m$. Then it is straightforward to find $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} Q_p(\boldsymbol{\alpha},\boldsymbol{\gamma})$ and $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha},\boldsymbol{\gamma})$ by solving

$$\frac{\partial}{\partial\boldsymbol{\alpha}}Q_p(\boldsymbol{\alpha},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{l=1}^{L}\hat{E}(\xi_{il})\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x}_i;\boldsymbol{\alpha}) = \mathbf{0}$$

and

$$\frac{\partial}{\partial \boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{l=1}^{L} \int_0^{t^*} \hat{E}(\xi_{il}) \left(\boldsymbol{z}_{il} - \frac{\sum_{j=1}^{n} \sum_{k=1}^{L} \hat{E}(\xi_{jk}) Y_j(u) \boldsymbol{z}_{jk} \exp(\boldsymbol{z}_{jk}^T \boldsymbol{\gamma})}{\sum_{j=1}^{n} \sum_{k=1}^{L} \hat{E}(\xi_{jk}) Y_j(u) \exp(\boldsymbol{z}_{jk}^T \boldsymbol{\gamma})} \right) dN_i(u) = \mathbf{0}$$

It is straightforward to show that solving $\frac{\partial}{\partial \boldsymbol{\alpha}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$ reduces to fitting a weighted multinomial logistic regression with weights $\hat{E}(\boldsymbol{\xi})$, which can be easily implemented by R package `VGAM` (Yee et al., 2010). In addition, equation $\frac{\partial}{\partial \boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$ is equivalent to a weighted partial score equation for the proportional hazard model with weights $\hat{E}(\boldsymbol{\xi})$. We choose not to use existing Cox regression software to solve $\frac{\partial}{\partial \boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$, which would automatically account for the pseudo ties caused by repeatedly counting each observed event (indexed by i) for multiple latent classes (indexed by l), making the resulting estimates not accurately based on the estimating equation $\frac{\partial}{\partial \boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$. Instead, we implement an efficient Newton-Raphson algorithm under `Rcpp` environment (Eddelbuettel et al., 2011) to ensure that the estimator is a rigorous solution of $\frac{\partial}{\partial \boldsymbol{\gamma}} Q_p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$.

We initialize the EM algorithm with an initial guess of $\hat{E}(\boldsymbol{\xi})$, which can be obtained from random guess or informative ways such as K-means clustering of $\tilde{T}$. Then we repeat the M-step and E-step until the stopping criterion is satisfied. We propose to use an Aitken acceleration-based stopping criterion as described in McLachlan and Peel (2000, page 52). Denote $l^{(k)}$ as the logarithm of the observed-data likelihood (3.3.1) evaluated using the parameter estimation at the kth iteration. Define $a^{(k)} = (l^{(k+1)} - l^{(k)})/(l^{(k)} - l^{(k-1)})$ and $l_A^{(k+1)} = l^{(k)} + (l^{(k+1)} - l^{(k)})/(1 - a^{(k)})$. The algorithm is stopped when $|l_A^{(k+1)} - l_A^{(k)}| < tol$, where tol is the tolerance parameter. In practice, we let $tol = 10^{-7}$ to ensure convergence to a local optimum.

3.3.3 Asymptotic properties and variance estimation

In this section, We establish the consistency and asymptotic normality using NPMLE arguments similar to those used in Zeng and Lin (2006) and Mao and Lin (2017). First we give the following regularity conditions:

(C1) There exists $t^* > 0$ such that $\Pr(C = t^*) > 0$ and $\Pr(C > t^*) = 0$;

(C2) For $l = 1, \ldots, L$, $\Pr(\xi_l = 1|\boldsymbol{x}; \boldsymbol{\alpha}) \in (0, 1)$.

(C3) $||\boldsymbol{\alpha}_0|| < \infty$; $||\boldsymbol{\gamma}_0|| < \infty$; $||z_l|| < \infty$ for $l = 1, \ldots, L$; Λ_0 is continuously differentiable with $\Lambda_0'(t) > 0$ on $[0, t^*]$, where $||\cdot||$ denotes the Euclidean norm.

Conditions (C1)-(C3) are reasonable in practical applications. Condition (C1) is commonly satisfied by administrative censoring, which also helps prove the uniform consistency of $\Lambda(\cdot)$ on $[0, t^*]$. Condition (C2) ensures that the latent class membership probabilities $p_l(\boldsymbol{x}; \boldsymbol{\alpha})$ is greater than zero, which further guarantees that $\log p_l(\boldsymbol{x}; \boldsymbol{\alpha})$ has a finite lower bound. Condition (C3) assumes the smoothness of $\Lambda(\cdot)$ and the boundedness of $\boldsymbol{\alpha}_0$, $\boldsymbol{\gamma}_0$ and baseline covariates $\boldsymbol{x}$. Proofs of the following two theorems are provided in Appendix sections 3.7.1 and 3.7.2.

Theorem 3.3.1. *Under regularity conditions (C1)-(C3),* $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ *and* $\hat{\Lambda}$ *are strongly consistent. That is,* $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0|| + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| + \sup_{t\in[0,t^*]} |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0$ *almost surely.*

Theorem 3.3.2. *Under regularity conditions (C1)-(C3),* $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ *and* $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ *converges to multivariate zero-mean Gaussian distributions;* $\sqrt{n}\{\hat{\Lambda}(t) - \Lambda_0(t)\}$ *converges to a univariate zero-mean Gaussian process on* $t \in [0, t^*]$. *In addition,* $\hat{\boldsymbol{\alpha}}$ *and* $\hat{\boldsymbol{\gamma}}$ *are semiparametric efficient as defined in Bickel et al. (1993).*

Variance estimation can be conducted based on the information matrix of the observed-data profile log-likelihood (Murphy and van der Vaart, 2000), defined by

$$\mathrm{pl}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \equiv \ell\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \hat{\Lambda}(\boldsymbol{\alpha}, \boldsymbol{\gamma}); \boldsymbol{O}\},$$

where $\hat{\Lambda}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{argmax}_\Lambda \ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{O})$. Given the point estimates $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\gamma}}^T)^T$, we obtain $\hat{\Lambda}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ by running the aforementioned EM algorithm with $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ fixed, and only updating $\hat{\Lambda}(\cdot)$ by formula (3.3.4) and $\hat{E}$ by formula (3.3.3) until convergence. Then it follows an estimation of the profile log-likelihood $\hat{\mathrm{pl}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \ell\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}); \boldsymbol{O}\}$. Let $\hat{\mathrm{pl}}_j(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ be the subject j's contribution to $\hat{\mathrm{pl}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. The covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\gamma}}^T)^T \in \mathbb{R}^r$, where

$r = (p+1) \times (L-1) + q \times L + L - 1$, can be estimated by the inverse of

$$\sum_{j=1}^{n} \begin{pmatrix} \frac{\hat{\text{pl}}_j(\hat{\boldsymbol{\theta}}+h_n\boldsymbol{\epsilon}_1)-\hat{\text{pl}}_j(\hat{\boldsymbol{\theta}}-h_n\boldsymbol{\epsilon}_1)}{2h_n} \\ \vdots \\ \frac{\hat{\text{pl}}_j(\hat{\boldsymbol{\theta}}+h_n\boldsymbol{\epsilon}_r)-\hat{\text{pl}}_j(\hat{\boldsymbol{\theta}}-h_n\boldsymbol{\epsilon}_r)}{2h_n} \end{pmatrix}^{\otimes 2},$$

where $\boldsymbol{\epsilon}_k$ is the kth canonical vector in $\mathbb{R}^r$, $\boldsymbol{d}^{\otimes 2} = \boldsymbol{d}\boldsymbol{d}^T$, and h_n is a constant of order $n^{-1/2}$. In the numerical studies, we used $h_n = 5n^{-1/2}$ as used by Gao and Chan (2019). Unlike the numerical approximation of Hessian matrix as used in Murphy and van der Vaart (2000), we utilize the outer product of the first order numerical differences, which is computationally more affordable and guarantees that the resulting covariance matrix estimator is positive definite. Alternatively, an analytical consistent variance estimator can be constructed based on similar arguments as in Zeng and Lin (2006), which allows inference for $\hat{\Lambda}(t)$ in addition to $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$. Details about the analytical variance estimator are provided in Appendix Section 3.7.3. Compared to the numerical variance estimator based on profile likelihood, the analytical variance estimator typically requires inverse matrix computation for a covariance matrix with much higher dimension due to the inclusion of cumulative hazard function, which might cause less stable numerical performance. Thus, we report inference for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ based on the profile likelihood approach in our simulation and real application analysis. In contrast, inference for $\hat{\Lambda}(t)$ is based on the analytical approach.

3.3.4 Selecting the number of latent classes

In practice, it is usually of interest to determine the number of latent classes, L, using data-driven criteria. Standard model selection criteria for likelihood-based latent class methods include the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). It is also common to use entropy-based criteria, such as integrated complete-data likelihood (Biernacki et al., 2000, ICL-BIC) and classification entropy extended BIC (Hart et al., 2020, CE-BIC). A standardized entropy index (Muthén et al., 2002), defined as

$$1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{L} \hat{E}(\xi_{il}|\boldsymbol{O}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda})\{-\log \hat{E}(\xi_{il}|\boldsymbol{O}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda})\}}{n \log L},$$

is another commonly used metric to assess the level of uncertainty of latent classes in a fitted model. When the latent classes are well separated, the estimated posterior class membership probability $\hat{E}(\xi_{il}|\boldsymbol{O}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda})$ is close to either one or zero, such that the corresponding standardized entropy index is close to one. According to our simulation analysis detailed in Section 3.4.2, BIC is the most effective criterion to determine L for the proposed method.

3.3.5 Assessing the prediction performance

Define $S(t|\boldsymbol{x},\xi_l=1)=\Pr(T\geq t|\boldsymbol{x},\xi_l=1), l=1,\ldots,L$ as the class-specific survival function, and $G(u)=\Pr(C\geq u)$ as the survival function of the censoring at time u. We evaluate the prediction performance of the proposed latent class model by the Brier Score, defined as $E[\{I(T\geq t)-\hat{S}(t|\boldsymbol{x})\}^2]$, where

$$\hat{S}(t|\boldsymbol{x})=\sum_{l=1}^{L}\hat{\Pr}(\xi_l=1|\boldsymbol{x})\hat{S}(t|\boldsymbol{x},\xi_l=1)=\sum_{l=1}^{L}p_l(\boldsymbol{x};\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}(t)\exp(\boldsymbol{z}_l^T\hat{\boldsymbol{\gamma}})\} \tag{3.3.6}$$

is the predicted survival probability at time t given baseline covariates $\boldsymbol{x}$. Here the predicted survival probability $\hat{S}(t|\boldsymbol{x})$ can be interpreted as a weighted summation of predicted class-specific survival probabilities $\hat{S}(t|\boldsymbol{x},\xi_l=1)=\exp\{-\hat{\Lambda}(t)\exp(\boldsymbol{z}_l^T\hat{\boldsymbol{\gamma}})\}$, with estimated class membership probabilities $\hat{\Pr}(\xi_l=1|\boldsymbol{x})=p_l(\boldsymbol{x};\hat{\boldsymbol{\alpha}})$ as weights. In practice, we observe $Y(t)=I(\tilde{T}\geq t)$ instead of $I(T\geq t)$. To account for the censoring status of $\tilde{T}$, we adapt the two types of estimators of the Brier Score as defined by formulae 12 and 13 in Proust-Lima et al. (2014), namely data-based Brier Score

$$\hat{\mathrm{BS}}_1(t)=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(\tilde{T}_i>t)}{\hat{G}(t)}\{1-\hat{S}(t|\boldsymbol{x}_i)\}^2+\frac{\Delta_i I(\tilde{T}_i\leq t)}{\hat{G}(\tilde{T}_i)}\{0-\hat{S}(t|\boldsymbol{x}_i)\}^2\right\}$$

and model-based Brier Score

$$\begin{aligned}\hat{\mathrm{BS}}_2(t)=&\frac{1}{n}\sum_{i=1}^{n}\Bigg[I(\tilde{T}_i>t)\{1-\hat{S}(t|\boldsymbol{x}_i)\}^2+\Delta_i I(\tilde{T}_i\leq t)\{0-\hat{S}(t|\boldsymbol{x}_i)\}^2\\&+(1-\Delta_i)I(\tilde{T}_i\leq t)\left\{\{1-\hat{S}(t|\boldsymbol{x}_i)\}^2\frac{\hat{S}(t|\boldsymbol{x}_i)}{\hat{S}(\tilde{T}_i|\boldsymbol{x}_i)}+\{0-\hat{S}(t|\boldsymbol{x}_i)\}^2\left(1-\frac{\hat{S}(t|\boldsymbol{x}_i)}{\hat{S}(\tilde{T}_i|\boldsymbol{x}_i)}\right)\right\}\Bigg].\end{aligned}$$

Here an estimate $\hat{G}(\cdot)$ of the survival function for censoring can be obtained by either Kaplan-Meier or regression models.

In numerical analysis, we conduct 5-fold cross validation, fit models on the training set, and estimate the Brier Score $\hat{\text{BS}}_j^{(f)}(t), j = 1, 2, f = 1, \ldots, 5$ for the testing set of the fth cross-validation fold for a given range of t. Then we report the average Brier score $\overline{\hat{\text{BS}}}_j(t) = \frac{1}{5}\sum_{f=1}^{5} \hat{\text{BS}}_j^{(f)}(t)$ among folds to assess the prediction performances. We use Kaplan-Meier estimator to estimate $\hat{G}(\cdot)$ in our estimation of Brier Scores.

3.4 Simulation study

We conducted simulation studies to evaluate the finite-sample performance of the proposed method in terms of parameter estimation, and selecting the number of classes L. In addition, we compared the proposed method and the standard proportional hazard model in terms of goodness-of-fit and prediction. With $L = 2$ or 3, we generated a two-dimensional baseline covariate vector $\boldsymbol{x} = (x_1, x_2)$, where x_1 is a binary $Bernoulli(0.5)$ random variable and x_2 is a continuous $Uniform(0, 1)$ random variable. Then the latent class label vector $\boldsymbol{\xi}$ was generated from a $Multinomial(1, \{p_1(\boldsymbol{x}; \boldsymbol{\alpha}), \ldots, p_L(\boldsymbol{x}; \boldsymbol{\alpha})\}^T)$ distribution following model (3.2.1). Given latent classes, the time-to-event T was generated from class-specific distribution function $F_T(t|\xi_l = 1) = 1 - \exp\{0.1(1 - e^t)\exp(\boldsymbol{z}_l^T\boldsymbol{\gamma})\}, l = 1, \ldots, L$ derived from model (3.2.2) with $\lambda_0(t) = 0.1(e^t - 1)$. Then we generated independent censoring time C as the minimum of an $Exponential(r)$ variable and a $Uniform(5, 6)$ variable.

Table 3.1: Choices of parameters in the five simulation scenarios.

Simulation scenarios		Censoring parameter	Parameters in model (3.2.1) $\boldsymbol{\alpha}$		Parameters in model (3.2.2) $\boldsymbol{\gamma}$				
		r	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_3$	$\boldsymbol{\zeta}_1$	a_2	$\boldsymbol{\zeta}_2$	a_3	$\boldsymbol{\zeta}_{3,1}$
$L = 2$	scenario (I)	0.1	(log(2),0,0)	NA	(-2,0)	2	(2,2)	NA	NA
	scenario (II)	0.1	(log(2),0,0)		(-2,0)	0	(2,2)		
	scenario (III)	0.6	(log(2),0,0)		(-2,0)	2	(2,2)		
	scenario (IV)	0.1	(2,-4,0)		(0,-3)	0.5	(0,6)		
$L = 3$	scenario (V)	0.1	(0,-0.5,0)	(0,0,0.5)	(-2,-2)	2	(2,2)	4	(4,4)

Table 3.1 summarizes the choice of r, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ in five simulation scenarios. For scenarios with $L = 2$ (I,II,III,IV), scenario (I) served as a benchmark with relatively light censoring rate ($r = 0.1$) and less overlapped survival distributions ($a_2 = 2$) among the two classes. In contrast, scenario (II) created more overlapped survival distributions ($a_2 = 0$) while scenario (III) created heavy censoring ($r = 0.6$). Scenario (IV) considered a special situation where covariate x_1 had a large effect size ($\alpha_{2,1} = -4$) on class probability $p_l(\boldsymbol{x}; \boldsymbol{\alpha})$ but no covariate effect ($\zeta_{1,1} = \zeta_{2,1} = 0$) in survival submodel, while x_2 had zero covariate effect ($\alpha_{2,2} = 0$) on class probability but a large effect size ($\zeta_{1,2} = -3, \zeta_{2,2} = 6$) in survival submodel. In our description later, the scenario (IV) is refer to as the scenario with "separation of covariate effects in submodels". Compared to scenario (I), scenario (IV) had slightly heavier censoring with similar overlapped level of survival distributions among the two classes. With three latent classes, scenario (V) was comparable to scenario (I) in terms of censoring and the overlapping among class-specific survival distributions. Empirical metrics of censoring and overlapping among classes for the five scenarios can be found in Table 3.2.

Table 3.2: Convergence rate, median standardized entropy index and median censoring rate out of 10000 simulations for the five simulation scenarios with non-informative initialization.

Simulation scenarios		Sample size	Convergence	Median entropy	Median censoring
$L = 2$	scenario (I)	1000	97.66%	0.7667	11%
	scenario (II)	1000	97.38%	0.4348	17%
	scenario (III)	1000	96.06%	0.6228	38%
	scenario (IV)	1000	97.20%	0.7766	19%
$L = 3$	scenario (V)	1000	97.68%	0.7585	15%
		2000	98.14%	0.7660	15%
		3000	95.29%	0.7717	15%

3.4.1 Estimation of parameters

To evaluate parameter estimation, we conducted 10000 simulations, with sample size $n = 1000$ for scenarios (I)-(IV) and sample sizes $n = 1000$, 2000 and 3000 for scenario (V).To initialize the algorithm, we used a perturbed $\hat{E}(\boldsymbol{\xi})$ from the true latent class labels $\boldsymbol{\xi}$. In addition, the variance estimation for $\{\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\gamma}}^T\}^T$ was conducted using the profile likelihood approach, while the variance estimation for $\hat{\Lambda}(\cdot)$ was conducted using the observed-data log-

likelihood approach. We seldom observed non-convergent estimates defined as the outlying point estimates whose L_2 norms $\sqrt{|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0|^2 + |\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|^2}$ were greater than the median L_2 norm out of 10000 results plus 5 times median absolute deviation (MAD). Table 3.2 displays convergence rate, median standardized entropy index, and median censoring rate for different simulation scenarios out of 10000 simulations. Table 3.2 indicates that compared to the benchmark scenario (I), more mixed survival distributions (II), heavier censoring (III), or larger number of L (V) would result in more non-convergent results. In addition, heavier censoring (III) would also result in a lower standardized entropy index, suggesting that censoring intensified the fuzziness of the mixtures. Moreover, scenario (IV) displays similar level of mixture as scenario (I), with a slightly higher censoring rate.

The simulation results for four representative parameters, $\alpha_{2,2}$, $\zeta_{1,1}$, a_2 and $\Lambda(3)$, are shown in Table 3.3. Full results for all unknown parameters are available in Tables 3.6 and 3.7 in Appendix section 3.7.4. As observed, under scenario (I) and (IV) the proposed estimator achieved very small median biases and accurately estimated standard errors. The coverage probabilities of the 95% confidence intervals are close to 0.95 for both regression coefficient $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and infinite-dimensional $\hat{\Lambda}(t)$. Compared to scenario (I) and (IV), fuzzier mixture pattern in scenario (II) and heavier censoring in scenario (III) result in larger median biases for most parameters. In addition, a slight underestimation of the standard errors is observed for scenario (II) and scenario (III), such that the coverage probabilities are slightly lower than 0.95, in particular for $\hat{a}_2$. For simulation (V) with three latent classes, the estimation tends to be unstable with smaller sample size 1000, showing higher biases in the proportionality parameters $\hat{a}_2$ and $\hat{a}_3$ and regression parameters $\hat{\zeta}_{21}$ and $\hat{\zeta}_{22}$. This is probably due to insufficient sample size, which in particular damages estimation for the parameters corresponding to the second class which overlaps with both class 1 and class 3. As sample size grows to 2000 and 3000, an improvement in median biases and coverage probabilities is observed. However, compared to scenarios (I)-(IV) with two classes, the proposed method requires a larger average sample size from each class to detect the mixture pattern of time-to-event distribution.

Table 3.3: Median bias (M.Bias), standard deviation (SE), median standard error estimate (SEE), and coverage probability (CP) of parameters $\hat{\alpha}_{2,2}$, $\hat{\zeta}_{1,1}$, $\hat{a}_2$ and $\hat{\Lambda}(3)$ out of 10000 simulations with non-informative initialization..

n	Scenarios	$\hat{\alpha}_{2,2}$				$\hat{\zeta}_{1,1}$			
		M.Bias	SE	SEE	CP	Bias	SE	SEE	CP
1000	(I)	-0.015	0.304	0.296	0.949	-0.023	0.197	0.201	0.957
1000	(II)	0.005	0.517	0.500	0.945	-0.045	0.319	0.313	0.956
1000	(III)	-0.035	0.417	0.380	0.936	-0.069	0.429	0.404	0.963
1000	(IV)	-0.011	0.668	0.695	0.962	-0.013	0.204	0.206	0.951
1000	(V)	0.013	0.606	0.532	0.908	-0.063	0.256	0.229	0.941
2000	(V)	0.012	0.397	0.381	0.936	-0.036	0.157	0.152	0.947
3000	(V)	0.020	0.314	0.312	0.944	-0.024	0.122	0.120	0.952
		$\hat{a}_2$				$\hat{\Lambda}(3)$			
		M.Bias	SE	SEE	CP	Bias	SE	SEE	CP
1000	(I)	0.011	0.446	0.412	0.941	-0.008	0.352	0.344	0.951
1000	(II)	0.023	0.441	0.406	0.930	0.010	0.540	0.503	0.952
1000	(III)	0.010	0.729	0.616	0.918	0.000	0.772	0.675	0.940
1000	(IV)	0.003	0.311	0.309	0.952	0.018	0.488	0.450	0.941
1000	(V)	-0.868	1.229	0.783	0.675	0.196	0.643	0.566	0.904
2000	(V)	-0.515	0.884	0.653	0.789	0.060	0.414	0.410	0.935
3000	(V)	-0.346	0.679	0.562	0.863	0.019	0.322	0.328	0.946

3.4.2 Determining the number of latent classes

We further conducted 1000 simulations for each of the five simulation scenarios with sample size $n = 1000$. In each simulation, we fitted the proposed latent class model for $L \in \{2, 3, 4, 5\}$ with algorithms initialized by K-means clustering. Then we compared model selection criteria for the models with different choices of L.

As shown in Figure 3.1, BIC correctly selected L in all 1000 simulations when the two latent classes are well separated (I), even if heavily censored (III). BIC also performed well under heavy mixture (II), with separated covariate effects in submodels (IV), and three-class (V) scenarios. Compared to BIC, AIC tended to select a larger number of latent classes, particularly for the heavy mixture scenario (II). In terms of entropy-based criteria, we found that the standardized entropy index tended to select incorrect L, which also explained that the classification entropy extended BIC (CE-BIC) performed worse than the standalone BIC. Similar results were also observed when there were three latent classes in scenario (V).

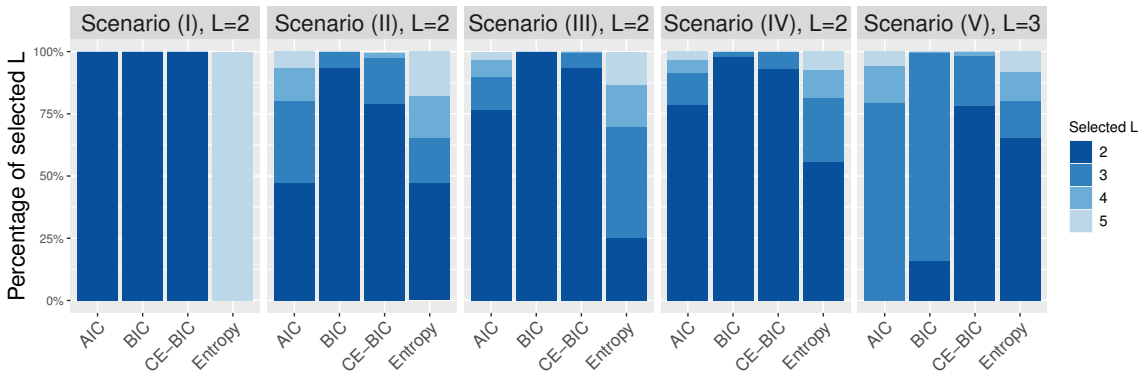

Figure 3.1: Percentage of latent classes selected by different model selection criteria out of 1000 simulations under simulation scenarios (I)-(V).

The superiority of BIC over entropy-based criteria can be explained by the fact that the proposed method is a likelihood-based method. According to the performance in the five scenarios, BIC is the most effective criterion in selecting L. We also utilized BIC to select L in our real data application in Section 2.5.

3.4.3 Goodness-of-fit and prediction

For each of the five simulation scenarios, we further simulated 1000 datasets with sample size 1000. For each simulated dataset, we conducted five-fold cross-validation as described in Section 3.3.5 to obtain the averaged estimates $\overline{\hat{\text{BS}}_1}(t)$ and $\overline{\hat{\text{BS}}_2}(t)$ of the Brier Score for a standard Cox regression model and the proposed latent class model. We set the upper bound of time interval $t^* = 5$ for scenarios (I) - (IV) and $t^* = 5.75$ for scenarios (V) to cover the support of time-to-event. Note that the Cox regression model is a special case of the latent class model with $L = 1$. Therefore, under the Cox regression model we have $\Pr(\xi_1 = 1|\boldsymbol{x}) = 1$ and the predicted survival function $\hat{S}(t|\boldsymbol{x}) = \hat{S}(t|\boldsymbol{x}, \xi_l = 1) = \exp\{-\hat{\Lambda}(t)\exp(\boldsymbol{x}^T\hat{\boldsymbol{\zeta}}_1)\}$ is solely based on the single class (or class 1) considered in the model.

Figure 3.2, and Figures 3.5-3.8 in Appendix section 3.7.4, shows the obtained time-dependent Brier Score estimates for scenarios (IV), (I), (II), (III) and (V), respectively. Overall, the proposed latent class model achieved consistently lower median average corss-validated Brier Score estimates than those obtained by the Cox model in all simulation scenarios. As shown

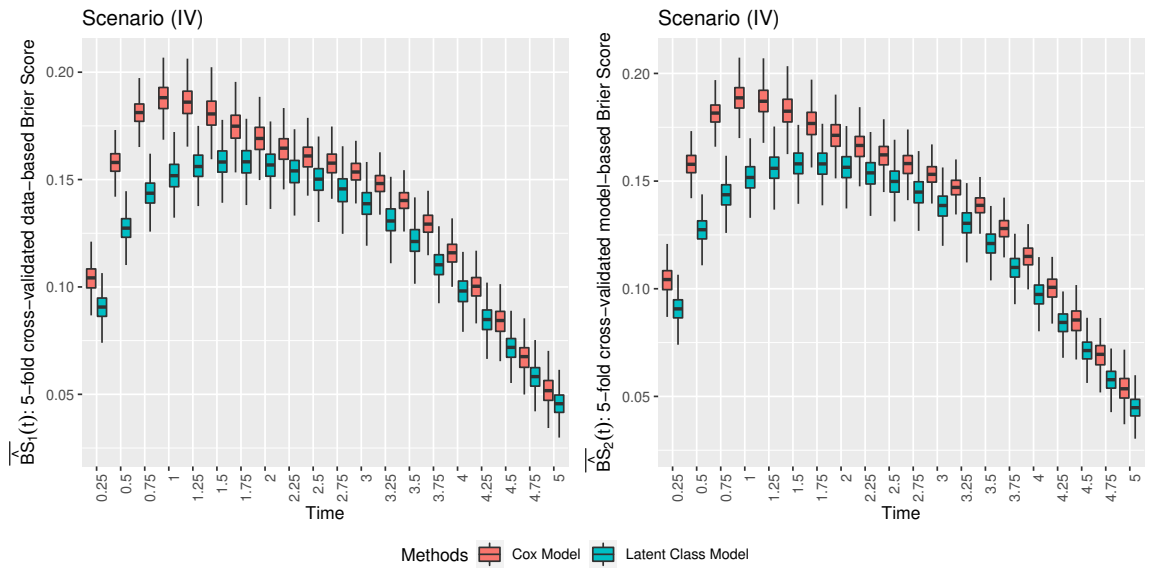

Figure 3.2: Boxplots for average cross-validated Brier Score $\overline{\hat{\mathrm{BS}}_1}(t)$ and $\overline{\hat{\mathrm{BS}}_2}(t)$, $t \in (0, 5]$, from 1000 simulations under scenario (IV) with sample size 1000, for the Cox model and the proposed latent class model with $L = 2$.

in Figures 3.5-3.8, however, only minor improvements can be recognized for scenarios (I), (II), (III) and (V), where baseline covariate effects are present in both class probability submodel and class-specific survival submodel. In contrast, the improvement is obvious under scenario (IV), where we have separation of covariate effects in submodels. Under this situation, covariates have different effects towards class membership probability and class-specific survival, which is difficult to be captured by a single-class standard Cox model.

3.5 Real data example

We applied our method to investigate the heterogeneity of mild cognitive impairment (MCI) using time-to-dementia data collected for 5348 patients in the Uniform Data Set between September 2005 and June 2015 by the U.S. National Alzheimer's Coordinating Center. 1501 patients developed dementia during the follow up, showing a high censoring rate of 72%. We incorporated patients' baseline cognitive characteristics as covariates, including overall cognition (Mini-mental state examination, MMSE), executive functions (Trail making test B , TB, and Digit symbol, DS), memory (logical memory delayed, LMD, and category

fluency, CF), language (Boston naming, BN), and attention (Trail making test A, TA, and digit span forward, DSF). In addition, patients' baseline number of impaired instrumental activities of daily living (IADLs), number of neuropsychiatric symptoms (NPI-Q), binary measure of depression (GDS), indicator of cerebrovascular disease (EH), and baseline age (AGE) were also included as baseline covariates. Detailed descriptions about the dataset and covariates were reported in Hanfelt et al. (2018).

High heterogeneity of the MCI population indicates that there exist MCI subgroups associated with a specific cognitive domain or domains. Thus, it is expected that the progression to dementia for different MCI subgroups are driven by their corresponding domain factors. We applied the proposed latent class model to investigate such heterogeneity in terms of the importance and effect sizes of baseline covariates.

We first decided the best number of classes L. Specifically, we fitted the proposed models with L classes with random initialization for multiple times, then selected the model with the smallest BIC as the best L-class model. We conducted the above procedure for $L \in \{2, 3, 4\}$. The 2-class model obtains the smallest BIC (24481) compared to the 3-class model (24625) and the 4-class model (24797), where the BIC shows an increasing trend as L increases from 2 to 4. Thus, we regard the 2-class model as the best latent class model.

3.5.1 Summary statistics of the obtained two latent classes

According to the fitted latent class model with two latent classes, we assign patients to the two classes by modal assignment. That is, we assign each patient to the class with the highest posterior membership probability $\hat{E}(\boldsymbol{\xi})$. As Table 3.4 shows, 69% of the patients are assigned to class 1, while 31% of the patients are assigned to class 2. Comparing the two classes, the first class had significantly smaller MMSE compared to the second class, showing better overall cognitive status. Moreover, class 1 was generally better than class 2 in most of the domain-specific scores, apart from the Boston Naming test associated with the language domain. In addition, patients in class 2 were older than those in class 1. In terms of time-to-event, patients in class 1 generally took longer than patients in class 2 to

Table 3.4: Summary statistics of the baseline covariates for the two latent classes, based on modal assignment of class identity.

Covariates	Class 1, N = 3714[1]	Class 2, N = 1634[1]	p-value[2]
$\tilde{T}$	1.83 (0.00, 3.42)	1.08 (0.00, 2.08)	<0.001
Δ[3]	683 (18%)	818 (50%)	<0.001
MMSE	-0.99 (-2.20, 0.00)	-2.09 (-3.78, -0.85)	<0.001
TB[4]	0.42 (-0.22, 1.42)	1.71 (0.52, 4.02)	<0.001
DS	-0.52 (-1.19, 0.11)	-1.38 (-2.01, -0.80)	<0.001
LMD	-1.23 (-2.05, -0.41)	-1.52 (-2.34, -0.65)	<0.001
CF	-0.75 (-1.35, -0.12)	-1.31 (-1.90, -0.73)	<0.001
BN	-0.61 (-1.88, 0.22)	-0.47 (-1.55, 0.29)	<0.001
TA[4]	0.12 (-0.44, 0.90)	0.70 (-0.07, 1.73)	<0.001
DSF	-0.29 (-0.88, 0.49)	-0.44 (-0.98, 0.39)	<0.001
EH	224 (6.0%)	104 (6.4%)	0.6
IADLs	1 (0, 2)	4 (2, 6)	<0.001
NPI-Q	1 (0, 2)	2 (1, 4)	<0.001
GDS	694 (19%)	279 (17%)	0.2
AGE	-0.20 (-0.81, 0.41)	0.22 (-0.38, 0.77)	<0.001

[1] Median (IQR); n (%)
[2] Wilcoxon rank sum test; Pearson's Chi-squared test
[3] Number of patients diagnosed with dementia
[4] Larger Trails B and Trails A scores indicate worse conditions.

reach dementia during the follow-up, where only 18% of patients developed dementia in class 1 but half of patients developed dementia in class 2.

3.5.2 Parameter estimation and interpretation

In order to demonstrate the utility of the proposed method in investigating the heterogeneity in covariate importance and effect sizes, we compare the point estimation, confidence interval and interpretations of the standard single-class Cox model and the proposed latent class model with $L = 2$ by Table 3.5. From the Cox model ($\hat{\boldsymbol{\zeta}}$ in Table 3.5), it is clear that patients with worse baseline conditions in different cognitive domains (executive function, memory, language and attention), functional abilities, behavioral scales and aging tended to have increased hazard, or earlier onset, of dementia. However, this overall picture revealed by the Cox regression model cannot conduct more detailed investigations on the

correspondence between the MCI subtypes and the associated domain factors.

In contrast, our proposed latent class model were able to capture the heterogeneous associations between baseline characteristics and dementia, with sensible clinical interpretations. According to the point estimates for the class membership probability submodel ($\hat{\boldsymbol{\alpha}}$ in Table 3.5), younger MCI patients with more severe problems in language domain (BN) were more likely to belong to the first latent class, while older MCI patients with worse executive functions (TB and DS) and impaired functional abilities (IADLs) were more likely to belong to the second class.

The class-specific survival submodel revealed further heterogeneity of covariate effects ($\hat{\boldsymbol{\zeta}}_1$ and $\hat{\boldsymbol{\zeta}}_2$ in Table 3.5) on survival. First of all, we found for both classes worse baseline overall cognition (MMSE) had statistically significant effect in increasing the hazard of dementia. In addition, memory loss (LMD) had significant effect for both classes but with fairly different effect sizes. In contrast, the effects of worse executive functions (TB and DS) were statistically significant only for the second class, while problems in language domain (BN), functional abilities (IADLs), behaviors (NPI-Q) and age (AGE) had significant effect only for the first class.

Combining our class probability submodel and class-specific survival submodel, we were able to correspond the two data-driven classes to meaningful clinical MCI subgroups. The first class were younger multi-domain amnestic MCI patients with early onset of language problem, which might be relevant to primary progressive aphasia occurring before memory related symptoms (Rogalski et al., 2016). In contrast, the second class were older multi-domain amnestic MCI patients with impaired executive functions, which appeared to be have more typical symptoms of Alzheimer's Disease.

3.5.3 Assessment of goodness-of-fit and prediction performances

We assessed the goodness-of-fit of our latent class model by comparing the Kaplan-Meier curve of time-to-dementia for the MCI population, and the estimated survival probability curve from the model, calculated by averaging the predicted survival probability (3.3.6) for

Table 3.5: Point estimates and 95% confidence intervals for the covariate effects obtained by Cox model and the latent class model with two classes.

		Cox model (1 class)		Latent class model (2 classes)					
				Class probability		Class-specific survival submodel			
Domains	Covariates	$\hat{\boldsymbol{\zeta}}$	95% CI	$\hat{\boldsymbol{\alpha}}$	95% CI	$\hat{\boldsymbol{\zeta}}_1$	95% CI	$\hat{\boldsymbol{\zeta}}_1 + \hat{\boldsymbol{\zeta}}_2$	95% CI
	Intercept	NA	NA	-2.94*	(-5.00,-0.88)	NA	NA	2.03*	(1.26,2.80)
Overall cognition	MMSE	-0.12*	(-0.15,-0.10)	-0.17	(-0.40, 0.07)	-0.14*	(-0.21,-0.07)	-0.09*	(-0.16,-0.03)
Executive functions	TB	0.08*	(0.05,0.12)	0.20*	(0.00, 0.40)	-0.01	(-0.14, 0.11)	0.12*	(0.04,0.19)
	DS	-0.11*	(-0.17,-0.05)	-0.77*	(-1.28,-0.26)	0.03	(-0.20, 0.26)	-0.17*	(-0.30,-0.03)
Memory	LMD	-0.41*	(-0.46,-0.35)	0.23	(-0.40, 0.86)	-0.63*	(-0.80,-0.46)	-0.27*	(-0.39,-0.15)
	CF	-0.21*	(-0.27,-0.14)	-0.74	(-1.73, 0.26)	-0.17	(-0.36, 0.02)	-0.13	(-0.29,0.03)
Language	BN	-0.03*	(-0.06,0.00)	0.35*	(0.15, 0.55)	-0.17*	(-0.26,-0.07)	0.04	(-0.07,0.15)
Attention	TA	-0.04*	(-0.08,0.00)	-0.15	(-0.49, 0.19)	-0.10	(-0.21, 0.01)	0.01	(-0.07,0.10)
	DSF	0.05	(-0.00,0.10)	0.06	(-0.32, 0.44)	0.01	(-0.11, 0.13)	0.07	(-0.10,0.25)
Cerebrovascular disease	EH	-0.02	(-0.23,0.18)	-1.10	(-2.53, 0.34)	0.39	(-0.16, 0.93)	-0.11	(-0.55,0.33)
Functional abilities	IADLs	0.12*	(0.10,0.14)	0.40*	(0.03, 0.76)	0.21*	(0.14, 0.28)	0.03	(-0.05,0.10)
Behavioral assessment	NPI-Q	0.06*	(0.04,0.09)	0.19	(-0.17, 0.55)	0.11*	(0.03, 0.19)	0.00	(-0.08,0.08)
	GDS	0.07	(-0.07,0.21)	-0.70	(-2.03, 0.62)	0.09	(-0.33, 0.50)	0.12	(-0.30,0.55)
Aging	AGE	0.27*	(0.20,0.33)	0.90*	(0.39, 1.42)	0.38*	(0.20, 0.56)	0.01	(-0.20,0.22)

*Statistically signficant covariate effect based on 95% confidence interval.
Higher scores on TB and TA indicated worse conditions.

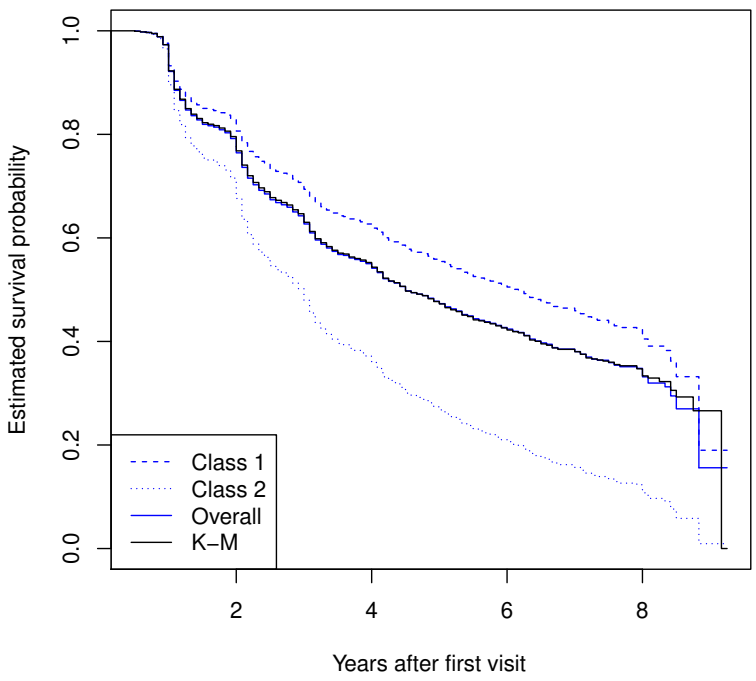

Figure 3.3: Blue dashed and dotted lines (Class 1 and Class 2): Predicted class-specific survival probabilities by the latent class model. Blue solid line (Overall): Predicted overall survival probability by the latent class model. K-M: Estimated Kaplan-Meier curve for overall survival probability.

all patients at each uncensored event time. As shown in Figure 3.3, the Kaplan-Meier curve (referred to as "K-M") is very close to the survival curve based on the proposed model (referred to as "Overall"), indicating reasonable goodness-of-fit. In Figure 3.3, we also plot the average class-specific survival probabilities for all patients at each observed event time. As observed, the survival curve for class 1 is higher than the curve for class 2, indicating that patients in class 1 had slower progression towards dementia.

As we did in simulation, we compared the proposed method and the Cox regression model in prediction by cross-validated average Brier Scores $\overline{\widehat{\mathrm{BS}}_1}(t)$ and $\overline{\widehat{\mathrm{BS}}_2}(t)$ for $t \in (0, 8]$. As shown in Figure 3.4, our proposed method achieved lower Brier Scores in five-fold cross validation, which is consistent with our observation in the simulation study. Similar to simulation scenario (IV), in real application we also observe separation of covariate effects (Table 3.5) in submodels for covariates MMSE, LMD, and NPI-Q, which explains the big

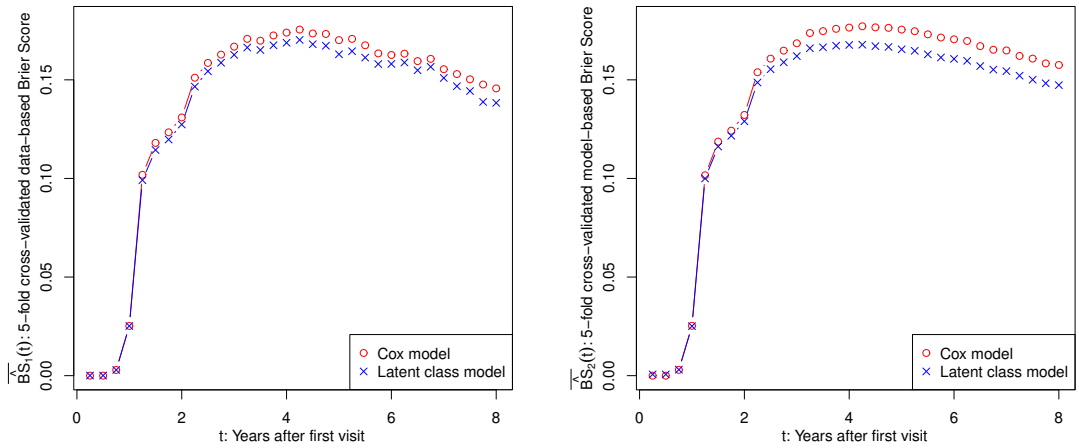

Figure 3.4: Average of 5-fold cross-validated Brier Scores, $\overline{\widehat{\mathrm{BS}}}_j(t), j = 1, 2$, obtained by the Cox model and the proposed latent class model with $L = 2$, for the UDS data application.

improvement in Brier Scores made by the latent class model. These further demonstrated that the prediction of the survival outcome can be improved by capturing the mixture structure of a heterogeneous population.

3.6 Discussion

In this article, we propose a semi-parametric approach to jointly modeling the latent class structure and the time-to-event outcome. By utilizing non-parametric maximum likelihood estimator (NPMLE) technique, the proposed method facilitates valid inference for both covariate effects and hazard functions following rigorous asymptotic theory, and is expected to be more robust than fully parametric methods. Our method also flexibly captures class-specific covariate effects in both latent class membership probabilities and class-specific hazard functions.

Instead of including both longitudinal and time-to-event information in the joint framework, we only consider time-to-event outcome in our method. Our treatment circumvents the popular but unreliable conditional independence assumption. Based on a similar finite mixture structure as used in the proposed method, further extensions can be studied to

account correlated structure of longitudinal data and survival data, while keeping the robust semi-parametric submodels developed in this method and for longitudinal observations (Hart et al., 2020, for example).

Computationally, we develop a stable EM algorithm which ensures increasing observed data likelihood in each iteration. The algorithm is efficiently implemented in `Rcpp` (Eddelbuettel and Sanderson, 2014) format and is publicly available as an R package.

3.7 Appendices

3.7.1 Proof of Theorem 3.3.1

Let $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\Lambda}$ be the maximum likelihood estimator corresponding to the observed data log-likelihood. Now define $N(t) = I(\tilde{T} \leq t, \Delta = 1)$, $\tilde{N}(t) = I(\tilde{T} \leq t, \Delta = 0)$ and let $\mathbb{P}_n$, P denote the empirical measure and probability measure, respectively. Then the log-likelihood ℓ satisfies $n^{-1}\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{O}) \equiv \ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)$, where

$$\begin{aligned}\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) =& \mathbb{P}_n \int_0^{t^*} \left[\log \left\{ \sum_{l=1}^{L} p_l(\boldsymbol{x}; \boldsymbol{\alpha}) e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} \exp\left(- \int_0^t e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \right) \right\} + \log \Lambda\{t\} \right] dN(t) \\ &+ \mathbb{P}_n \int_0^{t^*} \log \left\{ \sum_{l=1}^{L} p_l(\boldsymbol{x}; \boldsymbol{\alpha}) \exp\left(- \int_0^t e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \right) \right\} d\tilde{N}(t).\end{aligned}$$

Let $\mathcal{W}$ denote the space of functions on $[0, t^*]$ that are uniformly bounded by 1 and with total variation bounded by 1. Define $\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{(p+1)\times(L-1)} : ||\boldsymbol{u}|| \leq 1\}$ and $\mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{q\times L+L-1} : ||\boldsymbol{v}|| \leq 1\}$. Let $\boldsymbol{u} \in \mathcal{U}$, $\boldsymbol{v} \in \mathcal{V}$, and $h \in \mathcal{W}$. Then $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)$ can be identified as elements in the space of bounded functions on $\mathcal{U} \times \mathcal{V} \times \mathcal{W}$, $\ell^\infty(\mathcal{U} \times \mathcal{V} \times \mathcal{W})$, by $\boldsymbol{u}^T\boldsymbol{\alpha} + \boldsymbol{v}^T\boldsymbol{\gamma} + \int_0^{t^*} h d\Lambda$. Similarly, $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0, \hat{\Lambda} - \Lambda_0)$ can also be identified in $\ell^\infty(\mathcal{U} \times \mathcal{V} \times \mathcal{W})$ by

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0, \hat{\Lambda} - \Lambda_0)[\boldsymbol{u}, \boldsymbol{v}, h] = \sqrt{n}\{\boldsymbol{u}^T(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \boldsymbol{v}^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \int_0^{t^*} h d(\hat{\Lambda} - \Lambda_0)\}.$$

Proof. Step 1. We show by contradiction that $\hat{\Lambda}(t^*) < \infty$. Condition (C1) indicates that

for large n, there exists an observation with probability one such that $\tilde{T} = t^*$ and $\Delta = 0$. If $\hat{\Lambda}(t^*) = \infty$, then

$$\mathbb{P}_n \int_0^{t^*} \log \Big\{ \sum_{l=1}^{L} p_l(\boldsymbol{x}; \boldsymbol{\alpha}) \exp\Big(- \int_0^t e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big) \Big\} d\tilde{N}(t) = -\infty$$

and thus $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) = -\infty$. Therefore, it must satisfy $\hat{\Lambda}(t^*) < \infty$ to maximize ℓ_n.

Step 2. We show that $\limsup_n \hat{\Lambda}(t^*) < \infty$ by contradiction. By conditions (C2) and (C3) there exists a constant M such that $|\boldsymbol{z}_l^T \boldsymbol{\gamma}| \leq M$ for any $\boldsymbol{\gamma}$ and $\boldsymbol{z}_l$, and $p_l(\boldsymbol{x}; \boldsymbol{\alpha}) \in (0, 1)$ for any $\boldsymbol{x}$ and $\boldsymbol{\alpha}$. Define $\bar{\Lambda}(t) = [\hat{\Lambda}(t) \wedge \tilde{M}] \vee \tilde{M}/2$, where $\tilde{M} = e^{-M}\{\log(\epsilon_0)\}^{-1}$ for a chosen $\epsilon_0 \in (0, 1)$.

By definition of MLE $\ell_n(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda}) \geq \ell_n(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \bar{\Lambda})$. Assuming that $\limsup_n \hat{\Lambda}(t^*) = \infty$, then by the following inequality

$$\log(\sum_{l=1}^{L} a_l) \leq \sum_{l=1}^{L} \log a_l + \log L,$$

we have

$$\begin{aligned}
\ell_n(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda}) \leq & \mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \Big\{ \log p_l(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}) - \int_0^t e^{\boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}}} d\hat{\Lambda}(s) \Big\} d\{N(t) + \tilde{N}(t)\} \\
& + \mathbb{P}_n \sum_{l=1}^{L} \boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}} dN(t^*) + \mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \log \hat{\Lambda}\{t\} dN(t) + \log L \mathbb{P}_n\{N(t^*) + \tilde{N}(t^*)\} \\
\leq & - \mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \int_0^t e^{\boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}}} d\hat{\Lambda}(s) d\{N(t) + \tilde{N}(t)\} + LM\mathbb{P}_n dN(t^*) \\
& + \mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \log \hat{\Lambda}\{t\} dN(t) + \log L \mathbb{P}_n\{N(t^*) + \tilde{N}(t^*)\} \to -\infty.
\end{aligned}$$

On the other hand, by the following inequality

$$\log(\sum_{l=1}^{L} a_l) \geq \frac{1}{L} \sum_{l=1}^{L} \log a_l$$

we have

$$\begin{aligned}
\ell_n(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \bar{\Lambda}) \geq & \frac{1}{L}\Bigg[\mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \Big\{ \log p_l(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}) - \int_0^t e^{\boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}}} d\bar{\Lambda}(s) \Big\} d\{N(t) + \tilde{N}(t)\} \\
& + \mathbb{P}_n \sum_{l=1}^{L} \boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}} dN(t^*) + \mathbb{P}_n \sum_{l=1}^{L} \int_0^{t*} \log \bar{\Lambda}\{t\} dN(t) \Bigg] \\
\geq & \frac{1}{L}\Bigg[\mathbb{P}_n \sum_{l=1}^{L} \Big\{ \log p_l(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}) - \tilde{M} e^M \Big\} d\{N(t^*) + \tilde{N}(t^*)\} \\
& + \mathbb{P}_n \sum_{l=1}^{L} \boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}} dN(t^*) + \sum_{l=1}^{L} \log \frac{\tilde{M}}{2} \mathbb{P}_n dN(t^*) \Bigg] \\
= & \frac{1}{L}\Bigg[\mathbb{P}_n \sum_{l=1}^{L} \Big\{ \log p_l(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}) + \{\log(\epsilon_0)\}^{-1} \Big\} d\{N(t^*) + \tilde{N}(t^*)\} \\
& + \mathbb{P}_n \sum_{l=1}^{L} \boldsymbol{z}_l^T \hat{\boldsymbol{\gamma}} dN(t^*) + \sum_{l=1}^{L} \{-M - \log(-\log(\epsilon_0)) - \log 2\} \mathbb{P}_n dN(t^*) \Bigg] > -\infty.
\end{aligned}$$

The above contradiction shows that $\limsup_n \hat{\Lambda}(t^*) < \infty$. By Helly's selection theorem there exists a converging subsequence such that $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}^*$, $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}^*$ and $\hat{\Lambda} \to \Lambda^*$.

Step 3. We show that the limit of the subsequence mentioned in the end of step 2 are $\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0$ and Λ_0. Define $\Lambda^\epsilon(t) = \int_0^t \{1 + \epsilon h(s)\} d\Lambda(s)$, where $h(t) \in \mathcal{W}$, the space of functions on $[0, t^*]$ that are uniformly bounded by 1 and with total variation bounded by 1. Then we obtain the derivative of log-likelihood $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda^\epsilon)$ with respect to ϵ at 0, denoted by $\dot{\ell}_{n,\Lambda}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)[h]$:

$$\begin{aligned}
\dot{\ell}_{n,\Lambda}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)[h] = & \mathbb{P}_n \int_0^{t^*} \Bigg[\sum_{l=1}^{L} \tau_l(t; \boldsymbol{O}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) \Big\{ - \int_0^t h(s) e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} + h(t) \Bigg] dN(t) \\
& + \mathbb{P}_n \int_0^{t^*} \Bigg[\sum_{l=1}^{L} \tilde{\tau}_l(t; \boldsymbol{O}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) \Big\{ - \int_0^t h(s) e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} \Bigg] d\tilde{N}(t),
\end{aligned} \tag{3.7.1}$$

where

$$\begin{aligned}
\tau_{il}(t; \boldsymbol{O}_i, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) = & \frac{p_l(\boldsymbol{x}_i; \boldsymbol{\alpha}) f_l(\tilde{T}_i = t, \Delta_i = 1 | \boldsymbol{x}_i; \boldsymbol{\gamma}, \Lambda)}{\sum_{d=1}^{L} p_d(\boldsymbol{x}_i; \boldsymbol{\alpha}) f_d(\tilde{T}_i = t, \Delta_i = 1 | \boldsymbol{x}_i; \boldsymbol{\gamma}, \Lambda)} \\
= & \frac{p_l(\boldsymbol{x}_i; \boldsymbol{\alpha}) \exp(\boldsymbol{z}_{il}^T \boldsymbol{\gamma}) \exp\{- \int_0^t e^{\boldsymbol{z}_{il}^T \boldsymbol{\gamma}} d\Lambda(s)\}}{\sum_{d=1}^{L} p_d(\boldsymbol{x}; \boldsymbol{\alpha}) \exp(\boldsymbol{z}_{id}^T \boldsymbol{\gamma}) \exp\{- \int_0^t e^{\boldsymbol{z}_{il}^T \boldsymbol{\gamma}} d\Lambda(s)\}}
\end{aligned}$$

and

$$\begin{aligned}\tilde{\tau}_{il}(t;\boldsymbol{O}_i,\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& \frac{p_l(\boldsymbol{x}_i;\boldsymbol{\alpha})f_l(\tilde{T}_i=t,\Delta_i=0|\boldsymbol{x}_i;\boldsymbol{\gamma},\Lambda)}{\sum_{d=1}^{L}p_d(\boldsymbol{x}_i;\boldsymbol{\alpha})f_d(\tilde{T}_i=t,\Delta_i=0|\boldsymbol{x}_i;\boldsymbol{\gamma},\Lambda)}\\ =& \frac{p_l(\boldsymbol{x}_i;\boldsymbol{\alpha})\exp\{-\int_0^t e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}}d\Lambda(s)\}}{\sum_{d=1}^{L}p_d(\boldsymbol{x};\boldsymbol{\alpha})\exp\{-\int_0^t e^{\boldsymbol{z}_{il}^T\boldsymbol{\gamma}}d\Lambda(s)\}}.\end{aligned}$$

By changing the order of integration we have

$$\begin{aligned}\dot{\ell}_{n,\Lambda}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)[h] =& \mathbb{P}_n\int_0^{t^*}h(s)dN(s)-\mathbb{P}_n\sum_{l=1}^{L}\int_0^{t^*}h(s)e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}}\int_s^{t^*}\tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)dN(t)d\Lambda(s)\\ &-\mathbb{P}_n\sum_{l=1}^{L}\int_0^{t^*}h(s)e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}}\int_s^{t^*}\tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)d\tilde{N}(t)d\Lambda(s).\end{aligned}$$

By definition of the NPMLE, $\dot{\ell}_{n,\Lambda}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda})[h]=0$ for all $h\in\mathcal{W}$. By taking $h(\cdot)=I(\cdot\le t)$, we have

$$\hat{\Lambda}(t)=\int_0^t\frac{\mathbb{P}_n dN(s)}{\phi_n(s;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda})},$$

where

$$\phi_n(t;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)=\mathbb{P}_n\sum_{l=1}^{L}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}}\Big\{\int_t^{t^*}\tau_l(s;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)dN(s)+\int_t^{t^*}\tilde{\tau}_l(s;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)d\tilde{N}(s)\Big\}.$$

By definition and regularity conditions (C1) - (C3), $\tau(\cdot)\in(0,1)$ and $\tilde{\tau}(\cdot)\in(0,1)$ for all $t\in[0,t^*]$, thus both $\tau(\cdot)$ and $\tilde{\tau}(\cdot)$ are uniformly bounded away from zero. Then we can find neighborhoods $\mathcal{A}$ of $\boldsymbol{\alpha}^*$, Γ of $\boldsymbol{\gamma}^*$, $\mathcal{B}$ of Λ^*, such that $\{\tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda):t\in[0,t^*],l=1,\ldots,L,\boldsymbol{\alpha}\in\mathcal{A},\boldsymbol{\gamma}\in\Gamma,\Lambda\in\mathcal{B}\}$ and $\{\tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda):t\in[0,t^*],l=1,\ldots,L,\boldsymbol{\alpha}\in\mathcal{A},\boldsymbol{\gamma}\in\Gamma,\Lambda\in\mathcal{B}\}$ are Donsker thus Glivenko-Cantelli. Then by Glivenko-Cantelli theorem

$$\sup_{t\in[0,t^*],\boldsymbol{\alpha}\in\mathcal{A},\boldsymbol{\gamma}\in\Gamma,\Lambda\in\mathcal{B}}|\phi_n(t;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)-\phi^*(t;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)|\to 0,$$

where

$$\phi^*(t;\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)=P\sum_{l=1}^{L}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}}\Big\{\int_t^{t^*}\tau_l(s;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)dN(s)+\int_t^{t^*}\tilde{\tau}_l(s;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)d\tilde{N}(s)\Big\}.$$

Then by step 2 and the continuity of ϕ_n in $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and Λ, we have

$$\begin{aligned}\sup_{t\in[0,t^*]} &|\phi_n(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda}) - \phi^*(t;\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*)| \\ &\leq \sup_{t\in[0,t^*]} |\phi_n(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda}) - \phi_n(t;\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*)| \\ &+ \sup_{t\in[0,t^*]} |\phi_n(t;\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*) - \phi^*(t;\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*)| \to 0.\end{aligned}$$

In addition, we also have $\sup_{t\in[0,t^*]} |\mathbb{P}_n dN(t) - PdN(t)| \to 0$. Then we define

$$\tilde{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\phi_n(s;\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\Lambda_0)}.$$

Then by previous derivations,

$$\tilde{\Lambda}(t) \to \int_0^t \frac{PdN(s)}{\phi^*(s;\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\Lambda_0)} = \Lambda_0(t)$$

uniformly. Now define

$$\begin{aligned}\ell(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& P\int_0^{t^*} \left[\log\left\{\sum_{l=1}^L p_l(\boldsymbol{x};\boldsymbol{\alpha}) e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}} \exp\left(-\int_0^t e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}} d\Lambda(s)\right)\right\} + \log\Lambda\{t\}\right] dN(t) \\ &+ P\int_0^{t^*} \log\left\{\sum_{l=1}^L p_l(\boldsymbol{x};\boldsymbol{\alpha}) \exp\left(-\int_0^t e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}} d\Lambda(s)\right)\right\} d\tilde{N}(t).\end{aligned}$$

Then by definition of NPMLE, $\ell_n(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda}) - \ell_n(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\tilde{\Lambda}) \geq 0$, thus $\lim_n\{\ell_n(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda}) - \ell_n(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\tilde{\Lambda})\} \geq 0$. However, we can also show that

$$\lim_n\{\ell_n(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}},\hat{\Lambda}) - \ell_n(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\tilde{\Lambda})\} = \ell(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*) - \ell(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\Lambda_0) \leq 0.$$

Therefore, $\ell(\boldsymbol{\alpha}^*,\boldsymbol{\gamma}^*,\Lambda^*) = \ell(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0,\Lambda_0)$ and by regularity conditions (C1)-(C3), $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda^*(t) = \Lambda_0(t)$. Thus the consistency follows. □

3.7.2 Proof of Theorem 3.3.2

Proof. We use Theorem 19.26 from Van der Vaart (2000) to conduct the proof. In addition to the score function $\dot{\ell}_{n,\Lambda}$ for Λ derived in (3.7.1), we also derive the score functions for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ as following

$$\begin{aligned}
\dot{\ell}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& \mathbb{P}_n \int_0^{t^*} \sum_{l=1}^{L} \tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x};\boldsymbol{\alpha}) dN(t) \\
&+ \mathbb{P}_n \int_0^{t^*} \sum_{l=1}^{L} \tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x};\boldsymbol{\alpha}) d\tilde{N}(t); \\
\dot{\ell}_{n,\boldsymbol{\gamma}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& \mathbb{P}_n \int_0^{t^*} \sum_{l=1}^{L} \tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \Big\{ \boldsymbol{z}_l - \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} dN(t) \\
&- \mathbb{P}_n \int_0^{t^*} \sum_{l=1}^{L} \tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) d\tilde{N}(t).
\end{aligned}$$

Let

$$\begin{aligned}
\dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& \int_0^{t^*} \sum_{l=1}^{L} \tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x};\boldsymbol{\alpha}) dN(t) \\
&+ \int_0^{t^*} \sum_{l=1}^{L} \tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x};\boldsymbol{\alpha}) d\tilde{N}(t); \\
\dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) =& \int_0^{t^*} \sum_{l=1}^{L} \tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \Big\{ \boldsymbol{z}_l - \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} dN(t) \\
&- \int_0^{t^*} \sum_{l=1}^{L} \tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) d\tilde{N}(t); \\
\dot{\ell}_{\Lambda}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)[h] =& \int_0^{t^*} \Big[\sum_{l=1}^{L} \tau_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \Big\{ - \int_0^t h(s) e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} + h(t) \Big] dN(t) \\
&+ \int_0^{t^*} \Big[\sum_{l=1}^{L} \tilde{\tau}_l(t;\boldsymbol{O},\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda) \Big\{ - \int_0^t h(s) e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}} d\Lambda(s) \Big\} \Big] d\tilde{N}(t).
\end{aligned}$$

Then there exists $\delta > 0$ such that the class of functions

$$\Big\{ \dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda), \dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda), \dot{\ell}_{\Lambda}(\boldsymbol{\alpha},\boldsymbol{\gamma},\Lambda)[h] : \\ ||\boldsymbol{\alpha} - \boldsymbol{\alpha}_0|| + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| + \sup_{t \in [0,t^*]} |\Lambda(t) - \Lambda_0(t)| < \delta, h \in \mathcal{W} \Big\}$$

is Donsker. Define $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. Then by consistency of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda})$, the continuity of the score functions and the dominated convergence theorem,

$$\sup_{\boldsymbol{u},\boldsymbol{v},h} \Big| \mathbb{G}_n\{\boldsymbol{u}^T \dot{\ell}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda}) + \boldsymbol{v}^T \dot{\ell}_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda}) + \dot{\ell}_{\Lambda}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\Lambda})[h]\} \\ - \mathbb{G}_n\{\boldsymbol{u}^T \dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \boldsymbol{v}^T \dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \dot{\ell}_{\Lambda}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)[h]\} \Big| \to 0.$$

The next step is to show that the map $W : \ell^\infty(\mathcal{U}, \mathcal{V}, \mathcal{W}) \to \ell^\infty(\mathcal{U}, \mathcal{V}, \mathcal{W})$ defined by

$$W(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)[\boldsymbol{u}, \boldsymbol{v}, h] = P\{\boldsymbol{u}^T \dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) + \boldsymbol{v}^T \dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) + \dot{\ell}_{\Lambda}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)[h]\}$$

is Fréchet-differentiable at $(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)$ with a derivative $V(\boldsymbol{u}, \boldsymbol{v}, h)$ that has a continuous inverse. By direct calculation, we can show that

$$\frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} W(\boldsymbol{\alpha}_0 + \epsilon\tilde{\boldsymbol{u}}, \boldsymbol{\gamma}_0 + \epsilon\tilde{\boldsymbol{v}}, \Lambda_0 + \epsilon \int \tilde{h} d\Lambda_0)[\boldsymbol{u}, \boldsymbol{v}, h] \\ = \tilde{\boldsymbol{u}}^T \boldsymbol{B}_{\boldsymbol{\alpha}}[\boldsymbol{u}, \boldsymbol{v}, h] + \tilde{\boldsymbol{v}}^T \boldsymbol{B}_{\boldsymbol{\gamma}}[\boldsymbol{u}, \boldsymbol{v}, h] + \int_0^{t^*} B_\Lambda[\boldsymbol{u}, \boldsymbol{v}, h]\tilde{h}(s) d\Lambda_0(s),$$

where the operator $\boldsymbol{B}[\boldsymbol{u}, \boldsymbol{v}, h] \equiv (\boldsymbol{B}_{\boldsymbol{\alpha}}, \boldsymbol{B}_{\boldsymbol{\gamma}}, B_\Lambda)[\boldsymbol{u}, \boldsymbol{v}, h]$ can be rewritten as

$$- \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \phi^*(t; \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) h(t) \end{pmatrix} \\ + \begin{pmatrix} \boldsymbol{u}^T \varphi_1(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \boldsymbol{v}^T \vartheta_1(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \int \nu_1(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) h(t) d\Lambda_0(t) + \boldsymbol{u} \\ \boldsymbol{u}^T \varphi_2(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \boldsymbol{v}^T \vartheta_2(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \int \nu_2(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) h(t) d\Lambda_0(t) + \boldsymbol{v} \\ \boldsymbol{u}^T \varphi_3(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \boldsymbol{v}^T \vartheta_3(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \int \nu_3(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) h(t) d\Lambda_0(t) \end{pmatrix}. \tag{3.7.2}$$

Detailed calculations for $\boldsymbol{B}_{\boldsymbol{\alpha}}, \boldsymbol{B}_{\boldsymbol{\gamma}}$ and B_Λ can be found in the next subsection. We need to show that the operator $\boldsymbol{B}$ is invertible on its range.

By definition of $\phi^*(t; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)$, it is clear that $\phi^*(t; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda) > 0$ for any choice of $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda)$. Thus, the first term in (3.7.2) is an invertible operator. In addition, by conditions (C2) and (C3) the second term is a compact operator. Then we can show $\boldsymbol{B}$ is invertible by showing $\boldsymbol{B}$ is one-to-one. That is, if $\boldsymbol{B}(\boldsymbol{u}, \boldsymbol{v}, h) = \boldsymbol{0}$ then $(\boldsymbol{u}, \boldsymbol{v}, h) = \boldsymbol{0}$. Now assuming that

$\boldsymbol{B}(\boldsymbol{u}, \boldsymbol{v}, h) = \mathbf{0}$ for some $(\boldsymbol{u}, \boldsymbol{v}, h) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}$, it follows that

$$\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} W(\boldsymbol{\alpha}_0 + \epsilon\tilde{\boldsymbol{u}}, \boldsymbol{\gamma}_0 + \epsilon\tilde{\boldsymbol{v}}, \Lambda_0 + \epsilon \int \tilde{h} d\Lambda_0)[\boldsymbol{u}, \boldsymbol{v}, h] = 0,$$

which further indicates that the score function across the path $(\boldsymbol{\alpha}_0 + \epsilon\tilde{\boldsymbol{u}}, \boldsymbol{\gamma}_0 + \epsilon\tilde{\boldsymbol{v}}, \Lambda_0 + \epsilon \int \tilde{h} d\Lambda_0)$ is zero. That is, with probability one

$$\boldsymbol{u}^T \dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \boldsymbol{v}^T \dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \dot{\ell}_{\Lambda}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)[h] = 0.$$

By setting $dN(t) = 1$, we have for arbitrary t

$$\sum_{l=1}^{L} \tau_l(t; \boldsymbol{O}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) \Bigg\{ \boldsymbol{u}^T \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}; \boldsymbol{\alpha}_0) + \{\boldsymbol{v}^T \boldsymbol{z}_l + h(t)\} - \left(\int_0^t \{\boldsymbol{v}^T \boldsymbol{z}_l + h(t)\} e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}_0} d\Lambda_0(s) \right) \Bigg\} = 0.$$

The above equation holds only when $\boldsymbol{u} = 0$ and $\boldsymbol{v}^T \boldsymbol{z}_l + h(t) = 0$. Since $\boldsymbol{v}^T \boldsymbol{z}_l$ is a constant and $h(\cdot)$ is an arbitrary function in $\mathcal{W}$, the only solution which satisfies $\boldsymbol{v}^T \boldsymbol{z}_l + h(t) = 0$ for arbitrary t is $\boldsymbol{v} = \mathbf{0}$ and $h(\cdot) = 0$. Thus, $\boldsymbol{B}$ is one-to-one and consequently invertible, such that the derivative of W is also invertible. Now let $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{h}) = \boldsymbol{B}^{-1}(\boldsymbol{u}, \boldsymbol{v}, h)$ for some $(\boldsymbol{u}, \boldsymbol{v}, h) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}$, then it follows by Theorem 19.26 from Van der Vaart (2000) that uniformly in $(\boldsymbol{u}, \boldsymbol{v}, h)$,

$$\begin{aligned} &\sqrt{n}\{\boldsymbol{u}^T(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \boldsymbol{v}^T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \int_0^{t^*} h d(\hat{\Lambda} - \Lambda_0)\} \\ &\quad = -\mathbb{G}_n\{\tilde{\boldsymbol{u}}^T \dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \tilde{\boldsymbol{v}}^T \dot{\ell}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0) + \dot{\ell}_{\Lambda}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)[\tilde{h}]\} + o_p(1). \end{aligned}$$

Thus, $\sqrt{n}\{\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0, \hat{\Lambda} - \Lambda_0\}$ is asymptotically Gaussian. By similar semiparametric efficiency arguments (Bickel et al., 1993) as also used in (Zeng and Lin, 2006) and (Mao and Lin, 2017), the estimators $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ for the parametric component of the model are asymptotically semiparametric efficient.

□

3.7.3 Analytical variance estimator

By similar arguments as in Zeng and Lin (2006), a consistent variance estimator for $\sqrt{n}\{\boldsymbol{u}^T(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+\boldsymbol{v}^T(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)+\int_0^{t^*} hd(\hat{\Lambda}-\Lambda_0)\}$ can be constructed as

$$\hat{\boldsymbol{V}} = (\boldsymbol{u}^T, \boldsymbol{v}^T, \boldsymbol{H}^T)\hat{\mathcal{I}}_n^{-1} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{H} \end{pmatrix},$$

where $n\hat{\mathcal{I}}_n$ is the empirical information matrix of the observed-data log-likelihood $\ell(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \Lambda; \boldsymbol{O})$, which treats $\Lambda(\cdot)$ as a piecewise constant function, and $\boldsymbol{H}$ is a vector of length m with jth component equal to $h(t_j)$. Then it is straightforward to obtain variance estimations for $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ and $\hat{\Lambda}(\cdot)$ with appropriate choices of $\boldsymbol{u}$, $\boldsymbol{v}$ and h. Since $\Lambda(t)$ is positive, the 95% confidence interval is constructed by log-transformation

$$\left(\hat{\Lambda}(t)\exp\left\{\frac{-1.96\hat{SE}\{\hat{\Lambda}(t)\}}{\hat{\Lambda}(t)}\right\}, \hat{\Lambda}(t)\exp\left\{\frac{1.96\hat{SE}\{\hat{\Lambda}(t)\}}{\hat{\Lambda}(t)}\right\}\right),$$

where $\hat{SE}\{\hat{\Lambda}(t)\}$ is the estimated standard error of $\hat{\Lambda}(t)$.

Calculation related to the proof of Theorem 3.3.2

By direct calculation, we can show that

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=0} & W(\boldsymbol{\alpha}_0+\epsilon\tilde{\boldsymbol{u}}, \boldsymbol{\gamma}_0+\epsilon\tilde{\boldsymbol{v}}, \Lambda_0+\epsilon\int\tilde{h}d\Lambda_0)[\boldsymbol{u}, \boldsymbol{v}, h] \\ &= \tilde{\boldsymbol{u}}^T\boldsymbol{B}_{\boldsymbol{\alpha}}[\boldsymbol{u}, \boldsymbol{v}, h] + \tilde{\boldsymbol{v}}^T\boldsymbol{B}_{\boldsymbol{\gamma}}[\boldsymbol{u}, \boldsymbol{v}, h] + \int_0^s B_\Lambda[\boldsymbol{u}, \boldsymbol{v}, h]\tilde{h}(s)d\Lambda_0(s),\end{aligned}$$

where the operator $\boldsymbol{B}[\boldsymbol{u},\boldsymbol{v},h]=(\boldsymbol{B}_{\boldsymbol{\alpha}},\boldsymbol{B}_{\boldsymbol{\gamma}},B_\Lambda)[\boldsymbol{u},\boldsymbol{v},h]$ satisfies

$$
\begin{aligned}
\boldsymbol{B}_{\boldsymbol{\alpha}}[\boldsymbol{u},\boldsymbol{v},h] =& P\Big[\int_0^{t^*}\Big\{\sum_{l=1}^{L}\tau_{l0}(t)\frac{\partial^2}{\partial\boldsymbol{\alpha}^2}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)\boldsymbol{u}\\
&\quad+\boldsymbol{B}_{\boldsymbol{\alpha},l}(t)\Big(\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)^T\boldsymbol{u}+\boldsymbol{z}_l^T\boldsymbol{v}-\int_0^t\boldsymbol{z}_l^Te^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\boldsymbol{v}\Big)\Big\}dN(t)\\
&+\int_0^{t^*}\Big\{\sum_{l=1}^{L}\tilde{\tau}_{l0}(t)\frac{\partial^2}{\partial\boldsymbol{\alpha}^2}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)\boldsymbol{u}\\
&\quad+\tilde{\boldsymbol{B}}_{\boldsymbol{\alpha},l}(t)\Big(\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)^T\boldsymbol{u}-\int_0^t\boldsymbol{z}_l^Te^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\boldsymbol{v}\Big)\Big\}d\tilde{N}(t)\\
&+\sum_{l=1}^{L}\int_0^{t^*}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}\Big(-\int_s^{t^*}\boldsymbol{B}_{\boldsymbol{\alpha},l}(t)dN(t)-\int_s^{t^*}\tilde{\boldsymbol{B}}_{\boldsymbol{\alpha},l}(t)d\tilde{N}(t)\Big)h(s)d\Lambda_0(s)\Big];
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{B}_{\boldsymbol{\gamma}}[\boldsymbol{u},\boldsymbol{v},h] =& P\Big[\int_0^{t^*}\Big\{\sum_{l=1}^{L}\tau_{l0}(t)\Big(-\int_0^t\boldsymbol{z}_l^{\otimes 2}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\Big)\boldsymbol{v}\\
&\quad+\boldsymbol{B}_{\boldsymbol{\gamma},l}(t)\Big(\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)^T\boldsymbol{u}+\boldsymbol{z}_l^T\boldsymbol{v}-\int_0^t\boldsymbol{z}_l^Te^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\boldsymbol{v}\Big)\Big\}dN(t)\\
&+\int_0^{t^*}\Big\{\sum_{l=1}^{L}\tilde{\tau}_{l0}(t)\Big(-\int_0^t\boldsymbol{z}_l^{\otimes 2}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\Big)\boldsymbol{v}\\
&\quad+\tilde{\boldsymbol{B}}_{\boldsymbol{\gamma},l}(t)\Big(\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)^T\boldsymbol{u}-\int_0^t\boldsymbol{z}_l^Te^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(s)\boldsymbol{v}\Big)\Big\}d\tilde{N}(t)\\
&-\sum_{l=1}^{L}\int_0^{t^*}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}\Big(\int_s^{t^*}\{\boldsymbol{z}_l\tau_{l0}(t)+\boldsymbol{B}_{\boldsymbol{\gamma},l}(t)\}dN(t)\\
&\qquad+\int_s^{t^*}\{\boldsymbol{z}_l\tilde{\tau}_{l0}(t)+\tilde{\boldsymbol{B}}_{\boldsymbol{\gamma},l}(t)\}d\tilde{N}(t)\Big)h(s)d\Lambda_0(s)\Big];
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{B}_{\Lambda}[\boldsymbol{u},\boldsymbol{v},h] =& P\Big[\sum_{l=1}^{L}\int_0^{t^*}\Big\{\{B_{\Lambda,l}(s,1)+\tilde{B}_{\Lambda,l}(s,1)\}\frac{\partial}{\partial\boldsymbol{\alpha}}\log p_l(\boldsymbol{x};\boldsymbol{\alpha}_0)^T\boldsymbol{u}\\
&+B_{\Lambda,l}\{s,\boldsymbol{z}_l^T\boldsymbol{v}-\int_0^t\boldsymbol{z}_l^T\boldsymbol{v}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(\dot{t})\}+\tilde{B}_{\Lambda,l}\{s,-\int_0^t\boldsymbol{z}_l^T\boldsymbol{v}e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(\dot{t})\}\\
&+B_{\Lambda,l}\{s,-\int_0^t h(\dot{t})e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(\dot{t})\}+\tilde{B}_{\Lambda,l}\{s,-\int_0^t h(\dot{t})e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}d\Lambda_0(\dot{t})\}\\
&+h(s)e^{\boldsymbol{z}_l^T\boldsymbol{\gamma}_0}\Big(\int_s^{t^*}\tau_{l0}(t)dN(t)+\tilde{\tau}_{l0}(t)d\tilde{N}(t)\Big)\Big\}\tilde{h}(s)d\Lambda_0(s)\Big],
\end{aligned}
$$

where $\tau_{l0}(t) = \tau_l(t; \boldsymbol{O}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)$, $\tilde{\tau}_{l0}(t) = \tilde{\tau}_l(t; \boldsymbol{O}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \Lambda_0)$ and

$$\boldsymbol{B}_{\boldsymbol{\alpha},l}(t) = \tau_{l0}(t)\frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}; \boldsymbol{\alpha}_0) - \tau_{l0}(t) \sum_{d=1}^{L} \tau_{d0}(t) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_d(\boldsymbol{x}; \boldsymbol{\alpha}_0);$$

$$\tilde{\boldsymbol{B}}_{\boldsymbol{\alpha},l}(t) = \tilde{\tau}_{l0}(t)\frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}; \boldsymbol{\alpha}_0) - \tilde{\tau}_{l0}(t) \sum_{d=1}^{L} \tilde{\tau}_{d0}(t) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_d(\boldsymbol{x}; \boldsymbol{\alpha}_0);$$

$$\boldsymbol{B}_{\boldsymbol{\gamma},l}(t) = \tau_{l0}(t)\Big\{ \boldsymbol{z}_l - \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}_0} d\Lambda_0(s) \Big\} - \tau_{l0}(t) \sum_{d=1}^{L} \tau_{d0}(t) \Big\{ \boldsymbol{z}_d - \int_0^t \boldsymbol{z}_d e^{\boldsymbol{z}_d^T \boldsymbol{\gamma}_0} d\Lambda_0(s) \Big\};$$

$$\tilde{\boldsymbol{B}}_{\boldsymbol{\gamma},l}(t) = \tilde{\tau}_{l0}(t)\Big\{ - \int_0^t \boldsymbol{z}_l e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}_0} d\Lambda_0(s) \Big\} - \tilde{\tau}_{l0}(t) \sum_{d=1}^{L} \tilde{\tau}_{d0}(t) \Big\{ - \int_0^t \boldsymbol{z}_d e^{\boldsymbol{z}_d^T \boldsymbol{\gamma}_0} d\Lambda_0(s) \Big\};$$

$$\boldsymbol{B}_{\Lambda,l}(s, g(\dot{t})) = e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}_0} \Big\{ \int_s^{t^*} g(\dot{t}) \tau_{l0}(\dot{t}) dN(\dot{t}) - \sum_{d=1}^{L} \int_s^{t^*} g(\dot{t}) \tau_{l0}(\dot{t}) \tau_{d0}(\dot{t}) dN(\dot{t}) \Big\};$$

$$\tilde{\boldsymbol{B}}_{\Lambda,l}(s, g(\dot{t})) = e^{\boldsymbol{z}_l^T \boldsymbol{\gamma}_0} \Big\{ \int_s^{t^*} g(\dot{t}) \tilde{\tau}_{l0}(\dot{t}) dN(\dot{t}) - \sum_{d=1}^{L} \int_s^{t^*} g(\dot{t}) \tilde{\tau}_{l0}(\dot{t}) \tilde{\tau}_{d0}(\dot{t}) dN(\dot{t}) \Big\}.$$

Variance estimator components

$$\hat{B}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \sum_{i=1}^{n} \sum_{l=1}^{L} \Big[\hat{E}(\xi_{il}) \Big\{ \frac{\partial^2}{\partial \boldsymbol{\alpha}^2} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) + \Big(\frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \Big)^{\otimes 2} - \sum_{d=1}^{L} \hat{E}(\xi_{id}) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_d(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}})^T \Big\} \Big];$$

$$\hat{B}_{\boldsymbol{\alpha}\boldsymbol{\gamma}} = \sum_{i=1}^{n} \sum_{l=1}^{L} \Big[\hat{E}(\xi_{il}) \Big\{ \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \Big(\boldsymbol{z}_{il}^T \Delta_i - \boldsymbol{z}_{il}^T e^{\boldsymbol{z}_{il}^T \hat{\boldsymbol{\gamma}}} \hat{\Lambda}(\tilde{T}_i) \Big) - \sum_{d=1}^{L} \hat{E}(\xi_{id}) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \Big(\boldsymbol{z}_{id}^T \Delta_i - \boldsymbol{z}_{id}^T e^{\boldsymbol{z}_{id}^T \hat{\boldsymbol{\gamma}}} \hat{\Lambda}(\tilde{T}_i) \Big) \Big\} \Big];$$

$\hat{B}_{\boldsymbol{\alpha}\Lambda} = [\hat{B}_{\boldsymbol{\alpha}\Lambda_1}, \hat{B}_{\boldsymbol{\alpha}\Lambda_2}, \ldots, \hat{B}_{\boldsymbol{\alpha}\Lambda_m}]$, where

$$\hat{B}_{\boldsymbol{\alpha}\Lambda_j} = \sum_{i:\tilde{T}_i \geq t_j} \sum_{l=1}^{L} \Big[\hat{E}(\xi_{il}) \Big\{ \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \Big\{ \frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T \hat{\boldsymbol{\gamma}}} \Big\} - \sum_{d=1}^{L} \hat{E}(\xi_{id}) \frac{\partial}{\partial \boldsymbol{\alpha}} \log p_l(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \Big\{ \frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T \hat{\boldsymbol{\gamma}}} \Big\} \Big\} \Big], j = 1, \ldots, m;$$

$$\hat{B}_{\boldsymbol{\gamma\gamma}} = \sum_{i=1}^{n}\sum_{l=1}^{L}\Bigg[\hat{E}(\xi_{il})\Bigg\{ -\boldsymbol{z}_{il}^{\otimes 2}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i) + \Big(\boldsymbol{z}_{il}\Delta_i - \boldsymbol{z}_{il}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i)\Big)^{\otimes 2} - \sum_{d=1}^{L}\hat{E}(\xi_{id})\Big(\boldsymbol{z}_{il}\Delta_i - \boldsymbol{z}_{il}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i)\Big)\Big(\boldsymbol{z}_{id}^T\Delta_i - \boldsymbol{z}_{id}^Te^{\boldsymbol{z}_{id}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i)\Big)\Bigg\}\Bigg];$$

$\hat{B}_{\boldsymbol{\gamma\Lambda}} = [\hat{B}_{\boldsymbol{\gamma}\Lambda_1}, \hat{B}_{\boldsymbol{\gamma}\Lambda_2}, \ldots, \hat{B}_{\boldsymbol{\gamma}\Lambda_m}]$, where

$$\hat{B}_{\boldsymbol{\gamma}\Lambda_j} = \sum_{i:\tilde{T}_i\geq t_j}\sum_{l=1}^{L}\Bigg[\hat{E}(\xi_{il})\Bigg\{ -\boldsymbol{z}_{il}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}} + \Big(\boldsymbol{z}_{il}\Delta_i - \boldsymbol{z}_{il}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i)\Big)\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\} - \sum_{d=1}^{L}\hat{E}(\xi_{id})\Big(\boldsymbol{z}_{il}\Delta_i - \boldsymbol{z}_{il}e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\hat{\Lambda}(\tilde{T}_i)\Big)\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\}\Bigg\}\Bigg], j = 1, \ldots, m;$$

$\hat{B}_{\Lambda\Lambda}$ is a $m \times m$ matrix with diagonal entries

$$\hat{B}_{\Lambda_j\Lambda_j} = -\frac{1}{d_j^2} + \sum_{i:\tilde{T}_i\geq t_j}\sum_{l=1}^{L}\Bigg[\hat{E}(\xi_{il})\Bigg\{\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\}^2 - \sum_{d=1}^{L}\hat{E}(\xi_{id})\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\}\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{id}^T\hat{\boldsymbol{\gamma}}}\Big\}\Bigg\}\Bigg], j = 1, \ldots, m$$

and off-diagonal entries

$$\hat{B}_{\Lambda_j\Lambda_k} = \sum_{i:\tilde{T}_i\geq \max\{t_j, t_k\}}\sum_{l=1}^{L}\Bigg[\hat{E}(\xi_{il})\Bigg\{\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\}^2 - \sum_{d=1}^{L}\hat{E}(\xi_{id})\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{il}^T\hat{\boldsymbol{\gamma}}}\Big\}\Big\{\frac{I(\tilde{T}_i = t_j, \Delta_i = 1)}{d_i} - e^{\boldsymbol{z}_{id}^T\hat{\boldsymbol{\gamma}}}\Big\}\Bigg\}\Bigg], j = 1, \ldots, m.$$

3.7.4 Additional tables and figures for simulation results

Table 3.6: Simulation results for the simulation scenarios (I) - (IV) out of 10000 simulations with sample size $n = 1000$ and non-informative initialization. M.Bias: Median bias; SE: standard deviation; SEE: median standard error estimate; CP: coverage probability.

	Scenario (I)				Scenario (II)			
	M.Bias	SE	SEE	CP	M.Bias	SE	SEE	CP
$\hat{\alpha}_{2,0}$	0.014	0.267	0.255	0.941	0.003	0.580	0.560	0.947
$\hat{\alpha}_{2,1}$	0.007	0.201	0.192	0.943	0.000	0.503	0.477	0.945
$\hat{\alpha}_{2,2}$	-0.015	0.304	0.296	0.949	0.005	0.517	0.500	0.945
$\hat{a}_2$	0.011	0.446	0.412	0.941	0.023	0.441	0.406	0.930
$\hat{\zeta}_{11}$	-0.023	0.197	0.201	0.957	-0.045	0.319	0.313	0.956
$\hat{\zeta}_{12}$	0.009	0.256	0.258	0.953	-0.004	0.347	0.340	0.943
$\hat{\zeta}_{21}$	0.018	0.214	0.218	0.959	0.032	0.313	0.308	0.960
$\hat{\zeta}_{22}$	0.003	0.296	0.302	0.958	0.019	0.399	0.402	0.953
$\hat{\Lambda}(2)$	-0.004	0.163	0.165	0.956	-0.004	0.212	0.197	0.940
$\hat{\Lambda}(3)$	-0.008	0.352	0.344	0.951	0.010	0.540	0.503	0.952
$\hat{\Lambda}(4)$	0.034	1.146	1.051	0.944	0.151	1.606	1.345	0.950

	Scenario (III)				Scenario (IV)			
	M.Bias	SE	SEE	CP	M.Bias	SE	SEE	CP
$\hat{\alpha}_{2,0}$	0.018	0.381	0.340	0.924	0.018	0.488	0.504	0.960
$\hat{\alpha}_{2,1}$	0.021	0.258	0.234	0.936	-0.015	0.308	0.303	0.953
$\hat{\alpha}_{2,2}$	-0.035	0.417	0.380	0.936	-0.011	0.668	0.695	0.962
$\hat{a}_2$	0.010	0.729	0.618	0.918	0.003	0.311	0.309	0.952
$\hat{\zeta}_{11}$	-0.069	0.429	0.404	0.963	-0.013	0.204	0.206	0.951
$\hat{\zeta}_{12}$	0.010	0.542	0.534	0.950	-0.023	0.249	0.251	0.953
$\hat{\zeta}_{21}$	0.061	0.431	0.409	0.960	0.012	0.357	0.344	0.939
$\hat{\zeta}_{22}$	0.022	0.547	0.545	0.949	0.038	0.332	0.336	0.955
$\hat{\Lambda}(2)$	0.006	0.302	0.282	0.935	0.003	0.190	0.179	0.946
$\hat{\Lambda}(3)$	0.000	0.772	0.675	0.940	0.018	0.488	0.450	0.941
$\hat{\Lambda}(4)$	0.089	3.440	2.291	0.941	0.096	1.351	1.217	0.939

Table 3.7: Simulation results for scenario (V) out of 10000 simulations with different choices of sample size. M.Bias: Median bias; SE: standard deviation; SEE: median standard error estimate; CP: coverage probability.

	Scenario (IV), $n = 1000$				Scenario (IV), $n = 2000$				Scenario (IV), $n = 3000$			
	M.Bias	SE	SEE	CP	M.Bias	SE	SEE	CP	M.Bias	SE	SEE	CP
$\hat{\alpha}_{2,0}$	0.053	0.725	0.623	0.890	0.032	0.476	0.458	0.935	-0.001	0.383	0.376	0.940
$\hat{\alpha}_{2,1}$	-0.011	0.501	0.403	0.909	-0.006	0.303	0.284	0.942	0.005	0.239	0.230	0.943
$\hat{\alpha}_{2,2}$	0.013	0.606	0.532	0.908	0.012	0.397	0.381	0.936	0.020	0.314	0.312	0.944
$\hat{\alpha}_{3,0}$	0.058	0.449	0.359	0.889	0.042	0.291	0.264	0.927	0.021	0.235	0.217	0.928
$\hat{\alpha}_{3,1}$	-0.027	0.334	0.255	0.902	-0.024	0.204	0.182	0.930	-0.014	0.162	0.148	0.931
$\hat{\alpha}_{3,2}$	-0.011	0.386	0.346	0.922	-0.018	0.257	0.246	0.940	-0.002	0.208	0.200	0.937
$\hat{a}_2$	-0.868	1.229	0.783	0.675	-0.515	0.884	0.653	0.789	-0.346	0.679	0.562	0.863
$\hat{a}_3$	-0.656	1.008	0.717	0.711	-0.445	0.696	0.611	0.799	-0.309	0.550	0.533	0.857
$\hat{\zeta}_{11}$	-0.063	0.256	0.229	0.941	-0.036	0.157	0.152	0.947	-0.024	0.122	0.120	0.952
$\hat{\zeta}_{12}$	-0.108	0.385	0.355	0.936	-0.058	0.240	0.235	0.947	-0.036	0.187	0.187	0.950
$\hat{\zeta}_{21}$	0.228	0.472	0.353	0.829	0.142	0.314	0.247	0.835	0.093	0.246	0.203	0.866
$\hat{\zeta}_{22}$	0.333	0.707	0.555	0.841	0.200	0.455	0.384	0.869	0.120	0.356	0.312	0.896
$\hat{\zeta}_{31}$	0.078	0.298	0.276	0.940	0.046	0.187	0.187	0.948	0.034	0.151	0.150	0.947
$\hat{\zeta}_{32}$	0.129	0.450	0.421	0.936	0.069	0.286	0.284	0.950	0.041	0.227	0.228	0.952
$\hat{\Lambda}(2)$	0.344	0.497	0.360	0.673	0.179	0.336	0.255	0.762	0.115	0.245	0.203	0.803
$\hat{\Lambda}(3)$	0.196	0.643	0.566	0.904	0.060	0.414	0.410	0.935	0.019	0.322	0.328	0.946
$\hat{\Lambda}(4)$	0.278	1.399	1.334	0.961	0.135	0.843	0.888	0.969	0.080	0.655	0.694	0.972

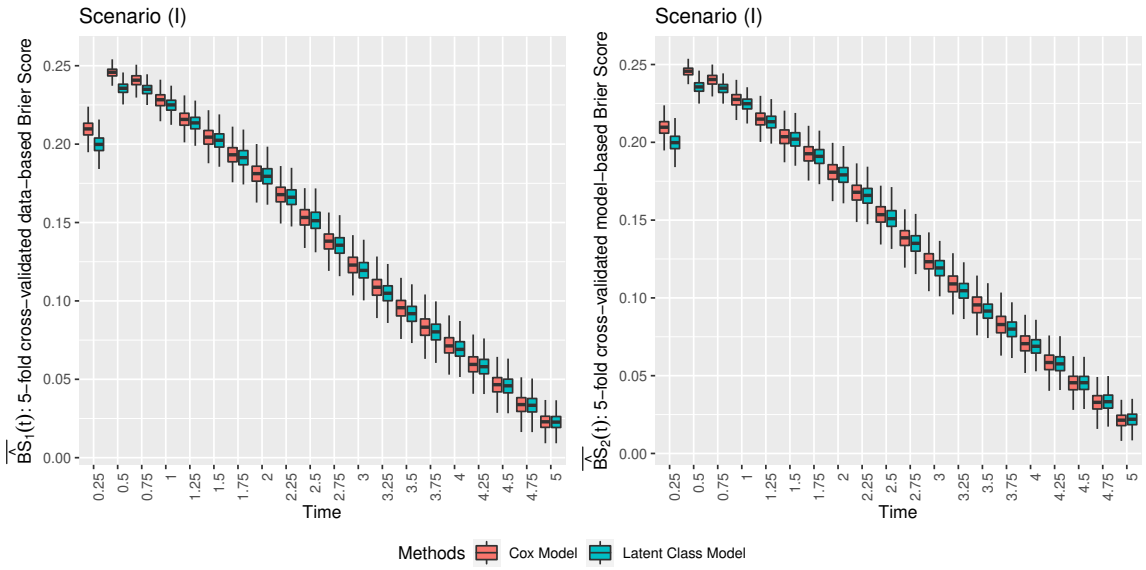

Figure 3.5: Boxplots for average cross-validated Brier Score $\overline{\hat{\mathrm{BS}}_1}(t)$ and $\overline{\hat{\mathrm{BS}}_2}(t)$, $t \in (0, 5]$, from 1000 simulations under scenario (I) with sample size 1000, for the Cox model and the proposed latent class model with $L = 2$.

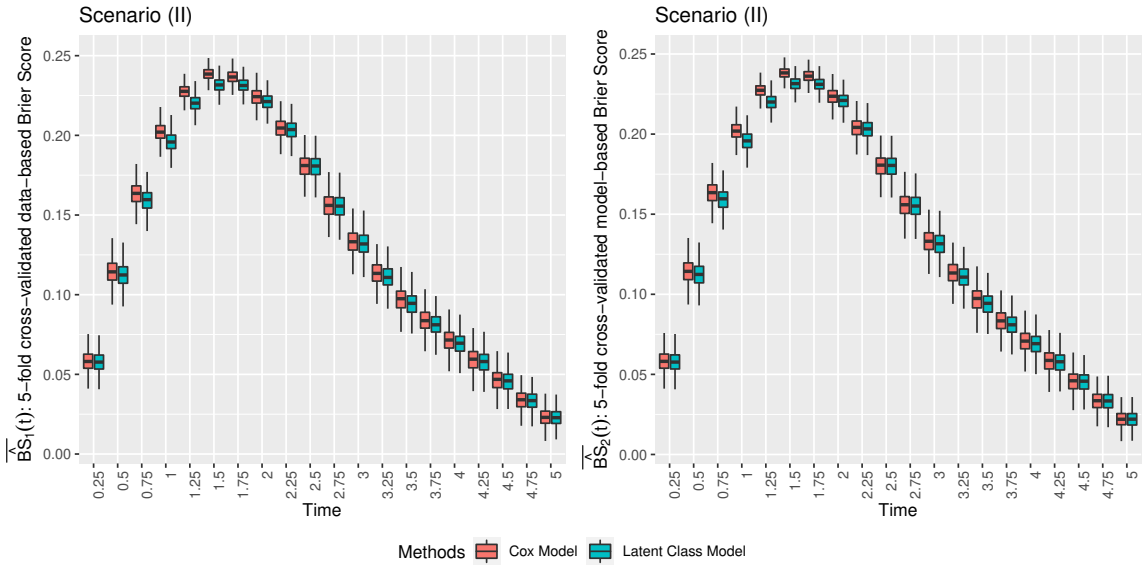

Figure 3.6: Boxplots for average cross-validated Brier Score $\overline{\hat{\mathrm{BS}}_1}(t)$ and $\overline{\hat{\mathrm{BS}}_2}(t)$, $t \in (0, 5]$, from 1000 simulations under scenario (II) with sample size 1000, for the Cox model and the proposed latent class model with $L = 2$.

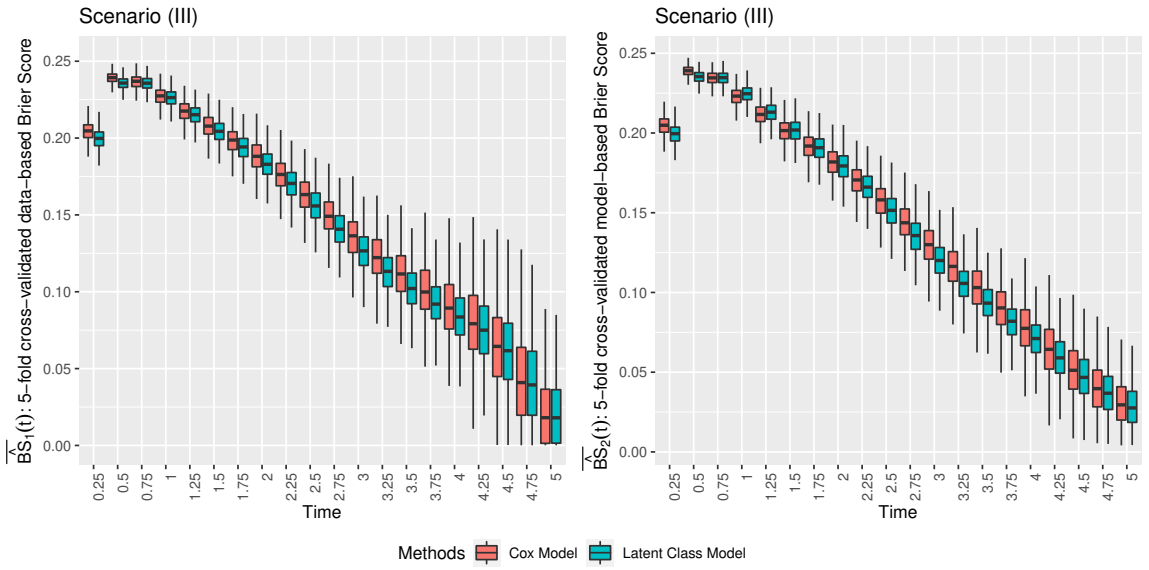

Figure 3.7: Boxplots for average cross-validated Brier Score $\overline{\hat{\mathrm{BS}}_1}(t)$ and $\overline{\hat{\mathrm{BS}}_2}(t)$, $t \in (0, 5]$, from 1000 simulations under scenario (III) with sample size 1000, for the Cox model and the proposed latent class model with $L = 2$.

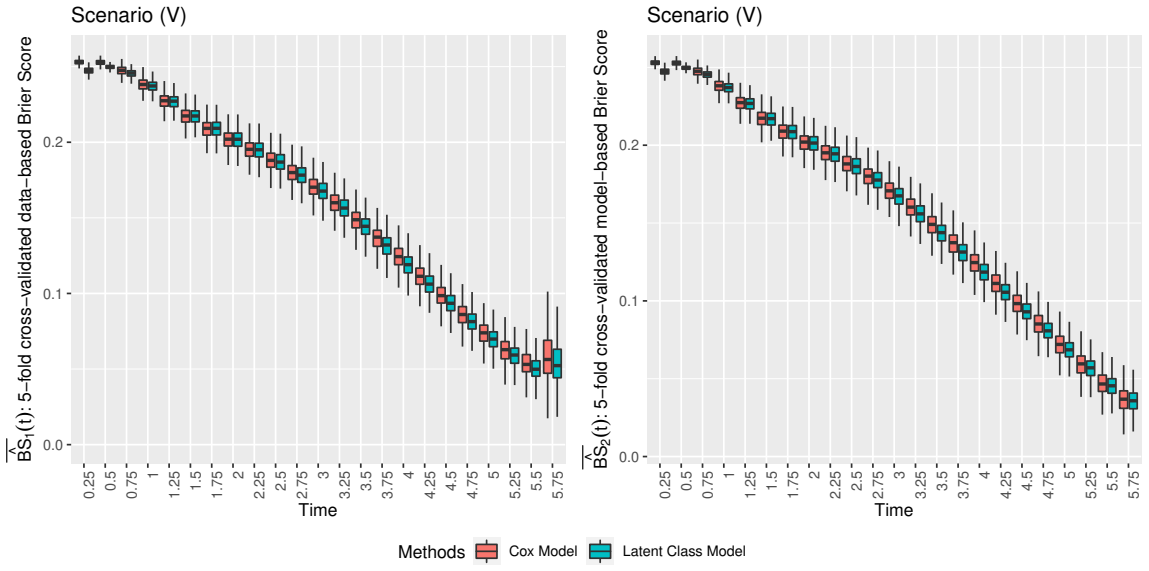

Figure 3.8: Boxplots for average cross-validated Brier Score $\overline{\hat{\mathrm{BS}}_1}(t)$ and $\overline{\hat{\mathrm{BS}}_2}(t)$, $t \in (0, 5]$, from 1000 simulations under scenario (V) with sample size 1000, for the Cox model and the proposed latent class model with $L = 3$.

Chapter 4

Semi-parametric Latent Class Analysis for Joint Longitudinal and Survival Data

4.1 Introduction

For diseases with heterogeneous subpopulations, such as mild cognitive impairment (MCI), it is of critical clinical interest to investigate disease subtypes underlying patients' longitudinal trajectories and time-to-event, such that the etiologies and phenotypes associated with different subtypes can be better understood and utilized for diagnosis and prediction. Data from large scale longitudinal studies create an ideal platform for the subtype investigation. Since 2005, the National Alzheimer's Coordinating Center (NACC) has been conducting annual neurological examinations for study participants with MCI (Weintraub et al., 2009; Beekly et al., 2007), and recording milestone events such as death and drop-out. Up to June 2015, the corresponding Uniform Data Set (UDS) of NACC has collected longitudinal and time-to-event information for more than 5000 participants, which contains rich information for a cohort of MCI patients. In particular, the UDS records patients' progressive history in multiple cognitive, functional and behavioral domains, which is highly relevant

to the etiologies of neurodegeneration and the corresponding subtypes. In addition, the observed time-to-death can be regarded as a consequence of disease progression reflected by the patients' longitudinal history, which may also provide valuable information about the disease heterogeneity. Therefore, it is appealing to jointly model the longitudinal and survival data, such that available information can be maximally utilized in the investigation of disease subtypes.

One common approach to conduct jointly modeling of longitudinal and survival data, while considering the heterogeneity of disease population, is the so-called shared random-effect model (Henderson et al., 2000; Xu and Zeger, 2001, for example). This model framework consists of a submodel for time-to-event and a submodel for longitudinal data, where the two submodels are linked by a shared random effect, which introduces a latent structure to capture the dependency structure between longitudinal and survival outcomes. One special case is that the longitudinal trajectories are regarded as time-dependent covariates of the survival submodel, such that different trajectory history will contribute to different survival probability curves. While the shared random-effect model accounts for the heterogeneity via a latent structure, it does not clearly define clinically interpretable disease subtypes. In addition, the effect size of covariates is assumed to be constant across the whole population, which may not fully capture the heterogeneity in terms of different effect sizes for different subpopulations. As mentioned in Proust-Lima et al. (2014), moreover, fitting the shared random-effect model can be computationally challenging due to the involvement of random effect terms and the implementation of numerical integration.

Alternatively, finite mixture structure (McLachlan and Peel, 2000) has been studied to construct latent class models of longitudinal and survival data (Lin et al., 2002; Larsen, 2004; Proust-Lima et al., 2017, for example). Different from the shared random-effect model, latent class models define a finite number of homogeneous classes, and assume that the longitudinal trajectories and time-to-event are independent of each other within a given class. Such conditional independence assumption usually leads to separate estimation procedure for the survival and longitudinal submodels in the M-step of expectation-maximization (EM) algorithm, which largely reduces the computational burden. Coincidentally, this as-

sumption has been widely adapted in the vast majority of existing latent class methods (Lin et al., 2002; Larsen, 2004; Proust-Lima et al., 2017, for example). While the conditional independence assumption brings computational convenience and becomes popular in the existing methods, the assumption seems less appropriate for the realistic setting of UDS. First, the longitudinal features collected in the UDS data are direct indicators of the participants' disease progression, which is likely to be correlated with time-to-death even within a homogeneous latent class. Furthermore, the longitudinal features can be censored by the terminal event of death. Given that the features and time-to-death are correlated, it is necessary to account for the informative censoring of longitudinal features by a likelihood structure similar to the shared random-effect model, or using inverse probability weighting (IPW) technique (Lin et al., 2004, for example) for the estimating equation of the longitudinal submodel.

In terms of distribution assumptions, both shared random-effect models and joint latent class models largely rely on parametric modeling and require specifying the distributions for longitudinal trajectories and time-to-event outcome. Although parametric modeling enables explicit forms of likelihood function, it also brings risks of model misspecification as the underlying true distribution is usually unknown in real applications. Research has been conducted to partially extend parametric modeling to semi-parametric modeling, aiming to relax restrictive assumptions. For example, Larsen (2004) proposed semi-parametric Cox model for the survival submodel of the finite mixture framework, while Hart et al. (2020) introduced semi-parametric generalized estimating equation (GEE) for the longitudinal submodel. However, it is still a challenging task to utilize semi-parametric modeling for both longitudinal and survival components in a concurrent manner.

To better identify MCI subtypes in the sense of longitudinal trajectories and associated distribution of time-to-death, we study a novel semi-parametric latent class framework for joint longitudinal and survival data. Our method absorbs the nice features of shared random-effect models and latent class models to enhance the estimating procedure with careful considerations of real world challenges reflected by the UDS data. Critically, the within-class dependency of longitudinal and survival outcomes is naturally considered by

incorporating longitudinal outcomes as time-dependent covariates in the survival submodel. In addition, we utilize IPW technique to account for the informative censoring of longitudinal observations due to death. Furthermore, the proposed method utilizes GEE submodel for longitudinal trajectory and semi-parametric proportional hazards submodel for time-to-event, which is highly flexible in handling non-Gaussian trajectory patterns and various survival distributions.

One major challenge for the proposed method is to appropriately handle the longitudinal outcomes as time-dependent covariates for the survival submodel. Since the observation of a patient's neurological examination scores implies that patient's survival beyond time of observation, the time-dependent covariates are "internal" as defined by Kalbfleisch and Prentice (2011). We apply sequential ignorability assumption similar to that used in Lin et al. (2004) to guarantee the interpretability of conditional survival probability at time of longitudinal observation with the presence of internal covariates. With our careful model specification, the class-specific proprotional hazards submodel with internal time-dependent covariates is still easy to fit using standard software. For the estimation of class-specific GEE model, in addition, sensible IPW terms can be obtained, based on the survival submodel, to account for the informative censoring of longitudinal observations.

4.2 Data, notation and models

Let T and C respectively denote the time to failure and censoring. Define $\tilde{T} = \min\{T, C\}$ and $\Delta = I(T \leq C)$. Let $\boldsymbol{Y}(t) = \{Y_1(t), \ldots, Y_J(t)\}$ denote a $J \times 1$ vector of longitudinal observations observed at time t for $0 \leq t < \tilde{T}$. Let $N(t) = I(\tilde{T} \leq t, \Delta = 1)$, $N^C(t) = I(\tilde{T} \leq t, \Delta = 0)$ and $N^V(t) = \sum_{k=1}^{\infty} I(V_k \leq t)$ be counting processes for failure, censoring and number of observed visits by time t. Let $R(t) = I(\tilde{T} \geq t)$ denote a left continuous at-risk process. The observed visit times are $V_1 < \ldots < V_K < \tilde{T}$, where $V_1 = 0$ and K is the random number of visits before failure, or censoring such as loss of follow-up. Let $\boldsymbol{X}$ denote a $p \times 1$ vector of the baseline covariates which are time-independent. Then the observed

longitudinal information from a participant up to time t can be represented by

$$\bar{\boldsymbol{Y}}^{\text{obs}}(t) = \{\boldsymbol{Y}(s) : 0 \leq s \leq t, dN^V(s) = 1, N(s) = 0, N^C(s) = 0\}.$$

That is, we observe longitudinal outcomes $\boldsymbol{Y}(\cdot)$ at each visit before failure or censoring occurs (i.e. $dN^V(s) = 1$, $N(s) = 0, N^C(s) = 0$). We assume that the visit process $N^V(\cdot)$ and censoring time C are independent of $\boldsymbol{Y}(\cdot)$ given baseline covariates $\boldsymbol{X}$ and class membership $\boldsymbol{\xi}$. We also assume that $dN(t)$, $dN^C(t)$ and $dN^V(t)$ are mutually independent given information $\{\bar{\boldsymbol{Y}}^{\text{obs}}(t-), \boldsymbol{X}, \boldsymbol{\xi}\}$, prior to time t. In the whole dataset, we observe n independent and identically distributed replicates of $\{\tilde{T}, \Delta, \bar{\boldsymbol{Y}}^{\text{obs}}(V_K), \boldsymbol{X}\}$, denoted by

$$\mathcal{O} = [\boldsymbol{O}_i = \{\tilde{T}_i, \Delta_i, \bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i}), \boldsymbol{X}_i\}, i = 1, \ldots, n].$$

We assume that there are L latent classes in the population. The unobserved latent class labels are denoted by a $L \times 1$ vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)^T$, where ξ_l equals one if the observation belongs to the lth latent class and zero otherwise.

4.2.1 Latent class probability submodel

Under the setting of clinical research, patients with different characteristics may have varied probabilities of belonging to certain latent disease subgroups, or latent classes. In addition, different disease subgroups may have varied relative frequencies given the same patient characteristics. To capture the above relationships between patient characteristics and latent class membership probabilities, we utilize the latent polytomous logistic regression model (Bandeen-Roche et al., 1997)

$$\Pr(\xi_l = 1|\boldsymbol{X}) = p_l(\boldsymbol{X}; \boldsymbol{\alpha}) = \frac{\exp(\tilde{\boldsymbol{X}}^T \boldsymbol{\alpha}_l)}{\sum_{d=1}^{L} \exp(\tilde{\boldsymbol{X}}^T \boldsymbol{\alpha}_d)}, l = 1, \ldots, L, \tag{4.2.1}$$

where $\tilde{\boldsymbol{X}} = (1, \boldsymbol{X}^T)^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_L^T)^T$ are unknown regression coefficients which represents class-specific baseline covariate effects on the relative frequency of latent classes.

For identifiability consideration, we set the first class as the reference class, where $\boldsymbol{\alpha}_1 = \mathbf{0}$.

4.2.2 Class-specific generalized estimating equation submodel

In the UDS data, longitudinal patient characteristics are collected as different data types, including continuous, count, and binary variables. The variety of data types motivates us to use class-specific generalized estimating equation (GEE) to model the longitudinal trajectories. Specifically, for class $l = 1, \ldots, L$ we have the marginal model

$$E\{Y_j(t)|\boldsymbol{X}, \xi_l = 1\} = \mu_{j,l}(t, \boldsymbol{X}; \boldsymbol{\beta}_{j,l}) = h_j(\boldsymbol{X}_t^T \boldsymbol{\beta}_{j,l}),\ j = 1, \ldots, J \tag{4.2.2}$$

for the first moment of $Y_j(t)$. Here $h(\cdot)$ is a prespecified link function corresponding to the types of variable $Y_j(t)$, the covariates $\boldsymbol{X}_t = (1, t, \boldsymbol{X}^T)^T$ consist of an intercept, a linear term of time t, and baseline covariates $\boldsymbol{X}$, and $\boldsymbol{\beta}_{j,l}$ is the unknown class-specific regression parameter. To capture non-linear changes in the h_j^{-1}-transformed trajectory, it is straightforward to include a higher-order polynomial of t in the covariates $\boldsymbol{X}_t$. Correspondingly to the mean model, the second moment of $Y_j(t)$ is formulated as

$$\sigma^2_{j,l}(t, \boldsymbol{X}; \phi_{j,l}, \boldsymbol{\beta}_{j,l}) = \phi_{j,l} \mathcal{V}_j\{\mu_{j,l}(t, \boldsymbol{X}; \boldsymbol{\beta}_{j,l})\},\ j = 1, \ldots, J, \tag{4.2.3}$$

where $\mathcal{V}_j(\cdot)$ is a data type-specific link function and $\phi_{j,l}$ is a class-specific scale parameter accounting for dispersion. To capture the dependency of $Y_j(\cdot)$ at different visit times, we further propose to use autoregressive correlation structure for the covariance model with class-specific correlation parameter $\rho_{j,l}$

$$\begin{aligned} Cov\{Y_j(s), Y_j(t)|\boldsymbol{X}, \xi_l = 1\} &= \eta_{j,l}(s, t, \boldsymbol{X}; \rho_{j,l}, \phi_{j,l}, \boldsymbol{\beta}_{j,l}) \\ &= \rho_{j,l}^{|t-s|} \phi_{j,l} \mathcal{V}_j^{1/2}\{\mu_{j,l}(s, \boldsymbol{X}; \boldsymbol{\beta}_{j,l})\} \mathcal{V}_j^{1/2}\{\mu_{j,l}(t, \boldsymbol{X}; \boldsymbol{\beta}_{j,l})\},\ j = 1, \ldots, J. \end{aligned} \tag{4.2.4}$$

The key parameter of interest is the regression parameter for the mean model (4.2.2), $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{j,l}, j = 1, \ldots, J, l = 1, \ldots, L\}$, which reflects the increasing or decreasing trend of

class-specific trajectories and the potential covariate effects. We denote $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\rho}\} = \{(\boldsymbol{\beta}_{j,l}, \phi_{j,l}, \rho_{j,l}), j = 1, \ldots, J, l = 1, \ldots, L\}$ as the collection of all unknown parameters

For the observed data for the jth longitudinal variable $\bar{\boldsymbol{Y}}_j^{\text{obs}}(V_K)$ with visit times $0 = V_1 < V_2 < \ldots < V_K < \tilde{T}$, we denote the class-specific mean vector across visit times as $\boldsymbol{\mu}_{j,l}(\boldsymbol{\beta}) = \{\mu_j(V_1, \tilde{\boldsymbol{X}}, \boldsymbol{\beta}_{j,l}), \ldots, \mu_j(V_K, \tilde{\boldsymbol{X}}, \boldsymbol{\beta}_{j,l})\}$ and the class-specific $k \times k$ variance-covariance matrix as $\boldsymbol{\Sigma}_{j,l}(\boldsymbol{\theta})$, where the (v, v) entry of $\boldsymbol{\Sigma}_{j,l}(\boldsymbol{\theta})$ is $\sigma_j^2(v, \tilde{\boldsymbol{X}}; \phi_{j,l}, \boldsymbol{\beta}_{j,l})$ and the (u, v) entry is $\eta_j(u, v, \tilde{\boldsymbol{X}}; \rho_{j,l}, \phi_{j,l}, \boldsymbol{\beta}_{j,l})$. For all longitudinal observations across J outcomes, $\bar{\boldsymbol{Y}}^{\text{obs}}(V_K) = \{\bar{\boldsymbol{Y}}_1^{\text{obs}}(V_K)^T, \ldots, \bar{\boldsymbol{Y}}_J^{\text{obs}}(V_K)^T\}^T$, the class specific mean vector and the variance-covariance matrix are denoted as $\boldsymbol{\mu}_l(\boldsymbol{\beta}) = \{\boldsymbol{\mu}_{1,l}(\boldsymbol{\beta})^T, \ldots, \boldsymbol{\mu}_{J,l}(\boldsymbol{\beta})^T\}^T$ and $\boldsymbol{\Sigma}_l(\boldsymbol{\theta}) = \text{blkdiag}\{\boldsymbol{\Sigma}_{1,l}(\boldsymbol{\theta}), \ldots, \boldsymbol{\Sigma}_{J,l}(\boldsymbol{\theta})\}$, where blkdiag refers to block diagonal matrix.

4.2.3 Class-specific Cox regression submodel

We assume a proportional hazards survival submodel with longitudinal history $\bar{\boldsymbol{Y}}^{\text{obs}}(t)$ as time-dependent covariates, which naturally accounts for the dependency between the longitudinal variables and time-to-event information. Let $\boldsymbol{Z}(t) = \{\boldsymbol{X}^T, \boldsymbol{Y}(t)^T\}^T$. We adapt the sequential ignorability assumption similar to that used in Lin et al. (2004),

$$\begin{aligned} \Pr\{dN(t) = 1 | \bar{\boldsymbol{Z}}^{\text{obs}}(t), \xi_l = 1\} &= \Pr\{dN(t) = 1 | \bar{\boldsymbol{Z}}^{\text{obs}}(t-), \xi_l = 1\} \\ &= R(t)\lambda_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\}dt, l = 1, \ldots, L, \end{aligned} \tag{4.2.5}$$

where $\lambda_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\}$ is the class-specific hazard of failure at time t based on the past history $\bar{\boldsymbol{Z}}^{\text{obs}}(t-)$ prior to t. The above assumption indicates that the hazard at time t is independent of the longitudinal observation $\boldsymbol{Y}(t)$ at time t given $\bar{\boldsymbol{Z}}^{\text{obs}}(t-)$. Emprically, this assumption implies that the hazard of failure for a participant on time interval $t \in (V_j, V_{j+1}]$ only depends on the history of longitudinal observations $\bar{Y}^{\text{obs}}(V_j)$ up to time V_j. Even at $t = V_{j+1}$, the hazard at the next infinitesimal interval is still independent of $Y(V_{j+1})$. Based on the sequential ignorability assumption, we formulate our class-specific Cox regression submodel as

$$\lambda_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\} = \lambda_{0l}(t) \exp\{\boldsymbol{H}(t)^T \boldsymbol{\gamma}_l\} \tag{4.2.6}$$

with unspecified infinite-dimensional class-specific baseline hazard $\lambda_{0l}(\cdot)$, and class-specific unknown regression parameters $\boldsymbol{\gamma}_l$. Here, $\boldsymbol{H}(t)$ is a vector of prespecified function of t and $\bar{\boldsymbol{Z}}^{\text{obs}}(t-)$. For example, one special case of $\boldsymbol{H}(t)$ is the set of baseline covariates and most recent observation of $\boldsymbol{Y}(\cdot)$ prior to t, $\boldsymbol{H}(t) = Z(t-)$. Note that the time-dependent covariates $\boldsymbol{H}(t)$ in model (4.2.6) are internal covariates as defined by (Kalbfleisch and Prentice, 2011). By sequential ignorability assumption and the definition of $\boldsymbol{H}(t)$, however, it is important to note that covariate for the hazard at time t, $\boldsymbol{H}(t)$, only implies observation of $\boldsymbol{Y}(\cdot)$ prior to time t. Empirically, this means the covariate $\boldsymbol{H}(V_j)$ for hazard at the jth visit time V_j only implies patient's survival beyond V_{j-1}, which is the time of the previous visit. While the internal covariates and sequential ignorability assumption introduce some complications as described above, the corresponding density function of $(\tilde{T}, \Delta)$ is not different from a standard formulation of time-to-event density

$$f_l\{\tilde{T}, \Delta | \bar{\boldsymbol{Z}}^{\text{obs}}(V_K); \boldsymbol{\zeta}\} = [\lambda_{0l}(\tilde{T}) \exp\{\boldsymbol{H}(\tilde{T})^T \boldsymbol{\gamma}_l\}]^{\Delta} \exp\left\{ - \int_0^{\tilde{T}} \lambda_{0l}(t) \exp\{\boldsymbol{H}(t)^T \boldsymbol{\gamma}_l\} dt \right\},$$

where $\boldsymbol{\zeta} = [\{\boldsymbol{\gamma}_l^T, \Lambda_{0l}(\cdot)\}^T, l = 1, \ldots, L]$ is the vector of unknown parameters with $\Lambda_{0l}(t) = \int_0^t \lambda_{0l}(s) ds$. Due to the presence of internal covariates $\boldsymbol{H}(\cdot)$, the exponentiated term $\exp[-\int_0^{\tilde{T}} \lambda_{0l}(t) \exp\{\boldsymbol{H}(t)^T \boldsymbol{\gamma}_l\} dt]$ can no longer be interpreted as the survival probability. Instead, it can be interpreted as a product integral of $1 - \lambda_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\}$ on $t \in [0, \tilde{T}]$.

4.3 Estimation

In this section, we derive estimating equations and the corresponding estimating procedures for the unknown parameters, $\boldsymbol{\psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T, \boldsymbol{\zeta}^T)^T$, from the three submodels introduced in Section 4.2. Since we do not have an explicit form of density function for the longitudinal observation $\bar{\boldsymbol{Y}}^{\text{obs}}(V_K)$, it is not straightforward to derive an estimation procedure for $\boldsymbol{\theta}$ from an approach based on likelihood.

4.3.1 Latent class probability submodel

Let $\tau_{il}\{\boldsymbol{\psi}; \boldsymbol{O}_i\} = \Pr\{\xi_{il} = 1 | \boldsymbol{O}_i; \boldsymbol{\psi}\}$ denote the posterior class membership probability conditioned on all observed data $\boldsymbol{O}_i$ for subject i. In the subsequent illustration, we denote $\tau_{il}(\boldsymbol{\psi}; \boldsymbol{O}_i)$ in short by τ_{il}. Then the estimating equation for the unknown parameter $\boldsymbol{\alpha}$ from the class probability submodel (4.2.1) can be constructed as

$$S_1(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{l=1}^{L} \tau_{il} \frac{\partial \log p_l(\boldsymbol{X}_i; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} (\boldsymbol{\tau}_i - \boldsymbol{p}_i) \otimes \tilde{\boldsymbol{X}} = 0, \tag{4.3.1}$$

where $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iL})^T$, $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iL})^T$ and $\otimes$ denotes the Kronecker product operator. It is easy to see

$$E\{S_1(\boldsymbol{\alpha})\} = \sum_{i=1}^{n} E\left\{ E\left((\boldsymbol{\xi}_i - \boldsymbol{p}_i) \otimes \tilde{\boldsymbol{X}} \middle| \boldsymbol{O}_i; \boldsymbol{\psi} \right) \right\} = \sum_{i=1}^{n} E\left\{ E\left((\boldsymbol{\xi}_i - \boldsymbol{p}_i) \otimes \tilde{\boldsymbol{X}} \middle| \boldsymbol{X}_i; \boldsymbol{\psi} \right) \right\} = 0,$$

which indicates $S(\boldsymbol{\alpha})$ is an unbiased estimating equation. Given $\boldsymbol{\tau}$ fixed, equation (4.3.1) can be solved by standard solver for weighted multinomial logistic regression, such as `VGAM` (Yee et al., 2010).

4.3.2 Class-specific Cox regression submodel

It directly follows from (4.2.5) and (4.2.6) that

$$E\left\{ dN_i(u) - R_i(u) \lambda_{0l}(u) \exp\{\boldsymbol{H}_i(u)^T \boldsymbol{\gamma}_l\} du \middle| \bar{\boldsymbol{Z}}^{\text{obs}}(u-), \xi_{il} = 1 \right\} = 0, \ l = 1, \ldots, L, \tag{4.3.2}$$

which further implies

$$\begin{aligned} 0 &= E\left\{ \xi_{il} \left(dN_i(u) - R_i(u) \lambda_{0l}(u) \exp\{\boldsymbol{H}_i(u)^T \boldsymbol{\gamma}_l\} du \right) \middle| \bar{\boldsymbol{Z}}^{\text{obs}}(u-) \right\} \\ &= E\left[E\left\{ \xi_{il} \left(dN_i(u) - R_i(u) \lambda_{0l}(u) \exp\{\boldsymbol{H}_i(u)^T \boldsymbol{\gamma}_l\} du \right) \middle| \boldsymbol{O}_i \right\} \middle| \bar{\boldsymbol{Z}}^{\text{obs}}(u-) \right] \\ &= E\left\{ \tau_{il} \left(dN_i(u) - R_i(u) \lambda_{0l}(u) \exp\{\boldsymbol{H}_i(u)^T \boldsymbol{\gamma}_l\} du \right) \middle| \bar{\boldsymbol{Z}}^{\text{obs}}(u-) \right\}, \ l = 1, \ldots, L. \end{aligned} \tag{4.3.3}$$

Then for $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_L^T)^T$ and $\boldsymbol{\Lambda}_0(\cdot) = \{\Lambda_{01}(\cdot), \ldots, \Lambda_{0L}(\cdot)\}$, we can construct unbiased estimating equations

$$\begin{aligned} S_2(\boldsymbol{\gamma}) &= \sum_{l=1}^{L}\sum_{i=1}^{n}\int_0^{t^*} \tau_{il}\boldsymbol{H}_i(u)\Big\{dN_i(u) - R_i(u)\lambda_{0l}(u)\exp\{\boldsymbol{H}_i(u)^T\boldsymbol{\gamma}_l\}du\Big\} = 0; \\ S_3\{d\boldsymbol{\Lambda}(u)\} &= \sum_{l=1}^{L}\sum_{i=1}^{n}\tau_{il}\Big\{dN_i(u) - R_i(u)\exp\{\boldsymbol{H}_i(u)^T\boldsymbol{\gamma}_l\}d\Lambda_l(u)\Big\} = 0. \end{aligned} \tag{4.3.4}$$

Here t^* is a finite constant that is larger than the maximum value of observed $\tilde{T}$. With $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ fixed, solving $S_3\{d\boldsymbol{\Lambda}(u)\} = 0$ at uncensored event times obtains a weighted Breslow's estimator of $\boldsymbol{\Lambda}(\cdot)$

$$\hat{\Lambda}_l(t;\boldsymbol{\tau},\boldsymbol{\gamma}) = \int_0^t \frac{\sum_{i=1}^n \tau_{il}dN_i(s)}{\sum_{i=1}^n \tau_{il}R_i(s)\exp\{\boldsymbol{H}_i(s)^T\boldsymbol{\gamma}_l\}},\ l = 1, \ldots, L. \tag{4.3.5}$$

Then by plugging in $d\hat{\boldsymbol{\Lambda}}(\cdot)$ to $S_2(\boldsymbol{\gamma})$ we obtain a weighted partial score for $\boldsymbol{\gamma}$

$$S_2\{\boldsymbol{\gamma}; \hat{\Lambda}_l(t;\boldsymbol{\tau},\boldsymbol{\gamma})\} = \sum_{l=1}^{L}\sum_{i=1}^{n}\int_0^{t^*} \tau_{il}\bigg(\boldsymbol{H}_i(u) - \frac{\sum_{j=1}^n \tau_{jl}R_j(u)\boldsymbol{H}_j(u)\exp\{\boldsymbol{H}_j(u)^T\boldsymbol{\gamma}_l\}}{\sum_{j=1}^n \tau_{jl}R_j(u)\exp\{\boldsymbol{H}_j(u)^T\boldsymbol{\gamma}_l\}}\bigg)dN_i(u). \tag{4.3.6}$$

Given $\boldsymbol{\tau}$, it is straightforward to use standard Cox regression solver, such as `coxph` (Therneau and Lumley, 2014), to solve (4.3.4) which reduce to (4.3.5) and (4.3.6).

4.3.3 Class-specific GEE submodel

Let $N_{\text{obs}}^V(t) = \sum_{k=1}^{\infty} I(V_k \le t, \tilde{T} \ge V_k)$ denote the observed visiting process. To estimate the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \boldsymbol{\rho}^T)^T$ from model (4.2.2), (4.2.3), it is critical to recognize that longitudinal observations $\boldsymbol{Y}(\cdot)$ is not independent of $N_{\text{obs}}^V(\cdot)$ given $\boldsymbol{X}$ and $\boldsymbol{\xi}$, although $\boldsymbol{Y}(\cdot)$ is independent of C and $N^V(\cdot)$ according to our assumptions. We use inverse probability weighting (IPW) technique to account for the dependency between $\boldsymbol{Y}(\cdot)$ and $N_{\text{obs}}^V(\cdot)$ conditioned on $\boldsymbol{X}$ and $\boldsymbol{\xi}$, such that the estimation for $\boldsymbol{\beta}$ in the mean model is unbiased. With unbiased estimation for $\boldsymbol{\beta}$, the scale parameter $\boldsymbol{\phi}$ and the correlation structure parameter $\boldsymbol{\rho}$ can also be estimated subsequently without bias.

By the mean model in (4.2.2), it is clear that $E\{\boldsymbol{Y}(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}; \boldsymbol{\beta})|\boldsymbol{X}, \xi_l = 1\} = 0$, where $\boldsymbol{Y}(t) = \{Y_1(t), \ldots, Y_J(t)\}^T$ and $\boldsymbol{\mu}_l(t, \boldsymbol{X}; \boldsymbol{\beta}) = \{\mu_{1,l}(t, \boldsymbol{X}; \boldsymbol{\beta}), \ldots, \mu_{J,l}(t, \boldsymbol{X}; \boldsymbol{\beta})\}^T$. Then the key equation which leads to the unbiased estimating function of $\boldsymbol{\beta}$ is

$$E\left\{ \int_0^{t^*} \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} dN_{\text{obs},i}^V(t) \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} = 0, \tag{4.3.7}$$

where the IPW term

$$S_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\} = \Pr(T \geq t | \bar{\boldsymbol{Z}}^{\text{obs}}(t-), \xi_l = 1)$$

is the class-specific survival probability beyond time t conditioned on the past history, $\bar{\boldsymbol{Z}}^{\text{obs}}(t-)$, prior to time t, at the subsequent visits. It is obvious that $S_l\{0, \bar{\boldsymbol{Z}}^{\text{obs}}(0-)\} = 1$. The proof of (4.3.7) can be found in the Appendix section 4.8.1. Denote $v_{i,1}, \ldots, v_{i,K_i}$ the observed visit times for the ith subject. We assume $\bar{\boldsymbol{Z}}^{\text{obs}}(t)$ is a piecewise constant function which only jumps at $v_{i,k} (k = 1, \ldots, K_i)$. Thus, $\bar{\boldsymbol{Z}}^{\text{obs}}(v_{i,k}-) = \bar{\boldsymbol{Z}}^{\text{obs}}(v_{i,k-1})$ for $k \geq 2$. Therefore, by (4.2.5), (4.2.6), and the property of internal covariates, we can derive that

$$\begin{aligned}
S_l\{t, \bar{\boldsymbol{Z}}^{\text{obs}}(t-)\} &= \Pr(T_i \geq v_{i,k} | \bar{\boldsymbol{Z}}_i^{\text{obs}}(v_{i,k}-), \xi_{il} = 1) \\
&= \Pr(T_i \geq v_{i,k} | \bar{\boldsymbol{Z}}_i^{\text{obs}}(v_{i,k-1}), T_i > v_{i,k-1}, \xi_{il} = 1) \\
&\quad \text{(Observation at } v_{i,k-1} \text{ indicates } T_i > v_{i,k-1}) \\
&= \exp\left\{ - \int_{v_{i,k-1}}^{v_{i,k}} \lambda_{0l}(s) \exp\{\boldsymbol{H}_i(v_{i,k-1})^T \boldsymbol{\gamma}_l\} ds \right\} \\
&= \exp\left\{ - \{\Lambda_{0l}(v_{i,k}) - \Lambda_{0l}(v_{i,k-1})\} \exp\{\boldsymbol{H}_i(v_{i,k-1})^T \boldsymbol{\gamma}_l\} \right\}.
\end{aligned}$$

Using similar steps in (4.3.3), we also show that equation (4.3.7) implies

$$E\left\{ \tau_{il} \int_0^{t^*} \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} \boldsymbol{c}(t, \boldsymbol{X}_i; \boldsymbol{\theta}) dN_{\text{obs},i}^V(t) \middle| \boldsymbol{X}_i \right\} = 0,$$

where $\boldsymbol{c}(t, \boldsymbol{X}_i; \boldsymbol{\theta})$ is a function for $t, \boldsymbol{X}$ and $\boldsymbol{\beta}$. With a proper choice of function $\boldsymbol{c}(t, \boldsymbol{X}_i; \boldsymbol{\theta})$,

we can construct a weighted GEE for $\boldsymbol{\beta}$

$$S_4(\boldsymbol{\beta}) = \sum_{l=1}^{L}\sum_{i=1}^{n} \tau_{il}\left(\frac{\partial \boldsymbol{\mu}_{i,l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_{i,l}^{-1}(\boldsymbol{X}_i;\boldsymbol{\theta})\boldsymbol{W}_i\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i}) - \boldsymbol{\mu}_{i,l}(\boldsymbol{\beta})\} = 0, \tag{4.3.8}$$

where $\boldsymbol{\Sigma}_{i,l}(\boldsymbol{X}_i;\boldsymbol{\beta})$ is the variance-covariance matrix of $\boldsymbol{Y}_i$ based on model (4.2.3) and (4.2.4), and $\boldsymbol{W}_i$ is a $K_iJ \times K_iJ$ diagonal weight matrix with $(k \times j)$th diagonal element equal to $S_l\{v_{i,k}, \bar{\boldsymbol{Z}}_i^{\text{obs}}(v_{i,k}-)\}, k = 1,\ldots,K_i, j = 1,\ldots,J$, corresponding to the kth visit and the jth longitudinal outcome. The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\rho}$ associated with the second moment of longitudinal outcomes can be estimated using standard procedures by solving the following estimating equations:

$$\begin{aligned} S_5(\boldsymbol{\phi}) &= \sum_{l=1}^{L}\sum_{i=1}^{n} \tau_{il}\left(\frac{\partial \boldsymbol{\sigma}_{i,l}(\boldsymbol{\beta},\boldsymbol{\phi})^2}{\partial \boldsymbol{\phi}}\right)^T \{\boldsymbol{s}_{i,l}^2 - \boldsymbol{\sigma}_{i,l}(\boldsymbol{\beta},\boldsymbol{\phi})^2\} = 0; \\ S_6(\boldsymbol{\rho}) &= \sum_{l=1}^{L}\sum_{i=1}^{n} \tau_{il}\left(\frac{\partial \boldsymbol{\eta}_{i,l}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}}\right)^T \{\boldsymbol{r}_{i,l} - \boldsymbol{\eta}_{i,l}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\rho})\} = 0, \end{aligned} \tag{4.3.9}$$

where

$$\begin{aligned} \boldsymbol{s}_{i,l}^2 &= \{\{[Y_j(V_k) - \mu_{j,l}(V_k,\boldsymbol{X}_i;\boldsymbol{\beta}_{j,l})]^2\}_{k=1}^{K_i}\}_{j=1}^{J}; \\ \boldsymbol{\sigma}_{i,l}^2(\boldsymbol{\beta},\boldsymbol{\phi}) &= E(\boldsymbol{s}_{i,l}^2|\boldsymbol{\xi}_{il}=1) = \{\{\sigma_j^2(V_k,\boldsymbol{X}_i;\phi_{j,l},\boldsymbol{\beta}_{j,l})\}_{k=1}^{K_i}\}_{j=1}^{J}; \\ \boldsymbol{r}_{i,l} &= \{\{[Y_j(V_p) - \mu_{j,l}(V_p,\boldsymbol{X}_i;\boldsymbol{\beta}_{j,l})][Y_j(V_q) - \mu_{j,l}(V_q,\boldsymbol{X}_i;\boldsymbol{\beta}_{j,l})]\}_{p<q}^{K_i}\}_{j=1}^{J}; \\ \boldsymbol{\eta}_{i,l}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\rho}) &= E(\boldsymbol{r}_{i,l}|\boldsymbol{\xi}_{il}=1) = \{\{\eta_j(p,q,\boldsymbol{X};\rho_{j,l},\phi_{j,l},\boldsymbol{\beta}_{j,l})\}_{p<q}^{K_i}\}_{j=1}^{J}. \end{aligned}$$

The estimating functions in (4.3.9) can be shown as unbiased following similar steps as in (4.3.3).

4.3.4 Posterior class membership probability

As we derive the estimating equations in the previous subsections, we assume that the posterior membership probability τ_{il} is well defined. In fact, due to the lack of explicit density function for the longitudinal observation, it is challenging to derive τ_{il} from a finite mixture likelihood. In this section we propose a formulation of τ_{il} under the semi-parametric

framework.

Let $g_l\{\bar{\boldsymbol{Y}}^{\text{obs}}(V_K)|\boldsymbol{X};\boldsymbol{\vartheta}\}$ denote the underlying true class-specific density of longitudinal data $\bar{\boldsymbol{Y}}^{\text{obs}}(V_K)$, for class l with parameters $\boldsymbol{\vartheta}$. Then it is straightforward to obtain the finite mixture density of observed data $\boldsymbol{O}$ as

$$f\{\boldsymbol{O}_i|\boldsymbol{X}_i\} = \sum_{l=1}^{L} p_l(\boldsymbol{X}_i;\boldsymbol{\alpha}) g_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X};\boldsymbol{\vartheta}\} f_l\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\text{obs}}(V_{i,K_i});\boldsymbol{\zeta}\}.$$

By Bayes rule, the posterior membership probability $\Pr(\xi_{il}=1|\boldsymbol{O}_i)$ can be derived as

$$\Pr(\xi_{il}=1|\boldsymbol{O}_i) = \frac{p_l(\boldsymbol{X}_i;\boldsymbol{\alpha})\exp[w_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\vartheta}\}] f_l\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\text{obs}}(V_{i,K_i})\}}{\sum_{d=1}^{L} p_d(\boldsymbol{X}_i;\boldsymbol{\alpha})\exp[w_d\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\vartheta}\}] f_d\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\text{obs}}(V_{i,K_i})\}}.$$

Here, $w_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\vartheta}\} = \log[g_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X};\boldsymbol{\vartheta}\}/g_1\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X};\boldsymbol{\vartheta}\}]$ is the log likelihood ratio of the density of $\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})$ between class l and class 1. In our GEE submodel, nevertheless, we only specify the first moment (4.2.2) and second moment (4.2.3), (4.2.4) of $\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})$, which means $g_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X};\boldsymbol{\vartheta}\}$ is not well defined in our model.

In order to utilize information from GEE submodel into the construction of posterior membership probability, we calculate the following linear projection (Li, 1993) of the log likelihood ratio

$$\tilde{w}_{il}\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\theta}\} = \frac{1}{2}\{\mu_{i,l}(\boldsymbol{\beta})-\mu_{i,1}(\boldsymbol{\beta})\}\Big\{\boldsymbol{\Sigma}_{i,l}^{-1}(\boldsymbol{X}_i;\boldsymbol{\theta})(\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})-\mu_{i,l}(\boldsymbol{\beta})) + \boldsymbol{\Sigma}_{i,1}^{-1}(\boldsymbol{X}_i;\boldsymbol{\theta})(\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})-\mu_{i,1}(\boldsymbol{\beta}))\Big\}$$

as an approximation of $w_{il}\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\theta}\}$. In a special case, $\tilde{w}_{il}\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\theta}\}$ is the exact log likelihood ratio when $\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})$ follows Gaussian distribution with identical variance-covariance matrix $\boldsymbol{\Sigma}_{i,l}$ for all latent classes. Then the posterior class probability based on proposed submodels can be approximated by

$$\tilde{\tau}_{il} = \frac{p_l(\boldsymbol{X}_i;\boldsymbol{\alpha})\exp[\tilde{w}_l\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\theta}\}] f_l\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\text{obs}}(V_{i,K_i})\}}{\sum_{d=1}^{L} p_d(\boldsymbol{X}_i;\boldsymbol{\alpha})\exp[\tilde{w}_d\{\bar{\boldsymbol{Y}}_i^{\text{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\boldsymbol{\theta}\}] f_d\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\text{obs}}(V_{i,K_i})\}}. \tag{4.3.10}$$

4.3.5 Algorithm

We propose an iterative expectation-solution algorithm to jointly solve the estimating equations (4.3.1), (4.3.4) and (4.3.8). The algorithm is inspired by the EM algorithm that is commonly applied in fully parametric modeling. Below the detailed steps are illustrated.

1. (Initialization) Initialize $\hat{\tau}_{il}^{(0)} = 1/L$ for all i and l. Set the number of iterations $i = 1$.

2. (Solution step, class probability submodel) Given $\hat{\boldsymbol{\tau}}^{(i-1)}$, solve (4.3.1) directly by weighted multinomial logistic regression solver to obtain $\hat{\boldsymbol{\alpha}}^{(i)}$.

3. (Solution step, class-specific Cox submodel) Given $\hat{\boldsymbol{\tau}}^{(i-1)}$, first solve the weighted partial score equation (4.3.6) for $\hat{\boldsymbol{\gamma}}^{(i)}$, then obtain the Breslow's estimator $\hat{\boldsymbol{\Lambda}}^{(i)}(\cdot)$ by plugging in $\hat{\boldsymbol{\tau}}^{(i-1)}$ and $\hat{\boldsymbol{\gamma}}^{(i)}$ in (4.3.5).

4. (Solution step, class-specific GEE submodel) Given $\hat{\boldsymbol{\tau}}^{(i-1)}$, $\hat{\boldsymbol{\gamma}}^{(i)}$ and $\hat{\boldsymbol{\Lambda}}^{(i)}(\cdot)$, solve (4.3.8) and (4.3.9) for $\hat{\boldsymbol{\theta}}^{(i)}$.

5. (Expectation step) Given $\hat{\boldsymbol{\psi}}^{(i)} = \{\hat{\boldsymbol{\alpha}}^{(i)T}, \hat{\boldsymbol{\gamma}}^{(i)T}, \hat{\boldsymbol{\Lambda}}^{(i)}(\cdot), \hat{\boldsymbol{\theta}}^{(i)^T}\}$, update $\hat{\boldsymbol{\tau}}^{(i)}$ by (4.3.10). Set $i = i + 1$.

6. (Convergence criterion) Repeat steps 2-5 until $\sum_i^n ||\hat{\boldsymbol{\tau}}^{(i)} - \hat{\boldsymbol{\tau}}^{(i-1)}||_2^2 < 0.001$.

Compared to the existing methods, the proposed algorithm does not require random guess or prior knowledge about the initial values. Instead, the initial posterior probabilities $\hat{\boldsymbol{\tau}}^{(0)}$ are equal for all classes. The convergence criterion based on $\hat{\boldsymbol{\tau}}^{(i)}$ ensures that the algorithm stops when the class membership probabilities remain stable enough between iterations.

4.4 Selecting the number of latent classes

To select the number of latent classes, L, it is common to use information criteria such as Akaike information criterion or Bayesian information criterion. For the proposed framework, however, the GEE submodel only defines the first two moments of the longitudinal

observations, such that the density for the longitudinal observations is not explicitly assumed. To account for the goodness-of-fit for the longitudinal submodel, we propose the following quasi-likelihood formulations

$$\mathrm{QL}(\hat{p}) = -2\sum_{i=1}^{n}\log\Bigg\{\sum_{l=1}^{L} p_l(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})\exp[Q_l^+\{\bar{\boldsymbol{Y}}_i^{\mathrm{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\}] \\ f_l\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\mathrm{obs}}(V_{i,K_i}),\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\Lambda}}\}\Bigg\},$$

$$\mathrm{QL}(\hat{\tau}) = -2\sum_{i=1}^{n}\log\Bigg\{\sum_{l=1}^{L}\hat{\tau}_{il}\exp[Q_l^+\{\bar{\boldsymbol{Y}}_i(\tilde{T}_i)|\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\}]f_l\{\tilde{T}_i,\Delta_i|\bar{\boldsymbol{Z}}_i^{\mathrm{obs}}(V_{i,K_i}),\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\Lambda}}\}\Bigg\},$$

where $Q_l^+\{\bar{\boldsymbol{Y}}_i^{\mathrm{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\}$ the class-specific extended log quasi-likelihood for the longitudinal submodel. For the observation of a specific longitudinal feature j at time t, $Y_j(t)$, the corresponding class-specific extended log quasi-likelihood is $Q_l^+\{Y_j(t)|\boldsymbol{X};\hat{\boldsymbol{\theta}}\} = Q_l\{Y_j(t)|\boldsymbol{X};\hat{\boldsymbol{\theta}}\} - \frac{1}{2}\log\hat{\phi}_{j,l}\mathcal{V}_j\{\mu_{j,l}(t,\boldsymbol{X},\hat{\boldsymbol{\beta}})\}$, where $Q_l\{Y_j(t)|\boldsymbol{X};\hat{\boldsymbol{\theta}}\}$ is the class-specific log quasi-likelihood and $\frac{1}{2}\log\hat{\phi}_{j,l}\mathcal{V}_j\{\mu_{j,l}(t,\boldsymbol{X},\hat{\boldsymbol{\beta}})\}$ is the extended component containing information of variance functions. Therefore, we have

$$Q_l^+\{\bar{\boldsymbol{Y}}_i^{\mathrm{obs}}(V_{i,K_i})|\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\} = \sum_{j=1}^{J}\sum_{k=1}^{K_i} Q_l^+\{Y_{i,j}(V_{i,k})|\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\}.$$

The difference between $\mathrm{QL}(\hat{p})$ and $\mathrm{QL}(\hat{\tau})$ is whether using baseline membership probability $p_l(\boldsymbol{X};\hat{\boldsymbol{\alpha}})$ or posterior membership probability $\hat{\tau}_l$ in the finite mixture structure. Compared to $\mathrm{QL}(\hat{p})$, $\mathrm{QL}(\hat{\tau})$ is able to assess the goodness-of-fit with posterior knowledge about the class membership. With the defined quasi-likelihood QL, we further define the following candidates of information criteria

$$\begin{aligned}\mathrm{QAIC}(\cdot) &= \mathrm{QL}(\cdot) + 2p,\\ \mathrm{QBIC}(\cdot) &= \mathrm{QL}(\cdot) + p\log n,\\ \text{CE-QBIC}(\cdot) &= \mathrm{QL}(\cdot) + p\log n - 2\sum_{i=1}^{n} n\sum_{l=1}^{n}\hat{\tau}_{il}\log\hat{\tau}_{il},\end{aligned} \tag{4.4.1}$$

where p is the number of unknown parameters, and $\mathrm{QL}(\cdot)$ can be either $\mathrm{QL}(\hat{p})$ or $\mathrm{QL}(\hat{\tau})$. In addition to the criteria defined in (4.4.1), we also consider standardized entropy index

defined by

$$\text{ENT} = 1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{L}\hat{\tau}_{il}\log\hat{\tau}_{il}}{n\log L},$$

which is a commonly used criterion to evaluate the fuzziness of posterior classification.

4.5 Simulation study

4.5.1 Data generation procedure

Simulations were conducted to investigate the performance of the proposed LCA framework for joint longitudinal and survival data with two latent classes ($L = 2$). First, we generate a bivariate vector of baseline covariates $\boldsymbol{X} = (X_1, X_2)^T$, where X_1 follows $Bernoulli(0.5)$ distribution and X_2 follows $Uniform(0, 1)$ distribution. Then we generate the class membership probability $\boldsymbol{p} = (p_1, p_2)^T$ by model (4.2.1) with $\boldsymbol{\alpha}_1 = \mathbf{0}$ and $\boldsymbol{\alpha}_2 = (0.5, 0, 0)^T$. Then the latent class membership $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$ is generated from $Multinomial\{1, (p_1, p_2)^T\}$ distribution. Given $\boldsymbol{\xi}$, we further generate longitudinal trajectories and time-to-event data.

Table 4.1: An overview of longitudinal features considered in simulation studies, including data types of features, link functions, and the associated regression coefficients for marginal model ($\beta_{j,l,0}, \beta_{j,l,1}$) and proportional hazards (γ).

			Class (l)					
			Class 1			Class 2		
$Y_j(t)$	*Distribution*	$h_j(x)$	$\beta_{j,1,0}$	$\beta_{j,1,1}$	$\gamma_{j,1}$	$\beta_{j,2,0}$	$\beta_{j,2,1}$	$\gamma_{j,2}$
$Y_1(t)$	Poisson	$\exp(x)$	0	0.5	0.1	0	0	0
$Y_2(t)$	Poisson	$\exp(x)$	1	0	0	1	-0.5	0.05
$Y_3(t)$	Binary	$\frac{\exp(x)}{1+\exp(x)}$	-0.5	1	0.05	-0.5	0	0
$Y_4(t)$	Binary	$\frac{\exp(x)}{1+\exp(x)}$	0.5	0	0	0.5	-1	0.1
$Y_5(t)$	Normal	x	0	0	0	0	1	0.5
$Y_6(t)$	Normal	x	0	1	0.25	0	0	0

We assume that the longitudinal outcome at time t, $\boldsymbol{Y}(t) = \{Y_1(t), \ldots, Y_6(t)\}$, is a vector with six independent elements, where $Y_1(t)$ and $Y_2(t)$ are count data, $Y_3(t)$ and $Y_4(t)$ are binary, and $Y_5(t)$ and $Y_6(t)$ are continuous. We consider the GEE model $E\{Y_j(t)|\xi_l = 1\} = \mu_{j,l}(t; \boldsymbol{\beta}_{j,l}) = h_j(\beta_{j,l,0} + r_\beta\beta_{j,l,1}t),\ j = 1, \ldots, J$, which is a simplified version of the

marginal model in (4.2.2). Here r_β is used to adjust the slope of trajectories for all features, where larger slopes will generate heavier differences in trajectories among classes. Table 4.1 lists the corresponding link functions and parameters $(\beta_{j,l,0}, \beta_{j,l,1})^T$ for each feature in each class. Auto-regressive process AR(1) is assumed for the correlation structure for each longitudinal feature with $\rho_{j,l} = 0.3$ for all j and l. For all l, the scale parameter $\phi_{j,l}$ is also set to be 1 for $j = 1, \ldots, 4$, and ϕ for $j = 5, 6$, where ϕ is used to adjust the variation from trajectory mean for the two continuous features. A larger ϕ can cause more severe overlapping of the trajectories among classes. With selected r_{β_1} and ϕ, we generate $\boldsymbol{Y}(t)$ for $t = 0, 1, 2, \ldots, 20$. We utilize techniques developed by Qaqish (2003) and Dalthorp and Madsen (2007) to generate correlated binary and count features, respectively. Correlated continuous features are generated from multivariate normal distribution.

With generated longitudinal outcomes, we have the time-varying covariates for the class-specific proportional hazards submodel of the terminal event (4.2.6). We let $\boldsymbol{H}(t) = \{\boldsymbol{X}^T, \boldsymbol{Y}(V_t)^T\}^T$, where $V_t = \max\{V_k : V_k < t\}$ is the latest visit time prior to t. The corresponding regression coefficient is zero for $\boldsymbol{X}$ and $r_\gamma \gamma_{j,l}$ for $\boldsymbol{Y}(V_t)$, where r_γ adjust the effect size and $\gamma_{j,l}$ is shown in Table 4.1. Baseline hazard function is set to be $\lambda_l(t) = \frac{1}{\eta_l(1+t)}$ where $\eta_2 = 6$ and η_1 adjusts the shape of the function for the first class. We choose this form of hazard function to ensure that multiple longitudinal features can be observed before the terminal event occurs. Piecewise exponential distribution technique (Hendry, 2014) is used to generate the terminal event time T with time-varying covariates $\boldsymbol{H}(t)$ at $t = 0, 1, \ldots, 20$. Independent censoring C is generated as the minimum of a $Uniform(5, 6)$ random variable and a $Exponential(0.1)$ random variable.

With the generated quantities as described above, we construct the dataset for our simulation with baseline covariate $\boldsymbol{X}$, time-to-event $(\tilde{T}, \Delta)$ where $\tilde{T} = \min\{T, C\}$ and $\Delta = I(T \leq C)$, and longitudinal features $\bar{\boldsymbol{Y}}^{\text{obs}}(V_K) = \{\boldsymbol{Y}(t) : t \in (0, 1, \ldots, 20), t < \tilde{T}\}$, where $V_K = \max\{t : t \in (0, 1, \ldots, 20), t < \tilde{T}\}$ is the largest visit time prior to $\tilde{T}$.

4.5.2 Simulation scenarios

We consider a number of settings regarding r_β, ϕ, r_γ and η_2 mentioned in Section 4.5.1, displayed in Table 4.2, to generate datasets with different characteristics. We expect that the algorithm works better when the two latent classes are more separated from each other.

Table 4.2: Simulation settings and the corresponding interpretations.

Settings	r_β	trajectory slope
(1a)	1	steep
(1b)	0.5	moderately steep
(1c)	0.25	gentle

Settings	ϕ	trajectory variation
(2a)	1	mild
(2b)	4	moderate
(2c)	9	heavy

Settings	r_γ	trajectory effect size
(3a)	1	moderate
(3b)	2	strong

Settings	η_2	Shape of $\Lambda_{02}(t)$
(4a)	8	gentle
(4b)	4	steep

Based on above settings, we identify four sets of investigations listed in Table 4.3. For each scenario, we generate 1000 datasets, collected data characteristics in terms of fuzziness (entropy) and censoring rate, and run the proposed algorithm to obtain point estimation for all unknown parameters. As observed in Table 4.3, the fuzziness of the latent class pattern is severely affected by different trajectory slopes in investigation (A), where the pattern is much more fuzzy with gentle slope (1c, median entropy=0.66) compared to steep slope (1a, median entropy=0.87). In contrast, heavier trajectory variation (2c) in investigation (B) does not dramatically affect median entropy but causes larger differences in the class-specific range of time-to-event and lower censoring rates. A lower censoring rate indirectly implies less visits observed for the population, such that less information is available for the trajectory. In investigation (C), larger effect size of longitudinal trajectory on survival in (3b) accelerates the failure process, which also causes lower censoring rate, fewer longitudinal observations per subject, and larger difference in the support of time-to-event for the two latent classes. In investigation (D), the difference in the shape of baseline hazard function (4b) causes apparent differences in the class-specific support for the two classes.

Table 4.3: Sets of investigations, the corresponding simulation scenarios, and empirical data characteristics from simulated datasets. $\Delta\{\text{Median}(\tilde{T})\}$ represents the difference between the class-specific median time-to-event.

Investigations	Scenarios†	Entropy‡	Censoring (%)‡	$\Delta\{\text{Median}(\tilde{T})\}$‡
(A) trajectory slope (2a)+(3a)+(4a)	(1a)	0.87	57	0.001
	(1b)	0.81	65	-0.005
	(1c)	0.66	63	-0.006
(B) trajectory variation (1a)+(3a)+(4a)	(2a)	0.87	57	0.001
	(2b)	0.85	52	0.324
	(2c)	0.85	45	0.591
(C) trajectory effect size (1a)+(2a)+(4a)	(3a)	0.87	57	0.001
	(3b)	0.88	25	0.305
(D) shape of $\Lambda_{02}(t)$ (1a)+(2a)+(3a)	(4a)	0.87	57	0.001
	(4b)	0.88	44	1.072

† Refer to Table 4.2.
‡ Median from 1000 simulations

4.5.3 Point estimation

Table 4.4 shows the average empirical bias and mean square error (MSE) for point estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ across 1000 simulations for each scenario. Under investigation (A), it is clear that gentle trajectory slope (1c) causes heavier fuzziness of latent class patterns compared to steep trajectory slope (1a), resulting in larger bias and MSE of $\hat{\boldsymbol{\alpha}}$ for the latent class probability model. In contrast, estimation for the class-specific time-to-event submodel ($\hat{\boldsymbol{\gamma}}$) and GEE submodel ($\hat{\boldsymbol{\beta}}$) is not heavily affected, showing solid robustness of the method. In contrast, point estimation under investigations (B), (C) and (D) appears more stable due to less fuzziness associated with higher entropy index, which also implies the method's high robustness under different data characteristics.

4.5.4 Selecting the number of latent classes

We fit the model for the 1000 simulated datasets for each scenario with $L \in \{2, 3, 4\}$ and apply model selection criteria introduced in Section 4.4 to check how different model selection criteria perform under different scenarios. Figure 4.1 shows the percentages of selected

Table 4.4: Average bias (mean square error) for point estimates in the class membership probability submodel ($\hat{\boldsymbol{\alpha}}$), time-to-event submodel $\hat{\boldsymbol{\gamma}}$, and longitudinal submodel (intercept $\hat{\boldsymbol{\beta}}_0$ and slope $\hat{\boldsymbol{\beta}}_1$) for all simulation scenarios.

Investigation	Scenario	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\gamma}}$	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$
(A) trajectory slope	(1a)	-0.023 (0.040)	-0.002 (0.049)	-0.002 (0.007)	-0.001 (0.001)
	(1b)	-0.123 (0.148)	-0.012 (0.107)	-0.003 (0.010)	-0.001 (0.057)
	(1c)	-0.185 (0.195)	-0.007 (0.180)	-0.001 (0.015)	-0.008 (0.024)
(B) trajectory variation	(2a)	-0.023 (0.040)	-0.002 (0.049)	-0.002 (0.007)	-0.001 (0.001)
	(2b)	-0.053 (0.041)	-0.013 (0.028)	-0.002 (0.010)	-0.008 (0.003)
	(2c)	-0.067 (0.042)	-0.010 (0.019)	-0.001 (0.018)	-0.009 (0.006)
(C) trajectory effect size	(3a)	-0.023 (0.040)	-0.002 (0.049)	-0.002 (0.007)	-0.001 (0.001)
	(3b)	-0.036 (0.033)	-0.001 (0.016)	-0.001 (0.007)	0.000 (0.003)
(D) shape of $\Lambda_{02}(t)$	(4a)	-0.023 (0.040)	-0.002 (0.049)	-0.002 (0.007)	-0.001 (0.001)
	(4b)	-0.072 (0.084)	-0.015 (0.056)	-0.005 (0.012)	0.000 (0.002)

number L of latent classes for all scenarios studied in Table 4.3 by seven different criteria. It can be seen that the entropy index (ENT) works well for most scenarios apart from the two scenarios with smaller trajectory slopes and consequently high fuzziness. Compared to ENT, both CE-QBIC(p) and CE-QBIC(τ) performs well under high fuzziness setting, where CE-QBIC(τ) obtains better performances over CE-QBIC(p) for most cases. In contrast, QAICs and QBICs did not account for the entropy in their formulation and perform poorer. Our practical recommendation, therefore, is to use CE-QBIC(τ) which robustly selects correct number of latent classes under all simulation scenarios.

4.6 Real data application

We applied the proposed method to analyze the UDS data to investigate the MCI subtypes indicated by both longitudinal patient characteristics and the distribution of time-to-death. The data consists of follow-up information for 5348 patients between September 2005 and June 2015, including baseline covariates such as smoking status and history of cerebrovascular disease, longitudinal scores on patients' cognitive, neuropsychiatric, and functional characteristics, and time-to-death since the first visit. Out of the 5348 patients, 667 died during the follow-up period, showing a high censoring rate of observing death. More than

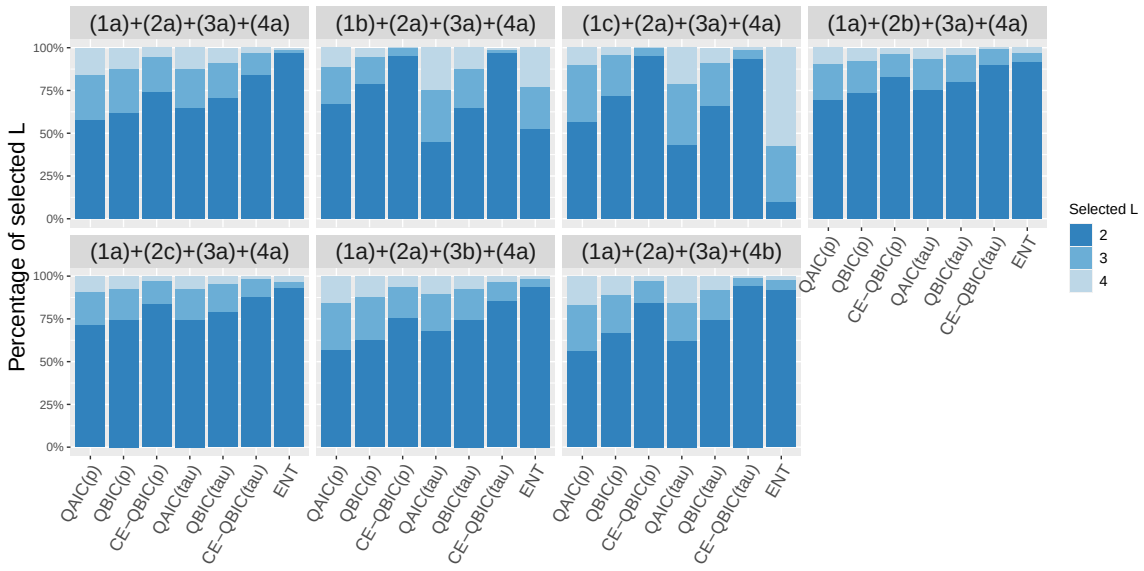

Figure 4.1: Percentage of latent classes selected by different model selection criteria listed in Section 4.4 out of 1000 simulations under simulation scenarios listed in Table 4.3. Greek letter τ is denoted by "tau" in the plot. Entropy index is denoted by "ENT".

1000 patients had at least five clinical visits, providing rich information about the longitudinal trajectory of MCI patients.

In our analysis, we let the decades of smoking and elevated Hachinski score (an indicator of cerebrovascular disease) as baseline covariates $\boldsymbol{X}$. The longitudinal features $\bar{\boldsymbol{Y}}(V_K)$ are the same as used in the analysis conducted by Hart et al. (2020), including one binary measure of depression, two count variables of patients' functional abilities and neuropsychiatric symptoms, and ten continuous cognitive scores. We assume that the longitudinal features follows the "last-value-carried-forward" (LCVF) principle, which is essentially piecewise constant. The terminal event death is jointly modeled with the longitudinal features, where the longitudinal features serve as time-dependent covariates of the time-to-event submodel. We fitted the proposed joint model for $L \in \{2, 3, 4, 5\}$ with non-informative initialization of the expectation-solution algorithm. The 4-class model obtained the smallest CE-QBIC(τ), 288776, compared to the 2-class model (305649), 3-class model (300454) and 5-class model (292896).

As shown in Figure 4.2, the resulting classes form distinct class-specific trajectories and survival probability curves. According to the estimated trajectory, class 3 (25% of study

population) is the benign non-amnestic MCI class with the mildest impairment at baseline and almost no progression of dementia in the follow-up. Class 2 (33% of study population) is the benign multi-domain anmestic MCI class with slight impairment at baseline in multiple cognitive domains but no obvious progression over years. In contrast, class 1 (mildly progressive amnestic multi-domain, 20% of study population) shows clear memory impairment at baseline but develops impairment in other domains such as the executive functions (Trails B score) during follow-up. As the amnestic class with the most rapid progression of dementia, class 4 (22% of study population) shows multi-domain impairment at baseline and rapidly deteriorating longitudinal trajectories.

The class-specific Kaplan-Meier curves by modal assignment (Figure 4.2) agree with our interpretation for the trajectories. The rapidly progressive amnestic multi-domain class (class 4) has the lowest survival probability in the eight years of follow-up, while the mildly progressive amnestic multi-domain class (class 1) has the second lowest survival probability curve. On the other hand, the benign classes (class 2 and class 3) have similarly higher survival probability curves. The concordance of longitudinal trajectories and survival curves indicates that the longitudinal features and time-to-death have high dependency in the UDS data, which further justifies our considerations on accounting for the within-class dependency of longitudinal and survival outcomes.

4.7 Discussion

In this project we propose a semi-parametric finite mixture latent class analysis framework for joint longitudinal and survival data. Our approach is flexible to define latent subtypes but still account for within-class dependency between longitudinal trajectories and time-to-event. The proposed algorithm effectively utilizes existing popular software for conduct iterative updates in a similar style as EM-algorithm. Our simulation shows that the algorithm can achieve good performance with non-informative initial values, which is critically important for real applications where prior knowledge may be unavailable about the underlying true latent class patterns.

We adapted semi-parametric submodels for the survival and longitudinal data to flexibly circumvent restrictive parametric assumptions imposed by fully parametric models. Our estimating procedure is carefully designed to address informative censoring of longitudinal observations by a terminal event. The estimating equation for the GEE marginal model (4.3.8) can be easily extended to account for informative censoring of trajectories caused by drop-out and visit time by introducing extra submodels for the drop-out process and visiting process.

There are several important aspects which require further investigation in the future. First, we are studying the asymptotic properties and inference procedures for the proposed estimators, which can crucially improve the practical utility in real applications. Second, we plan to develop approaches to conducting dynamic prediction of patient survival based on the proposed latent class framework, which may be an improvement over the existing methods due to our ability to account for the within-class correlation of longitudinal and survival information. We are hopeful that the above investigations can contribute significantly to the MCI research.

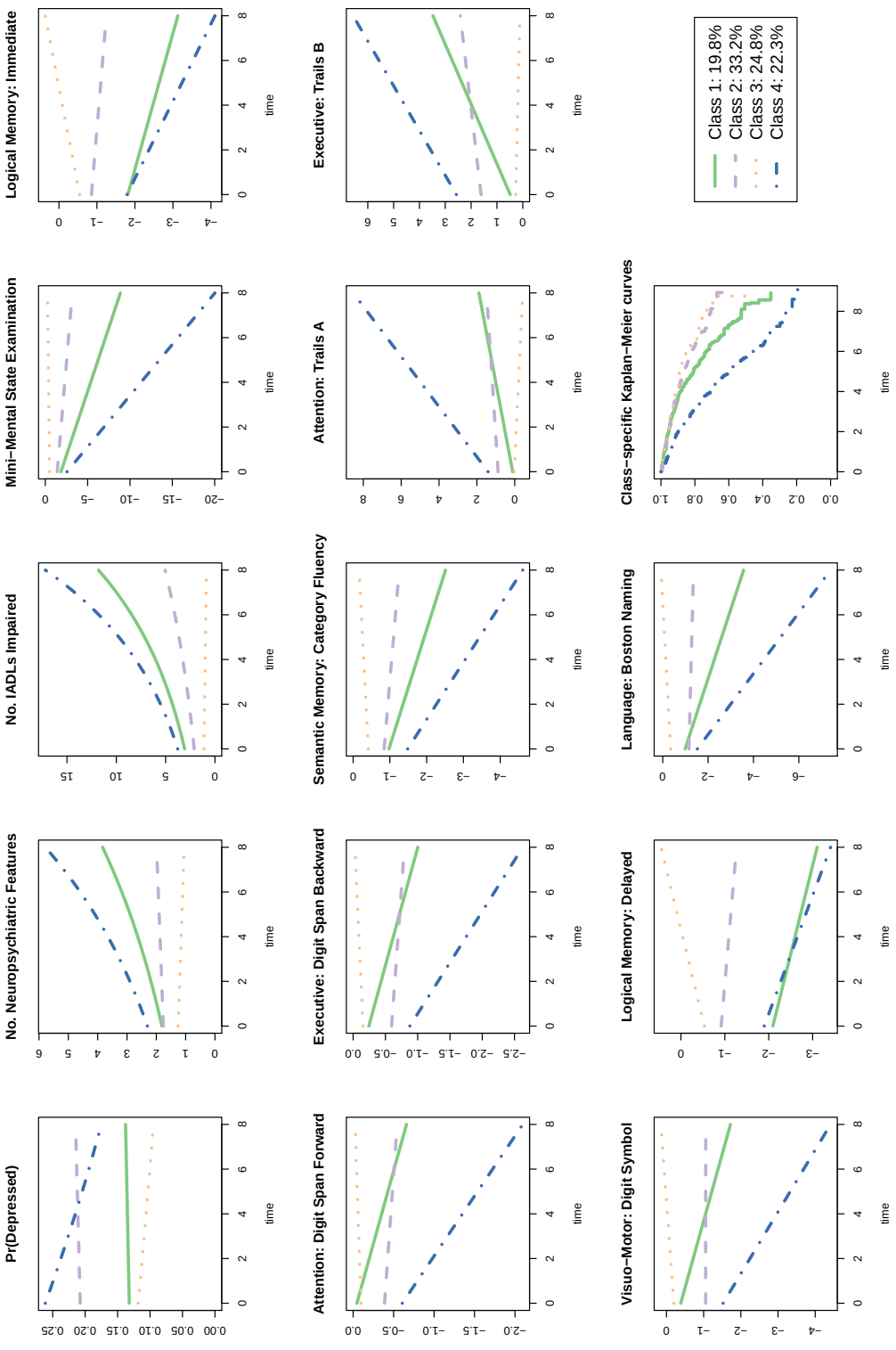

Figure 4.2: Estimated trajectories and Kaplan-Meier curves based on modal class assignment rule for the fitted four-class joint latent class model.

4.8 Appendices

4.8.1 Proof of Equation (4.3.7)

$$
\begin{aligned}
&E\left\{ \int_0^{t^*} \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} dN_{\text{obs},i}^V(t) \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
=&E\left\{ E\left\{ \int_0^{t^*} \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} I(\tilde{T}_i \geq t) dN_i^V(t) \middle| \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1 \right\} \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
=&E\left\{ E\left\{ \int_0^{t^*} \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} I(T_i \geq t, C_i \geq t) dN_i^V(t) \middle| \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1 \right\} \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
=& \int_0^{t^*} E\left\{ E\left\{ \frac{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} I(T_i \geq t) \middle| \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1 \right\} \right. \\
&\qquad \left. \Pr(C_i \geq t | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1) E\{dN_i^V(t) | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1\} \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
&\qquad\qquad (N(t) \perp N^C(t) \perp N^V(t) | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \boldsymbol{\xi}) \\
=& \int_0^{t^*} E\left\{ \frac{E\{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta}) | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1\}}{S_l\{t, \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-)\}} \Pr(T_i \geq t | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1\} \right. \\
&\qquad \left. \Pr\{C_i \geq t | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1\} E\{dN_i^V(t) | \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1\} \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
&\qquad\qquad \text{(Sequential ignorability assumption)} \\
=& \int_0^{t^*} E\left\{ E\left\{ \{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})\} I(C_i \geq t) dN_i^V(t) \middle| \bar{\boldsymbol{Z}}_i^{\text{obs}}(t-), \xi_{il} = 1 \right\} \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
=& \int_0^{t^*} E\left\{ \{\boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta})\} I(C_i \geq t) dN_i^V(t) \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \\
=& \int_0^{t^*} E\left\{ \boldsymbol{Y}_i(t) - \boldsymbol{\mu}_l(t, \boldsymbol{X}_i; \boldsymbol{\beta}) \middle| \boldsymbol{X}_i, \xi_{il} = 1 \right\} \Pr(C_i \geq t | \boldsymbol{X}_i, \xi_{il} = 1) A_l(t; \boldsymbol{X}_i) dt = 0 \\
&\qquad\qquad (A_l(t; \boldsymbol{X}_i) dt = E\{dN_i^V(t) | \boldsymbol{X}_i, \xi_{il} = 1\})
\end{aligned}
$$

Bibliography

Adler, C. H., Caviness, J. N., Sabbagh, M. N., Shill, H. A., Connor, D. J., Sue, L., Evidente, V. G., Driver-Dunckley, E., and Beach, T. G. (2010). Heterogeneous neuropathological findings in parkinson's disease with mild cognitive impairment. *Acta neuropathologica* **120,** 827–828.

Agresti, A. (2003). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Ardia, D., Boudt, K., Carl, P., Mullen, K., and Peterson, B. G. (2011). Differential evolution with deoptim: an application to non-convex portfolio optimization. *The R Journal* **3,** 27–34.

Bakk, Z. and Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika* **83,** 871–892.

Bakk, Z., Tekle, F. B., and Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology* **43,** 272–311.

Bakk, Z. and Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal* **23,** 20–31.

Bakoyannis, G., Zhang, Y., and Yiannoutsos, C. T. (2020). Semiparametric regression and

risk prediction with competing risks data under missing cause of failure. *Lifetime data analysis* **26,** 659–684.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92,** 1375–1386.

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., Hubbard, J. L., Koepsell, T. D., Morris, J. C., Kukull, W. A., et al. (2007). The national alzheimer's coordinating center (nacc) database: the uniform data set. *Alzheimer Disease & Associated Disorders* **21,** 249–258.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **22,** 719–725.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis* **12,** 3–27.

Boldea, O. and Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association* **104,** 1539–1549.

Bučar, T., Nagode, M., and Fajdiga, M. (2004). Reliability approximation using finite weibull mixture distributions. *Reliability Engineering & System Safety* **84,** 241–251.

Clogg, C. C. (1995). Latent class models. *Handbook of statistical modeling for the social and behavioral sciences* pages 311–359.

Cooper, D., Lacritz, L. H., Weiner, M., Rosenberg, R., and Cullum, C. (2004). Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. *Alzheimer Disease & Associated Disorders* **18,** 120–122.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202.

Dalthorp, D. and Madsen, L. (2007). Generating correlated count data. *Environmental and Ecological Statistics* **14,** 129–148.

Dias, J. G. and Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics* **23,** 643–659.

Dugger, B. N., Davis, K., Malek-Ahmadi, M., Hentz, J. G., Sandhu, S., Beach, T. G., Adler, C. H., Caselli, R. J., Johnson, T. A., Serrano, G. E., et al. (2015). Neuropathological comparisons of amnestic and nonamnestic mild cognitive impairment. *BMC neurology* **15,** 146.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software* **40,** 1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis* **71,** 1054–1063.

Elliott, M. R., Zhao, Z., Mukherjee, B., Kanaya, A., and Needham, B. L. (2020). Methods to account for uncertainty in latent class assignments when using latent classes as predictors in regression models, with application to acculturation strategy measures. *Epidemiology (Cambridge, Mass.)* **31,** 194.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94,** 496–509.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* **12,** 189–198.

Fraley, C. and Raftery, A. E. (2006). Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.

Gao, F. and Chan, K. C. G. (2019). Semiparametric regression analysis of length-biased interval-censored data. *Biometrics* **75,** 121–132.

Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92,** 875–891.

Guillozet, A. L., Weintraub, S., Mash, D. C., and Mesulam, M. M. (2003). Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Archives of neurology* **60,** 729–736.

Hanfelt, J. J., Peng, L., Goldstein, F. C., and Lah, J. J. (2018). Latent classes of mild cognitive impairment are associated with clinical outcomes and neuropathology: Analysis of data from the national alzheimer's coordinating center. *Neurobiology of disease* **117,** 62–71.

Hanfelt, J. J., Wuu, J., Sollinger, A. B., Greenaway, M. C., Lah, J. J., Levey, A. I., and Goldstein, F. C. (2011). An exploration of subgroups of mild cognitive impairment based on cognitive, neuropsychiatric and functional features: Analysis of data from the national alzheimer's coordinating center. *The American Journal of Geriatric Psychiatry* **19,** 940–950.

Hart, K. R., Fei, T., and Hanfelt, J. J. (2020). Scalable and robust latent trajectory class analysis using artificial likelihood. *Biometrics* .

Heinly, M. T., Greve, K. W., Bianchini, K. J., Love, J. M., and Brennan, A. (2005). Wais digit span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment* **12,** 429–444.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480.

Hendry, D. J. (2014). Data generation for the cox proportional hazards model with time-dependent covariates: a method for medical researchers. *Statistics in medicine* **33,** 436–454.

Hilton, R. P., Zheng, Y., and Serban, N. (2018). Modeling heterogeneity in healthcare utilization using massive medical claims data. *Journal of the American Statistical Association* **113,** 111–121.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79,** 531–541.

Lambert, P. C., Dickman, P. W., Weston, C. L., and Thompson, J. R. (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59,** 35–55.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics* **60,** 85–92.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80,** 741–753.

Lin, H., Scharfstein, D. O., and Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 791–813.

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. *Journal of the American Statistical Association* **97,** 53–65.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57,** 1191–1197.

Ma, H., Peng, L., Zhang, Z., and Lai, H. J. (2018). Generalized accelerated recurrence time model for multivariate recurrent event data with missing event type. *Biometrics* **74,** 954–965.

Mair, P. and Hudec, M. (2009). Multivariate weibull mixtures with proportional hazard restrictions for dwell-time-based session clustering with incomplete data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58,** 619–639.

Mao, L. and Lin, D. (2017). Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 573–587.

McLachlan, G. and McGiffin, D. (1994). On the role of finite mixture models in survival analysis. *Statistical methods in medical research* **3,** 211–226.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., Cline, J., et al. (2011). Deoptim: An r package for global optimization by differential evolution. *Journal of Statistical Software* **40,**.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–465.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S.-T., Yang, C.-C., Wang, C.-P., Kellam, S. G., Carlin, J. B., and Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3,** 459–475.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* **86,** 770–778.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology* **56,** 303–308.

Pfeffer, R. I., Kurosaki, T. T., Harrah Jr, C., Chance, J. M., and Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of gerontology* **37,** 323–329.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational statistics & data analysis* **53,** 1142–1154.

Proust-Lima, C., Philipps, V., Liquet, B., et al. (2017). Estimation of extended mixed models using latent classes and latent processes: The r package lcmm. *Journal of Statistical Software* **78,**.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research* **23,** 74–90.

Qaqish, B. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Journal of the American Statistical Association* **90,** 455–463.

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and motor skills* **8,** 271–276.

Rogalski, E., Sridhar, J., Rader, B., Martersteck, A., Chen, K., Cobia, D., Thompson, C. K., Weintraub, S., Bigio, E. H., and Mesulam, M.-M. (2016). Aphasic variant of alzheimer disease: clinical, anatomic, and genetic features. *Neurology* **87,** 1337–1343.

Rowley, M., Garmo, H., Van Hemelrijck, M., Wulaningsih, W., Grundmark, B., Zethelius, B., Hammar, N., Walldius, G., Inoue, M., Holmberg, L., et al. (2017). A latent class model for competing risks. *Statistics in medicine* **36,** 2100–2119.

Schaubel, D. E. and Cai, J. (2006). Multiple imputation methods for recurrent event data with missing event category. *Canadian Journal of Statistics* **34,** 677–692.

Scheike, T. H. and Zhang, M.-J. (2011). Analyzing competing risk data using the r timereg package. *Journal of statistical software* **38,**.

Scheike, T. H., Zhang, M. J., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95,** 205–220.

Sheikh, J. I. and Yesavage, J. A. (1986). Geriatric depression scale (gds): recent evidence and development of a shorter version. *Clinical Gerontologist: The Journal of Aging and Mental Health* .

Therneau, T. M. and Lumley, T. (2014). Package 'survival'. *Survival analysis Published on CRAN* **2,** 3.

van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis* **18,** 450–469.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society* **54,** 426–482.

Wang, C.-P., Hendricks Brown, C., and Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association* **100,** 1054–1076.

Wechsler, D. (1945). Wechsler memory scale.

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., Cummings, J., DeCarli, C., Foster, N. L., Galasko, D., et al. (2009). The alzheimer's disease

centers' uniform data set (uds): The neuropsychological test battery. *Alzheimer disease and associated disorders* **23,** 91.

Welsh, K., Butters, N., Hughes, J., Mohs, R., and Heyman, A. (1991). Detection of abnormal memory decline in mild cases of alzheimer's disease using cerad neuropsychological measures. *Archives of neurology* **48,** 278–281.

Williams, B. W., Mack, W., and Henderson, V. W. (1989). Boston naming test in alzheimer's disease. *Neuropsychologia* **27,** 1073–1079.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., Nordberg, A., Bäckman, L., Albert, M., Almkvist, O., et al. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *Journal of internal medicine* **256,** 240–246.

Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50,** 375–387.

Yee, T. W. et al. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software* **32,** 1–34.

Yin, G. and Cai, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91,** 801–818.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93,** 627–640.

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