Semiparametric Latent Variable Models for Chronic Diseases with Responses of Multiple Types and Scales

by

Yu-Pu Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2018

Doctoral Committee:

Professor Alexander Tsodikov, Chair
Professor Daniel Clauw
Associate Professor Brisa Sánchez
Professor Douglas Schaubel
Professor David Williams

Yu-Pu Chen

yupuchen@umich.edu

ORCID ID: 0000-0002-3555-5191

ACKNOWLEDGEMENTS

I would like to greatly thank my academic advisor Alex Tsodikov, for his unlimited patience, intelligence and support during my PhD study. Alex is one of the smartest and nicest person I've ever met. He has the most solid theoretical background and the best methodological ideas. Alex's enthusiasm and integrity towards research have provided a great role model that makes me a better scholar. In addition, Alex made my doctoral journey full of inspirations and joys. Alex helped me tremendously I can't possibly thank him enough. He encouraged me every time I was frustrated and sat down with me to investigate possible issues in our research. He attended all of my conference presentations and discussed audience's comments and questions with me. The day before my dissertation proposal was his birthday, but he spent four hours on his birthday listening to my practice talk and made my presentation awesome! Alex made me a better researcher, statistician, and collaborator. I am extremely fortunate to have Alex as my academic advisor and I would like to keep doing research with Alex after I graduate.

I would like to express my great appreciation to my dissertation committee members, Daniel Clauw, Brisa Sánchez, Douglas Schaubel, and David Williams, for contributing many insightful and constructive suggestions toward the dissertation. Thanks to Suzie As-Sanie, Daniel Clauw and David Williams for providing pelvic pain patient data in Chapter III and IV, and providing valuable comments in pain centrality construction. I would like to sincerely thank Douglas Schaubel, for his kindness in spending precious time to meet with me and providing valuable suggestions long before I was admitted to the PhD program. I took two courses taught by Doug, Linear Regression and Generalized

Linear Models. These two courses are the best courses I've ever had. Doug established solid foundation for my knowledge of linear and nonlinear models that not only aspires me to pursue PhD research in biostatistics, but also greatly facilitates my theoretical and applied work in both research and GSRA/consultation work. Thanks to Doug for being my friend along these years. Thanks to Lu Wang for teaching me probability and distribution theory, thanks to Jack Kalbfleisch for teaching me survival analysis, thanks to Tom Braun for teaching me nonparametric statistics, thanks to Ananda Sen for teaching me longitudinal data analysis, these are all best courses I've ever had and tremendously established the foundation of my statistical knowledge. The Biostatistics Department at the University of Michigan has awesome faculty, countless research opportunities, healthy environment, and the most positive and encouraging atmosphere. I am so fortunate to pursue my doctoral research here.

I would like to gratefully thank Consulting for Statistics, Computing and Analytics Research (CSCAR) and all the colleagues. I worked as a graduate student consultant at CSCAR for three years during my PhD years. This is one of the most valuable experience I have in my life. I was exposed to a wide variety of statistical methods and collaborated with researchers from various disciplines. CSCAR made me a great collaborator, statistician, and consultant. I established long term relationship with researchers of other fields and had great collaborative publications. I learned a great deal of communication with people of different fields. Thanks to Kerby Shedden, for your unlimited patience, broad knowledge, and best communication skills. Kerby also provided valuable career advice. I obtained helpful suggestions every time I talked with Kerby during the time I was looking for a job. I feel I know better what I should do after talking to him. Thanks to Brenda Gillespie, for her help in survival analysis, longitudinal studies, clinical trial and many valuable inputs. Thanks Brenda for introducing a great collaborator, MaryCarol Hunter, from School of Natural Resources & Environment. I learned so much from the nature pill project. And I gained two sincere friends, Brenda and MaryCarol. The weekly staff

meeting at CSCAR is very informative and I really love the environment that all talented CSCAR consultants gather together to discuss and help each other on current research questions encountered during consultation. My colleagues at CSCAR are so intelligent and creative. There are too much I can learn from them. Thanks to my wonderful colleagues, Yumeng Li, Josh Errickson, Chris Andrew, Brady West, Corey Powell, Shyamala Nagaraj, Myra Kim, Michael Clark, Kim Ward, Marcio, Manish Verma, Alex Giessing, Alex Cao, Alex Gaenko. Thanks to Dr. John Nicklas, Dr. Robert O'Rourke, Jonathan Morris, Eric Krawczyk, for being a great collaborator and long-term client. The training and experience at CSCAR got me an offer from my dream job.

I would like to express my great appreciation to former GSRA advisor, Niko Kaciroti, Julie Lumeng, and Alison Miller, for their supports and guidance in my early PhD study. They provided me great collaborative experience on meaningful medical researches dedicated to understand various biological and behavioral mechanisms of obesity among preschool children. The high-quality collaborative work and publications made profound impacts in researches on early age children health.

I would like to show my deepest appreciation to all my fellow students and friends, Keng-Han Lin, Hsuan-Jung Chen, Wan-Chu Chang, Kam Chung Wong, Yu-Jui Huang, Xi Lu, Chenghao Chu (Indiana University), Kin Yau Wong (UNC), Andrew Raim (US Census Bureau), Sai H Dharmarajan, Qui Tran, Lili Wang, Yebin Tao, Yun-Jhong Wu for providing me with both invaluable moral support as well as academic assistance with my research, for making statistical conference so much fun, for sharing wonderful memories with me, and making my life here so colorful and enjoyable.

Thanks to my parents, Liang-An Chen and Mei-Hua Lin, and my brother, Yu-Chi Chen, for your endless love, belief, and support. Thanks to my bear, BeiBei, for accompanying me all the way since I was 9 years old.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

In chronic diseases, research often centers on discovering a latent trait trajectory that manifests itself through multiple response variables on different measurement scales. In longitudinal studies, it is common to collect multivariate response data consisting of mixtures of continuous, survival, ordinal, count and multinomial variables. Development of the methodology was motivated by situations when measuring and predicting the latent trait can provide important insights for managing the observed phenotype.

In Chapter II, we study survival models of cancer where a latent trait is responsible for the cure process. Traditional cure models assume that the cure status is determined at the beginning of the follow up. However, patients often receive treatments during the follow up time that may affect their chance of cure. We propose a dynamic joint cure model where a cure process is affected by time-dependent covariates. Therapeutic interventions and prognostic factors can follow two causal paths affecting survival directly or through the latent cure process.

Chapter III addresses the challenge of latent trait measurement through multiple outcomes of different scales, which are often collected when the construct of interest cannot be measured directly. We proposed a shared latent variable model where a logistic link is used to accommodate nonparametrically transformed continuous, ordinal, count, multinomial and survival outcomes. The proposed model avoids restrictive normality assumptions and allows for negative correlation among outcomes. The model provides a subject-specific measure of the latent trait.

Chapter IV extends the method of Chapter III to allow for longitudinal responses of mixed types. We proposed a joint modeling approach for nonparametrically transformed

multivariate longitudinal responses of mixed scales. Multivariate longitudinal responses of mixed continuous, ordinal, count and multinomial outcomes and a time-to-event outcome are linked through a shared latent trait trajectory measurement. The model is used to provide a subject-specific measure of the latent trait trajectory through multiple correlated responses observed repeatedly on the subject.

CHAPTER I

Introduction

In chronic diseases, research often centers on discovering a latent trait trajectory that manifests itself through multiple response variables on different measurement scales. In longitudinal studies, it is common to collect multivariate response data consisting of mixtures of continuous, survival, ordinal, count and multinomial variables. Development of the methodology was motivated by situations when measuring and predicting the latent trait can provide important insights for managing the observed phenotype. Joint models are commonly used to provide an efficient and flexible framework to model correlated longitudinal and survival data. A useful feature of a joint model is that it provides subject-specific latent trait trajectories and enables survival risk predictions. Motivated by the need to develop a general modeling framework for longitudinal responses of mixed types and survival times, we proposed a semiparametric joint model for survival outcomes with a latent dynamic cure process in Chapter II. In addition, we proposed a semiparametric shared latent trait joint model for cross-sectional and longitudinal observed outcomes of mixed types in Chapter III and Chapter IV, respectively.

In Chapter II, we study survival models of cancer where a latent trait is responsible for the cure process. Cure with time-to-event data refers to an unobserved event when subject is no longer at risk of death from the disease of interest. Two major classes of cure models were developed to allow for a subgroup of non-susceptible subjects. One class is two-

component mixture cure models that explicitly model survival as a mixture of cured and susceptible patients (*Berkson and Gage* (1952); *Farewell* (1982); *Kuk and Chen* (1992); *Peng and Dear* (2000); *Sy and Taylor* (2000); *Li and Taylor* (2002); *Othus et al.* (2012); *Wang et al.* (2012)). The other class of cure models is a Cox proportional hazards model that allows for a cure fraction (*Tsodikov* (1998); *Broet et al.* (2001); *Tsodikov* (2002); *Tsodikov et al.* (2003); *Chen et al.* (1999); *Yin and Ibrahim* (2005)). This class of model is also referred to as non-mixture cure model or promotion time cure model. Current work on cure models assume that, although unobserved, the cure status is determined at the beginning of the follow up ($t = 0$). However, patients often receive treatments during the follow up time that may affect their chance of cure. Also, the event of cure may not be an immediate consequence of treatment and may include a period when the immune system struggles to achieve it. In this chapter, we propose a dynamic joint cure model where a cure process is affected by time-dependent covariates. The proposed model considers cure as an outcome (a stopping point) of a latent stochastic process as it touches an absorbing boundary at zero. Fundamental to the model is the mechanistic competing nature of cure and failure, with time to cure representing a latent competing risk. As mentioned in *Fine and Gray* (1999), time to cure is unobservable, so the estimation of overall survival is tantamount to estimation of the subdistribution for failure. At the latent level the events of cure and failure are competing. The latent cure event intercepts the failure process and vice versa. We model the time to cure and time to failure with the proportional hazard model. Two separate baseline hazards are estimated nonparametrically in the model to allow different time scales for the time to cure and time to failure processes. The model can easily be extended to other link functions, if needed. The proposed model is a new class of cure models that allows the cure rate to change over time by introducing time-dependent covariates into the cure process. Therapeutic interventions and prognostic factors can follow two causal pathways affecting survival directly or through the latent cure process. The proposed model is applied to study the effect of secondary cancer on

primary prostate cancer-specific survival using data from the SEER program. The joint modeling approach allows us to make subject-specific prediction of patient's cure process given the current follow-up time and the trajectory of patient's characteristics.

There are previous works on promotion time cure model (2.1) that allow the cure probability to change instantaneously at a pre-specified time point. For example, *Tsodikov* (1998) used a promotion time cure model to study leukemia induced by the primary and relapse treatments of Hodgkin's disease (HD). The proposed model specifies a time-dependent covariate $\theta(z(t)) = \exp(\beta_0 - \beta_1 \mathbb{1}(t \geq T_R))$, where T_R is the time of HD relapse. In this case, the probability of cure changes instantaneously at T_R. The difference between our dynamic cure model proposed in Chapter II and the model used in *Tsodikov* (1998) is that our dynamic cure model models cure as a **process**. Cancer treatment initiates a cure process that may or may not result in cure in the follow up. On the contrary, the time-dependent promotion time cure model assumes cure happens immediately at the time of the treatment. Considering the biological mechanism of how a medical treatment functions on a human body, it is more reasonable to consider cure as a process instead of an instantaneous event.

Chapter III addresses the challenge of latent trait measurement through multiple outcomes of mixed categorical and continuous types. Multiple outcomes of different scales are often collected when the construct of interest cannot be measured directly. A popular approach to latent variable models maps observed continuous and ordinal outcomes to underlying Gaussian continuous responses and is limited to mixed continuous and ordinal outcomes (*Muthén* (1984); *Shi and Lee* (2000); *Murray et al.* (2013); *Lin et al.* (2014); *Snavely et al.* (2014)). This approach does not accommodate nominal scale outcomes. Another popular approach is the parametric generalized linear models (*Sammel et al.* (1997); *Moustaki and Knott* (2000); *Dunson and Herring* (2005); *Skrondal and Rabe-Hesketh* (2004)). This class of approach explicitly specifies transformation function for the measurable outcomes of different scales and introduce dependence among mixed out-

comes through shared latent variables. We propose a shared latent variable model where a logistic link is used to accommodate nonparametrically transformed continuous, ordinal, count, multinomial and survival outcomes. The model is used to provide a subject-specific measure of the latent trait, given the information observed on the subject. The modeling framework is generic with respect to the parametric distribution assumed for the trait. The proposed model avoids restrictive normality assumptions and allows for negative correlation among outcomes. We developed the model under the case of univariate and multivariate measurable outcomes. An EM-DCA algorithm is developed to estimate the nonparametric transformation functions for each observed outcome. Covariates are modeled parametrically and their effects are estimated using the profile likelihood. In the univariate model, the logistic link provides closed form conditional expectations that yields computational efficient estimating procedure. The proposed method is applied to measure the pain centrality trait of patients undergoing hysterectomy as a treatment for pelvic pain and to explain the heterogeneity of patients reported outcomes. The method is compared with the ad-hoc 2011 Fibromyalgia (FM) Survey Criteria instrument designed to characterize a similar construct.

In Chapter IV, we extend the methods of Chapter III to allow for longitudinal responses of mixed types. Multidimensional longitudinal data of mixed types are collected to fully explore the latent trait trajectory that is often of main interest but cannot be measured directly. In addition, time-to-event data is often considered if the occurrence of the terminal event is dependent on the latent trait of interest. Statistical approaches were developed to jointly modeling longitudinal responses of mixed scales and the event time data to improve inference for latent trait trajectory, and to account for the dependency the two correlated processes. The previous work on the joint model of multivariate longitudinal responses either maps the discrete outcomes to latent continuous variables (*Gueorguieva and Sanacora* (2006)) or rely on explicitly specified link functions based on exponential family (*Dunson* (2003); *Jaffa et al.* (2016)). Further, there is not a single

longitudinal model that accommodates all continuous, ordinal, count and multinomial outcome types. As for the joint model between longitudinal and survival outcomes, there is no single joint model that allows for nonparametric transformation of the longitudinal outcomes, and no single model accommodate all continuous, ordinal, count and multinomial outcome types. Motivated by the needs to develop a general statistical framework for longitudinal responses of mixed types and survival times, we proposed a flexible joint modeling approach for nonparametrically transformed multivariate longitudinal responses of mixed scales. Multivariate longitudinal responses of mixed continuous, ordinal, count and multinomial outcomes and a time-to-event outcome are linked through a shared latent trait trajectory measurement. The model is used to provide a subject-specific measure of the latent trait trajectory through multiple correlated responses observed repeatedly on the subject. An EM-DCA algorithm is developed to estimate the nonparametric transformation functions for each response and to estimate the population average latent trait trajectory. The maximum likelihood estimators are consistent and asymptotically normal. The proposed method is applied to measure pain centrality trajectory of pelvic pain patients undergoing hysterectomy. For each of the patient, multiple pain-related responses of different scales were collected longitudinally. The model-based centrality trajectory is more closely aligned to longitudinal pain-related outcomes compared with the Fibromyalgia (FM) Survey Criteria trajectory across all the time points.

Overall, the dissertation provides a statistical framework for joint modeling multiple outcomes of a variety of scales, to assess the effect of dynamic factor on the latent cure process, and to provide prediction of subject-specific latent trait of interest, with the scientific goal of understanding the heterogeneity among study population. These methods can potentially be useful in areas of medical research, psychological research and social research. We hope the application of our work can lead to a more effective subject-specific latent trait construction and a better understanding in fundamental differences among study subjects.

CHAPTER II

A Semiparametric Joint Survival Model with A Time-Dependent Cure Process

2.1 Introduction

Due to advances in modern medical practice and therapy, a substantial proportion of patients may never experience the event of interest even after extended follow up. The improvement in cause-specific survival may result in heavy censoring at the end of the follow-up period. The need to account for long-term survivors has led to the development of cure models. Cure models have the flexibility to estimate the cure rate and at the same time incorporate non-proportional effects into the model through short-term and long-term effects.

Two major classes of cure models were developed to allow for a subgroup of non-susceptible subjects. One class is two-component mixture cure models that explicitly model survival as a mixture of cured and susceptible patients

$$S(t|\mathbf{z},\ \mathbf{x}) = p(\mathbf{x})S_0(t|\mathbf{z}) + ((1 - p(\mathbf{x})).$$

Berkson and Gage (1952) first introduced the idea of two-component mixture cure model in which the probability of cure $p(\mathbf{x})$ was assumed an unknown constant and the survival function for noncured patients, $S_0(t|\mathbf{z})$, was assumed to follow a parametric form. *Farewell*

(1982) extended the model to allow covariate effects in the cure probability $p(\mathbf{x})$ through a logistic regression. *Kuk and Chen* (1992), *Peng and Dear* (2000), *Sy and Taylor* (2000), *Li and Taylor* (2002), *Peng* (2003), *Lu* (2010), *Zhang et al.* (2013) further proposed semiparametric models for the susceptible survival function $S_0(t|\mathbf{z})$. *Othus et al.* (2009) allowed for dependent censoring. Most recently, *Othus et al.* (2012) proposed a change-point mixture cure model that allows hazard rate and cure rate to jump at an unknown covariate value. *Wang et al.* (2012) proposed a model with nonparametric forms for both the cure probability and the hazard rate function.

The other class of cure models is a Cox proportional hazards model that allows for a cure fraction *Tsodikov* (1998),

$$S(t|\mathbf{x}) = \exp[-\theta(\mathbf{x})F(t)], \tag{2.1}$$

where $F(t)$ has the form of a cumulative distribution function and θ incorporates an intercept term. The model was first proposed as a mechanistic promotion time model *Yakovlev and Tsodikov* (1996), motivated by biological processes associated with development of cancer. Therefore this model is also known as promotion time cure model. The model was extended by *Broet et al.* (2001); *Tsodikov* (2002); *Tsodikov et al.* (2003) to allow the latent distribution $F(t)$ to be dependent on covariates. Bayesian formulations were also considered *Chen et al.* (1999); *Yin and Ibrahim* (2005). Note that if $\theta(x) = \exp(\beta x)$, then β represents the log hazard ratio. The hazard function is $\lambda(t|x) = \exp(\beta x)f(t)$. Therefore, the ratio of hazard for one unit increase in x is $\frac{\lambda(t|x+1)}{\lambda(t|x)} = \frac{\exp(\beta(x+1))f(t)}{\exp(\beta x)f(t)} = \exp(\beta)$. Thus, promotion time cure model is a proportional hazard survival model.

Both classes of cure models can be interpreted as part of the univariate frailty model family $S(t|\mathbf{x}) = \mathbb{E}\left[e^{-VH(t)}\right]$, where H is a cumulative hazard, and $V \sim P(t|\mathbf{x})$ is a non-negative frailty random variable whose distribution P has a mass at zero (*Tsodikov et al.* (2003)). There is an extensive literature on developing cure models exploiting different

frailty schemes (*Cooner et al.* (2007)). When V is a binary variable, $S(t|\mathbf{x})$ recovers a two-component mixture cure model. When V follows a Poisson distribution with mean θ, $S(t|\mathbf{x})$ recovers a promotion time cure model.

It should be noted that the standard Cox model just like most other semi-parametric models naturally incorporates cure. Indeed, in semi-parametric models, the baseline survival function is arbitrary, and this includes functions that plateau with time. With the exception of some rank-based methods that explicitly rely on the zero tail defect assumption (Tsodikov (1998)), semiparametric MLE theory remains valid regardless of cure. Exposing the cure rate as an explicit parameter in the model is a matter of reparameterization that keeps the MLEs invariant. Semiparametric models in their non-cure form provide as estimate of the probability of cure as the last value of the predicted survival function. While initially specialized cure models may have been developed out of underappreciation for the generality of semi-parametric survival models, they opened the field to consideration of mechanistic models of heterogeneity and a more meaningful interpretation of the data. In most of the previous work on cure models, including the above formulations, it is assumed that the cure status is determined, if unknown, at the beginning of the follow up ($t = 0$). In practice though, patients often receive treatments or experience intermediate events during the follow up. In this case it is natural to expect the chance of cure to change in response to dynamic factors. To account for this situation, one could consider models where survival times are based on stopping times for some stochastic processes. A number of so-called first hit models (FHT) were proposed that define the failure as a stochastic process reaching a boundary *Lee and Whitmore* (2006). The event of cure appeared as an incidental finding when the process never reaches a boundary with non-zero probability under certain conditions ensuring that the process drifts away from boundary. The idea was operationalized by (*Balka et al.*, 2009) using a Wiener process. It it possible, within this framework, to have time-dependent covariates affecting the process and the associated stopping time properties.

In this paper, we propose a framework where cure is an outcome (a stopping point) of a latent stochastic process as it touches an absorbing boundary at zero. Fundamental to the model is the mechanistic competing nature of cure and failure, with time to cure representing a latent competing risk. The latent cure event intercepts the failure process and vice versa. We model the time to cure and time to failure with the proportional hazard model. Two separate baseline hazards are estimated nonparametrically in the model to allow different time scales for the time to cure and time to failure processes. The model can easily be extended to other link functions, if needed. The proposed model is a new class of cure models that allows the cure rate to change over time by introducing time-dependent covariates into the cure process. Therapeutic interventions and prognostic factors can follow two causal pathways affecting survival directly or through the latent cure process.

Asymptotic properties are established using empirical process (*Kosorok* (2008)) and martingale theory (*Andersen et al.* (1993)) with details in the Appendix A.6.

The paper is organized as follows. In section 2.2, we describe the framework of the proposed cure model. Section 2.3 describes the likelihood and the corresponding martingale properties, as well as the EM algorithm. Asymptotic theory is presented in Section 2.4. Section 2.5 provides simulation results. Section 2.6 gives an example of real data analysis. In Section 2.7 we develop a prediction for the probability of cure over time, given observed information. Section 2.9 provides conclusions and discussion.

2.2 Statistical Framework

Consider two time to event processes, time to failure and time to cure. We observe a failure event if time to cure is longer than time to failure. However, time to cure is not observed directly. If time to cure precedes the time to failure, we would eventually observe a censored event. With this in mind, the observed failure time for a subject, denoted by

T, can be defined as

$$T = \mathbb{1}(T^* < T_u)T^* + \mathbb{1}(T^* \geq T_u)\infty,$$

where $T^* < \infty$, denotes the potential time to failure in the absence of cure; T^* is only observed if failure precedes cure; T_u is the potential time to cure; $\mathbb{1}(\cdot)$ is an indicator function taking value of 1 if $(\cdot)$ is true, and 0 otherwise. $P(T^* = T_u) = 0$ for continuous random variables. Let $\mathbf{z}(t)$ be a set of possibly time-dependent fully observed covariates associated with T and T_u, and let $\overline{\mathbf{z}}(t) = \{\mathbf{z}(s), s \leq t\}$ denote the covariate path associated with $\mathbf{z}(\cdot)$. Let C be the censoring time which is independent of (T^*, T_u), given $\mathbf{z}(t)$. Define $X = \min(T, C)$ and $\delta = \mathbb{1}(T \leq C)$. Cure, once it happens, is irreversible. Cured subjects are always censored but some censored subjects may experience failures beyond their follow-up period. Because the cure event is unobserved, the survival function is a conditional average over the cure process as explained in the next section.

2.2.1 Model Specification

Define a conditional hazard function for the failure event (given the cure process) as a non-negative stochastic process $d\Lambda_T = d\mathcal{U}(t|\mathbf{z}(t))$ with an absorbing boundary of 0. When cure event happens, the random process $d\mathcal{U}$ touches 0 and remains at 0 thereafter. So cure is defined as a stopping point of the process $d\mathcal{U}$.

The observed (marginal) survival function $S(t)$ can be obtained by taking the expectation over the trajectory $\overline{\mathcal{U}}(t)$ of the stochastic process $\mathcal{U}$ from time 0 to t, that is,

$$S(t|\overline{\mathbf{z}}(t)) = \mathbb{E}\left[e^{-\int_0^t d\mathcal{U}(s)}\right]. \tag{2.2}$$

The marginal density function is then $f(t) = \mathbb{E}\left[d\mathcal{U}(t)e^{-\int_0^t d\mathcal{U}(s)}\right]/dt$, and the marginal hazard function is a conditional expectation, given survival up to t, $\lambda(t) = \mathbb{E}\left[\mathcal{U}(t)|T > t\right] = f(t)/S(t)$, see *Gjessing et al.* (2003).

Specifically, as an example of this paper, consider a change point conditional hazard process $d\Lambda_T(t|T_u, \mathbf{z}) = d\mathcal{U}(t|\mathbf{z}) = \mathbb{1}(T_u > t)\theta(t|\mathbf{z}(t))dH_1(t)$ that models the conditional hazard function for the failure event (given time to cure T_u). The rationale behind $\mathbb{1}(T_u > t)$ is that a subject is at risk of failure only if cure event has not yet occurred. Conditional on the cure event, we assume a time-dependent Cox model for the failure time. In addition, we specify a marginal hazard function for the cure event $d\Lambda_{T_u}$ as another time-dependent Cox model. Specifically,

$$d\Lambda_{T_u}(t|\mathbf{z}(t)) = \lim_{\Delta\to 0} \frac{P(T_u \in [t, t+\Delta)|T_u \geq t, \mathbf{z}(t))}{\Delta} = \eta(t|\mathbf{z}(t))dH_2(t), \tag{2.3}$$

$$d\Lambda_T(t|T_u, \mathbf{z}(t)) = \lim_{\Delta\to 0} \frac{P(T \in [t, t+\Delta)|T \geq t, T_u, \mathbf{z}(t))}{\Delta} = \mathbb{1}(T_u > t)\theta(t|\mathbf{z}(t))dH_1(t), \tag{2.4}$$

where $\mathbf{z}(t)$ is a vector of possibly time-dependent covariates of dimension p, $\eta(t|\mathbf{z}(t)) = e^{\beta_\eta \mathbf{z}(t)}, \theta(t|\mathbf{z}(t)) = e^{\beta_\theta \mathbf{z}(t)}$ and $\beta = (\beta_\eta, \beta_\theta)$ is the combined vector of regression coefficients. The predictor η models covariate effects on the time to cure T_u and the predictor θ models covariate effects on the failure time T. For the rest of the paper, we omit $\mathbf{z}(t)$ as an argument for brevity and denote $\theta(t|\mathbf{z}(t))$ by $\theta(t)$ and $\eta(t|\mathbf{z}(t))$ by $\eta(t)$.

To account for potentially different time scales for the time to cure and the time to failure processes, separate nonparametrically specified cumulative baseline hazard functions H_1, H_2 were used. The baseline cumulative hazard function $H_1(t)$ summarizes the underlying disease progression time pattern leading to a failure event. Another unspecified cumulative baseline hazard function $H_2(t)$ summarizes the immune function process leading to a cure event.

The conditional survival function of the time to event T given T_u is

$$S(t|T_u) = e^{-\int_0^t \mathbb{1}(T_u > s)\theta(s)dH_1(s)}. \tag{2.5}$$

The conditional probability density function (*pdf*) of time to event T given T_u is

$$f(t|T_u) = \mathbb{1}(T_u > t)\theta(t)h_1(t)e^{-\int_0^t \theta(s)dH_1(s)}. \tag{2.6}$$

2.2.2 Marginal Distribution of Time to The Failure Event

According to (2.2), for our model, the marginal survival function of the time to failure event T, can be obtained by taking expectation of (2.5) over latent time to cure T_u

$$S(t) = \mathbb{E}[S(t|T_u)] = \\ e^{-\int_0^t \theta(s)dH_1(s)}e^{-\int_0^t \eta(s)dH_2(s)} + \int_0^t \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s). \tag{2.7}$$

The marginal probability density function of time to event T, can be obtained by taking expectation of (2.6):

$$f(t) = \theta(t)h_1(t)e^{-\int_0^t \theta(s)dH_1(s)}e^{-\int_0^t \eta(s)dH_2(s)}. \tag{2.8}$$

Details can be found in Appendix A.1.

The proposed model can be considered as a stochastic process frailty model with a latent stochastic process $\mathcal{V}(t) = \mathbb{1}(t < T_u)\theta(t|\mathbf{z}(t))$ acting multiplicatively on the baseline hazard of failure.

2.2.3 A Special Case: Static Cure Model

Excluding time-dependent covariates $\mathbf{z}(t)$ from the cure part of the model, and making the time to cure a degenerate improper distribution with a single mass at $t = 0$ makes it a two-component mixture model. Another way to get a two-component mixture model as a nested special case would be to impose a proportionality assumption as follows.

If the cure process intensity is proportional to failure process, that is,

$$\frac{\eta(t|\mathbf{x}, \mathbf{z}(t))dH_2(t)}{\theta(t|\mathbf{x}, \mathbf{z}(t))dH_1(t)} = a(\mathbf{x})$$

uniformly over t in the observation period, where $a(\mathbf{x})$ depends only on time-independent covariate set $\mathbf{x}$, then the model can be formally written as a static two-component mixture cure model with the logistic probability of cure $p(\mathbf{x}) = \dfrac{a(\mathbf{x})}{1+a(\mathbf{x})}$. The corresponding survival function is

$$S(t|\mathbf{x}, \overline{\mathbf{z}}(t)) = \frac{a(\mathbf{x})}{1+a(\mathbf{x})} + \frac{1}{1+a(\mathbf{x})} e^{-(1+a(\mathbf{x}))\int_0^t \theta(s|\mathbf{x}, \mathbf{z}(s))dH_1(s)}.$$

Any such model can also be represented as a promotion time cure model
$S(t|\mathbf{x}) = \exp[-\gamma(\mathbf{x})F(t|\mathbf{x}, \overline{\mathbf{z}}(t))]$, where $\gamma = -\log(a/(1+a))$, and $F = -(\log S)/\gamma$.

In general, unlike the two-component mixture model, the promotion time cure model does not loose its formal validity if all of its predictors contain time-dependent covariates, $S(t|\overline{\mathbf{z}}(t)) = \exp[-\gamma(\mathbf{z}(t))F(t|\overline{\mathbf{z}}(t))]$. However, in this context $\exp(-\gamma)$ cannot be interpreted as the probability of cure. Also, to ensure a cure model, γ may be restricted to be bounded resulting in a non-zero probability of cure regardless of the behavior of $\mathbf{z}(t)$.

2.3 Estimation

We assume each patient is subjected to random right censoring and the censoring time C is independent of T, given $\mathbf{z}(t)$. The observed time is $X = \min(T, C)$, and let $\delta = \mathbb{1}(T \leq C)$ be the censoring indicator. The observed time-to-event data for subject $i = 1, \cdots, n$ consist of i.i.d. $\{X_i, \delta_i, \mathbf{z}_i(t) : 0 < t \leq X_i\}, i = 1, ..., n$. For a subject with observed data $(X, \delta, \mathbf{z}(t))$ and unobserved time to cure T_u, the contribution to complete

data likelihood is

$$L_0(\beta|X,\delta,\mathbf{z}(\cdot),T_u) = [\mathbb{1}(T_u > X)\theta(X)dH_1(X)]^{\delta} e^{-\int_0^X \mathbb{1}(T_u>s)\theta(s)dH_1(s)}. \tag{2.9}$$

The subject's contribution to the marginal likelihood can be obtained by taking the expectation of complete data likelihood (2.9) over T_u

$$\mathcal{L}(\beta|X,\delta,\mathbf{z}(\cdot)) = \left[\theta(X)dH_1(X)e^{-\int_0^X \theta(s)dH_1(s)}e^{-\int_0^X \eta(s)dH_2(s)}\right]^{\delta}$$
$$\left[e^{-\int_0^X \theta(s)dH_1(s)}e^{-\int_0^X \eta(s)dH_2(s)} + \int_0^X \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)\right]^{1-\delta}. \tag{2.10}$$

2.3.1 Martingale Theory

In counting process notation, let $Y_i(t) = \mathbb{1}(X_i \geq t)$ be the observed at-risk process for failure in subject i, and $N_i(t) = \delta_i\,\mathbb{1}(X_i \leq t)$ be the counting process that records the number of events that occurred by time t for subject i. Define the filtration as $\mathcal{F}_{t-} = \sigma\{N_i(x), Y_i(x), \mathbf{z}_i(x) : x \in [0,t), i = 1,\cdots,n\}$, and consider the continuous (orthogonal) case where no two counting processes can jump simultaneously and the process $\mathbf{z}_i(x)$ is predictable.

Our model is semiparametric in the sense that the baseline cumulative hazard for failure and cure processes $H_1(\cdot)$ and $H_2(\cdot)$ are unspecified non-decreasing step functions with jumps dH_1 and dH_2 at the times where failure events are observed. The full parameter set is $\Omega = (\beta, H_1(\cdot), H_2(\cdot))$, where $\beta = (\beta_\theta, \beta_\eta)$ is finite-dimensional parameter vector and $H_1(\cdot)$ and $H_2(\cdot)$ are infinite-dimensional. We use the EM algorithm (*Tsodikov* (2003)) to derive estimation procedures for our joint model. Following *Taylor* (1995) and *Sy and Taylor* (2000), for our model to be identifiable, and to obtain stable MLEs, we impose a zero-tail constraint on our joint model, namely, $S(t_{(k)}|T_u > t_{(k)}) = 0$, or equivalently, $dH_1(t_{(k)}) = \infty$, where $t_{(k)}$ is the last event time. Consistency is established based on empirical processes (*Zeng and Lin* (2007); *Kosorok* (2008)). Weak convergence is proved

based on the martingale structure of the score equations elucidated by *Chen* (2009), and used by *Chen* (2010), *Chen* (2012), *Hu and Tsodikov* (2014a) and *Rice and Tsodikov* (2017).

We can write the marginal log-likelihood (2.10) in counting process form:

$$\ell(\Omega) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \left[\log \gamma_i \left(t; \beta, \overline{H}_1(t), \overline{H}_2(t) \right) + \log dH_1(t) \right] dN_i(t) \right.$$
$$\left. - Y_i(t) \gamma_i \left(t; \beta, \overline{H}_1(t), \overline{H}_2(t) \right) dH_1(t) \right\}, \quad (2.11)$$

where

$$\gamma_i = \frac{\theta(t) e^{-\int_0^t \theta(s) dH_1(s)} e^{-\int_0^t \eta(s) dH_2(s)}}{e^{-\int_0^t \theta(s) dH_1(s)} e^{-\int_0^t \eta(s) dH_2(s)} + \int_0^t \eta(s) e^{-\int_0^s \theta(y) dH_1(y)} e^{-\int_0^s \eta(y) dH_2(y)} dH_2(s)}$$

and τ is the maximum follow-up time in the study such that $\tau = \inf\{t : P(T > t) = 0\}$. $\overline{H}_1(t)$ and $\overline{H}_2(t)$ represent the trajectory of the hazard functions H_1 and H_2 from time 0 to time t. Under independent censoring,

$$\mathbb{E}[dN_i(t) | \mathcal{F}_{t-}] = Y_i(t) P(dN_i(t) = 1 | \mathcal{F}_{t-}) = Y_i(t) \gamma_i \left(t; \beta, \overline{H}_1(t), \overline{H}_2(t) \right) dH_1(t),$$

and the process $dM_i(t) = dN_i(t) - Y_i(t) \gamma_i \left(t; \beta, \overline{H}_1(t), \overline{H}_2(t) \right) dH_1(t)$ where we assume $\gamma \left(t; \beta, \overline{H}_1(t), \overline{H}_2(t) \right)$ is predictable, is a martingale under the true model.

2.3.2 Functional Derivatives and Score Equations

As in *Hu and Tsodikov* (2014a), for a functional $J(f), f = f(x)$, the local functional derivative at s is defined as

$$\frac{\partial J(f)}{\partial df(s)} = \left. \frac{\partial J(f + \epsilon g)}{\partial \epsilon} \right|_{\epsilon = 0, g = \mathbb{1}(x > s)}.$$

The above functional derivative corresponds to differentiating over the "jumps" of f function in both discrete and continuous cases. Detailed definition of functional derivative is described in Appendix A.3.

Denote the partial derivatives of γ_i with respect to its $dH_1(t), dH_2(t)$ and β arguments as

$$\dot{\gamma}_{i,dH_1(s)}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right) = \frac{\partial\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\partial dH_1(s)}$$
$$\dot{\gamma}_{i,dH_2(s)}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right) = \frac{\partial\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\partial dH_2(s)}$$
$$\dot{\gamma}_{i,\beta}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right) = \frac{\partial\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\partial\beta}$$

Because $\dfrac{\partial \log dH(t)}{\partial dH(s)} = \dfrac{1}{dH(t)}\dfrac{\partial dH(t)}{\partial dH(s)} = \dfrac{1}{dH(t)}\,\mathbb{1}(t<s)$, applying the functional derivative to the log-likelihood (2.11) gives the score equations for the infinite-dimensional parameters $\{dH_1(s) : U_{dH_1(s)} = 0\}$ and $\{dH_2(s) : U_{dH_2(s)} = 0\}$, uniformly over t. We obtain the score functions for dH_1 and dH_2

$$U_{dH_1(s)} = \sum_{i=1}^{n}\int_s^\tau \left[\frac{\dot{\gamma}_{i,dH_1(s)}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}dM_i(t) + \frac{dM_i(s)}{dH_1(s)}\right], \tag{2.12}$$

$$U_{dH_2(s)} = \sum_{i=1}^{n}\int_s^\tau \frac{\dot{\gamma}_{i,dH_2(s)}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}dM_i(t). \tag{2.13}$$

The score function for the finite-dimensional parameter β is

$$U_\beta = \sum_{i=1}^{n}\int_0^\tau \frac{\dot{\gamma}_{i,\beta}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}dM_i(t). \tag{2.14}$$

Given β, the iterative EM algorithm solves the estimating equations $U_{dH_1(s)} = 0$ and $U_{dH_2(s)} = 0$ uniformly over s, giving the profile likelihood of β as discussed in the next section.

2.3.3 Nonparametric Maximum Likelihood Estimation (NPMLE)

We estimate β and $\{dH_1, dH_2\}$ jointly using the profile likelihood approach. This is accomplished by applying an EM algorithm to obtain implicit estimators for hazards $\{d\hat{H}_1(\beta), d\hat{H}_2(\beta)\}$ that depend on β being held fixed. Replacing $\{dH_1, dH_2\}$ in the marginal log-likelihood $\ell(\beta, dH_1, dH_2)$ with $\{d\hat{H}_1(\beta), d\hat{H}_2(\beta)\}$ we obtain the profile log-likelihood

$$\ell_{pr}(\beta) = \ell\left(\beta, d\hat{H}_1(\beta), d\hat{H}_2(\beta)\right).$$

The estimate of $\hat{\beta}$ is obtained by maximizing the profile likelihood over a finite-dimensional Euclidean space. The derivation of the EM algorithm for our model is shown in Appendix A.4. For a single observation data (X, δ), it's contribution to the joint complete-data likelihood of time to failure and time to cure can be expressed as

$$L_0\left(\{dH_1\}, \{dH_2\} \mid X, \delta, T_u\right) f(T_u) = \\ \left[\mathbb{1}(T_u > X)\theta(X)dH_1(X)\right]^{\delta} e^{-\int_0^X \mathbb{1}(T_u > s)\theta(s)dH_1(s)} \eta(T_u) dH_2(T_u) e^{-\int_0^{T_u} \eta(x)dH_2(x)}.$$

The joint likelihood is given by L_0 marginalized over T_u. In the spirit of EM (Appendix 2.3), L_0 is parameterized by the next iteration $(k+1)$ parameters, and the expectation $\mathbb{E}\left[\sum_i \log L_{0i} \,|\, \text{Observed data}\right]$ is taken, where the expectation is parameterized by the current iteration (k) parameters. On differentiation of the result, we obtain the score functions for dH_1 and dH_2 (at iteration k+1):

$$U_{dH_1^{(k+1)}(s)}\left(dH_1^{(k)}\right) = \sum_{i=1}^{n} \left\{ \frac{dN_i(s)}{dH_1^{(k+1)}(s)} - Y_i(s)\theta_i(s)p_i^{(k)}(s) \right\}, \tag{2.15}$$

$$U_{dH_2^{(k+1)}(s)}\left(dH_2^{(k)}\right) = \sum_{i=1}^{n} \left\{ -\eta_i(s)\Psi_i^{(k)}(s) + \left[\frac{dH_2^{(k+1)}(s)}{dH_2^{(k)}} - 1 \right] \eta_i(s)\mu_i^{(k)}(s) \right\}, \tag{2.16}$$

where

$$p_i^{(k)}(s) = \left[\frac{G_{1i}^{(k)}(X_i, X_i) + G_{2i}^{(k)}(s^+, X_i)}{G_{1i}^{(k)}(X_i, X_i) + G_{2i}^{(k)}(0, X_i)}\right]^{1-\delta_i},$$

$$\Psi_i^{(k)}(s) = Y_i(s)\left[\frac{G_{1i}^{(k)}(X_i, X_i) - G_{1i}^{(k)}(s, s) + G_{2i}^{(k)}(s, X_i)}{G_{1i}^{(k)}(X_i, X_i) + G_{2i}^{(k)}(0, X_i)}\right]^{1-\delta_i},$$

$$\mu_i^{(k)}(s) = \left[\frac{Y_i(s)G_{1i}^{(k)}(s, s) + (1 - Y_i(s))G_{1i}^{(k)}(X_i, s)}{G_{1i}^{(k)}(X_i, X_i) + G_{2i}^{(k)}(0, X_i)}\right]^{1-\delta_i}\left[(1 - Y_i(s))\frac{G_{1i}^{(k)}(X_i, s)}{G_{1i}^{(k)}(X_i, X_i)}\right]^{\delta_i},$$

$$G_{1i}^{(k)}(u, v) = e^{-\int_0^u \theta_i(y)dH_1^{(k)}(y)}e^{-\int_0^v \eta_i(y)dH_2^{(k)}(y)},$$

$$G_{2i}^{(k)}(u, v) = \int_u^v \eta_i(x)e^{-\int_0^x \theta_i(y)dH_1^{(k)}(y)}e^{-\int_0^x \eta_i(y)dH_2^{(k)}(y)}dH_2^{(k)}(x).$$

Equations (2.15) and (2.16) constitute self-consistency equations that can be solved iteratively $k = 0, 1, 2, \ldots$, given a suitable initial model at $k = 0$. Setting score equations (2.15) and (2.16) to zero, we obtain the Breslow-type estimators

$$dH_1^{(k+1)}(s) = \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\theta_i(s)p_i^{(k)}(s)}, \tag{2.17}$$

$$dH_2^{(k+1)}(s) = \frac{\left[\sum_{i=1}^n \eta_i(s)\mu_i^{(k)}(s)\right]dH_2^{(k)}(s)}{\sum_{i=1}^n \eta_i(s)\left[\mu_i^{(k)}(s) + \Psi_i^{(k)}(s)\right]} \tag{2.18}$$

Note that $p_i^{(k)}(s)$ may be thought of as the imputed subject-specific probability of staying uncured for subjects in the risk set at time s. The numerator of equation (2.18) may be thought of as the imputed subject-specific probability of cure happening exactly at time s for subjects at risk. At the convergence when $dH_2^{(k+1)}(s) = dH_2^{(k)}(s)$, the second term in (2.16) disappears, and we have $\sum_{i=1}^n \eta_i(s)\Psi_i^{(k)}(s) = 0$. The term $\Psi_i^{(k)}(s)$ can be interpreted as the imputed probability of failing in the future $[s, X_i]$ for censored subjects in the risk set at time s, and this term is equal to 1 for failed subjects in the risk set. We can also

observe the following self-consistency principle. The solution to $\sum_{i=1}^{n} \eta_i(s)\Psi_i^{(k)}(s) = 0$ is achieved if for the risk set at time s, the sum of cure intensities $\eta_i(s)$ for failed subjects is equal to the sum of imputed cure intensities among censored subjects weighted by the imputed probability of failing in the future $[s, X_i]$.

By replacing $dH_1(s)$ and $dH_2(s)$ in the log-likelihood by the point of convergence $\{d\hat{H}_1(\beta), d\hat{H}_2(\beta)\}$ of the above EM algorithm, we obtain the profile log-likelihood $\ell_{pr}(\beta) = \ell\left(\beta, d\hat{H}_1(\beta), d\hat{H}_2(\beta)\right)$. Asymptotic properties of the NPMLE estimators are established in Section 2.4 and in Appendix A.6.

2.3.4 Estimation Procedure

The estimation procedure consists of two nested parts, maximize the likelihood over $H(\beta)$, given β (inner loop), and maximize the profile log-likelihood over β (outer loop). Specifically, we proceed with the following procedure for estimation.

Part 1. Maximize the likelihood over $H(\beta)$, given β:

(1) Set $k = 0$. Initialize $d\hat{H}_1^{(k)}(s)$ as Nelson-Aalen estimates. $d\hat{H}_2^{(k)}(s)$ is initialized as $d\hat{H}_1^{(k)}(s)/1000$.

(2) Given β fixed, calculate $d\hat{H}_1^{(k+1)}(s)$ and $d\hat{H}_2^{(k+1)}(s)$ using (2.15) and (2.16).

(3) Keep updating $d\hat{H}_1^{(k+1)}(s)$ and $d\hat{H}_2^{(k+1)}(s)$ as in previous step until convergence $\left\|d\hat{H}_1^{(k+1)}(s) - d\hat{H}_1^{(k)}(s)\right\|^2 < \epsilon$ and $\left\|d\hat{H}_2^{(k+1)}(s) - d\hat{H}_2^{(k)}(s)\right\|^2 < \epsilon$.

Part 2. Maximize the profile likelihood $\ell_{pr}(\beta) = \ell\left(\beta, d\hat{H}_1(\beta), d\hat{H}_2(\beta)\right)$ over β:

(1) Set $j = 0$. Set $\hat{\beta}^{(j)} = 0$ to start.

(2) Find $\hat{\beta}^{(j+1)}$ by taking one step toward maximizing the profile likelihood with respect to β using a general optimization routine.

(3) This step is nested within (2). Update $d\hat{H}_1^{(j+1)}(t) = d\hat{H}_1^{(j+1)}(\beta^{(j+1)})(t)$ and $d\hat{H}_2^{(j+1)}(t) = d\hat{H}_2^{(j+1)}(\beta^{(j+1)})(t)$ using steps in Part 1.

(4) Repeat steps (2) and (3) until convergence $\left\|\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}\right\|^2 < 10\epsilon$.

Note that Part 1 steps represent the inner loop nested within Part 2 step (2). The convergence tolerance for inner loop (Part 1) has to be stricter than for the outer loop in Part 2.

2.4 Asymptotic Properties

The proposed NPMLE is shown to be consistent and asymptotically normal by making use of the empirical process (*Kosorok* (2008); *Van Der Vaart and Wellner* (1996)) and martingale theory following a general line of *Zeng and Lin* (2007, 2010), *Chen* (2009, 2010), *Hu and Tsodikov* (2014b,a), *Rice and Tsodikov* (2017)). Regularity conditions are listed in Appendix A.6.

By integrating the score functions (2.12) and (2.13) over time s, we obtain the alternative form of the score functions for cumulative baseline hazards $H_1(s)$ and $H_2(s)$ in martingale form:

$$U_{H_1(s)} = \sum_{i=1}^{n} \int_0^{\tau} \left[\frac{\dot{\gamma}_{i,dH_1}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)} H_1(t\wedge s) + \mathbb{1}(t<s) \right] dM_i(t), \tag{2.19}$$

$$U_{H_2(s)} = \sum_{i=1}^{n} \int_0^{\tau} \frac{\dot{\gamma}_{i,dH_2}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)} H_2(t\wedge s) dM_i(t) \tag{2.20}$$

Define $\epsilon_{1i}(t,s;H_1,H_2,\beta) = \dfrac{\dot{\gamma}_{i,dH_1}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)} H_1(t\wedge s) + \mathbb{1}(t<s)$ and $\epsilon_{2i}(t,s;H_1,H_2,\beta) = \dfrac{\dot{\gamma}_{i,dH_2}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)} H_2(t\wedge s)$. As we show in Appendix A.5, since $\epsilon_{ki}(t,s;H_1,H_2,\beta), k=1,2$, does not depend on s for $t<s$, the linear transform $\int_0^{\tau} \epsilon_{ki}(t,s;H_1,H_2,\beta) dM_i(t), k=1,2$, is a martingale.

As for the score function for β (2.14), since $\dfrac{\dot{\gamma}_{i,\beta}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}$ is a predictable process, the linear transform $\int_0^{\tau} \dfrac{\dot{\gamma}_{i,\beta}\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)} dM_i(t)$ is also a martingale. There-

fore, under the true model the score functions for $H_1(s), H_2(s)$ and for β are martingales.

The following propositions present the consistency and weak convergence for the proposed NPMLE $\hat{\Omega} = \left(\hat{\beta}, \{d\hat{H}_1\}, \{d\hat{H}_2\}\right)$ with details given in Appendix A.6.

Proposition II.1. *Let β^0 and $H^0(t) = (H_1^0(t), H_2^0(t))$ be the true values of $\hat{\beta}$ and $\hat{H}(t) = \left(\hat{H}_1(t), \hat{H}_2(t)\right)$, respectively. Assuming regularity conditions hold, then with probability one, $\hat{\beta}$ converges to β^0, $\hat{H}(t)$ converges to $H^0(t)$ uniformly in the interval $[0, \tau]$.*

Consider a linear functional of the NPMLE $\hat{\Omega}$

$$n^{1/2}\left\{a^T(\hat{\beta} - \beta^0) + \int_0^\tau b(t)^T d\left(\hat{H}(t) - H(t)^0\right)\right\}, \tag{2.21}$$

where a is real vector, $b(t) = \{b_1(t), b_2(t)\}$ is in $BV[0,\tau] \times BV[0,\tau]$, where $BV[0,\tau]$ is the space of functions with bounded total variation in $[0,\tau]$. Let $B^T = (B_1^T, B_2^T)$, and $\mathcal{E}^T = (a^T, B^T)$, where B_k is the vector consisting of the values of $b_k(t)$ evaluated at the observed failure times corresponding to the jumps of $\hat{H}_k$, and $\{d\hat{H}_k\}$ is the vector of jump sizes at the observed failure times, for k=1,2, respectively.

Proposition II.2. *Assuming regularity conditions hold, $n^{1/2}\{\hat{\beta} - \beta^0, \hat{H}(t) - H^0(t)\}$ converges weakly to a zero-mean Gaussian process. In addition, the linear functional (2.21) converges weakly to a zero-mean Gaussian process with variance-covariance matrix $\mathcal{E}^T(\mathcal{I}^0)^{-1}\mathcal{E}$ which can be consistently estimated by $n\mathcal{E}^T(\mathcal{I}_n)^{-1}\mathcal{E}$, where $\mathcal{I}_n$ is the negative Hessian matrix of the observed log-likelihood function with respect to $\hat{\Omega} = \left(\hat{\beta}, \{d\hat{H}_1\}, \{d\hat{H}_2\}\right)$ and $\mathcal{E}^T = (a^T, B^T)$.*

For a Hadamard differentiable functional $F(\Omega)$ of Ω, based on the functional delta method (*Andersen et al.* (1993) Section II.8), $n^{1/2}\{F(\hat{\Omega}) - F(\Omega)\}$ converges weakly to a zero-mean Gaussian process with variance-covariance function $\dot{F}(\Omega)^T(\mathcal{I}^0)^{-1}\dot{F}(\Omega)$, where $\dot{F}(\Omega)$ is the gradient of $F(\Omega)$ with respect to Ω. The information operator $\mathcal{I}^0$ can be consistently estimated by $n^{-1}\mathcal{I}_n$. The observed information matrix $\mathcal{I}_n$ is obtained by

evaluating the negative Hessian matrix $\mathcal{I}$ at $\hat{\Omega}$. The explicit expression of $\mathcal{I}_n$ is derived in Appendix A.6.

Proposition II.3. *Assuming regularity conditions hold, the variance-covariance matrix of β from profile likelihood is a consistent estimator of true variance-covariance matrix of β.*

The variance-covariance matrix of β from profile likelihood can be obtained from inverse profile information matrix. We proved Proposition II.3 in Appendix A.6.3.

2.5 Simulation Study

To examine the finite-sample performance of the parameter estimates obtained by the proposed model, we conducted a Monte Carlo simulation study. We generated the survival times T from the marginal survival function with settings as follows.

Table 2.1: Simulation results using the proposed method. β_1 describes the effect of time-dependent covariates $z_1(t) = \mathbb{1}(t > v_1)$ on failure process. β_3 describes the effect of time-dependent covariates $z_2(t) = \mathbb{1}(t > v_2)$ on cure process. β_2 and β_4 describe the effect of time-independent covariate $z_3 \sim B(0.5)$ on the failure and cure process, respectively. v_1 and v_2 are simulated from exponential distribution for each subject. The results are based on 500 simulated datasets with sample size of $n = 300$ and $n = 500$.

N	Process	β	Truth	Avg. est.	ESD	ASE	95% CP
300	failure	β_1	1	1.01	0.23	0.21	0.93
		β_2	-0.5	-0.45	0.23	0.22	0.91
	cure	β_3	1	1.13	0.42	0.38	0.95
		β_4	0.5	0.62	0.33	0.30	0.91
500	failure	β_1	1	0.97	0.17	0.16	0.93
		β_2	-0.5	-0.48	0.18	0.18	0.94
	cure	β_3	1	1.10	0.30	0.31	0.95
		β_4	0.5	0.55	0.25	0.24	0.94

Avg. est.: average of Monte Carlo estimates of the true parameter values over the 500 simulations
ESD: empirical standard deviation based on Monte Carlo estimates
ASE: average of numerically estimated standard errors
95% CP: 95% coverage probability

The true baseline cumulative hazard for failure process is specified as $H_1(t) = t^2/2$

and the true baseline cumulative hazard for cure process is specified as $H_2(t) = t^2/4$. The true failure intensity is $\theta(t|\mathbf{z}(t)) = e^{\beta_1 z_1(t)+\beta_2 z_3}$ and the true cure intensity is $\eta(t|\mathbf{z}(t)) = e^{\beta_3 z_2(t)+\beta_4 z_3}$. For simplicity, we consider two change-point binary covariates, $z_1(t) = \mathbb{1}(t > v_1)$ that changes its value at time v_1, and $z_2(t) = \mathbb{1}(t > v_2)$ that changes its value at time v_2 as time-dependent covariates. The covariate times v_1 and v_2 are both generated from an exponential distribution with rate 1. Additionally, we consider a time-independent binary covariate $z_3 \sim B(0.5)$. The true parameters were $\beta_1 = \beta_3 = 1$,$\beta_2 = -0.5$, and $\beta_4 = 0.5$. Censoring is simulated from a uniform distribution $U(0, 3)$, which yields 50% of censoring. We examined the performance of estimation for the proposed model under the sample size of $n = 300$ and $n = 500$; each was replicated 500 times. Initial values were chosen to be Nelson-Aalen estimates for the two baseline hazards and set to be 0 for all βs. Standard errors were obtained from the numerically evaluated Hessian matrix at the solution.

The results of the simulation study are summarized in Table 2.1. The proposed estimation and inference procedures perform well with diminishing bias as sample sizes increases, and coverage probability approaching 95% nominal level. With the larger sample size, we see better agreement between empirical standard deviations and asymptotic standard errors. This suggests that the asymptotic approximation of the covariance matrix for the profile likelihood is reasonable for the sample size of $n \sim 500$ or larger. Note that the standard errors for β associated with the cure process are substantially larger than for the β associated with failure process. This phenomenon is typical of models incorporating a latent component.

2.6 Real Data Analysis: The SEER Prostate Cancer Data

We applied the proposed time-dependent cure model to SEER registry data on prostate cancer patients with the aim to study if diagnosis and treatment of a secondary cancer affects the prostate cancer specific survival. Secondary cancer is a new cancer diagnosed

in a prostate cancer patient, i.e. a subject who already is a (primary) prostate cancer survivor. The primary cancer is a cancer that has been diagnosed in the subject for the first time in a lifetime. Both primary and secondary cancers are classified based on disease spread into three categories: localized (tumor confined to the organ), regional (regional spread beyond the organ; merged with localized category in prostate cancer SEER data), and distant (with distant metastases). In addition, the first primary cancer is coded based on the tumor grade into two groups: 0="Low grade" for well differentiated or moderately differentiated cells; 1="High grade" for poorly differentiated or undifferentiated anaplastic cells.

Specifically, we looked at males diagnosed with prostate cancer between 2000 and 2011 in the United States. We restricted our sample to males whose prostate cancer was their first cancer diagnosis and who survived for at least one month. If there was a secondary cancer diagnosis, we restricted our sample to males with secondary cancer diagnosis that is at least one month after the first primary cancer. To avoid potential biases associated with the variable use of PSA screening, we only consider cases diagnosed in and after year 2000. There are a total of 200,994 men in our sample; 8,516 of them died of prostate cancer. Among the 11,730 men who developed secondary cancer during the follow up period, 48% were diagnosed with localized, 23.7% with regional and 28.3% with distant stage secondary cancers. 488 of the secondary cancer patients died of prostate cancer, (180 localized, 113 regional, 113 distant stage secondary cancer patients). Characteristics of the studied patients are presented in Appendix Table A.1, A.2 and A.3.

Our main interest lies is estimating the effects of secondary cancer occurrence and treatment in patients who are primary cancer survivors. The survival time is defined to be the time from diagnosis of prostate cancer to death due to prostate cancer. In our models we include three time-dependent variables $\mathbb{1}(t > localized\ C_2)$, $\mathbb{1}(t > regional\ C_2)$ and $\mathbb{1}(t > distant\ C_2)$, where "$localized\ C_2$", "$regional\ C_2$", and "$distant\ C_2$" is the time to the localized, regional and distant stage secondary cancer diagnosis, respectively.

Additionally, we control for the stage and grade of the first primary (prostate) cancer by including them as time-independent covariates. The hypothesis regarding the effect of the time-dependent covariates is that of an incidental treatment effect on the primary cancer. Namely, diagnosis of a secondary cancer and its treatment in a primary prostate cancer survivor affects the time to death from prostate cancer. Also, it is interesting to study whether the stage of secondary cancer matters, because treatment of the distant stage disease is usually systemic and has a chance to affect latent primary cancer elsewhere.

We start by using a standard Cox model incorporating secondary cancer occurrence and its stage as time-dependent covariates potentially affecting the time to death specific to primary cancer (Table 2.2). We find that the risk of prostate cancer death increases with the degree of cancer spread. The effect may have to do with compromised immune system under a systemic treatment that is not prostate specific. And it affects growth of prostate tumor cells that managed to survive the secondary cancer treatment. Also, prostate cancer cells surviving additional treatment may become more aggressive on average as a survival of the fittest selection effect. Alternatively, in a non-treatment related pathway, occurrence of secondary cancer may identify the subject as having a compromised immune system or adverse genetics. However, this analysis does not shed any light on the hypothesized incidental therapeutic effect of secondary treatment on prostate cancer that is lost in the combined effects reported by the Cox model.

The proposed cure model allows a richer interpretation as it incorporates a more sophisticated mechanism of dynamic counteraction between failure and cure. In the cure model a similar set of effects is incorporated in the failure process and cure components (Table 2.3).

Regarding the failure process, the results are very similar to the Cox model presented earlier. Prostate cancer patients who were diagnosed with secondary cancer have elevated risk of death from prostate cancer compared to those who have one primary cancer only ($\beta = 0.22, 0.81, 1.70$ for localized, regional and distant stage secondary cancer, respec-

Table 2.2: Parameter estimates, standard error and p-values from analysis of SEER data on prostate cancer in the United States, using a Cox model with time-dependent secondary cancer effect. C_2 denotes the time to the secondary cancer from the diagnosis of the first primary cancer. Stage refers to the distant vs. localized/regional stage contrast of patient's first primary cancer. Grade refers to the high (poor or undifferentiated) vs. low (well or moderately differentiated) contrast of patient's first primary cancer. Localized, regional and distant in the time-dependent covariates refer to the stage of secondary cancer.

Parameter	Group	Estimate	Standard Error	p
$\mathbb{1}(t > localized\ C_2)$		0.23	0.08	0.003
$\mathbb{1}(t > regional\ C_2)$		0.69	0.10	< 0.0001
$\mathbb{1}(t > distant\ C_2)$		1.66	0.07	< 0.0001
Stage	1	3.22	0.02	< 0.0001
Grade	1	1.33	0.03	< 0.0001

Stage: 0=Local/Regional vs. 1=Distant
Grade: 0=Low Grade vs. 1=High Grade
$\mathbb{1}(t > localized\ C_2)$: indicator of whether there exists localized stage secondary cancer at time t
$\mathbb{1}(t > regional\ C_2)$: indicator of whether there exists regional stage secondary cancer at time t
$\mathbb{1}(t > distant\ C_2)$: indicator of whether there exists distant stage secondary cancer at time t

tively, vs. no secondary cancer).

As hypothesized, the cure process component reveals a possible incidental therapeutic effect of secondary cancer systemic treatment in its distant stage. Patients diagnosed with and treated for distant stage secondary cancer have shorter times to cure from prostate cancer ($\beta = 0.53$, $p = 0.0004$) compared to those without secondary cancer.

The primary prostate cancer stage and grade present no surprises and show higher risk of prostate cancer death with more advanced stage and with higher grade, in both failure and cure components, although the grade effect in the cure component does not reach significance.

We treated death from other causes as random censoring. This implies that we assume there is no relationship between death from prostate cancer and death from other causes. And we assume the censoring mechanism is random censoring.

Table 2.3: Parameter estimates, standard error and p-values from analysis of SEER data on prostate cancer in the United States, using the proposed dynamic cure model. C_2 denotes the time to the secondary cancer from the diagnosis of the first primary cancer. Stage refers to the distant vs. localized/regional stage contrast of patient's first primary cancer. Grade refers to the high (poor or undifferentiated) vs. low (well or moderately differentiated) contrast of patient's first primary cancer. For all the patients, their first primary cancer is prostate cancer. Localized, regional and distant in the time-dependent covariates refer to the stage of secondary cancer.

Process	Parameter	Group	Estimate	Standard Error	p
failure	$\mathbb{1}(t > localized\ C_2)$		0.22	0.08	0.009
	$\mathbb{1}(t > regional\ C_2)$		0.81	0.11	< 0.0001
	$\mathbb{1}(t > distant\ C_2)$		1.70	0.07	< 0.0001
	Stage	1	3.48	0.03	< 0.0001
	Grade	1	1.37	0.03	< 0.0001
cure	$\mathbb{1}(t > localized\ C_2)$		-0.07	0.17	0.69
	$\mathbb{1}(t > regional\ C_2)$		0.27	0.29	0.34
	$\mathbb{1}(t > distant\ C_2)$		0.53	0.15	0.0004
	Stage	1	0.59	0.08	< 0.0001
	Grade	1	-0.06	0.05	0.22

Stage: 0=Local/Regional vs. 1=Distant
Grade: 0=Low Grade vs. 1=High Grade
$\mathbb{1}(t > localized\ C_2)$: indicator of whether there exists localized stage secondary cancer at time t
$\mathbb{1}(t > regional\ C_2)$: indicator of whether there exists regional stage secondary cancer at time t
$\mathbb{1}(t > distant\ C_2)$: indicator of whether there exists distant stage secondary cancer at time t

2.7 Prediction

Although cure is an unobserved event, a benefit of our joint cure model is that it provides a tool for making predictions of the distribution of the time to cure, given current follow up time and secondary cancer information. The prediction is relevant for those who do not experience the failure event ($\delta = 0$). Derived in Appendix A.2 are the predicted conditional survival functions for onset of cure given patient's follow up time X, and first primary cancer and secondary cancer information. Specifically, we have the following conditional survival distribution for cure onset at time point t_u, given a follow

up of X:

$$S(t_u \mid X, \delta = 0) = \frac{e^{-\int_0^X \theta(y)dH_1(y)} e^{-\int_0^{X \vee t_u} \eta(y)dH_2(y)} + \int_{t_u}^{X \vee t_u} \eta(s) e^{-\int_0^s \theta(y)dH_1(y)} e^{-\int_0^s \eta(y)dH_2(y)} dH_2(s)}{e^{-\int_0^X \theta(y)dH_1(y)} e^{-\int_0^X \eta(y)dH_2(y)} + \int_0^X \eta(s) e^{-\int_0^s \theta(y)dH_1(y)} e^{-\int_0^s \eta(y)dH_2(y)} dH_2(s)}.$$

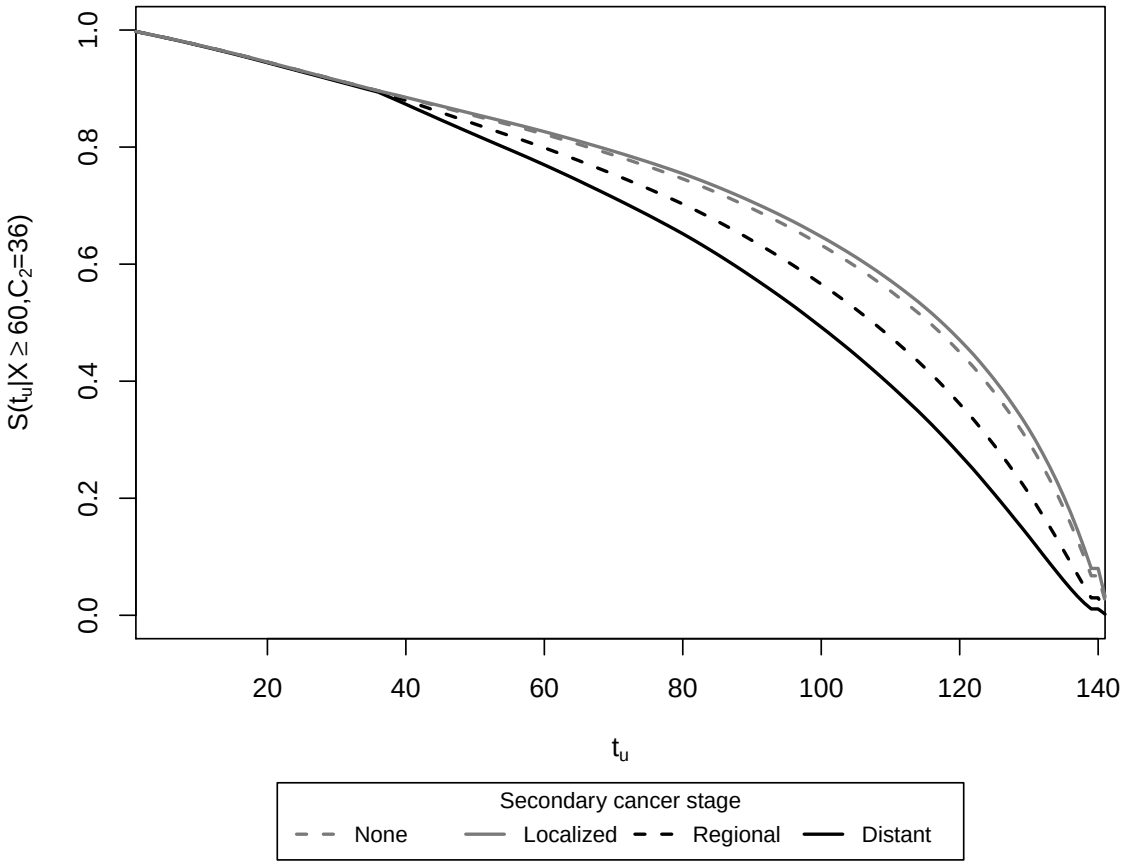

Figure 2.1: Conditional survival functions for onset of cure given 5 years of follow up ($X \geq 60$ months), localized/regional and low grade first primary cancer, and secondary cancer diagnosed at 3rd year ($C_2 = 36$ months, if exist) by secondary cancer stage

Shown in Figure 2.1 are the conditional survival functions for onset of cure for a local/regional and low grade first primary cancer patient followed up for 5 years ($X \geq 60$ months) with localized, regional or distant stage secondary cancer diagnosed at the 3rd

year post the diagnosis of a prostate cancer (first primary), compared to a similar patient without a secondary cancer diagnosis.

A feature to notice is that these four curves are identical until the 36^{th} month when the secondary cancer is diagnosed, indicating the cure rate changes in response to the discovery and/or treatment of secondary cancer. The ordering of these curves gives us an idea of how the cure rate changes as the stage of the secondary cancer increases. Compared to those who are not diagnosed with secondary cancer, patients with distant and regional stage secondary cancer have lower curves, indicating an earlier onset of cure. While for a localized secondary cancer, a rather mild treatment effect is observed, and we don't see much of a difference in predicted cure rate between those who are diagnosed with a localized secondary cancer and those who are not diagnosed with secondary cancer.

2.8 Practical Concept of Cure: Death from Other Causes

The prediction of the time to cure in Section 2.7 is based on the biological concept of cure, that is, cure in absence of death from other causes. In practice, patients with prostate cancer diagnosed at older age are more likely to die of other causes compared to younger patients. Applying treatment to prostate cancer for older patients may not be so relevant as for those who are younger because older patients are more likely to die of other causes, not of prostate cancer itself. Therefore, practical concept of cure considers death from other causes as another source of cure, on top of biological concept of cure.

Modeling practical concept of cure requires modeling residual survival for other causes from the age of prostate cancer diagnosis. Denote T_{oc} as the time to death from other causes from birth and $S_{oc}(t|Z)$ as the survival function of T_{oc}. Then the residual survival for other causes is

$$P(T_{oc} > t + a|T_{oc} > a, Z) = \frac{P(T_{oc} > t + a|Z)}{P(T_{oc} > a|Z)} = \frac{S_{oc}(t + a|Z)}{S_{oc}(a|Z)} = S_{oc}(t|a, Z)$$

where a is the age at prostate cancer diagnosis. Assume age at prostate cancer diagnosis is proportional hazard covariate for $S_{oc}(t|a, Z)$, then this survival function must follow the form a Gompertz survival model. The derivation is based on differential equation. Details can be found in Appendix A.8. This is called "characterization of Gompertz distribution". Gompertz survival model is a popular survival model used in studies of longevity. Using the SEER registry data as in Section 2.6, we fitted a Gompertz survival model to the time to death from other causes, with age at prostate cancer diagnosis as a covariate, adjusting for prostate cancer stage and grade. Death from prostate cancer is treated as censored. Age was centered at 47 years old and scaled by 10 year. The Gompertz model survival function is

$$S_{oc}(t|a, Z) = e^{-\frac{\lambda}{\gamma}(e^{\gamma t}-1)},$$

and the hazard function is

$$h_{oc}(t|a, Z) = \lambda e^{\gamma t}$$

where $\lambda = e^{\beta_0+\beta_1 age+\beta_2 stage+\beta_3 grade}$ describes the covariate effects.

The parameter estimates of the fitted Gompertz survival model to death from other causes are: $\hat{\gamma} = 0.012, \hat{\beta}_0 = -7.28, \hat{\beta}_1 = 1.05, \hat{\beta}_2 = 0.68, \hat{\beta}_3 = 0.15$. Figure 2.2 shows the probability of death from other causes within the interval of last follow up (143 months), as a function of age at diagnosis. Figure 2.2 shows the probability of death from other causes within the interval of total follow up is $1 - S_{oc}(t = 143|a, Z)$. The probability of death from other causes increases rapidly after age of 60. Patients with distant, high grade prostate cancer are more likely to die from other causes, compared to localized/regional, low grade prostate cancer patients. This may be due to the side effects of the more aggressive therapy applied to distant prostate cancer. According to the practical concept of cure, patients with distant, high grade prostate cancer has a higher probability of

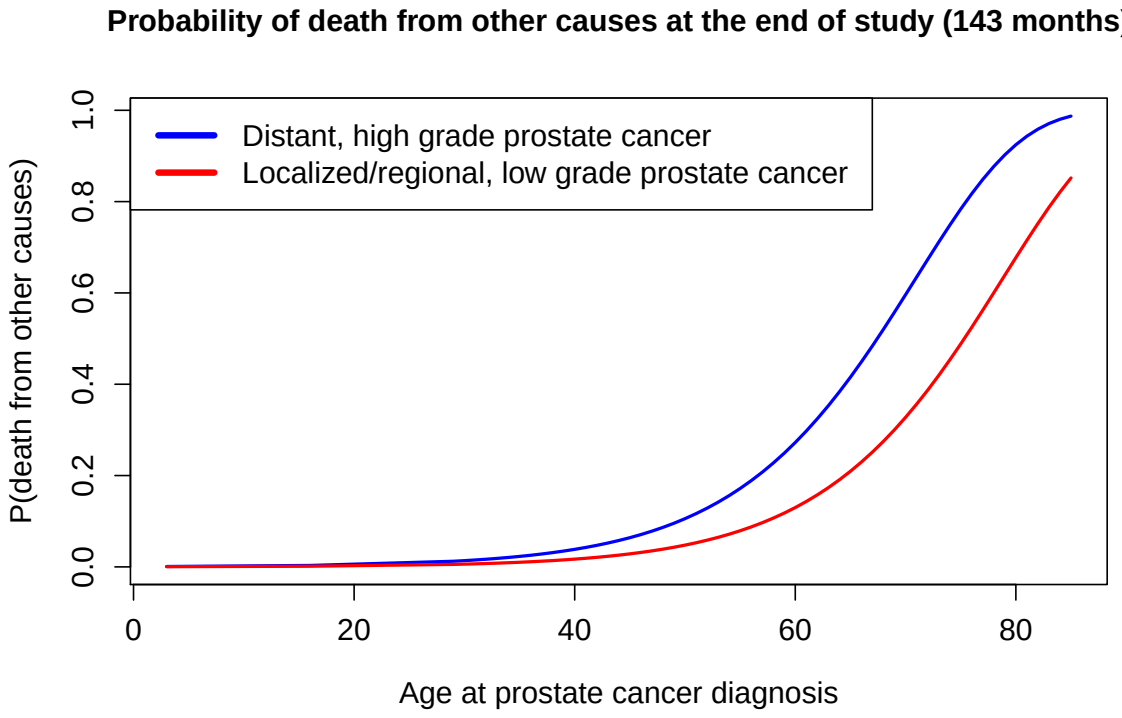

Figure 2.2: Probability of death from other causes within the interval of last follow up (143 months) for distant high grade prostate cancer (blue) and for localized/regional low grade prostate cancer (red)

cure compare to localized/regional low grade prostate cancer patients due to a higher probability of death from other causes for distant high grade prostate cancer patients.

The biological cure from prostate cancer is $S(\infty|Z)$, which can be obtained from estimated survival function from Section 2.6. The probability of die from other causes at age a given patient is at risk (have not died from prostate cancer) is

$$P(\text{die from other causes}|a, Z) = \int_0^\infty f_{oc}(t|a, Z)S(t|Z)dt$$

where $f_{oc}(t|a, Z) = h_{oc}(t|a, Z)S_{oc}(t|a, Z)$ is the density function for time to death from

other causes. Probability of practically cured is $S(\infty|Z) + \int_0^\infty f_{oc}(t|a, Z)S(t|Z)dt$. Figure 2.3 and 2.4 shows the probability of being biologically cured (blue) versus the probability of being practically cured (red), as a function of age at diagnosis. The biological cure is basically the same as the practical cure before age of 40. After age of 60, there is a greater chance of being practically cured by the source of death from other causes. Therefore, treatment for a prostate cancer is not that relevant for people of older age. Caution should be applied to the use of practical concept of cure when choosing the type of treatment for prostate cancer. Choosing the treatment based on practical cure may favor a therapy that has serious side effects leading to higher probability of dying from other causes. It is not appropriate to use practical concept of cure that consider death from other causes as a source of cure when making decision on the type of treatment.

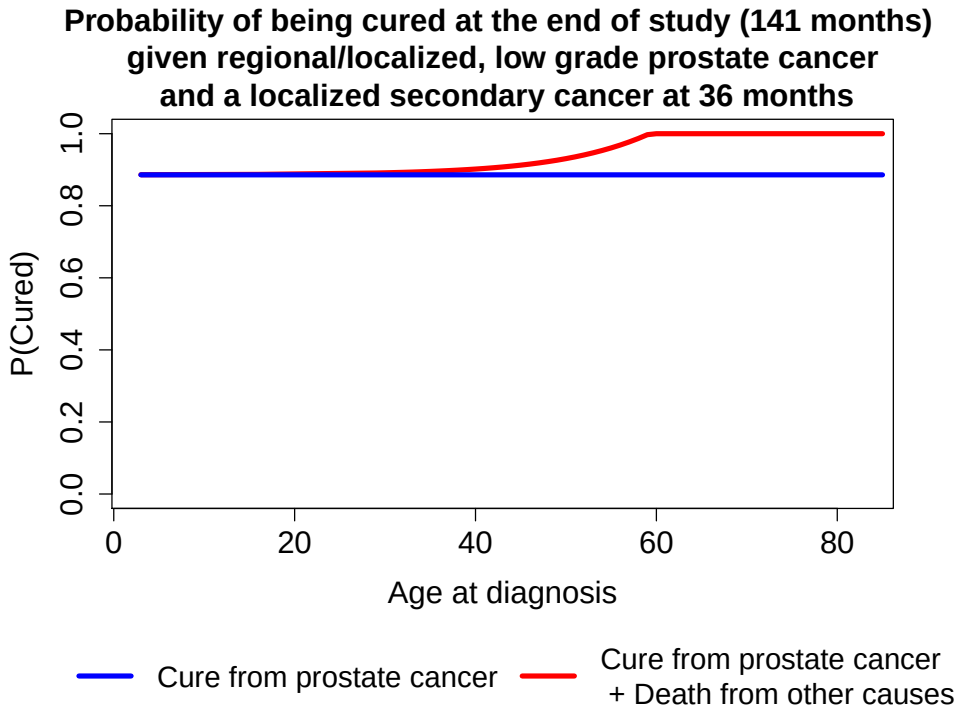

Figure 2.3: Probability of being biologically cured (blue) and probability of being practically cured (red) within interval of last follow up (141 months), for regional/localized low grade prostate cancer patients.

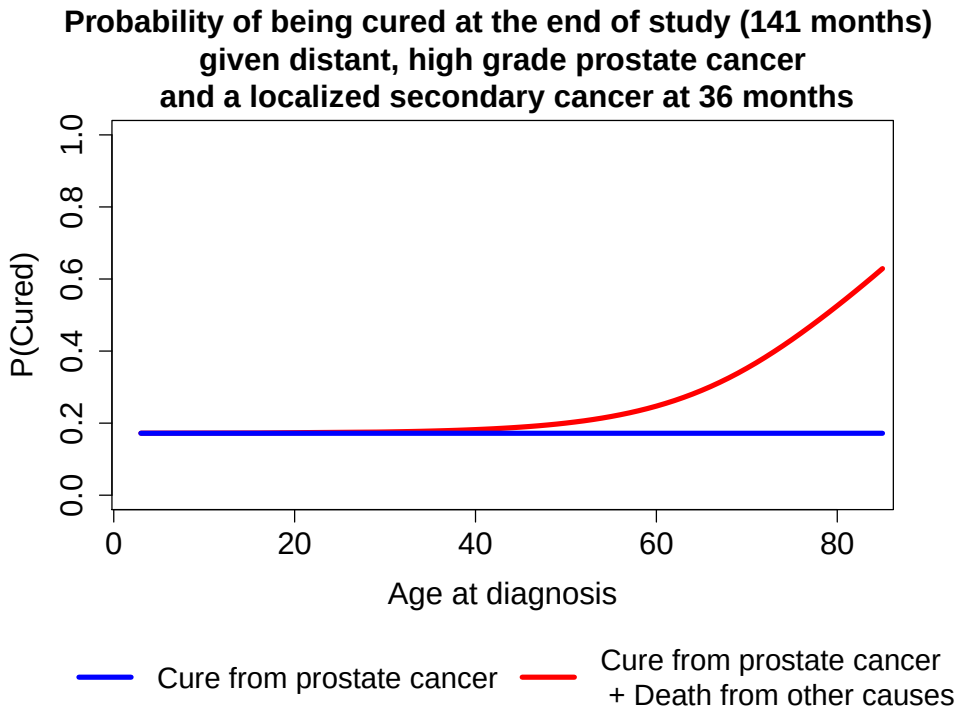

Figure 2.4: Probability of being biologically cured (blue) and probability of being practically cured (red) within interval of last follow up (141 months), for distant high grade prostate cancer patients.

2.9 Discussion

Using a mechanistic competing nature of cure and failure within the subject, we constructed a new class of cure models driven by a latent stochastic cure process that allowed us to incorporate time dependent covariates into the cure. Unlike other typical cure models, we do not assume that the cure status is predetermined at time zero. As an example, we modeled the conditional hazard function for terminal event as a change-point function driven by the latent event of cure. In general, any stochastic hazard process $\mathcal{U}(t)$ that has an absorbing boundary of 0 leads to a cure model. The proposed model framework has the flexibility to incorporate a wide variety of dynamic cure models. We model the time to cure and time to failure with a proportional hazard model. Other link functions can be naturally incorporated, for example, using artificial frailties (*Tsodikov* (2003)).

If the time to cure model is itself a cure model, the baseline hazard for cure H_2 is

bounded, and the cure event may not happen. The PH model just like most other semi-parametric model families includes the corresponding family of cure models. Cure model becomes a way to reparameterize the corresponding non-cure model to expose an explicit parameter responsible for the cure rate, when this parameter is of main interest. However, in our model, the cure rate for the time to cure distribution is a nuisance and does not bear direct relevance to the probability of cure for the observed time to failure. The time to cure is modeled by a PH model in its traditional non-cure parameterization accordingly. If p_c is the probability of cure in the time to cure model, then $1 - p_c$ represents the upper boundary for the probability of cure in the population that would be achieved if all cures in the $1 - p_c$ fraction happened at time 0. The predicted probability of cure based on the data in the example is smaller than $1 - p_c$, because the time to cure distribution is not degenerate and failure has a chance to happen before cure. Figure 2.1 shows predictions of p_c as the value of the time to cure survival functions at the right extreme of the data at 141 months. They are quite small so time to cure is estimated as a virtually proper distribution.

Predicting something that is fully unobserved, such as the time to cure, should be treated with caution because the model typically has less power for parameters associated with latent components. However, the model-based predictions represent a useful tool to generate hypotheses on the latent effects and to guide further confirmatory studies pursuing more explicit measurements.

A variety of stochastic processes $\mathcal{U}$ can be chosen to generate various rich classes of cure models by imposing an absorbing boundary. These include squared Gaussian (*Yashin and Manton* (1997)) and Lévy processes (*Gjessing et al.* (2003)) and their extensions (*Putter and van Houwelingen* (2015)).

Many patients with prostate cancer who are detected by screening are not expected to die from the disease in their lifespan. Screening leads to some cancer diagnoses that would not occur during the subject's lifespan in the absence of screening. Even though

such overdiagnosed patients have a latent disease, they present as cured for all practical purposes. Separation of biological cure vs. overdiagnosis is a challenge that requires joint modeling of diagnosis and survival. Because overdiagnosed patients are not expected to be affected by secondary cancer as far as their prostate cancer survival goes, the interesting effects found in this paper are conservative.

In the data analysis, the terminal event of interest is prostate cancer specific death. In clinical practice, sometimes patients die with prostate cancer instead of die of prostate cancer. The effects found in this paper depend on how the cause of death was assigned and recorded in SEER registry. If misattributed causes of death is a concern, additional adjustment for misclassification error should be applied. Methods to adjust for misclassified cause of failure was addressed by *Ha and Tsodikov* (2015).

Acknowledgment

This research is supported by the grant U01CA199338 (CISNET) and P50CA186786 (SPORE) from the National Cancer Institute. A manuscript of this paper won the 2017 John Van Ryzin award presented by the Eastern North American Region (ENAR) of the International Biometric Society.

CHAPTER III

A Semiparametric Latent Trait Model for Multiple Mixed Continuous, Categorical, and Time-to-event Outcomes

3.1 Introduction

In chronic diseases, research often centers on discovering a latent trait that manifests itself through a variety of observable responses and covariates. Multiple outcomes (phenotype) are often collected when the construct of interest cannot be measured directly. Mixed continuous, binary, ordinal and survival outcomes are commonly collected in studies of complex health conditions in order to capture different aspects of patient specific latent trait. Generalized latent variable models with mixed outcomes are developed to accommodate outcomes of mixed measurement scales and estimate the latent trait of interest.

Existing research on generalized latent variable models for mixed outcomes focuses on two strategies. The first approach is parametric and utilizes joint Gaussian framework by linking observed categorical outcomes to underlying continuous normally distributed latent responses (*Muthén* (1984); *Shi and Lee* (2000)). These models are related to Gaussian copula model of dependence between the phenotypes. *Hoff* (2007) proposed a semiparametric Gaussian copula model, leaving marginal distributions of mixed scale

outcomes unspecified. Associations among mixed-scale outcomes are induced by correlations among the latent Gaussian variables. *Murray et al.* (2013) and *Gruhl et al.* (2013) extend this approach to incorporate a latent factor model with mixed continuous and ordinal outcomes under Bayesian framework. *Lin et al.* (2014) proposed a semiparametric normal transformation latent variable model for continuous and ordinal outcomes. *Snavely et al.* (2014) extended this model to further allow for censored outcomes. This approach maps observable outcomes to underlying Gaussian continuous responses and is limited to mixed continuous and ordinal outcomes. This type of models require estimating a set of unknown thresholds for ordinal outcomes. For the semiparametric models proposed by *Lin et al.* (2014) and *Snavely et al.* (2014), the parameter estimates come from maximizing a pseudo-likelihood through a set of estimating equations, an approach potentially less efficient than the standard maximum likelihood. Standard errors of estimates are estimated through nonparametric bootstrap. In addition, this approach does not accommodate the multinomial scale outcomes.

The second approach is the shared latent variable models that induce dependence among mixed outcomes through shared latent variables (*Sammel et al.* (1997); *Moustaki and Knott* (2000); *Dunson* (2000); *Dunson and Herring* (2005)). This class of models are mainly parametric generalized linear models. *Skrondal and Rabe-Hesketh* (2004) extended generalized latent trait model to accommodate censored outcomes. This approach is more flexible in modeling non-normal outcomes, however, existing work on these models requires explicitly specified transformation function for the measurable outcomes of different scales.

In this paper, we propose a semiparametric shared latent variable model where a logistic link is used to accommodate continuous, ordinal, count, multinomial and survival outcomes. The model is used to provide a subject-specific measure of the latent trait, given the information observed on the subject. The proposed model avoids the restrictive multivariate normal assumption of the underlying continuous latent responses. Furthermore, the proposed model does not require implicitly or explicitly estimating a

set of unknown thresholds inherent in latent normal models for categorical data. Our model incorporates a nonparametric transformation of the outcomes of different scales. Unlike current work on the shared latent variable models that rely on explicit specification of the baseline distributions of the measureable outcomes on different scales, we use non-parametric transformation functions that are completely data-driven. Since our model is a joint model of outcomes of mixed types, we can estimate the subject-specific shared latent trait even when some of the outcomes are missing. Unlike *Lin et al.* (2014) and *Snavely et al.* (2014), our model parameters are estimated by maximizing the full-likelihood so likelihood based standard errors can be used for inference and the estimates are asymptotically efficient. Furthermore, our model accommodates measurable outcomes of multinomial scale. The modeling framework is also generic with respect to the parametric distribution assumed for the latent trait, allowing a choice of the distribution of the trait. Our model has the potential to avoid direct specification of the distribution for shared latent variable as long as the marginal model is specified in terms of a Laplace transform, as described in Section 3.2. Unlike the gamma frailty model that implicitly assumes positive correlation between outcomes, our model allows for both positive and negative correlation between outcomes through the sign of factor loadings. And the proposed model needs not assume Gaussian distribution for the latent variable to allow for negative correlation among outcomes.

The proposed method is applied to measure pain centrality trait of patients undergoing hysterectomy as a treatment for pelvic pain and to explain the heterogeneity of patients' reported outcomes. The method is compared with the ad-hoc 2011 Fibromyalgia (FM) Survey Criteria instrument designed to characterize a similar construct. Difference of convex functions algorithm (DCA) is used to estimate the nonparametric data-driven link functions of the model. Covariate parameters and distributional parameter for latent variable are estimated by maximizing the profile likelihood using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The rest of the article is organized as

follows. Section 3.2 introduces the logistic link function. In Section 3.3, we show how the logistic link can accommodate measurements of a variety of scales. In addition, we show how it enables a closed form conditional expectations in a typical EM algorithm under univariate case. Section 3.4 extends the framework to the multivariate case. The proposed multivariate model framework and the likelihood function are presented in Section 3.4. The estimation procedures are described in Section 3.5. The asymptotic properties are shown in Section 3.6. Section 3.7 shows the simulation study. In Section 3.8 we applied the proposed model to construct latent pain centrality score from pelvic pain patients. Section 3.9 concludes the proposed method and discussion.

3.2 Logistic Link Function

We propose using the logistic link function to accommodate outcomes of a variety of scales. Consider a logistic link function L given a non-negative random variable U,

$$G(s|U) = \frac{U^\alpha}{1 + U^\alpha s},$$

and its expectation over U,

$$\mathcal{G}(s) = \mathbb{E}_U \left[G(s|U) \right] = \mathbb{E}_U \left[\frac{U^\alpha}{1 + U^\alpha s} \right]$$

where U is a non-negative random variable that represents the latent trait of interest and α is a fix constant that represents the association across outcomes, similar to the factor loading in the factor analysis. We have the following general propositions:

Proposition 1. The function G is a special function such that the product of the functions can be expressed as a first order difference/ratio. For $a \neq b$,

$$G(a|U)G(b|U) = \frac{G(a|U) - G(b|U)}{b - a}.$$

By the linearity of the expectation operator, the expectation of the product over U can be expressed as

$$\mathbb{E}_U\left[G(a|U)G(b|U)\right] = \frac{\mathcal{G}(a) - \mathcal{G}(b)}{b-a}.$$

Proposition 2. The squared function $G^2(s)$ can be expressed as the derivative of the function:

$$G^2(s|U) = -G'(s|U)$$

By the Leibniz's rule for differentiation under the integral sign, the expectation of the squared function $G^2(s)$ can be expressed as

$$\mathbb{E}_U\left[G^2(s|U)\right] = -\mathcal{G}'(s).$$

In general, the power function $G^{n+1}(s|U)$ can be expressed as the n^{th} order derivative of the function:

$$G^{n+1}(s|U) = \frac{(-1)^n}{n!}G^{(n)}(s|U)$$

And the expectation of the power function $G^{n+1}(s|U)$ can be expressed as

$$\mathbb{E}\left[G^{n+1}(s|U)\right] = \frac{(-1)^n}{n!}\mathcal{G}^{(n)}(s).$$

Proposition 3. The function $1 - sG(s|U)$ is a Laplace transform. Consider a random variable W that has a unit exponential distribution and $W \perp\!\!\!\perp U$. Then $1 - sG(s|U)$ is a Laplace transform of the random variable W:

$$\mathbb{E}_W\left(e^{-WU^\alpha s} \mid U\right) = \frac{1}{1+U^\alpha s} = 1 - sG(s|U)$$

And the expectation of $1 - sG(s|U)$ is a Laplace transform of the random variable WU^{α}.

$$\mathbb{E}_U \mathbb{E}_W \left(e^{-WU^{\alpha} s} \right) = \mathbb{E}_U \left(\frac{1}{1 + U^{\alpha} s} \right) = 1 - s\mathcal{G}(s).$$

The above propositions yield a convenient form of the likelihood amenable to the EM algorithm applied to profile out the infinite dimensional parameters, similar to the role of the Laplace transform in shared frailty Archimedian copula models. As a particular case of the above framework, we propose unified treatment of univariate models for responses of a variety of scales in the next section. The univariate model framework is generic with respect to the distribution for the latent variable U. We do not need to specify the distribution for U as long as we know the form of $\mathcal{G}(\cdot)$.

3.3 The Univariate Model

3.3.1 Model Framework

Under the case of univariate outcome, α is not identifiable, so we set $\alpha = 1$ throughout this section. Consider an observed outcome Y which can be a continuous, ordinal, count, or time-to-event outcome. Let U be a latent variable which represents the latent trait of interest but cannot be measured directly. Let $\bar{F}(y) = P(Y > y)$ denotes the tail/survival distribution of the variable Y. We define the conditional tail/survival distribution function for Y as

$$\bar{F}(Y \mid U) = \frac{1}{1 + U\gamma H(y)} = 1 - \gamma H(y) G\left(\gamma H(y) | U\right),$$

where $H(y)$ is an unspecified non-decreasing function of y, and $\gamma = \exp(\mathbf{Z}^T \beta)$ where $\mathbf{Z}$ is the covariate matrix and β is the vector of regression coefficients. The conditional

probability mass/density function for Y is

$$P(Y = y \mid U) = \bar{F}(y^- \mid U) - \bar{F}(y \mid U) = \frac{1}{1 + U\gamma H(y^-)} - \frac{1}{1 + U\gamma H(y)}$$
$$= \gamma H(y) G\left(\gamma H(y)|U\right) - \gamma H(y^-) G\left(\gamma H(y^-)|U\right),$$

where $y^- = \lim_{\Delta \to 0}(y - \Delta)$.

The marginal tail/survival distribution Y is the expectation of $\bar{F}(y \mid U)$ over U and by Proposition 3 is a Laplace transform. Specifically,

$$\bar{F}(y) = \mathbb{E}_U \left(\frac{1}{1 + U\gamma H(y)} \right) = 1 - \gamma H(y) \mathcal{G}\left(\gamma H(y)\right).$$

The marginal probability mass/density function for Y can be expressed as

$$\mathbb{E}\left[\bar{F}(y^- \mid U)\right] - \mathbb{E}\left[\bar{F}(y \mid U)\right] = \gamma H(y) \mathcal{G}\left(\gamma H(y)\right) - \gamma H(y^-) \mathcal{G}\left(\gamma H(y^-)\right).$$

Suppose there are n independent subjects with observed outcome $(y_1, \cdots, y_n)$ and covariate matrix $(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$. For the time-to-event outcome, we assume for each subject i, the censoring time C_i^* is independent of true event time T_i, given covariate set $\mathbf{Z}_i$. The observed event time is $Y_i = (T_i \wedge C_i)$ and $C_i = (C_i^* \wedge \tau)$. Let $\delta_i = \mathbb{1}(T_i \leq C_i)$ be the censoring indicator. Here $\mathbb{1}(\cdot)$ is the indicator function, and τ denotes the time to the end of the study. The latent variable U_i is the latent trait of interest for subject i. The complete data likelihood for the $i$the subject can be written using the properties of the logistic link function as

$$L_{0i} = [\gamma_i H(y_i) G(\gamma_i H(y_i)|U_i) - \gamma_i H(y_i) G(\gamma_i H(y_i^-)|U_i)]^{\delta_i} [1 - \gamma_i H(y_i) G(\gamma_i H(y_i)|U_i)]^{1-\delta_i},$$

and the complete data likelihood for all n subject is $L_0 = \prod_{i=1}^n L_{0i}$.

The observed data likelihood for the ith subject can be obtained by taking expectation

of L_{0i},

$$L_i = \mathbb{E}[L_{0i}] = [\gamma_i H(y_i)\mathcal{G}(\gamma_i H(y_i)) - \gamma_i H(y_i)\mathcal{G}(\gamma_i H(y_i^-))]^{\delta_i}[1 - \gamma_i H(y_i)\mathcal{G}(\gamma_i H(y_i))]^{1-\delta_i}$$

and the observed data likelihood for all n subject is $L = \prod_{i=1}^{n} L_i$.

Notice L has a closed form representation given the form of $\mathcal{L}$. The complete data log-likelihood is

$$\ell_0 = \sum_{i=1}^{n} \left\{ \delta_i \log \left(\frac{U_i \gamma_i dH(y_i)}{[1 + U_i \gamma_i H(y_i^-)][1 + U_i \gamma_i H(y_i)]} \right) + (1 - \delta_i) \log \left(\frac{1}{1 + U_i \gamma_i H(y_i)} \right) \right\}.$$

Here we assume H is an arbitrary step function that only jumps at the set of observed values of y_i and we denote the jump of H at s as $dH(s)$. Notice H is similar to a cumulative hazard function in a survival model, can be treated as an unknown outcome/time transformation.

If the observed outcome is nominal, then the modeling framework is as follows. For a K-category multinomial observable outcome with an observed response category c, we define the conditional probability mass function as

$$p(Y = c \mid U) = \begin{cases} \frac{U\theta_c}{1+U\sum_{k=2}^{K}\theta_k} = \theta_c G\left(\sum_{k=2}^{K}\theta_k | U\right), c \geq 1, \\ \frac{1}{1+U\sum_{k=2}^{K}\theta_k} = 1 - \sum_{k=2}^{K}\theta_k G\left(\sum_{k=2}^{K}\theta_k | U\right), c = 1, \end{cases}$$

where θ_k is the covariate effect for category c versus the reference category 1.

The marginal probability mass function can be expressed as

$$p(Y = c) = \begin{cases} \theta_c \mathcal{G}\left(\sum_{k=2}^{K}\theta_k\right), c \geq 1, \\ 1 - \sum_{k=2}^{K}\theta_k \mathcal{G}\left(\sum_{k=2}^{K}\theta_k\right), c = 1. \end{cases}$$

Estimation of multinomial outcomes follows the methods described in *Tsodikov and Chefo* (2008).

3.3.2 Estimation Procedure

Let $\Omega = \{\beta, \eta\}$ where η is the parameter characterizing the distribution of the latent variable U. We estimate Ω and $\{dH\}$ jointly using the profile likelihood approach. This is accomplished by applying an EM algorithm (*Tsodikov* (2003)) to obtain implicit profile likelihood estimators $dH(\cdot|\Omega)$ for $dH(\cdot)$ that depends on Ω. To obtain stable MLEs, we impose a zero-tail constraint on non-survival outcome, namely, $\bar{F}(y_{(n)}|U) = 0$, or equivalently, $dH(y_{(n)}) = \infty$. Replacing $dH(\cdot|\Omega)$ in the marginal log-likelihood we obtain the profile log-likelihood $\ell_{pr}(\Omega) = \ell(\Omega, \{dH(\cdot|\Omega)\})$. On differentiation of the conditional log-likelihood with respect to $dH(s)$, we obtain the conditional score function for $dH(s)$ as

$$\mathcal{U}_{0,dH}(s) = \frac{\partial \ell_0}{\partial dH(s)} = \frac{\sum_{i=1}^n \mathbb{1}(y_i = s)}{dH(y_i)} - \sum_{i=1}^n \delta_i \frac{\mathbb{1}(y_i^- \geq s) U_i \gamma_i}{1 + U_i \gamma_i H(y_i^-)} - \sum_{i=1}^n \frac{\mathbb{1}(y_i \geq s) U_i \gamma_i}{1 + U_i \gamma_i H(y_i)}$$
$$= \frac{\sum_{i=1}^n \mathbb{1}(y_i = s)}{dH(y_i)} - \sum_{i=1}^n \delta_i \, \mathbb{1}(y_i^- \geq s) \gamma_i G(\gamma_i H(y_i^-)|U_i) - \sum_{i=1}^n \mathbb{1}(y_i \geq s) \gamma_i G(\gamma_i H(y_i)|U_i)$$

Define the conditional expectation operator $\mathbb{E}[f\|g] = \frac{\mathbb{E}[f \cdot g]}{\mathbb{E}[g]}$. We use the EM-DCA algorithm to estimate the NPMLE. The derivation of the EM algorithm and the conditional expectation operator $\mathbb{E}[\cdot\|L_0]$ that is used in the rest of the article are presented in Appendix B.2. The EM algorithm involves iteratively update model parameters by maximizing the conditional expectation of the complete data likelihood $\mathbb{E}[\ell_0\|L_0]$. The model parameters are updated by maximizing $\mathbb{E}[\ell_0\|L_0]$. This can be achieved by setting the derivative of $\mathbb{E}[\ell_0\|L_0]$ with respect to model parameters to 0. Specifically, we calculate the conditional expectation of the conditional score function given the observed data, latent variable U and the model parameters at the kth iteration. For the following derivation of the EM algorithm, we use the expression $\mathbb{E}\left[\mathcal{U}_0 \middle\| L_0^{(k)}\right]$ to represent the conditional expectation of the conditional score function given the observed data, latent variable U and the model parameters at the kth iteration. Note that $\mathbb{E}[\mathcal{U}_0\|L_0]$ is the marginal score

function as derived in Appendix B.2. Specifically, at the kth iteration, the conditional expectation of the conditional score function for dH can be expressed as

$$\mathbb{E}\left[\mathcal{U}_{0,dH}(s)\middle\|L_0^{(k)}\right] = \frac{\mathbb{E}\left[\mathcal{U}_{0,dH}(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} = \frac{\sum_{i=1}^{n}\mathbb{1}(y_i = s)}{dH_j^{(k+1)}(s)}$$
$$-\sum_{i=1}^{n}\mathbb{1}(y_i^- \geq s)\gamma_i\mathbb{E}\left[G(\gamma_i H^{(k+1)}(y_i^-)|U_i)\middle\|L_0^{(k)}\right] - \sum_{i=1}^{n}\mathbb{1}(y_i \geq s)\gamma_i\mathbb{E}\left[G(\gamma_i H^{(k+1)}(y_i)|U_i)\middle\|L_0^{(k)}\right]. \tag{3.1}$$

where $L_0^{(k)}$ notation is used to indicate that the complete data likelihood L_0 is parameterized by the kth iteration copy of $\{dH\}$.

Notice the complete data log-likelihood ℓ_0 can be represented as a difference between two concave functions. Consequently, the conditional score equation $\mathcal{U}_{0,dH}(s)$ has a representation of a difference between derivatives of two concave functions. Since the conditional expectation operator $\mathbb{E}\left[\quad\cdot\quad\middle\|L_0^{(k)}\right]$ is a linear operator, it does not alter convexity properties. Therefore, the unconditional score function $\mathbb{E}\left[\mathcal{U}_{0,dH}(s)\middle\|L_0^{(k)}\right]$ is also a difference between derivatives of two concave functions.

For the M-step, we employ the difference of convex functions algorithm (DCA) to iteratively maximize log-likelihood with respect to dH_j. DCA was first introduced by Pham Dinh Tao in their preliminary form in 1985. DCA is a version of the MM-algorithm (*Lange et al.* (2000)) that has been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao for nonconvex optimization problems (*Tao and An* (1997, 1998); *An and Tao* (2005)). DCA is particularly efficient when the target function to be minimized/maximized can be represented as a difference between two convex/concave functions. Equation (3.1) equals to zero is a self-consistency equation that can be solved iteratively. Solving the conditional expectation of the score equation (3.1) equals to zero

using DCA, we obtain a Breslow-type estimator for $dH(s)$ at the $(k+1)^{th}$ iteration

$$dH^{(k+1)}(s) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = s)}{\sum_{i=1}^{n} \delta_i \, \mathbb{1}(y_i^- \geq s)\gamma_i \mathbb{E}\left[G(\gamma_i H^{(k)}(y_i^-)|U_i)\middle\| L_0^{(k)}\right] + \sum_{i=1}^{n} \mathbb{1}(y_i \geq s)\gamma_i \mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i)\middle\| L_0^{(k)}\right]} \tag{3.2}$$

Iterations proceed until $\|dH^{(k+1)}(s) - dH^{(k)}(s)\|_2 < \epsilon$. The derivation of EM-DCA algorithm is presented in Appendix B.3 when $J = 1$ and $\alpha_J = 1$, with the conditional expectation terms calculated as follows.

Equation (3.2) is computationally efficient since the conditional expectations $\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i^-)|U_i)\middle\| L_0^{(k)}\right]$ and $\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i)\middle\| L_0^{(k)}\right]$ have closed form solutions given $\mathcal{G}$. Specifically,

$$\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i)\middle\| L_0^{(k)}\right] = \frac{\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i) L_{0i}^{(k)}\right]}{\mathbb{E}[L_{0i}^{(k)}]}$$

The denominator is the observed data likelihood for the ith subject evaluated at the kth iteration,

$$\mathbb{E}[L_{0i}^{(k)}] = [\gamma_i H^{(k)}(y_i)\mathcal{G}(\gamma_i H^{(k)}(y_i)) - \gamma_i H^{(k)}(y_i)\mathcal{G}(\gamma_i H^{(k)}(y_i^-))]^{\delta_i}[1 - \gamma_i H^{(k)}(y_i)\mathcal{G}(\gamma_i H^{(k)}(y_i))]^{1-\delta_i} \tag{3.3}$$

For the numerator, the term inside the expectation is

$$
\begin{aligned}
& G(\gamma_i H^{(k)}(y_i)|U_i) L_{0i}^{(k)} \\
=& \{G(\gamma_i H^{(k)}(y_i)|U_i)[\gamma_i H(y_i) G(\gamma_i H^{(k)}(y_i)|U_i) - \gamma_i H^{(k)}(y_i) G(\gamma_i H^{(k)}(y_i^-)|U_i)]\}^{\delta_i} \\
& \{G(\gamma_i H^{(k)}(y_i)|U_i)[1 - \gamma_i H^{(k)}(y_i) G(\gamma_i H^{(k)}(y_i)|U_i)]\}^{1-\delta_i} \\
& \text{By proposition 1 and 2,} \\
=& \left\{ \gamma_i H^{(k)}(y_i) \left[-G'(\gamma_i H^{(k)}(y_i)|U_i) - \frac{G(\gamma_i H^{(k)}(y_i)|U_i) - G(\gamma_i H^{(k)}(y_i^-)|U_i)}{\gamma_i H^{(k)}(y_i^-) - \gamma_i H^{(k)}(y_i)} \right] \right\}^{\delta_i} \\
& \left\{ G(\gamma_i H^{(k)}(y_i)|U_i) + \gamma_i H^{(k)}(y_i) G'(\gamma_i H^{(k)}(y_i)|U_i) \right\}^{1-\delta_i}
\end{aligned}
$$

Therefore, the numerator can be expressed as

$$
\begin{aligned}
\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i) L_{0i}^{(k)} \right] = & \left\{ \gamma_i H^{(k)}(y_i) \left[-\mathcal{G}'(\gamma_i H^{(k)}(y_i)) - \frac{\mathcal{G}(\gamma_i H^{(k)}(y_i)) - \mathcal{G}(\gamma_i H^{(k)}(y_i^-))}{\gamma_i H^{(k)}(y_i^-) - \gamma_i H^{(k)}(y_i)} \right] \right\}^{\delta_i} \\
& \left\{ \mathcal{G}(\gamma_i H^{(k)}(y_i)) + \gamma_i H^{(k)}(y_i) \mathcal{G}'(\gamma_i H^{(k)}(y_i)) \right\}^{1-\delta_i} \qquad (3.4)
\end{aligned}
$$

Notice that (3.3) and (3.4) are closed form expressions given $\mathcal{G}$. Therefore, $\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i)|U_i) \middle\| L_0^{(k)} \right]$ has a closed form expression. Similarly, applying the same techniques, the other conditional expectation $\mathbb{E}\left[G(\gamma_i H^{(k)}(y_i^-)|U_i) \middle\| L_0^{(k)} \right]$ has a closed form expression. Therefore, the Breslow type estimator (3.2) for updating $dH(s)$ in the EM-DCA algorithm has a closed form expression, which results in a computationally efficient estimating algorithm for the nonparametric transformation function H. In addition, the modeling framework is generic with respect to the distribution for the latent variable U. There is no need to specify the distribution for U as long as we know the form of $\mathcal{G}(\cdot)$.

The estimation procedure consists of two nested loops, maximize $\{dH(\cdot|\Omega)\}$, given $\Omega = (\beta, \eta)$, and then maximize the profile log-likelihood over Ω. The two nested loops are described in Section 3.5.2 when $J = 1$ and $\alpha_J = 1$.

The next section describes the statistical setting of the model specified using a logis-

tic link and the corresponding likelihood under the case of multivariate outcomes with multiple scales.

3.4 The Multivariate Model

Suppose there are m distinct continuous, ordinal, count or time-to-event outcomes with observed responses $Y = (Y_1, \ldots, Y_m)$ with values in domain space $D = (D_1, \cdots, D_m)$ respectively. For any j, if the jth outcome Y_j is a time-to-event outcome, we assume for each subject, the censoring time C_j^* is independent of true event time T_j, given covariate set Z_j. The observed event time is $Y_j = (T_j \wedge C_j)$ and $C_j = (C_j^* \wedge \tau_j)$. Let $\delta_j = \mathbb{1}(T_j \leq C_j)$ be the censoring indicator. Here $\mathbb{1}(\cdot)$ is the indicator function, and τ_j denotes the time to the end of the study for the jth outcome. The domain space for Y_j is $D_j = [0, \tau_j]$. For any j, if the jth outcome Y_j is non-censored continuous, ordinal, or count outcome, then $\delta_j = 1$ for every such observation. We develop the main framework for continuous, ordinal, count and survival outcomes first. Such outcome distributions can be represented by a tail/survival function. Later at the end of this section, we show how the approach is extended to include multinomial responses.

Let U be the latent variable that is shared by all observable outcomes. The observed outcomes are assumed to be conditionally independent given U. Let $\bar{F}(x) = P(X > x)$ denote the tail/survival distribution of the variable x. We define the conditional tail/survival distribution function for the jth outcome Y_j as a semiparametric transformation model through a logistic link

$$\bar{F}_j(y \mid U) = \frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y)}, \tag{3.5}$$

where $H_j(y)$ is an unspecified non-negative and non-decreasing function of y that ranges from 0 to ∞. The covariate effect is $\gamma_j = \exp(Z_j^T \beta_j)$, where Z_j is the covariate vector and β_j is the vector of regression coefficients for the jth outcome. α_j is the factor loading

for the jth outcome. Notice that β_j is the proportional odds ratio for one unit increase in Z_j on Y_j, given U. See Appendix B.1 for derivation. The conditional probability mass/density function for Y_j is

$$P(Y_j = y \mid U) = \bar{F}_j(y^- \mid U) - \bar{F}_j(y \mid U) = \frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y^-)} - \frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y)}, \quad (3.6)$$

where $y^- = \lim_{\Delta \to 0}(y - \Delta)$.

Consider a subject with observed outcomes $y_1, \ldots, y_m$, then the conditional joint probability distributional function is

$$\begin{aligned}
&\prod_{j=1}^{m} \left[\bar{F}_j(y_j^- \mid U) - \bar{F}_j(y_j \mid U)\right] \\
=&\prod_{j=1}^{m} \left[\frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y_j^-)} - \frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y_j)}\right]^{\delta_j} \left[\frac{1}{1 + U^{\alpha_j}\gamma_j H_j(y_j)}\right]^{1-\delta_j}.
\end{aligned}$$

For identifiability, Z_j do not contain constant terms, $\alpha_1 = 1$ and $\gamma_1 = 1$. The factor loading α_j determines the dependence between Y_j and Y_1.

Suppose there are n subjects with m distinct observed outcomes. Let $i = 1, \ldots, n$ denote the ith participant, and $j = 1, \ldots, m$ denote the jth outcome. Let y_{ij} denote the observed response of participant i on outcome j. The latent variable U_i is the latent trait of interest for the ith participant. For participant $i = 1, \ldots, n$, we observe the covariate matrix $Z_i = (Z_{i1}, \ldots, Z_{im})$, each corresponding to a vector of outcomes $Y_i = (Y_{i1}, \ldots, Y_{im})$ and censoring status $\delta_i = (\delta_{i1}, \cdots, \delta_{im})$. The observed data for subject $i = 1, \ldots, n$ consist of i.i.d. $\{Y_i, \delta_i, Z_i\}$ observations. Note that for non-survival outcomes, δ_{ij} is always 1.

Let $H = (H_1, \ldots, H_m)$. The complete data likelihood for the observed data (Y_i, Z_i, U_i) for $i = 1, \ldots, n$ is

$$L_0(\beta, \alpha, H|Y, Z, U) = \prod_{i=1}^{n}\prod_{j=1}^{m} \left[\frac{1}{1 + U_i^{\alpha_j}\gamma_{ij} H_j(y_{ij}^-)} - \frac{1}{1 + U_i^{\alpha_j}\gamma_{ij} H_j(y_{ij})}\right]^{\delta_{ij}} \left[\frac{1}{1 + U_i^{\alpha_j}\gamma_{ij} H_j(y_{ij})}\right]^{1-\delta_{ij}},$$

and the complete data log-likelihood can be written as

$$
\begin{aligned}
&\ell_0(\beta, \alpha, H|Y, Z, U) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\delta_{ij}\log\left[\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)} - \frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right] + (1-\delta_{ij})\log\left[\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right]\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij}\log\left(\frac{U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij})}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]}\right) + (1-\delta_{ij})\log\left(\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij}\log U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij}\log[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})] \\
&- \sum_{i=1}^{n}\sum_{j=1}^{m}(1-\delta_{ij})\log[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})].
\end{aligned}
$$

Here we assume H_j is an outcome-specific arbitrary step function that only jumps at the set of observed values of $y_{ij}, i = 1, \ldots, n$. We denote the jump of H_j at value s as $dH_j(s)$. Notice H_j is similar to a cumulative hazard function in a survival model, and can be treated as an unknown outcome/time transformation.

Let $\Omega = (\alpha, \beta, \eta)$, where η is the parameter characterizing the distribution of the latent variable U. Denote $f_U(u;\eta)$ as the distribution of the shared latent variable U. Note that each subject may have subject-specific distributional parameters θ_i that depend on subject-specific covariates Z_i. For example, in Appendix B.3 and the real data application in Section 3.8, we consider $\theta_i(\eta) = (\exp(\eta_1 Z_i), \exp(\eta_2 Z_i))$ and we estimate $\eta = (\eta_1, \eta_2)$ that characterizes the subject-specific distribution $f_U(u;\theta_i(\eta))$ of shared latent variable U_i. The observed data likelihood is the expectation of the complete data likelihood over U,

$$
L(\Omega, H|Y, Z) = \prod_{i=1}^{n}\mathbb{E}_\eta\left\{\prod_{j=1}^{m}\left[\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)} - \frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right]^{\delta_{ij}}\left[\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right]^{1-\delta_{ij}}\right\}.
$$

Our model has the flexibility to accommodate multinomial outcomes. For a K-category multinomial observable outcome Y with an observed response category c, we

define the conditional probability mass function as

$$p(Y = c \mid U) = \begin{cases} \frac{U^{\alpha}\theta_c}{1+U^{\alpha}\sum_{k=2}^{K}\theta_k}, c \geq 1, \\ \frac{1}{1+U^{\alpha}\sum_{k=2}^{K}\theta_k}, c = 1, \end{cases}$$

where θ_k is the covariate effect for category k versus the reference category 1.

Notice that since our model is a joint model of multiple outcomes, the model parameter and subject-specific latent variable can be estimated even if some outcomes are missing.

3.5 Estimation

3.5.1 Nonparametric Maximum Likelihood Estimation (NPMLE)

The full parameter sets include finite-dimensional parameter vectors $\Omega = (\alpha, \beta, \eta)$ and infinite-dimensional $H = (H_1(\cdot), \ldots, H_m(\cdot))$. We estimate Ω and $\{dH_j\}_{j=1,\ldots,m}$ jointly using the profile likelihood approach. This is accomplished by applying an EM algorithm (*Tsodikov* (2003)) to obtain implicit profile likelihood estimators $dH_j(\cdot|\Omega)$ for $dH_j(\cdot)$ that depend on Ω. To obtain stable MLEs and maintain proper density functions for each outcome, we impose a zero-tail constraint on non-survival outcomes, namely, $\bar{F}_j(y_{(n)} \mid U) = 0$, or equivalently, $dH_j(y_{(n)}) = \infty$, where $y_{(n)} = \max\{y_{1j}, \ldots, y_{nj}\}$, the maximum observed value for the j^{th} outcome.

Replacing $dH_j(\cdot|\Omega)$ in the marginal log-likelihood we obtain the profile log-likelihood $\ell_{pr}(\Omega) = \ell\left(\Omega, \{dH_j(\cdot|\Omega)\}_{j=1,\ldots,m}\right)$. On differentiation of the log-likelihood with respect to

$dH_j(s)$, we obtain the conditional score function for $dH_j(s)$ as

$$\begin{aligned}\mathcal{U}_{0,dH_j}(s) =& \frac{\partial \ell_0}{\partial dH_j(s)} \\ =& \frac{\sum_{i=1}^n \delta_{ij}\, \mathbb{1}(y_{ij} = s)}{dH_j(y_{ij})} - \sum_{i=1}^n \delta_{ij} \frac{\mathbb{1}(y_{ij}^- \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij}^-)} - \sum_{i=1}^n \frac{\mathbb{1}(y_{ij} \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})} \\ =& \frac{\sum_{i=1}^n dN_{ij}(s)}{dH_j(y_{ij})} - \sum_{i=1}^n \delta_{ij} \frac{\mathbb{1}(y_{ij}^- \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij}^-)} - \sum_{i=1}^n \frac{\mathbb{1}(y_{ij} \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})}.\end{aligned}$$

where $\sum_{i=1}^n dN_{ij}(s) = \sum_{i=1}^n \delta_{ij}\, \mathbb{1}(y_{ij} = s)$ records the number of observations in outcome Y_j are of value s.

Define the conditional expectation operator $\mathbb{E}[f \| g] = \frac{\mathbb{E}[f \cdot g]}{\mathbb{E}[g]}$. We use the EM-DCA algorithm to estimate the NPMLE. The derivation of the EM algorithm and the conditional expectation operator $\mathbb{E}[\cdot \| L_0]$ that is used in the rest of the article are presented in Appendix B.2. The EM algorithm involves iteratively update model parameters by maximizing the conditional expectation of the complete data likelihood $\mathbb{E}[\ell_0 \| L_0]$. The model parameters are updated by maximizing $\mathbb{E}[\ell_0 \| L_0]$. This can be achieved by setting the derivative of $\mathbb{E}[\ell_0 \| L_0]$ with respect to model parameters to 0. Specifically, we calculate the conditional expectation of the conditional score function given the observed data, latent variable U and the model parameters at the kth iteration. For the following derivation of the EM algorithm, we use the expression $\mathbb{E}\left[\mathcal{U}_0 \middle\| L_0^{(k)}\right]$ to represent the conditional expectation of the conditional score function given the observed data, latent variable U and the model parameters at the kth iteration. Note that $\mathbb{E}[\mathcal{U}_0 \| L_0]$ is the marginal score function as derived in Appendix B.2. Specifically, at the kth iteration, the conditional expectation of the conditional score function for dH_j can be expressed as

$$\mathbb{E}\left[\mathcal{U}_{0,dH_j}(s)\middle\|L_0^{(k)}\right] = \frac{\mathbb{E}\left[\mathcal{U}_{0,dH_j}(s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]}$$
$$= \frac{\sum_{i=1}^n dN_{ij}(s)}{dH_j^{(k+1)}(s)} - \sum_{i=1}^n \delta_{ij}\,\mathbb{1}(y_{ij}^- \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k+1)}(y_{ij}^-)}\middle\|L_0^{(k)}\right]$$
$$- \sum_{i=1}^n \mathbb{1}(y_{ij} \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k+1)}(y_{ij})}\middle\|L_0^{(k)}\right], \tag{3.7}$$

where $L_0^{(k)}$ notation is used to indicate that the complete data likelihood L_0 is parameterized by the kth iteration version $dH_j^{(k)}(\cdot|\Omega)$ of the function $dH_j(\cdot|\Omega)$.

Notice the complete data log-likelihood ℓ_0 can be represented as a difference between two concave functions. Consequently, the conditional score equation $\mathcal{U}_{0,dH_j}(s)$ has a representation of a difference between derivatives of two concave functions. Since the conditional expectation operator $\mathbb{E}\left[\quad\cdot\quad\middle\|L_0^{(k)}\right]$ is a linear operator, it does not alter convexity properties. Therefore, the unconditional score function $\mathbb{E}\left[\mathcal{U}_{0,dH_j}(s)\middle\|L_0^{(k)}\right]$ is also a difference between derivatives of two concave functions.

For the M-step, we employ the difference of convex functions algorithm (DCA) to iteratively maximize log-likelihood with respect to dH_j. DCA was first introduced by Pham Dinh Tao in their preliminary form in 1985. DCA is a version of the MM-algorithm (*Lange et al.* (2000)) that has been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao for nonconvex optimization problems (*Tao and An* (1997, 1998); *An and Tao* (2005)). DCA is particularly efficient when the target function to be minimized/maximized can be represented as a difference between two convex/concave functions. Equation (3.7) equals to zero is a self-consistency equation that can be solved iteratively. Solving the conditional expectation of the score equation (3.7) equals to zero

using DCA, we obtain a Breslow-type estimator for $dH_j(s)$ at the $(k+1)^{th}$ iteration

$$dH_j^{(k+1)}(s) = \frac{\sum_{i=1}^n dN_{ij}(s)}{\sum_{i=1}^n \delta_{ij}\, \mathbb{1}(y_{ij}^- \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right] + \sum_{i=1}^n \mathbb{1}(y_{ij} \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right]} \tag{3.8}$$

Iterations proceed until $\|dH_j^{(k+1)}(s) - dH_j^{(k)}(s)\|_2 < \epsilon$.

The conditional expectations $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right]$ and $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right]$ are computed by Laplace approximation (*Laplace* (1986)). The derivation of the EM algorithm along with DCA and Laplace approximation is shown in Appendix B.3. If a subject has only one non-missing outcome, the conditional expectations have closed form solution using the expression described in Section 3.3.

3.5.2 Estimation Procedure

The estimation procedure consists of two nested parts; maximize the full likelihood over $\{dH_j(\cdot|\Omega)\}_{j=1,\ldots,m}$, given $\Omega = \{\alpha, \beta, \eta\}$, and then maximize the profile log-likelihood over Ω. Specifically, we proceed with the following procedure for estimation.

Part 1. Maximize the full likelihood over $\{dH_j(\cdot|\Omega)\}_{j=1,\ldots,m}$, given Ω:

(1) Set $k = 0$. For each of the m outcomes, initialize $d\hat{H}_j^{(0)}(s)$ as Breslow estimates $\frac{\sum_{i=1}^n dN_{ij}(s)}{\sum_{i=1}^n \mathbb{1}(y_{ij}\geq s)\gamma_{ij}}$.

(2) Given Ω fixed, calculate $d\hat{H}_j^{(k+1)}(s)$ using the Breslow-type estimator (3.8).

(3) Keep updating $d\hat{H}_j^{(k+1)}(s)$ as in previous step until convergence $\left\|d\hat{H}_j^{(k+1)}(s) - d\hat{H}_j^{(k)}(s)\right\|_2 < \epsilon$ for all $j = 1, \ldots, m$.

Part 2. Maximize the profile likelihood $\ell_{pr}(\Omega) = \ell\left(\Omega, \{d\hat{H}_j(\cdot|\Omega)\}_{j=1,\ldots,m)}\right)$ over $\Omega = \{\alpha, \beta, \eta\}$, when $d\hat{H}_j(\cdot|\Omega)$ is obtained using Part 1.

(1) Set $r = 0$. Set $\hat{\alpha}^{(0)} = 1, \hat{\beta}^{(0)} = 0$ to start.

(2) Find $\hat{\alpha}^{(r+1)}, \hat{\beta}^{(r+1)}$ and $\hat{\eta}^{(r+1)}$ by taking one step towards maximizing the profile likelihood with respect to α, β and η using an optimization routine (e.g., BFGS).

(3) Repeat step (2) until convergence $\left\|\hat{\Omega}^{(r+1)} - \hat{\Omega}^{(r)}\right\|_2 < 10\epsilon$.

Note that Part 1 represents the inner loop nested within Part 2. The convergence tolerance for inner loop in Part 1 has to be stricter than for the outer loop in Part 2.

3.6 Asymptotic Properties

The proposed NPMLE is shown to be consistent and asymptotically normal by making use of the empirical process (*Murphy* (1995), *Zeng et al.* (2005), *Zeng and Lin* (2007), *Zeng and Lin* (2010)). The following regularity conditions are required to establish asymptotic properties of NPMLE.

1. If the k^{th} outcome is continuous or time-to-event, the true function $H_{0k}(y)$ of $H_k(y)$, is strictly increasing and continuously differentiable. If the k^{th} outcome is discrete, the true function $H_{0k}(y)$ of $H_k(y)$, is increasing. The true value of parameter set $\Omega_0 = (\beta_0, \alpha_0, \eta_0)$ and $\{H_0\}$ fall in the interior of a compact convex set $\mathcal{H}$.

2. For survival outcome, there exists a positive constant ν_0 such that $P(C^* \geq \tau | \mathbf{z}(t)) \geq \nu_0$ almost surely.

3. The number of non-missing outcomes for subject i, denoted as m_i, is bounded by some positive integer m_0 and $P(m_i \geq 2) > 0$ with probability 1.

4. All the covariate set Z_{ij} are bounded. Further, if there exist a constant vector $\mathbf{c}$ such that $[1, \mathbf{Z}_{ij}^T]\mathbf{c} = 0$ almost surely, then $\mathbf{c} = 0$.

5. If $\int \prod_j u^{\alpha_j} f(u;\eta)du = \int \prod_j u^{\alpha_{0j}} f(u;\eta_0)du$ for any subset of $j = 1, \cdots, m$, then $\eta = \eta_0, \alpha_j = \alpha_{0j}$ for all $j = 1, \cdots, m$.

6. The score operator for (Ω, H) is Fréchet differentiable at (Ω_0, H_0) with a continuously invertible derivative $-\mathcal{I}_0$. The hessian matrix $\mathcal{I}_n$ evaluated at the true values of H and Ω is positive definite, and converges in probability to a deterministic and invertible operator $\mathcal{I}_0$.

The following theorems present the consistency and weak convergence for the proposed NPMLE $\hat{\Omega} = \left(\hat{\beta}, \hat{\alpha}, \hat{\eta}\right)$ and $\hat{H} = (\hat{H}_1(\cdot), \cdots, \hat{H}_m(\cdot))$ with details given in the Appendix Section B.4.

Theorem III.1. *Let $\Omega_0 = (\beta_0, \alpha_0, \eta_0)$ and $H_0(y) = (H_{01}(y_1), \cdots, H_{0m}(y_m))$ be the true values of $\hat{\Omega} = \left(\hat{\beta}, \hat{\alpha}, \hat{\eta}\right)$ and $\hat{H}(y) = (\hat{H}_1(y_1), \cdots, \hat{H}_m(y_m))$, respectively. Under regularity conditions, $||\hat{\Omega} - \Omega_0|| \to 0$ and $\sum_{k=1}^m \sup_{y_k \in D_k} |\hat{H}_k(y_k) - H_{0k}(y_k)| \to 0$ almost surely.*

Theorem III.2. *Assuming regularity conditions hold, $n^{1/2}\{\hat{\Omega} - \Omega_0, \hat{H}(y) - H_0(y)\}$ converges weakly to a zero-mean Gaussian process in $R^d \times l^\infty(\prod_{k=1}^m \mathcal{Q}_k)$, where $\mathcal{Q}_k = \{h(t) : \|h\|_{BV[D_k]} \leq 1\}$. Furthermore, $\hat{\Omega}$ is asymptotically efficient.*

Consider a linear functional of the NPMLE $\hat{\Omega}$ and $\hat{H}$

$$n^{1/2}\left\{v^T(\hat{\Omega} - \Omega_0) + \sum_{k=1}^m \int w_k d(\hat{H}_k - H_{0k})\right\}, \tag{3.9}$$

where v is real vector, $\mathbf{h}_k$ is the vector consisting of the values of $w_k(\cdot)$ evaluated at the observed outcome values corresponding to the jumps of for $\hat{H}_k$, and $\{d\hat{H}_k\}$ is the vector of jump sizes at the observed outcome values, for $k = 1, \cdots, m$ respectively. $w_k(y)$ is in $BV[D_k]$, where $BV[D_k]$ is the space of functions with bounded total variation in D_k.

Theorem III.3. *Under regularity conditions, the linear function (3.9) converges weakly to a zero-mean Gaussian process with variance-covariance matrix*

$(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)\mathcal{I}_0^{-1}(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)^T$ *which can be consistently estimated by* $n(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)\mathcal{I}_n^{-1}(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)^T$*, where* $\mathcal{I}_n$ *is the negative Hessian matrix of the observed log-likelihood function with respect to* $\hat{\Omega}$ *and the jump sizes of* $(\hat{H}_1, \cdots, \hat{H}_m)$*.*

If we are primarily interested in Ω, the profile likelihood method (*Murphy and Van der Vaart* (2000)) can be used. Let $l_{pr}(\Omega) = \ell_n(\Omega, \hat{H}(\Omega)|Y, Z)$ be the profile log-likelihood function for Ω.

Theorem III.4. *Assuming the regularity conditions hold, for* $\varepsilon_n = O_p(n^{-1/2})$ *and any vector* v*,* $-\{l_{pr}(\hat{\Omega}+\varepsilon_n v) - 2l_{pr}(\hat{\Omega}) + l_{pr}(\hat{\Omega}-\varepsilon_n v)\}/n\varepsilon_n^2$ *converges in probability to* $v^T\Sigma^{-1}v$*, where* Σ *is the asymptotic covariance matrix of* $\sqrt{n}(\hat{\Omega} - \Omega_0)$*.*

The profile likelihood $l_{pr}(\Omega)$ can be calculated via the EM-DCA algorithm by holding Ω fixed. The negative second-order numerical difference of the profile log-likelihood function at Ω is used to estimate the inverse covariance matrix. Specifically, the (i, j)th element of the inverse covariance matrix can be consistently estimated by

$$-\{l_{pr}(\hat{\Omega} + \varepsilon_n e_i + \varepsilon_n e_j) - l_{pr}(\hat{\Omega} + \varepsilon_n e_i - \varepsilon_n e_j) - l_{pr}(\hat{\Omega} - \varepsilon_n e_i + \varepsilon_n e_j) + l_{pr}(\hat{\Omega})\}/\varepsilon_n^2$$

where e_i and e_j are the i^{th} and j^{th} canonical basis vectors respectively. Theorem III.4 can be verified by closely following the lines of *Zeng and Lin* (2010) Section 9 and *Murphy and Van der Vaart* (2000).

3.7 Simulation Study

To examine the finite-sample performance of the parameter estimates obtained by the proposed model, we conducted a Monte Carlo simulation study. We generated two continuous outcome Y_2, Y_6 and four ordinal outcomes Y_1, Y_3, Y_4 and Y_5, each with five levels, from the conditional cumulative probability function. The continuous outcome was

generated using inverse CDF transform sampling. The ordinal outcomes were generated by random sampling with level-specific probability. To show the performance of the proposed model under random missing outcomes, 10% of the subjects have part of their outcome values missing. Specifically, 2% missing one outcome; 2% missing two outcomes; 2% missing three outcomes; 2% missing four outcomes and 2% missing five outcomes.

We consider three covariates $X_{i1} \sim N(0,1), X_{i2} \sim \text{Binom}(0.5)$ and $X_{i3} \sim \text{Binom}(0.5)$. The shared latent variable U_i was generated from a Gamma$(a_i = e^{\eta_1 X_{i3}}, b_i = e^{\eta_2 X_{i3}})$ distribution where the true parameter values are specified as $\eta_1 = 1$ and $\eta_2 = 1.7$. The true $H_j(y) = \frac{1}{2}y^2$ for $j = 2, 6$ and the true $H_j(y) = \frac{1}{4}y^2$ for $j = 1, 3, 4, 5$. The covariate effect for each outcome is of the following forms: $\gamma_{i1} = 1, \gamma_{i2} = e^{\beta_2 X_{i2}}, \gamma_{i3} = e^{\beta_3 X_{i3}}, \gamma_{i4} = e^{\beta_4 X_{i1}}, \gamma_{i5} = e^{\beta_5 X_{i2}}, \gamma_{i6} = e^{\beta_6 X_{i3}}$, where the true parameters are $\beta_2 = 2, \beta_3 = \beta_4 = 0.5, \beta_5 = 1$ and $\beta_6 = 1.5$. The true factor loading parameters are $\alpha_1 = 1, \alpha_2 = 0.8, \alpha_3 = 0.6, \alpha_4 = 0.3, \alpha_5 = 0.5$ and $\alpha_6 = 1$.

We examined the performance of estimation for the proposed model under the sample size of $n = 200$ and $n = 500$; each was replicated 500 times. Standard errors were obtained from the numerically evaluated Hessian matrix at the solution.

The results of the simulation study are summarized in Table 3.1. The proposed estimation and inference procedures perform well with diminishing bias as sample sizes increases, and coverage probability at 95% nominal level. With the larger sample size, we see better agreement between empirical standard deviation and asymptotic standard errors. This suggests that the asymptotic approximation of the covariance matrix for the profile likelihood is reasonable for the sample size of $n = 200$ or larger.

Next, we examine the finite-sample performance of the parameter estimates obtained by the proposed model under the situation when one of the outcome is time-to-event data. We generated two continuous outcome Y_2, Y_6 and six ordinal outcomes Y_1, Y_3, Y_4, Y_5, Y_7 and Y_8, each with five levels, from the conditional cumulative probability function. The continuous outcome was generated using inverse CDF transform sampling. We generate one

Table 3.1: Simulation results using proposed model. β_1 to β_6 are regression coefficients describe the covariate effects on the outcome Y_1 to Y_6, respectively. α_1 to α_6 are factor loadings for the outcome Y_1 to Y_6, respectively; η_1, η_2 are coefficient effects on the Gamma distribution scale and rate parameters for the shared latent variable U. The results are based on 500 simulated datasets with sample size of $n = 200$ and $n = 500$.

n	Outcome	Type	parameter	Truth	Avg. est.	ESD	ASE	95% CP
200	Y_1	ordinal	β_1	0				
	Y_2	continuous	β_2	2	2.06	0.33	0.32	0.93
	Y_3	ordinal	β_3	0.5	0.51	0.33	0.33	0.90
	Y_4	ordinal	β_4	1	1.02	0.16	0.16	0.95
	Y_5	ordinal	β_5	0.5	0.52	0.29	0.28	0.92
	Y_6	continuous	β_6	1.5	1.54	0.38	0.38	0.92
	Y_1	ordinal	α_1	1				
	Y_2	continuous	α_2	0.8	0.88	0.22	0.23	0.90
	Y_3	ordinal	α_3	0.6	0.66	0.24	0.24	0.91
	Y_4	ordinal	α_4	0.3	0.35	0.20	0.19	0.88
	Y_5	ordinal	α_5	0.5	0.56	0.22	0.22	0.91
	Y_6	continuous	α_6	1	1.04	0.23	0.27	0.92
	U	shape	η_1	1	1.16	0.55	0.73	0.86
	U	rate	η_2	1.7	1.85	0.65	0.85	0.85
500	Y_1	ordinal	β_1	0				
	Y_2	continuous	β_2	2	2.03	0.19	0.19	0.95
	Y_3	ordinal	β_3	0.5	0.52	0.20	0.20	0.94
	Y_4	ordinal	β_4	1	1.02	0.10	0.10	0.95
	Y_5	ordinal	β_5	0.5	0.51	0.17	0.17	0.94
	Y_6	continuous	β_6	1.5	1.51	0.22	0.22	0.92
	Y_1	ordinal	α_1	1				
	Y_2	continuous	α_2	0.8	0.85	0.13	0.14	0.92
	Y_3	ordinal	α_3	0.6	0.65	0.15	0.15	0.91
	Y_4	ordinal	α_4	0.3	0.34	0.12	0.12	0.93
	Y_5	ordinal	α_5	0.5	0.55	0.13	0.13	0.93
	Y_6	continuous	α_6	1	1.03	0.14	0.16	0.93
	U	shape	η_1	1	1.25	0.42	0.50	0.90
	U	rate	η_2	1.7	1.96	0.49	0.58	0.91

Avg. est.: average of Monte Carlo estimates of the true parameter values over the 500 simulations
ESD: empirical standard deviation based on Monte Carlo estimates
ASE: average of numerically estimated standard errors
95% CP: 95% coverage probability

time-to-event outcome Y_9 using inverse CDF transform sampling with random censoring time following Uniform$(5, 20)$ distribution. This leads to around 5% censoring for Y_9. The ordinal outcomes were generated by random sampling with level-specific probability.

We consider three covariates $X_{i1} \sim N(0, 1), X_{i2} \sim \text{Binom}(0.5)$ and $X_{i3} \sim \text{Binom}(0.5)$. The shared latent variable U_i was generated from a Gamma$(a_i = e^{\eta_1 X_{i3}}, b_i = e^{\eta_2 X_{i3}})$ distribution where the true parameter values are specified as $\eta_1 = 1$ and $\eta_2 = 1.7$. The true $H_j(y) = \frac{1}{2}y^2$ for $j = 2, 6, 9$ and the true $H_j(y) = \frac{1}{4}y^2$ for $j = 1, 3, 4, 5, 7, 8$. The covariate effect for each outcome is of the following forms: $\gamma_{i1} = 1$, and $\gamma_{ik} = e^{\beta_k X_{ik}}$ for $k = 2, \cdots, 9$, where the true parameters are $\beta_2 = \beta_8 = 2, \beta_3 = \beta_5 = \beta_7 = 0.5, \beta_4 = 1, \beta_6 = 1.5$ and $\beta_9 = 0.3$. The true factor loading parameters are $\alpha_1 = 1, \alpha_2 = \alpha_4 = \alpha_8 = 0.8, \alpha_3 = \alpha_9 = 0.6, \alpha_5 = 0.9$ and $\alpha_6 = \alpha_7 = 1$.

The results of the simulation study based on 200 simulated datasets, each with $n = 100$ subjects, are summarized in Table 3.2. From our simulation experience, joint modeling of multivariate outcomes including survival outcome requires either (1) more observed outcomes per subjects (bigger m) or (2) larger sample size (bigger n), compared to joint modeling non-censored observed outcomes. This is due to the fact that survival outcome has random censoring, which in general contains less information available compared to a regular non-censored continuous outcome. To achieve stable estimation, the contribution of (1), increasing number of observed outcomes per subject, is more effective than (2), increasing sample size n.

3.8 Real Data Analysis: Pain Centrality Measurement on Pelvic Pain Patients

The proposed method was applied to measure pain centrality trait of patients undergoing hysterectomy as a treatment for pelvic pain and explain the heterogeneity of patients reported outcomes. The proposed joint shared variable model uses ad-hoc 2011

Table 3.2: Simulation results using proposed model. β_1 to β_9 are regression coefficients describe the covariate effects on the outcome Y_1 to Y_9, respectively. α_1 to α_9 are factor loadings for the outcome Y_1 to Y_9, respectively; η_1, η_2 are coefficient effects on the Gamma distribution scale and rate parameters for the shared latent variable U. The results are based on 200 simulated datasets with sample size of $n = 100$.

n	Outcome	Type	parameter	Truth	Avg. est.	ESD	ASE	95% CP
100	Y_1	ordinal	β_1	0				
	Y_2	continuous	β_2	2	2.01	0.43	0.43	0.94
	Y_3	ordinal	β_3	0.5	0.52	0.40	0.42	0.96
	Y_4	ordinal	β_4	1	1.07	0.26	0.25	0.94
	Y_5	ordinal	β_5	0.5	0.52	0.36	0.42	0.93
	Y_6	continuous	β_6	1.5	1.53	0.40	0.45	0.95
	Y_7	ordinal	β_7	0.5	0.51	0.24	0.21	0.92
	Y_8	ordinal	β_8	2	2.12	0.48	0.47	0.95
	Y_9	survival	β_9	0.3	0.25	0.42	0.36	0.89
	Y_1	ordinal	α_1	1				
	Y_2	continuous	α_2	0.8	0.78	0.27	0.27	0.90
	Y_3	ordinal	α_3	0.6	0.61	0.32	0.28	0.88
	Y_4	ordinal	α_4	0.8	0.90	0.33	0.33	0.90
	Y_5	ordinal	α_5	0.9	0.90	0.34	0.33	0.89
	Y_6	continuous	α_6	1	0.95	0.28	0.29	0.90
	Y_7	ordinal	α_7	1	1.08	0.44	0.38	0.90
	Y_8	ordinal	α_8	0.8	0.87	0.35	0.33	0.89
	Y_9	survival	α_9	0.6	0.44	0.26	0.23	0.77
	U	shape	η_1	1	1.22	0.63	0.67	0.81
	U	rate	η_2	1.7	1.93	0.74	0.75	0.80

Avg. est.: average of Monte Carlo estimates of the true parameter values over the 200 simulations
ESD: empirical standard deviation based on Monte Carlo estimates
ASE: average of numerically estimated standard errors
95% CP: 95% coverage probability

Fibromyalgia (FM) Survey Criteria (designed to characterize a similar latent construct) as the baseline instrument (Y_1) and extracts information from five other pain centrality relevant outcomes. FM is an ad-hoc construct of pain centrality measure in the medical field. FM is used for diagnosis of Fibromyalgia, a central pain disorder characterized by widespread musculoskeletal pain accompanied by fatigue, sleep, memory and mood issues. We intended to use our model to provide a model-based analogue to FM that would measure a degree of pain centrality.

The study sample consists of 225 female pelvic pain patients. We consider six cross-sectional responses of mixed scale collected prior to hysterectomy. Fibromyalgia (FM) Survey Criteria score was included as the baseline instrument (Y_1). Opioid use (OME), BPI pain severity score, BPI surgical pain score, HADS depression score and HADS anxiety score are the other five pain centrality relevant outcomes included in the model. There are 9 missing values in the FM score, 2 missing values in OME, 6 missing values in BPI pain severity score, 5 missing values in BPI surgical pain score, and 14 missing values in HADS depression and anxiety scores. There are 5 patients with one missing outcome, 8 patients with two missing outcomes, 2 patients with three and four missing outcomes, and 3 patients with 5 missing outcomes.

The distribution of each response is presented in Figure 3.1. All of them are right-skewed. We can see the six pain responses are on very different scales. Notice that 80% of the patients have 0 opioid use and one patient has extremely heavy opioid use of 120 (which is 80 higher than the second highest opioid use in the sample. Our model is robust to outliers because of nonparametric transformation of the observable outcomes. While analysis results from traditional latent variable models with pre-specified link may be dominated by influential outliers.

We include age centered at 47 years old as a covariate in our model. The unit for age is per 20 years. The subject specific latent trait, pain centrality U_i, is assumed to follow a Gamma(a_i, b_i) distribution with $a_i = e^{\eta_1 Age_i}$ and $b_i = e^{\eta_2 Age_i}$. The estimation results of

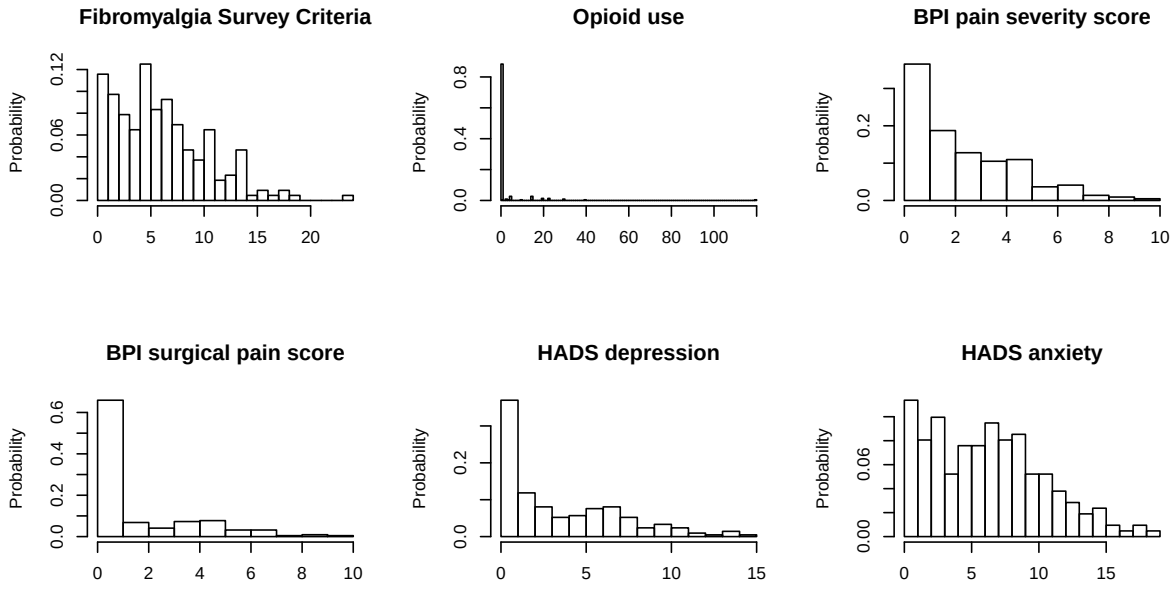

Figure 3.1: Distribution of pain responses: Fibromyalgia (FM) Survey Criteria score, Opioid use, BPI pain severity score, BPI surgical pain score, HADS depression score and HADS anxiety score. All of them are right-skewed. Notice that 80% of the patients have 0 opioid use and one patient has extremely heavy opioid use of 120 (which is 80 higher than the second highest opioid use in the sample.

the proposed joint latent trait model are shown in Table 3.3.

Table 3.3: Parameter estimates, factor loadings, standard error and p-value from analysis of $n = 225$ female pelvic pain patients. The unit for age is per 20 years. Age is centered at 47 years old.

Response	Covariate	Param Est	Standard Error	p
Fibromyalgia survey criteria	Age	0		
Opioid use	Age	0.037	0.431	0.932
BPI pain severity	Age	-1.311	0.151	< 0.0001
BPI surgical pain	Age	0.086	0.110	0.434
HADS depression	Age	-0.148	0.183	0.419
HADS anxiety	Age	0.261	0.219	0.233
Response		Factor loading	Standard Error	p
Fibromyalgia survey criteria		1		
Opioid use		0.871	0.009	< 0.0001
BPI pain severity		2.529	0.134	< 0.0001
BPI surgical pain		2.531	0.169	< 0.0001
HADS depression		0.754	0.037	< 0.0001
HADS anxiety		0.530	0.030	< 0.0001
Latent trait distribution parameter	Covariate	Param Est	Standard Error	p
η_1	Age	0.130	0.013	< 0.0001
η_2	Age	-0.893	0.099	< 0.0001

From the estimation results, we found that age has a larger effect on pain severity relative to FM survey score, and younger people feel more severe pain. The proportional odds ratio for every 20 years increase in age on BPI pain severity level is 0.27 ($e^{-1.311}$), given the model-based centrality is held constant. Thus, for every 20 years increase in age, the odds of having higher pain severity decreases by 73%, given the model-based centrality is held constant. In addition, age has significant effects on both the shape and rate parameters of latent trait distribution. Based on the factor loadings we can see that BPI pain severity and BPI surgical pain contribute the most to the construct of latent trait.

Figure 3.3 shows the relationship of estimated pain centrality with each of the six pain responses. Since the model-predicted $\log(U)$ represents how pain-uncentralized a person is, $-\log(U)$ represents pain centrality. From Figure 3.2 we can see the model-

based pain centrality score $-\log(U)$ is positively correlated with all six pain responses. The model-based pain centrality score is highly correlated with BPI pain severity, BPI surgical pain and Fibromyalgia Survey Criteria score. Age is negatively correlated with the model-based pain centrality score, as illustrated in Figure 3.4. This implies that younger patients are more pain centralized. In fact, age is negatively correlated with all six pain responses (see Appendix Table B.1 for Pearson correlations between age and all six pain responses).

Figure 3.3 shows the relationship between Fibromyalgia survey criteria score with each of the six pain responses, including our model-based pain centrality score. We can see that Fibromyalgia survey criteria score is much less correlated with overall pain and surgical pain compared to model-based pain centrality score by our model. Fibromyalgia survey criteria score is also slightly less correlated with depression and anxiety compared to model-based pain centrality score by our model. By examining the difference between Figure 3.2 and Figure 3.3, we can see that the model-based pain centrality score by our model picked up more information related to pain responses compared to the ad-hoc Fibromyalgia survey criteria score.

3.9 Discussion

In this article, we proposed a new class of shared latent variable models where a logistic link is used to accommodate nonparametrically transformed continuous, ordinal, count, multinomial and survival outcomes. The resulting parametric and nonparametric estimators are $n^{-1/2}$ consistent and asymptotically normal. The proposed model has independent nonparametrically specified transformations of different scales. Furthermore, the proposed model does not require estimating unknown thresholds in latent normal models for categorical data. Since our latent variable model is a joint model, we can estimate subject-specific shared latent trait even when some of the outcomes are missing. If a subject has only one non-missing outcome, the E-step integration for this subject can

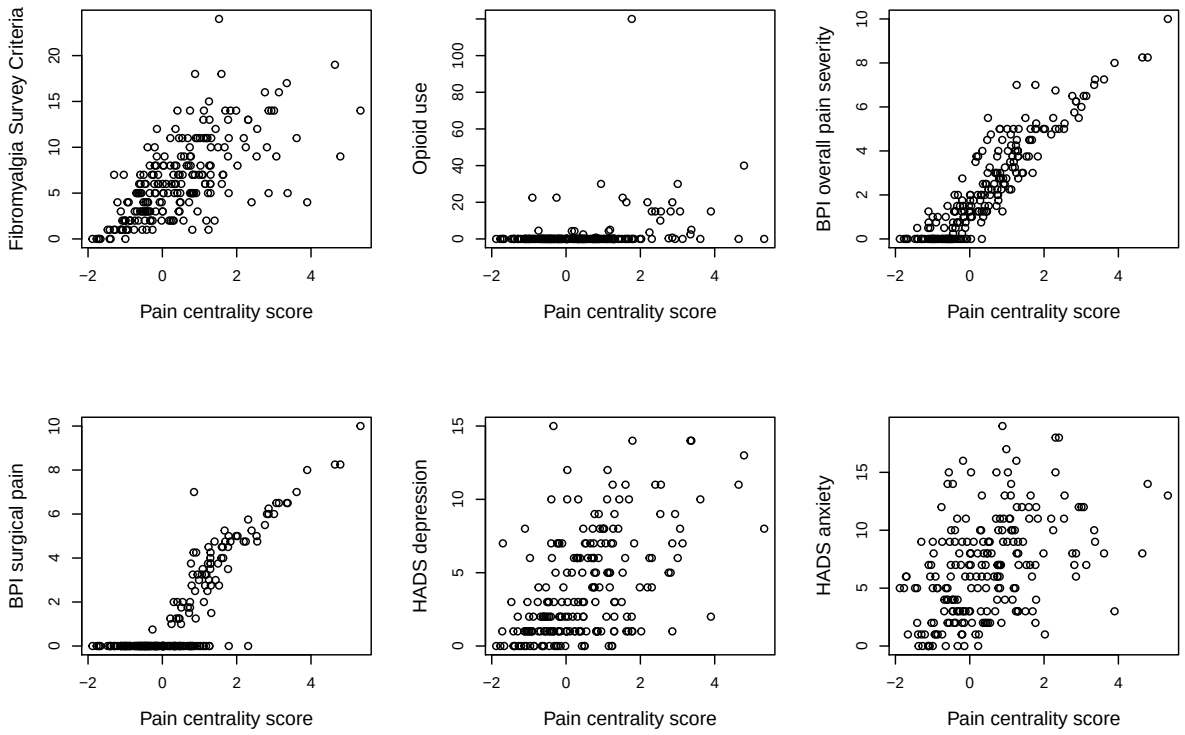

Figure 3.2: Model-based latent pain centrality score $-\log(U)$ versus each of six pain responses

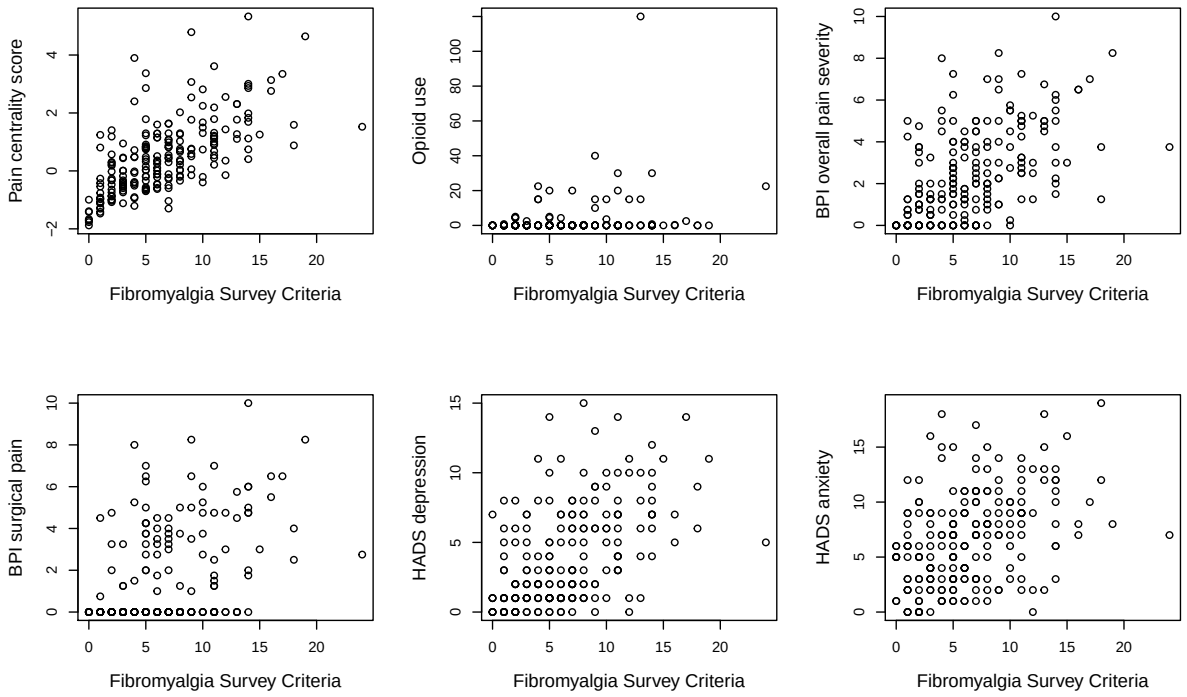

Figure 3.3: Fibromyalgia survey criteria score $-\log(U)$ versus each of six pain responses

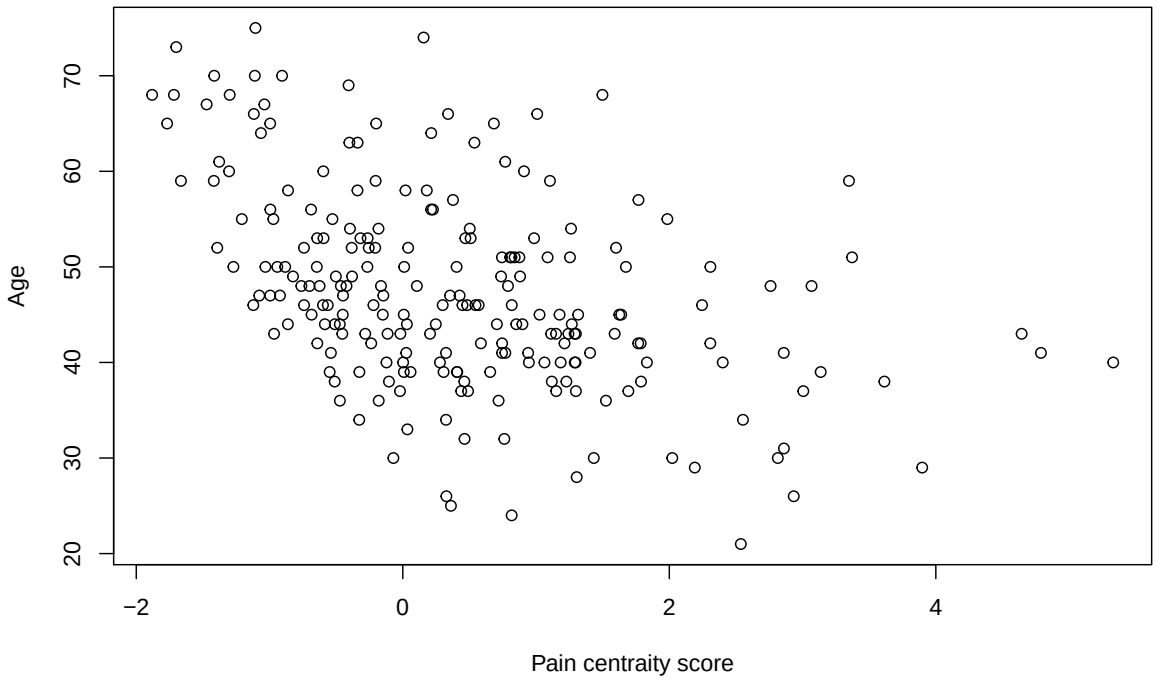

Figure 3.4: Model-based latent pain centrality score $-\log(U)$ versus age

be expressed in a closed form solution as described in Section 3.3.

Since our model assumes $\alpha_1 = 1$, in order for the latent variable U to be meaningful and represent the latent trait of interest, Y_1 has to be a latent trait relevant measurement and other outcome measurements are correlated with Y_1 through U.

Unlike gamma frailty models, the proposed model allows for both positive and negative correlation between outcomes, and unlike Gaussian frailty models it does not have to make Gaussian distributional assumption on the latent variable.

In the current framework, we consider only one latent factor in our model. However, our model can be extended to allow for multiple factors. Extra care should be exercised regarding model identifiability and factor selections.

Our method can also be extended to accommodate clustered data and longitudinal data where additional correlation is introduce across time.

Application to hysterectomy patient data indicated that the proposed method can offer an improved model-based measurement of the latent trait that better utilizes the information encoded in the multivariate multi-scale observed phenotype.

Acknowledgments

This work is supported by grants 1R01HD088712-01A1 "Peripheral and central nervous system correlates of persistent post-hysterectomy pain" and 5P50AR070600-02 "University of Michigan Fibromialgia Center of Research Translation (CORT)" from the National Institutes of Health.

CHAPTER IV

A Semiparametric Joint Latent Trait Model for Multiple Mixed Longitudinal Continuous, Categorical Outcomes and Time-to-event Data

4.1 Introduction

In biomedical studies, multivariate response data consisting of mixtures of continuous and discrete variables are often collected repeatedly over time. Multidimensional longitudinal data of mixed types are collected to fully explore the latent trait trajectory that is often of main interest but cannot be measured directly. In addition, time-to-event data is often considered if the occurrence of the terminal event is dependent on the latent trait of interest. Statistical approaches were developed to jointly modeling longitudinal responses of mixed scales and the event time data to improve inference for latent trait trajectory, and to account for the dependency the two correlated processes. Joint multivariate modeling avoids the issue of multiple testing and substantially improves the efficiency of estimation if the responses are correlated. Proper analysis of longitudinal responses of mixed scales needs to account for dependency across responses and the dependency across time points.

A number of research articles addressed the analysis of multivariate longitudinal data that incorporates the latent variable. *Dunson* (2003) proposed dynamic latent trait models to account for serial correlations using Gaussian latent variable. There are joint multi-

variate models that incorporate random effects (*Gueorguieva and Sanacora* (2006), *Jaffa et al.* (2016)). *Ghosh and Hanson* (2010)) proposed a semiparametric approach with a mixture of Polya trees for random effect distributions. *Proust-Lima et al.* (2013) implemented a latent process model for multivariate mixed longitudinal outcomes. *Kunihama et al.* (2016) proposed a nonparametric Bayes models for mixed scale longitudinal surveys which models the latent continuous variable through a Dirichlet process mixture of Gaussian factor models and model the subject specific trajectory by time-varying latent factors via Gaussian processes.

There are a couple of recent research on joint modeling longitudinal measurements and survival data. *Hickey et al.* (2016) gave a comprehensive review of recent developments and issues in joint modeling of time-to-event and multivariate longitudinal outcomes. The majority of previous studies focus on longitudinal measurements of the same scale. Regarding joint models incorporating multiple outcomes of mixed types with a time-to-event data, *Rizopoulos and Ghosh* (2011) proposed a semiparametric joint model for continuous and binary longitudinal outcomes and a time-to-event data, in which the latent variable is modelled using a Dirichlet Process prior formulation. *He and Luo* (2016) developed a joint model for continuous and ordinal longitudinal outcomes and a terminal event time, linked through shared random effects. *Proust-Lima et al.* (2016) developed a joint model for multiple longitudinal responses of different scales and competing risks, in which a latent process model was used to describe the latent trait trajectory, and a latent class structure links the longitudinal and cause-specific survival models.

The previous work on the joint model of multivariate longitudinal responses either maps the discrete outcomes to latent continuous variables or rely on pre-specified link functions based on exponential family. Further, there is not a single longitudinal model that accommodates all continuous, ordinal, count and multinomial outcome types. As for the joint model between longitudinal and survival outcomes, there is no single joint model that allows for nonparametric transformation of the longitudinal outcomes, and no single

model accommodate all continuous, ordinal, count and multinomial outcome types.

Motivated by the needs to develop a general statistical framework for longitudinal responses of mixed types and survival times, we propose a semiparametric joint model with shared latent trait trajectory for multiple longitudinal responses of mixed scales and time-to-event data. A logistic link is used to accommodate continuous, ordinal, count, and multinomial repeated measurements. A proportional odds survival model is used to model the time-to-event data. The time-to-event data is linked with longitudinal responses through subject-specific random effects. The model is used to provide a subject-specific measure of the latent trait trajectory over time. The proposed model uses a nonparametric transformation function on longitudinal responses of different scales. The proposed model avoids the restrictive multivariate normal assumption of the underlying continuous latent responses. In addition, the modeling framework is also generic with respect to the parametric distribution assumed for the latent trait trajectory, allowing a flexible choice of the process of the trait. Our model allows for negative correlation between multivariate outcomes through factor loadings, therefore, we do not need to assume Gaussian distribution/process for the latent trait trajectory to allow for negative correlation among outcomes. Our model parameters are estimated by maximizing the full-likelihood so likelihood based standard errors can be used for inference and the estimates are asymptotically efficient.

The proposed method is applied to measure the pain centrality trajectory of patients undergoing hysterectomy as a treatment for pelvic pain with longitudinal pain-related responses measured prior to the surgery, one month after the surgery and three months after the surgery. The method is compared with the ad-hoc 2011 Fibromyalgia (FM) Survey Criteria instrument designed to characterize a similar construct. Difference of convex functions algorithm (DCA) is used to estimate the nonparametric transformation functions of the model. Covariate parameters and distributional parameters for latent variable are estimated by maximizing the profile likelihood using the Broyden–Fletcher–Goldfarb–Shanno

(BFGS) algorithm. The rest of the article is organized as follows. The proposed joint model framework and likelihood function are presented in Section 4.2. The estimation procedures are described in Section 4.3. The asymptotic properties are discussed in Section 4.4. Section 4.5 shows the simulation study. In Section 4.6 we applied the proposed joint longitudinal model to characterize latent pain centrality trajectories for pelvic pain patients over the three month period. Section 4.7 concludes the proposed method and discussion.

4.2 Joint Model Framework

The proposed joint model consists of three linked submodels: (1) a proportional odds model for the longitudinal continuous, ordinal, and count measurements; (2) a proportional odds model for the time-to-event data; (3) a multinomial logistic model for nominal responses.

4.2.1 Proportional Odds Model for Longitudinal Continuous, Ordinal and Count Responses

Suppose there are J distinct continuous, ordinal, or count outcomes measured for n participants at K follow up times. Let y_{ijk} be the jth observed outcome for participant i at time t_k, where $i = 1, \cdots, n, j = 1, \cdots, J$, and $k = 1, \cdots, K$. Therefore, for participant i at time t_k, we observe outcome vector $\mathbf{y}_{ik} = (y_{i1k}, \ldots, y_{iJk})^T$ with values in domain space $\mathbf{D} = (D_1, \cdots, D_J)^T$ respectively. Let $U_i(t)$ be the latent trait function that is shared by all observable outcomes at time t. $U_i(t)$ represents the underlying latent trait score for participant i at time t. The observed outcomes are assumed to be conditionally independent given $U_i(t)$ for all times. Let $\bar{F}(x) = P(X > x)$ denotes the tail/survival distribution of the variable x. We define the conditional tail/survival distribution function

for Y_{ijk} as a semiparametric transformation model through a logistic link

$$\bar{F}_j(y_{ijk} \mid U_i(t_k)) = \frac{1}{1 + U_i^{\alpha_j}(t_k)\gamma_{ij}H_j(y_{ijk})}, \tag{4.1}$$

where $H_j(y)$ is an arbitrary (nonparametrically specified) non-negative and non-decreasing function of y that ranges from 0 to ∞. The covariate effect is $\gamma_{ij} = \exp(\mathbf{Z}_{ij}^T\beta_j)$, where $\mathbf{Z}_{ij}$ is the covariate vector for participant i for the jth outcome, and β_j is the vector of regression coefficients. α_j is the factor loading for the jth outcome. The conditional probability mass/density function for Y_{ijk} is

$$P(Y_{ijk} = y_{ijk} \mid U_i(t_k)) = \frac{1}{1 + U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)} - \frac{1}{1 + U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})},$$

where $y^- = \lim_{\Delta \to 0}(y - \Delta)$.

Let $\mathbf{y}_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{iK})^T$ be the outcome vector across time for participant i. Let $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \cdots, \mathbf{Z}_{iJ})^T$. Let $U_i(\mathbf{t}) = (U_i(t_1), \cdots, U_i(t_K))^T$ be the trajectory of latent trait over time for participant i. The conditional likelihood of the multiple mixed longitudinal outcomes for participant i is

$$L_Y(\mathbf{y}_i|U_i(\mathbf{t})) = \prod_{j=1}^{J}\prod_{k=1}^{K}\left[\frac{1}{1 + U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)} - \frac{1}{1 + U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})}\right] \tag{4.2}$$

The variable t_k is the time of measurement with $t_1 = 0$ as the baseline. For identifiability, $\mathbf{Z}_{ij}$ do not contain constant terms, $\alpha_1 = 1$ and $\gamma_{i1} = 1$ for all i. The factor loading α_j determines the dependence between Y_j and Y_1. The latent trait function $U_i(t)$ can take a flexible functional form of t or can be a process over time. As an example, we assume the latent trait function takes the form $U_i(t) = U(t)e^{a_i + b_i t}$. The function $U(t)$ represents population average disease trajectory over time. The random variable a_i and b_i represent subject-specific disease severity at baseline and disease progression rate, respectively, relative to the population average. The function $U(t)$ satisfies the constraints

that $U(0) = 1$ and $U(t) > 0$ for all t. The shared random variables (a_i, b_i) follow a joint distribution $f(a, b|\theta_i)$. Note that each subject may have subject-specific distributional predictors θ_i that depend on subject-specific covariates Z_i. For example, in Appendix C.1 and the real data application in Section 4.6, we consider $f(a, b|\theta_i) = f(a|\theta_{1i}, \theta_{2i})f(b|\eta_3)$, where $f(a|\theta_{1i}, \theta_{2i})$ is a log-Gamma density function with shape $\theta_{1i} = \exp(\eta_1 Z_i)$, and rate $\theta_{2i} = \exp(\eta_2 Z_i)$. $f(b|\eta_3)$ is a log-Gamma density function with both shape and rate being η_3. For the rest of the article, we denote the distribution of (a_i, b_i) as $f_{ab}(a_i, b_i|\boldsymbol{\eta})$ where $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T$.

4.2.2 Proportional Odds Model for Time-to-event Data

For the time-to-event data, we use a proportional odds survival model that shares the subject-specific random variables (a_i, b_i) in (4.2). Assume for each participant $i = 1, \cdots, n$, the censoring time C_i^* is independent of true event time T_i^*, given covariate set Z_{is}. The observed event time is $T_i = (T_i^* \wedge C_i)$ and $C_i = (C_i^* \wedge \tau)$. Let $\delta_i = \mathbb{1}(T_i \leq C_i)$ be the censoring indicator. Here $\mathbb{1}(\cdot)$ is the indicator function, and τ denotes the time to the end of the study. The domain space for T_i is $(0, \tau]$. Under the proportional odds survival model, the conditional survival function for T_i is

$$\bar{F}_s(T_i|a_i, b_i) = \frac{1}{1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s(T_i)},$$

where v_0 and v_1 measure the association between the longitudinal sub-model and the survival sub-model. $H_s(t)$ is an unspecified non-negative and non-decreasing function in $[0, \tau]$ with $H_s(0) = 0$ and $H_s(\infty) = \infty$. The conditional density function of having a terminal event at time T_i is

$$f_s(T_i|a_i, b_i) = \frac{e^{v_0 a_i + v_1 b_i}\gamma_{is} h_s(T_i)}{[1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s(T_i)]^2},$$

where $h_s(t) = dH_s(t)/dt$.

Therefore, the conditional likelihood for a participant i with observed data (T_i, δ_i, Z_{is}) is

$$L_s(T_i, \delta_i | a_i, b_i) = \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right)$$

Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)^T$ be the regression coefficient matrix. Let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_J)^T$ be the vector of factor loadings. Let the functional parameter $\mathbf{H} = (H_1, \cdots, H_J, H_s)$ be the set of transformation functions for the 1st to the Jth outcome and the survival data, respectively. The observed data for subject $i = 1, \ldots, n$ consist of i.i.d. $\{\mathbf{Y}_i, T_i, \delta_i, \mathbf{Z}_i, Z_{is}\}$ observations. The complete data likelihood for the observed data is

$$\begin{aligned} L_0(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{H} | \mathbf{Y}, \mathbf{Z}, U(t), a_i, b_i) &= \prod_{i=1}^{n} L_Y(\mathbf{y}_i | U_i(\mathbf{t})) L_s(T_i, \delta_i | a_i, b_i) \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} \left[\frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)} - \frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})} \right] \right. \\ &\qquad \left. \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right) \right\}, \end{aligned}$$

and the complete data log-likelihood is

$$\begin{aligned} &\ell_0(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{H} | \mathbf{Y}, \mathbf{Z}, U(t), a_i, b_i) \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K} \log \left(\frac{U_i(t_k)^{\alpha_j} \gamma_{ij} dH_j(y_{ijk})}{[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})]} \right) \right. \\ &\left. + \delta_i \log \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right) - \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \{ \log U_i(t_k)^{\alpha_j} \gamma_{ij} dH_j(y_{ijk}) - \log[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})] \} \\ &+ \sum_{i=1}^{n} \{ \delta_i (v_0 a_i + v_1 b_i + \log \gamma_{is} dH_s(T_i)) - (1 + \delta_i) \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] \}. \end{aligned}$$

Here we assume H_j is an outcome-specific step function that only jumps at the set of observed values of $y_{ijk}, i = 1, \ldots, n, k = 1, \cdots, K$. We denote the jump of H_j at value x as $dH_j(x)$. H_s is assumed to be a step function that only jumps at the observed event times.

Let $\Omega = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{v})$ be the set of model parameters, where $\mathbf{v} = (v_0, v_1)$ and $\boldsymbol{\eta}$ is the parameter characterizes the distribution of the latent variable $U_i(t)$ and (a_i, b_i). The marginal likelihood for the observed data is the expectation of the complete data likelihood over (a_i, b_i),

$$L(\Omega, \mathbf{H}|\mathbf{Y}, \mathbf{Z}) = \prod_{i=1}^{n} \mathbb{E}_{\eta} \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} \left[\frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)} - \frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})} \right] \right.$$
$$\left. \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right) \right\}. \quad (4.3)$$

4.2.3 Multinomial Logistic Model for Nominal Responses

Our model has the flexibility to accommodate multinomial outcomes. For a M-category multinomial observable outcome Y_{ijk} with an observed response category c, we define the conditional probability mass function as

$$p(Y_{ijk} = c \mid U_i(t_k)) = \begin{cases} \frac{U_i(t_k)^{\alpha_j} \theta_c}{1 + U_i(t_k)^{\alpha_j} \sum_{m=2}^{M} \theta_m}, c \geq 1, \\ \frac{1}{1 + U_i(t_k)^{\alpha_j} \sum_{m=2}^{M} \theta_m}, c = 1, \end{cases}$$

where θ_m is the covariate effect for category m versus the reference category 1. Since the multinomial part of the likelihood does not include infinite dimensional parameters, we omit it for brevity in the following exposition.

Notice that since our model is a joint model of multiple outcomes, the model parameters and subject-specific latent variable can be estimated even if some outcomes are missing.

4.3 Estimation

4.3.1 Nonparametric Maximum Likelihood Estimation (NPMLE)

The full parameter sets are finite-dimensional parameter vectors $\Omega = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{v})$, infinite-dimensional $\mathbf{H} = (H_1(\cdot), \ldots, H_J(\cdot), H_s(\cdot))$ and $U(\cdot)$. Denote $\{U\} = (U(t_1), \cdots, U(t_K))$ and $\{dH_j\}$ as the vector of jumps for H_j at the observed values of the jth outcome. We estimate $\Omega, \{U\}$ and $\{dH_s\}, \{dH_j\}_{j=1,\ldots,J}$ jointly using the profile likelihood approach. This is accomplished by applying an EM algorithm (*Tsodikov* (2003)) to obtain implicit estimators $\{\hat{U}(\Omega)\}, \{d\hat{H}_s(\Omega)\}$ and $\{d\hat{H}_j(\Omega)\}$ that depend on Ω being held fixed, when $\{U\}, \{dH_s\}$ and $\{dH_j\}$ are profiled out. To obtain stable MLEs and maintain proper density functions for each outcome, we impose a zero-tail constraint on longitudinal outcomes, namely, $\bar{F}_j(y_{j(n)} \mid U_i(t)) = 0$, or equivalently, $dH_j(y_{j(n)}) = \infty$, where $y_{j(n)} = \max\{y_{ijk}, i = 1, \cdots, n, k = 1, \cdots, K\}$, the maximum observed value for the jth outcome over all participants and across all times.

Replacing $\{U(\Omega)\}$, $\{dH_s(\Omega)\}$ and $\{dH_j(\Omega)\}_{j=1,\ldots,J}$ in the marginal log-likelihood we obtain the profile log-likelihood $\ell_{pr}(\Omega) = \ell\left(\Omega, \{U(\Omega)\}, \{dH_s(\Omega)\}, \{dH_j(\Omega)\}_{j=1,\ldots,J}\right)$. To obtain the conditional score function for dH_j, we differentiate the log-likelihood with respect to $dH_j(x)$

$$
\begin{aligned}
&\mathcal{U}_{0,dH_j(x)} = \frac{\partial \ell_0}{\partial dH_j(x)} \\
=&\frac{\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)}{dH_j(x)} - \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)} - \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})} \\
=&\frac{\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x)}{dH_j(x)} - \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)} - \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})}.
\end{aligned}
$$

where $\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x) = \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)$ records the multiplicity of observations in the jth outcome having value x.

Let $Y_{is}(t) = \mathbb{1}(T_i \geq t)$ be the at-risk process for time-to-event data for subject i, and $N_{is}(t) = \delta_i \, \mathbb{1}(T_i \leq t)$ be the counting process that records the number of events that have occurred by time t for subject i. On functional differentiation of the log-likelihood with respect to $dH_s(x)$, we obtain the conditional score function for dH_s

$$\mathcal{U}_{0,dH_s(x)} = \frac{\partial \ell_0}{\partial dH_s(x)} = \frac{\sum_{i=1}^n dN_{is}(x)}{dH_s(x)} - \sum_{i=1}^n \frac{Y_{is}(x)(1+\delta_i)e^{v_0 a_i + v_1 b_i}\gamma_{is}}{1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s(T_i)}$$

The conditional score function for $\{U\}$ can be obtained by functionally differentiating the log-likelihood with respect to $U(x)$:

$$\mathcal{U}_{0,U(x)} = \frac{\partial \ell_0}{\partial U(x)} = \frac{\sum_{i=1}^n \sum_{j=1}^J \alpha_j \, \mathbb{1}(H_j(y_{ijx}) \neq \infty)}{U(x)} - \sum_{i=1}^n \sum_{j=1}^J \frac{\alpha_j U(x)^{\alpha_j - 1} e^{\alpha_j(a_i + b_i x)} \gamma_{ij} H_j(y_{ijx}^-)}{1 + U_i^{\alpha_j}(x)\gamma_{ij} H_j(y_{ijx}^-)}$$
$$- \sum_{i=1}^n \sum_{j=1}^J \frac{\mathbb{1}(H_j(y_{ijx}) \neq \infty)\alpha_j U(x)^{\alpha_j - 1} e^{\alpha_j(a_i + b_i x)} \gamma_{ij} H_j(y_{ijx})}{1 + U_i^{\alpha_j}(x)\gamma_{ij} H_j(y_{ijx})}$$

Define the conditional expectation operator $\mathbb{E}[f \| g] = \frac{\mathbb{E}[f \cdot g]}{\mathbb{E}[g]}$. We use the EM-DCA algorithm to estimate the NPMLE. The derivation of the EM algorithm and the conditional expectation operator $\mathbb{E}[\cdot \| L_0]$ that is used in the rest of the article are presented in Appendix B.2. The EM algorithm involves iteratively update model parameters by maximizing the conditional expectation of the complete data likelihood $\mathbb{E}[\ell_0 \| L_0]$. The model parameters are updated by maximizing $\mathbb{E}[\ell_0 \| L_0]$. This can be achieved by setting the derivative of $\mathbb{E}[\ell_0 \| L_0]$ with respect to model parameters to 0. Specifically, we calculate the conditional expectation of the conditional score function given the observed data, latent variable a_i, b_i and the model parameters at the kth iteration. For the following derivation of the EM algorithm, we use the expression $\mathbb{E}\left[\mathcal{U}_0 \middle\| L_0^{(m)}\right]$ to represent the conditional expectation of the conditional score function given the observed data, latent variable a_i, b_i and the model parameters at the mth iteration. Note that $\mathbb{E}[\mathcal{U}_0 \| L_0]$ is the marginal score function as derived in Appendix B.2.

The E-step involves taking conditional expectation $\mathbb{E}[\cdot \| L_0]$ of the conditional score

functions over a_i and b_i. At the m^{th} iteration,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{U}_{0,dH_j(x)}\middle\|L_0^{(m)}\right] = \frac{\mathbb{E}\left[\mathcal{U}_{0,dH_j(x)}L_0^{(m)}\right]}{\mathbb{E}\left[L_0^{(m)}\right]} \\
&= \frac{\sum_{i=1}^n\sum_{k=1}^K dN_{ijk}(x)}{dH_j^{(m+1)}(x)} - \sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x)\gamma_{ij}U_i^{\alpha_j}(t_k)}{1+U_i^{\alpha_j}(t_k)\gamma_{ij}H_j^{(m+1)}(y_{ijk}^-)}\middle\|L_0^{(m)}\right] \\
&- \sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x)\gamma_{ij}U_i^{\alpha_j}(t_k)}{1+U_i^{\alpha_j}(t_k)\gamma_{ij}H_j^{(m+1)}(y_{ijk})}\middle\|L_0^{(m)}\right], \qquad (4.4)
\end{aligned}$$

$$\mathbb{E}\left[\mathcal{U}_{0,dH_s(x)}\middle\|L_0^{(m)}\right] = \frac{\sum_{i=1}^n dN_{is}(x)}{dH_s^{(m+1)}(x)} - \sum_{i=1}^n \mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0a_i+v_1b_i}}{1+e^{v_0a_i+v_1b_i}\gamma_{is}H_s^{(m+1)}(T_i)}\middle\|L_0^{(m)}\right], \qquad (4.5)$$

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{U}_{0,U(x)}\middle\|L_0^{(m)}\right] \\
&= \frac{\sum_{i=1}^n\sum_{j=1}^J \alpha_j\,\mathbb{1}(H_j(y_{ijx}) \neq \infty)}{U^{(m+1)}(x)} - \sum_{i=1}^n\sum_{j=1}^J \mathbb{E}\left[\frac{\alpha_j U^{(m+1)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U_i^{(m+1)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx}^-)}\middle\|L_0^{(m)}\right] \\
&- \sum_{i=1}^n\sum_{j=1}^J \mathbb{E}\left[\frac{\mathbb{1}(H_j(y_{ijx}) \neq \infty)\alpha_j U^{(m+1)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx})}{1+U_i^{(m+1)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx})}\middle\|L_0^{(m)}\right], \qquad (4.6)
\end{aligned}$$

where $L_0^{(m)}$ indicates that the complete data likelihood L_0 is parameterized by the mth iteration copy of the parameters $\{U\}, \{dH_s\}, \{dH_j\}_{j=1,\cdots,J}$.

The complete data log-likelihood ℓ_0 can be represented as a difference between two concave functions of $\{dH_j\}$ for each j, holding all other variables fixed. Consequently, the conditional score equation $\mathcal{U}_{0,dH_j(x)}$ has a representation of a difference between derivatives of two concave functions. Since the imputation operator $\mathbb{E}[f\|g]$ is a linear operator, it does not alter convexity properties. Therefore, the unconditional score function $\mathbb{E}\left[\mathcal{U}_{0,dH_j(x)}\middle\|L_0^{(m)}\right]$ is also a difference between derivatives of two concave functions. Sim-

ilarly, ℓ_0 can be represented as a difference between two concave functions of $\{dH_s\}$, holding all other variables fixed. And ℓ_0 can be represented as a difference between two concave functions of $\{U\}$, holding all other variables fixed. Therefore, the unconditional score functions $\mathbb{E}\left[\mathcal{U}_{0,dH_s(x)}\middle\|L_0^{(m)}\right]$ and $\mathbb{E}\left[\mathcal{U}_{0,U(x)}\middle\|L_0^{(m)}\right]$ each is a difference between the derivatives of two concave functions.

For the M-step, we employ the difference of convex functions algorithm (DCA) to iteratively maximize log-likelihood with respect to $\{dH_j\}, \{dH_s\}$, and $\{U\}$. DCA was first introduced by Pham Dinh Tao in its preliminary form in 1985. DCA is a version of MM-algorithm (*Lange et al.* (2000)) that has been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao for nonconvex optimization problems (*Tao and An* (1997, 1998); *An and Tao* (2005)). DCA is particularly efficient when the target function to be minimized/maximized can be represented as a difference between two convex/concave functions. Equations (4.4), (4.5) and (4.6) are a set of self-consistency equations that can be solved iteratively. Solving the conditional expectation of the score equations (4.4), (4.5) and (4.6) equal to zero, respectively, and employing DCA, we obtain Breslow-type estimators for $dH_j(x), dH_s(x)$, and an updating equation for $U(x)$, at the $(m+1)$th iteration

$$dH_j^{(m+1)}(x) = \frac{\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x)}{\sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x)\gamma_{ij} U_i^{(m)}(t_k)^{\alpha_j}}{1+U_i^{(m)}(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ij}^-)}\middle\|L_0^{(m)}\right] + \sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x)\gamma_{ij} U_i^{(m)}(t_k)^{\alpha_j}}{1+U_i^{(m)}(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijk})}\middle\|L_0^{(m)}\right]}, \tag{4.7}$$

$$dH_s^{(m+1)}(x) = \frac{\sum_{i=1}^n dN_{is}(x)}{\sum_{i=1}^n \mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0 a_i + v_1 b_i}}{1+e^{v_0 a_i + v_1 b_i}\gamma_{is}H_s^{(m)}(T_i)}\middle\|L_0^{(m)}\right]}, \tag{4.8}$$

$$U^{(m+1)}(x)$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\alpha_j\,\mathbb{1}(H_j(y_{ijx})\neq\infty)}{\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{E}\left[\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j^{(m)}(y_{ijx}^-)}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijx}^-)} - \frac{\mathbb{1}(H_j(y_{ijx})\neq\infty)\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j^{(m)}(y_{ijx})}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijx})}\middle\| L_0^{(m)}\right]} \tag{4.9}$$

Iterations proceed until the following convergence criteria are satisfied: $\|dH_j^{(m+1)}(x) - dH_j^{(m)}(x)\|_2 < \epsilon$, $\|dH_s^{(m+1)}(x) - dH_s^{(m)}(x)\|_2 < \epsilon$ and $\|U^{(m+1)}(x) - U^{(m)}(x)\|_2 < \epsilon$ for some small $\epsilon > 0$.

The conditional expectations $\mathbb{E}\left[\quad\cdot\quad\middle\| L_0^{(m)}\right]$ in (4.7), (4.8) and (4.9) were computed by Laplace approximation (*Laplace* (1986)). The derivation of the EM algorithm along with DCA and Laplace approximation is shown in Appendix C.1.

4.3.2 Estimation Procedure

The estimation procedure consists of two nested loops. The inner loop maximizes $\{dH_j(\Omega)\}_{j=1,\ldots,J}$, $\{dH_s(\Omega)\}$ and $\{U\}$, given Ω. The outer loop then maximizes the profile log-likelihood over Ω. Specifically, we proceed with the following procedure for estimation.

Inner loop. Maximize $\{dH_j(\Omega)\}_{j=1,\ldots,J}$, $\{dH_s(\Omega)\}$ and $\{U\}$, given Ω:

(1) Set $m = 0$. For each of the J outcomes, initialize $\{dH_j\}$ as the Breslow estimates $d\hat{H}_j^{(0)}(x) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}dN_{ijk}(x)}{\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(y_{ijk}\geq x)\gamma_{ij}}$. For the time-to-event outcome, initialize $\{dH_s\}$ as the Breslow estimates $d\hat{H}_s^{(0)}(x) = \frac{\sum_{i=1}^{n}dN_{is}(x)}{\sum_{i=1}^{n}Y_{is}(x)\gamma_{is}}$. In addition, $\hat{U}^{(0)}(t) = 1$ for all $t = t_1, \cdots, t_K$.

(2) Given fixed Ω, calculate $d\hat{H}_j^{(m+1)}(x)$ using the Breslow-type estimator (4.7) for all $j = 1, \cdots, J$; calculate $d\hat{H}_s^{(m+1)}(x)$ using the Breslow-type estimator (4.8); and calculate $U^{(m+1)}(x)$ using the equation (4.8).

(3) Keep updating $d\hat{H}_j^{(m+1)}(x), d\hat{H}_s^{(m+1)}(x)$ and $\hat{U}^{(m+1)}(x)$ as in previous step until all the convergence criteria are satisfied:
$\left\| d\hat{H}_j^{(m+1)}(\cdot) - d\hat{H}_j^{(m)}(\cdot) \right\|_2 < \epsilon$ for all $j = 1, \ldots, J$,
$\left\| d\hat{H}_s^{(m+1)}(\cdot) - d\hat{H}_s^{(m)}(\cdot) \right\|_2 < \epsilon$, and $\left\| \hat{U}^{(m+1)}(\cdot) - \hat{U}^{(m)}(\cdot) \right\|_2 < \epsilon$ for some small $\epsilon > 0$.

Outer loop. Maximize the profile likelihood $\ell_{pr}(\Omega) = \ell\left(\Omega, \{\hat{U}(\Omega)\}, \{d\hat{H}_s(\Omega)\}, \{d\hat{H}_j(\Omega)\}_{j=1,\ldots,J}\right)$ over Ω:

(1) Set $r = 0$. Set $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{1}, \hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}, \hat{\mathbf{v}}^{(0)} = \mathbf{1}, \hat{\boldsymbol{\eta}}^{(0)} = (0, 0, 1)^T$ to start.

(2) Find $\hat{\boldsymbol{\alpha}}^{(r+1)}, \hat{\boldsymbol{\beta}}^{(r+1)}, \hat{\mathbf{v}}^{(r+1)}$ and $\hat{\boldsymbol{\eta}}^{(r+1)}$ by taking one step towards maximizing the profile likelihood with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{v}$ and $\boldsymbol{\eta}$ using an optimization routine (e.g., BFGS). Note that when the parametric multinomial component is present, it is added to the profile log-likelihood, and the vector of parameters include the multinomial logistics parameter matrix.

(3) While executing (2), update $d\hat{H}_j^{(r+1)} = d\hat{H}_j(\Omega^{(r+1)})$, for $j = 1, \cdots, J$, $d\hat{H}_s^{(r+1)} = d\hat{H}_s(\Omega^{(r+1)})$ and $\hat{U}^{(r+1)} = \hat{U}(\Omega^{(r+1)})$ using steps in the inner loop.

(4) Repeat steps (2) and (3) until convergence $\left\| \hat{\Omega}^{(r+1)} - \hat{\Omega}^{(r)} \right\|_2 < 10\epsilon$.

Note that the convergence tolerance for the inner loop has to be stricter than for the outer loop.

4.4 Asymptotic Properties

The proposed NPMLE are shown to be consistent and asymptotically normal by making use of the empirical process (*Murphy* (1995), *Zeng et al.* (2005), *Zeng and Lin* (2007), *Zeng and Lin* (2010)). The following regularity conditions are required to establish asymptotic properties of NPMLE.

1. $H_s(\cdot)$ is a strictly increasing and continuously differentiable function. If the jth outcome is continuous, the true function $H_{0j}(\cdot)$ of $H_j(\cdot)$, is strictly increasing and continuously differentiable. The true value of parameter set $\Omega_0 = (\boldsymbol{\beta_0}, \boldsymbol{\alpha_0}, \boldsymbol{\eta_0}, \mathbf{v}_0)$, $\mathbf{H}_0 = (\mathbf{H}_{01}, \cdots, \mathbf{H}_{0J}, \mathbf{H}_{0s})$ and $U_0(\mathbf{t})$ fall in the interior of a compact convex set $\mathcal{H}$.

2. For the time-to-event outcome, there exists a positive constant ν_0 such that $P(C_i^* \geq \tau|\mathbf{z}) \geq \nu_0$ almost surely.

3. The number of non-missing outcomes for subject i, denoted as m_i, is bounded by some positive integer m_0 and $P(m_i \geq 2) > 0$ with probability 1.

4. All the covariate set Z_{ij} are bounded. Further, if there exist a constant vector $\mathbf{c}$ such that $[1, Z_{is}^T, \mathbf{Z}_{ij}^T]\mathbf{c} = 0$ almost surely, then $\mathbf{c} = 0$.

5. If

$$U(t_k)^{\alpha_j} = U_0(t_k)^{\alpha_{0j}}, j = 1, \cdots, J, k = 1, \cdots, K$$

and

$$\int\int e^{v_0 a + v_1 b + \sum_{k=1}^K \alpha_1 (a+bt_k)} f_{ab}(a, b|\eta) dadb = \int\int e^{v_0^0 a_i + v_1^0 b_i + \sum_{k=1}^K \alpha_{01}(a+bt_k)} f_{ab}(a, b|\eta_0) dadb$$

then $\alpha_j = \alpha_{0j}, v_0 = v_0^0, v_1 = v_1^0, \eta = \eta_0$ and $U(t_k) = U_0(t_k)$ for all $j = 1, \cdots, J, k = 1, \cdots, K$.

6. The score operator for $(\Omega, U, \mathbf{H})$ is Fréchet differentiable at $(\Omega_0, U_0, \mathbf{H}_0)$ with a continuously invertible derivative $-\mathcal{I}_0$. The hessian matrix $\mathcal{I}_n$ evaluated at the true values of Ω, U, and $\mathbf{H}$ is positive definite, and converges in probability to a deterministic and invertible operator $\mathcal{I}_0$.

The following theorems present the consistency and weak convergence for the proposed NPMLE $\hat{\Omega} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\mathbf{v}}\right)$, $\hat{\boldsymbol{H}} = (\hat{H}_1, \cdots, \hat{H}_J, \hat{H}_s)$ and U with details given in the

Appendix Section C.2.

Theorem IV.1. *Let* $\Omega_0 = (\boldsymbol{\beta_0}, \boldsymbol{\alpha_0}, \boldsymbol{\eta_0}, \mathbf{v}_0)$, $U_0(\mathbf{t})$ *and* $\mathbf{H}_0(y) = (H_{01}(y_1), \cdots, H_{0J}(y_J), H_{0s}(t))$ *be the true values of* $\hat{\Omega} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}, \hat{\mathbf{v}}\right)$, $\hat{U}(\mathbf{t})$ *and* $\hat{\mathbf{H}}(y) = (\hat{H}_1(y_1), \cdots, \hat{H}_J(y_J), \hat{H}_s(t))$, *respectively. Under regularity conditions,* $||\hat{\Omega} - \Omega_0|| \to 0, ||\hat{U} - U_0|| \to 0$, $\sup_{t\in[0,\tau]} |\hat{H}_s(t) - H_{0s}(t)| \to 0$ *and* $\sum_{j=1}^{J} \sup_{y_j \in D_j} |\hat{H}_j(y_j) - H_{0j}(y_j)| \to 0$ *almost surely.*

Theorem IV.2. *Assuming regularity conditions hold,* $n^{1/2}\{\hat{\Omega} - \Omega_0, \hat{U}(\mathbf{t}) - U_0(\mathbf{t}), \hat{\mathbf{H}}(y) - \mathbf{H}_0(y)\}$ *converges weakly to a zero-mean Gaussian process in* $R^d \times R^K \times l^\infty(\prod_{j=1}^{J+1} \mathcal{Q}_j)$, *where* $\mathcal{Q}_j = \{h(y) : \|h\|_{BV[D_j]} \leq 1\}$ *for* $j = 1, \cdots, J$, $\mathcal{Q}_{J+1} = \{h(t) : \|h\|_{BV[0,\tau]} \leq 1\}$, d *is the dimension of* Ω, *and* $\|h\|_{BV[D_j]}$ *denotes the total variation of* $h(\cdot)$ *in* D_j. *Furthermore,* $\hat{\Omega}$ *and* $\hat{U}$ *are asymptotically efficient.*

Consider a linear functional of the NPMLE $\hat{\Omega}, \hat{U}$ and $\hat{\mathbf{H}}$

$$n^{1/2}\left\{v^T(\hat{\Omega} - \Omega_0, \hat{U} - U_0) + \sum_{j=1}^{J} \int w_j d(\hat{H}_j - H_{0j}) + \int w_{J+1} d(\hat{H}_s - H_{0s})\right\}, \qquad (4.10)$$

where v is real vector, $\mathbf{w}_j$ is the vector consisting of the values of $w_j(\cdot)$ evaluated at the observed outcome values corresponding to the jumps of for $\hat{H}_j$, and $\{d\hat{H}_j\}$ is the vector of jump sizes at the observed outcome values, for $j = 1, \cdots, J$ respectively. For each of the j outcomes, $w_j(y)$ is in $BV[D_j]$; $w_{J+1}(t)$ is in $BV[0, \tau]$.

Theorem IV.3. *Under regularity conditions, the linear function (4.10) converges weakly to a zero-mean Gaussian process with variance-covariance matrix* $(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_{J+1}^T)\mathcal{I}_0^{-1}(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_{J+1}^T)^T$ *which can be consistently estimated by* $n(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_{J+1}^T)\mathcal{I}_n^{-1}(v^T, \mathbf{w}_1^T, \cdots, \mathbf{w}_{J+1}^T)^T$, *where* $\mathcal{I}_n$ *is the negative Hessian matrix of the observed log-likelihood function with respect to* $\hat{\Omega}, \hat{U}$ *and the jump sizes of* $(\hat{H}_1, \cdots, \hat{H}_J, \hat{H}_s)$.

If we are primarily interested in Ω, the profile likelihood method (*Murphy and Van der Vaart* (2000)) can be used. Let $l_{pr}(\Omega) = \ell_n(\Omega, \hat{U}(\Omega), \hat{\mathbf{H}}(\Omega)|\mathbf{Y}, \mathbf{Z})$ be the profile log-likelihood function for Ω.

Theorem IV.4. *Assuming the regularity conditions hold, for* $\varepsilon_n = O_p(n^{-1/2})$ *and any vector* v, $-\{l_{pr}(\hat{\Omega}+\varepsilon_n v) - 2l_{pr}(\hat{\Omega}) + l_{pr}(\hat{\Omega}-\varepsilon_n v)\}/n\varepsilon_n^2$ *converges in probability to* $v^T\Sigma^{-1}v$, *where* Σ *is the asymptotic covariance matrix of* $\sqrt{n}(\hat{\Omega}-\Omega_0)$.

The profile likelihood $l_{pr}(\Omega)$ can be calculated via the EM-DCA algorithm by holding Ω fixed. The negative second-order numerical difference of the profile log-likelihood function at Ω is used to estimate the inverse covariance matrix. Specifically, the (i,j)th element of the inverse covariance matrix can be consistently estimated by

$$-\{l_{pr}(\hat{\Omega}+\varepsilon_n e_i+\varepsilon_n e_j) - l_{pr}(\hat{\Omega}+\varepsilon_n e_i-\varepsilon_n e_j) - l_{pr}(\hat{\Omega}-\varepsilon_n e_i+\varepsilon_n e_j) + l_{pr}(\hat{\Omega})\}/\varepsilon_n^2$$

where e_i and e_j are the i^{th} and j^{th} canonical basis vectors respectively. Theorem IV.4 can be verified by closely following the lines of *Zeng and Lin* (2010) Section 9 and *Murphy and Van der Vaart* (2000).

4.5 Simulation Study

A Monte Carlo simulation study was conducted to examine the finite-sample performance of the parameter estimates obtained by the proposed model.

For each time point $t = 0, 1, 2, 3$, we generated two continuous outcome Y_2, Y_6, and four ordinal outcomes Y_1, Y_3, Y_4 and Y_5, each with five levels, from the conditional cumulative probability function. The continuous outcome was generated using inverse CDF transform sampling. The ordinal outcomes were generated by random sampling with level-specific probability.

The simulation settings are as follows. The true population average disease trajectory is specified as $U(t) = \exp(0.5t)$. The shared latent variable a_i was generated from a log-Gamma distribution with subject-specific shape $e^{\eta_1 Z_{i3}}$ and rate $e^{\eta_2 Z_{i3}}$ where the true parameter values are specified as $\eta_1 = 1$ and $\eta_2 = 1.7$. The shared latent variable b_i was

generated from a log-Gamma(η_3, η_3) distribution with the true $\eta_3 = 2$. We consider three covariates: $Z_{i1} \sim N(0,1), Z_{i2} \sim$ Binom(0.5) and $Z_{i3} \sim$ Binom(0.5). The true $H_j(y) = \frac{1}{2}y^2$ for $j = 2, 6$ and the true $H_j(y) = \frac{1}{4}y^2$ for $j = 1, 3, 4, 5$. The covariate effect for each outcome is of the following forms: $\gamma_{i1} = 1, \gamma_{i2} = e^{\beta_2 Z_{i2}}, \gamma_{i3} = e^{\beta_3 Z_{i3}}, \gamma_{i4} = e^{\beta_4 Z_{i1}}, \gamma_{i5} = e^{\beta_5 Z_{i2}}, \gamma_{i6} = e^{\beta_6 Z_{i3}}$, where the true parameters are $\beta_2 = 1, \beta_3 = -0.5, \beta_4 = -1, \beta_5 = 1$ and $\beta_6 = 0.5$. The true factor loading parameters are $\alpha_1 = 1, \alpha_2 = 0.8, \alpha_3 = 0.6, \alpha_4 = 1, \alpha_5 = 1$ and $\alpha_6 = 1$.

A single simulated dataset consists of longitudinal outcomes and corresponding covariates as $\{Y_{ijk}, Z_{ij}\}$ for $i = 1, \cdots, n, j = 1, \cdots, 6, k = 1, \cdots, 4$ with time points at $t_1 = 0, t_2 = 1, t_3 = 2, t_4 = 3$. We examined the performance of estimation for the proposed model under the sample size of $n = 100$ and $n = 200$; each was replicated 500 times. Standard errors were obtained from the numerically evaluated Hessian matrix at the solution.

The results of the simulation study are summarized in Table 4.1. The proposed estimation and inference procedures perform well with diminishing bias as sample sizes increases, and coverage probability at 95% nominal level. With the larger sample size, we see better agreement between empirical standard deviation (ESD) and asymptotic average standard errors (ASE). This suggests that the asymptotic approximation of the covariance matrix from the profile likelihood is reasonable for the sample size of $n = 100$ or larger.

4.6 Real Data Analysis: Pain Centrality Trajectories on Pelvic Pain Patients

The proposed longitudinal joint modeling approach was applied to measure pain centrality trait trajectory of patients undergoing hysterectomy as a treatment for pelvic pain. The proposed model allows us to estimate the population pain centrality trajectory over time for pelvic pain patients and further allows us to explain the heterogeneity of

Table 4.1: Simulation results from the proposed longitudinal joint model. β_1 to β_6 are regression coefficients representing the covariate effects on the outcomes Y_1 to Y_6, respectively. α_1 to α_6 are factor loadings for the outcome Y_1 to Y_6, respectively; η_1, η_2 are regression coefficient effects on the log-Gamma distribution shape and rate parameters for the shared latent variable a_i; η_3 is the shape and rate parameter of the log-Gamma distribution for the shared latent variable b_i. All the outcomes Y_1 to Y_6 are generated at time points $t = 0, 1, 2, 3$. The results are based on 500 simulated datasets with sample size of $n = 100$ and $n = 200$.

n	Outcome	Type	parameter	Truth	Avg. est.	ESD	ASE	95% CP
100	Y_1	ordinal	β_1	0				
	Y_2	continuous	β_2	1	1.01	0.20	0.20	0.94
	Y_3	ordinal	β_3	-0.5	-0.52	0.20	0.20	0.95
	Y_4	ordinal	β_4	-1	-1.01	0.13	0.13	0.95
	Y_5	ordinal	β_5	1	1.02	0.24	0.23	0.94
	Y_6	continuous	β_6	0.5	0.51	0.21	0.21	0.95
	Y_1	ordinal	α_1	1				
	Y_2	continuous	α_2	0.8	0.80	0.09	0.09	0.94
	Y_3	ordinal	α_3	0.6	0.61	0.09	0.09	0.95
	Y_4	ordinal	α_4	1	1.00	0.12	0.12	0.95
	Y_5	ordinal	α_5	1	1.01	0.13	0.12	0.94
	Y_6	continuous	α_6	1	1.00	0.12	0.11	0.94
	a_i	shape	η_1	1	1.14	0.50	0.46	0.97
latent variable		rate	η_2	1.7	1.85	0.59	0.56	0.97
	b_i	shape/rate	η_3	2	2.05	0.44	0.43	0.95
200	Y_1	ordinal	β_1	0				
	Y_2	continuous	β_2	1	1.01	0.14	0.14	0.96
	Y_3	ordinal	β_3	-0.5	-0.50	0.15	0.14	0.94
	Y_4	ordinal	β_4	-1	-1.01	0.09	0.09	0.96
	Y_5	ordinal	β_5	1	1.01	0.15	0.16	0.96
	Y_6	continuous	β_6	0.5	0.50	0.15	0.15	0.95
	Y_1	ordinal	α_1	1				
	Y_2	continuous	α_2	0.8	0.80	0.07	0.07	0.95
	Y_3	ordinal	α_3	0.6	0.60	0.06	0.06	0.94
	Y_4	ordinal	α_4	1	1.00	0.09	0.09	0.95
	Y_5	ordinal	α_5	1	1.01	0.09	0.09	0.95
	Y_6	continuous	α_6	1	1.00	0.08	0.08	0.96
	a_i	shape	η_1	1	1.06	0.30	0.29	0.97
latent variable		rate	η_2	1.7	1.76	0.38	0.35	0.95
	b_i	shape/rate	η_3	2	2.05	0.32	0.30	0.95

Avg. est.: average of Monte Carlo estimates of the true parameter values over the 500 simulations
ESD: empirical standard deviation based on Monte Carlo estimates
ASE: average of numerically estimated standard errors
95% CP: 95% coverage probability

patients' longitudinal outcomes by estimating patient-specific pain centrality disease progression measure over time. The proposed joint shared variable model uses ad-hoc 2011 Fibromyalgia (FM) Survey Criteria (designed to characterize a similar latent construct) as the baseline instrument (Y_1), and extracts information from three other pain centrality relevant outcomes.

The study sample consists of $n = 160$ female pelvic pain patients. We consider four longitudinal responses of mixed scales collected prior to hysterectomy, at one month after hysterectomy, and at three months after hysterectomy. Fibromyalgia (FM) Survey Criteria score was included as the baseline instrument (Y_1). Opioid use (OME), BPI pain severity score, and BPI surgical pain score are the other three pain centrality relevant outcomes included in the model.

The distribution of each outcome is presented in Figure 4.1. All of them are right-skewed, however, the degree of skewness is different among different outcomes. Notice that 90% of the longitudinal opioid use measurements are 0. One patient had extremely heavy opioid use of 120 at baseline and one month after hysterectomy, and the usage increased to 135 three months after hysterectomy. Our model is robust to outliers across all times because of nonparametric transformation of the observable outcomes. While analysis results from traditional latent variable models with pre-specified link may be dominated by influential outlier trajectories.

We include age centered at 47 years old as a covariate for all the responses in our model. The unit for age is per 20 years. The subject-specific latent trait at baseline, a_i, is assumed to follow a log-Gamma$\left(e^{\eta_1 Age_i}, e^{\eta_2 Age_i}\right)$ distribution. The subject-specific disease progression rate b_i is assumed to follow a log-Gamma(η_3, η_3) distribution. The estimation results of the proposed longitudinal joint latent trait model are shown in Table 4.2.

From the estimation results, we see that younger people feel more overall pain in this sample. In addition, age has significant effects on the rate parameters of the baseline latent trait distribution. The factor loadings are all close to 1, implying the FM score,

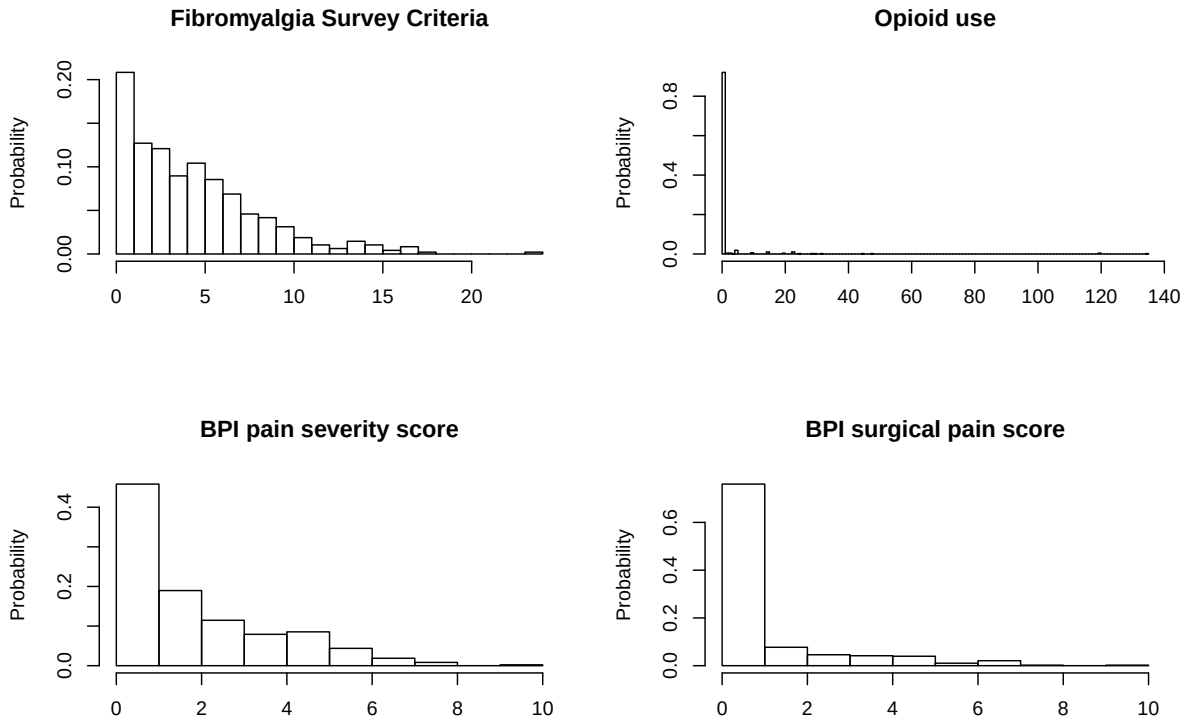

Figure 4.1: Distribution of longitudinal pain responses: Fibromyalgia (FM) Survey Criteria score, Opioid use, BPI pain severity score, and BPI surgical pain score. All of them are right-skewed but with different degrees of skewness. Notice that 90% of the opioid use measurements are 0. One patient has extremely heavy opioid use over time: 120 prior to and at one month after hysterectomy, and 135 at three month after hysterectomy

opioid usage, BPI pain severity and BPI surgical pain contribute approximately equally to the construct of latent trait trajectory.

The model predicted subject-specific latent variable $\log(U_i(t))$ represents how pain-uncentralized a person is at time t, with $-\log(U_i(t))$ representing pain centrality at time t. To see the relationship between the model-based pain centrality score and pain related responses in the model, Figure 4.2 shows a set of scatter plots of model-based pain centrality score predictions versus Fibromyalgia Survey Criteria (first row), opioid use (second row), BPI overall pain severity (third row), and BPI surgical pain (fourth row) at the baseline, at one month after hysterectomy, and at three months after hysterec-

Table 4.2: Parameter estimates, factor loadings, standard error and p-value from analysis of $n = 160$ female pelvic pain patients with responses measured prior to hysterectomy, one month after hysterectomy, and three months after hysterectomy. The unit for age is per 20 years. Age is centered at 47 years old.

Response	Covariate	Param Est	Standard Error	p
Fibromyalgia survey criteria	Age	0		
Opioid use	Age	-0.514	0.356	0.149
BPI pain severity	Age	-0.737	0.249	0.003
BPI surgical pain	Age	0.054	0.291	0.853
Response		Factor loading	Standard Error	p
Fibromyalgia survey criteria		1		
Opioid use		0.897	0.153	< 0.0001
BPI pain severity		1.182	0.131	< 0.0001
BPI surgical pain		1.055	0.139	< 0.0001
Latent trait distribution parameter	Covariate	Param Est	Standard Error	p
η_1	Age	0.020	0.299	0.947
η_2	Age	-0.828	0.401	0.039
η_3		2.934	0.716	

tomy. The model-based pain centrality score is positively correlated with Fibromyalgia Survey Criteria, opioid use, BPI overall pain severity, and BPI surgical pain across all time points. For people with low centrality and low pain, there is still a nice resolution to the positive correlation between model-based centrality and all four pain responses across all time points.

Figure 4.3 shows a set of scatter plots of Fibromyalgia Survey Criteria versus the model-based pain centrality score (first row), opioid use (second row), BPI overall pain severity (third row), and BPI surgical pain (fourth row) at the baseline, at one month after hysterectomy, and at three months after hysterectomy. The Fibromyalgia Survey Criteria is positively correlated with the model-based pain centrality score, opioid use, BPI overall pain severity, and BPI surgical pain across all time points, but the correlation with the observed phenotype outcomes is not as strong compared to the mode-based pain centrality. Compared to the Fibromyalgia Survey Criteria, the model-based centrality score in general is better aligned with BPI overall pain severity and BPI surgical pain.

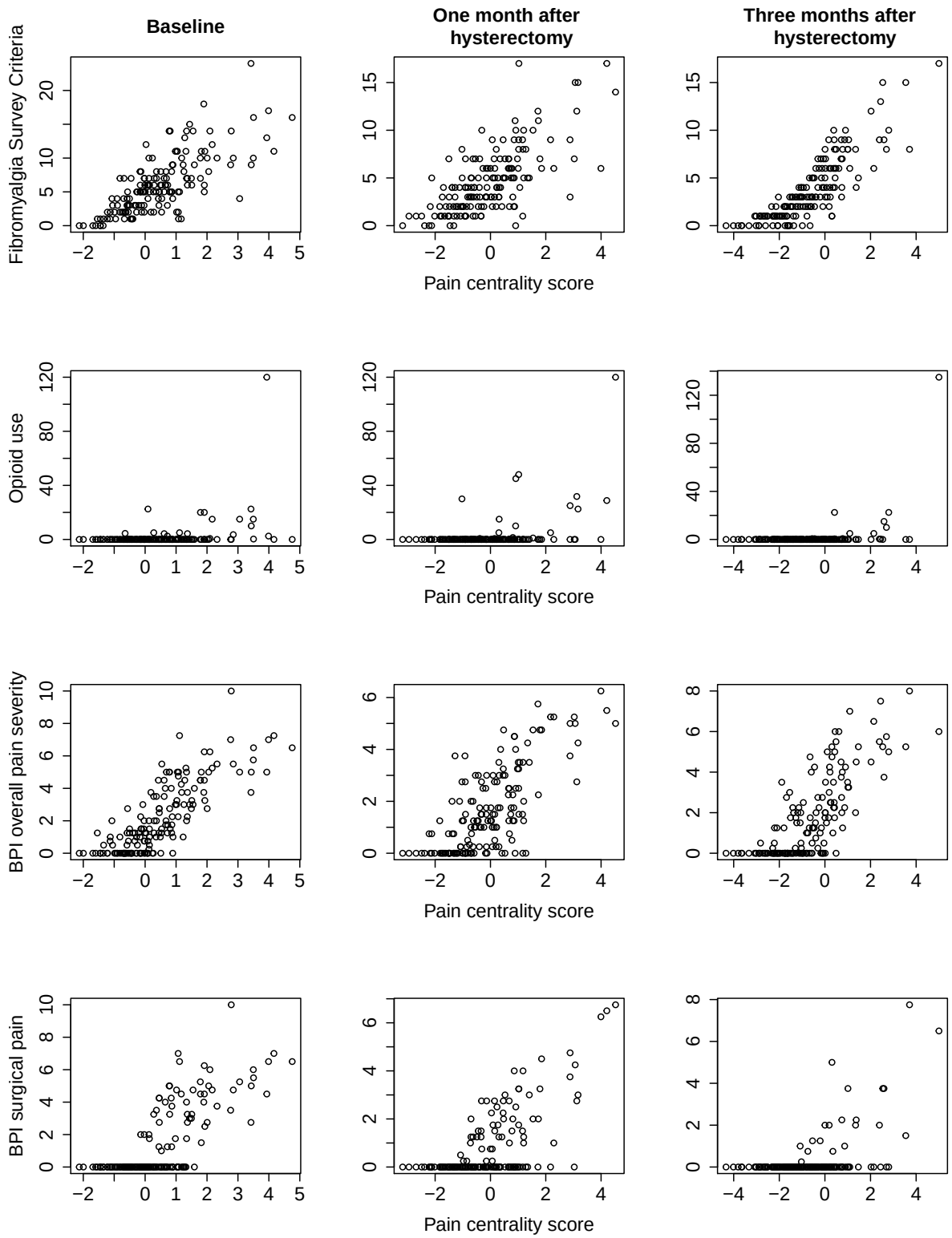

Figure 4.2: Scatterplots of the model-based pain centrality score $-\log(U_i(t))$ vs. Fibromyalgia Survey Criteria (first row), opioid use (second row), BPI overall pain severity (third row), and BPI surgical pain (fourth row) at the baseline, one month after hysterectomy, and three months after hysterectomy.

From both Figure 4.2 and Figure 4.3, we see that there is generally more variability and less correlation at one month after hysterectomy between centrality and pain related responses, as compared to other time points. This is probably because of the noise induced by transient acute pain effects soon after the surgery.

Figure 4.4 shows the trajectory of the model-based pain centrality score $-\log(U_i(t))$ over time grouped by quartiles of the predicted subject-specific baseline score $-a_i$. Patients were grouped into four equal sized clusters based on the predicted 1st quartile (Q_1), the median (Q_2) and the 3rd quartile (Q_2) of baseline scores $-a_i$. Groups are labeled based on baseline score as follows. "High": $-a_i \geq Q_3$; "Medium High": $Q2 \leq -a_i < Q_3$; "Medium Low": $Q1 \leq -a_i < Q_2$; "Low": $-a_i < Q_1$. The black triangle points represent the sample mean of the predicted measure at each time point. From Figure 4.4 we can see that on average pain centrality decreases over time. However, there is large variability in the pain centrality trajectory for people who are more central at baseline.

Figure 4.5 shows the trajectory of the Fibromyalgia Survey Criteria (FM) over time grouped by quartiles of the baseline FM. We can see that for baseline "High", "Medium High", and "Medium Low" groups, the FM decreases over time on average, which is consistent with the model-based centrality score. However, for the "Low" baseline group, on average the FM goes up at one month and decreases at three month after hysterectomy, while the mode-based centrality score decreases all the way for the "Low" group. This may imply that model-based centrality is more sensitive to sub-clinical centrality compared to FM. This is expected because FM is designed for the diagnosis of fibromyalgia where substantial widespread pain is a classic symptom. Similar to model-based centrality, we also see greater variability in the FM trajectory for people who have high FM at baseline. Even though Figure 4.4 and Figure 4.5 convey similar messages over time, the FM trajectory in Figure 4.5 is a lot noisier than the model-based trajectory in Figure 4.2, implying the model-based centrality trajectory summarizes the longitudinal pain information more precisely compared to the FM score.

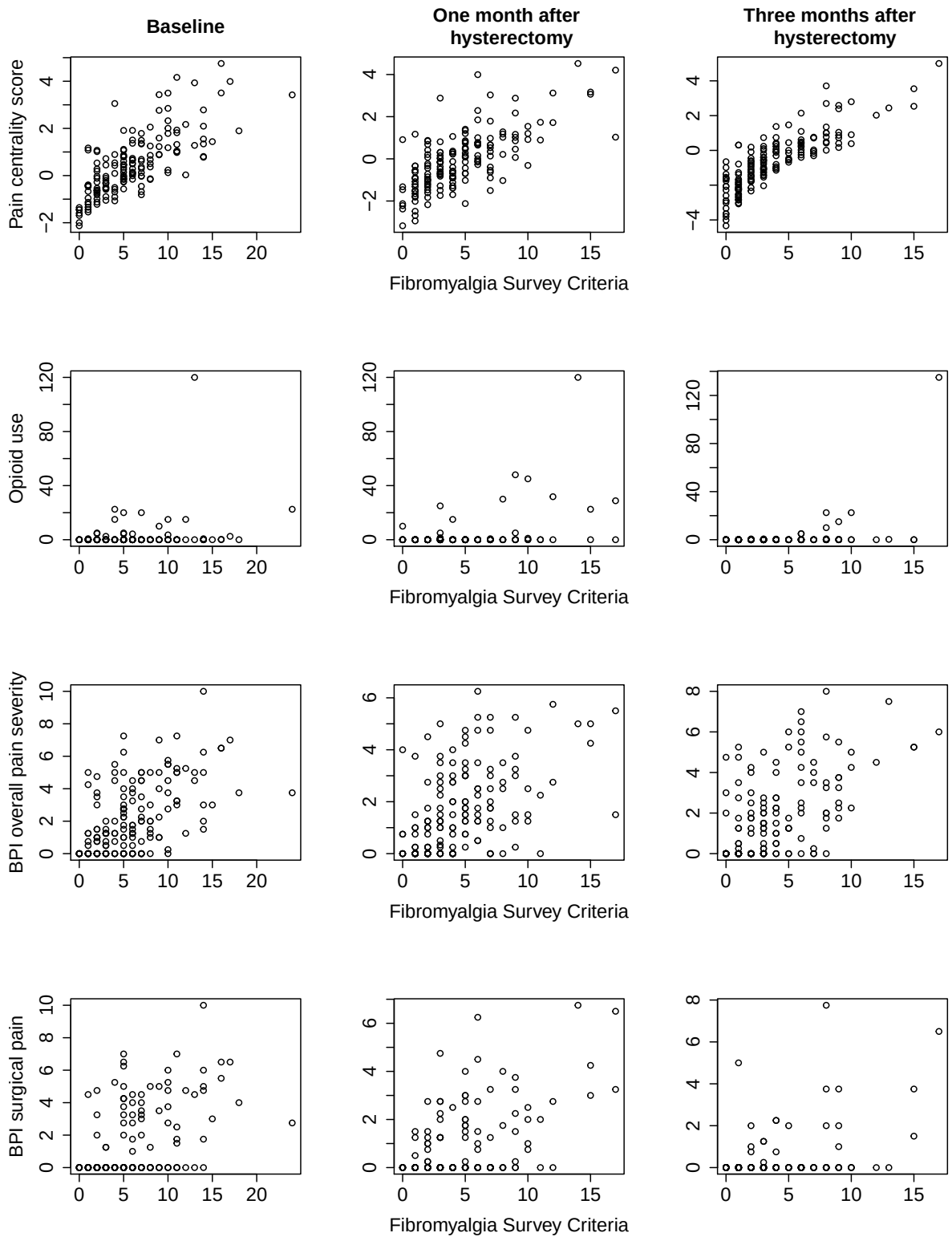

Figure 4.3: Scatterplots of Fibromyalgia Survey Criteria versus model-based pain centrality score (first row), opioid use (second row), BPI overall pain severity (third row), and BPI surgical pain (fourth row) at the baseline, one month after hysterectomy, and three months after hysterectomy.

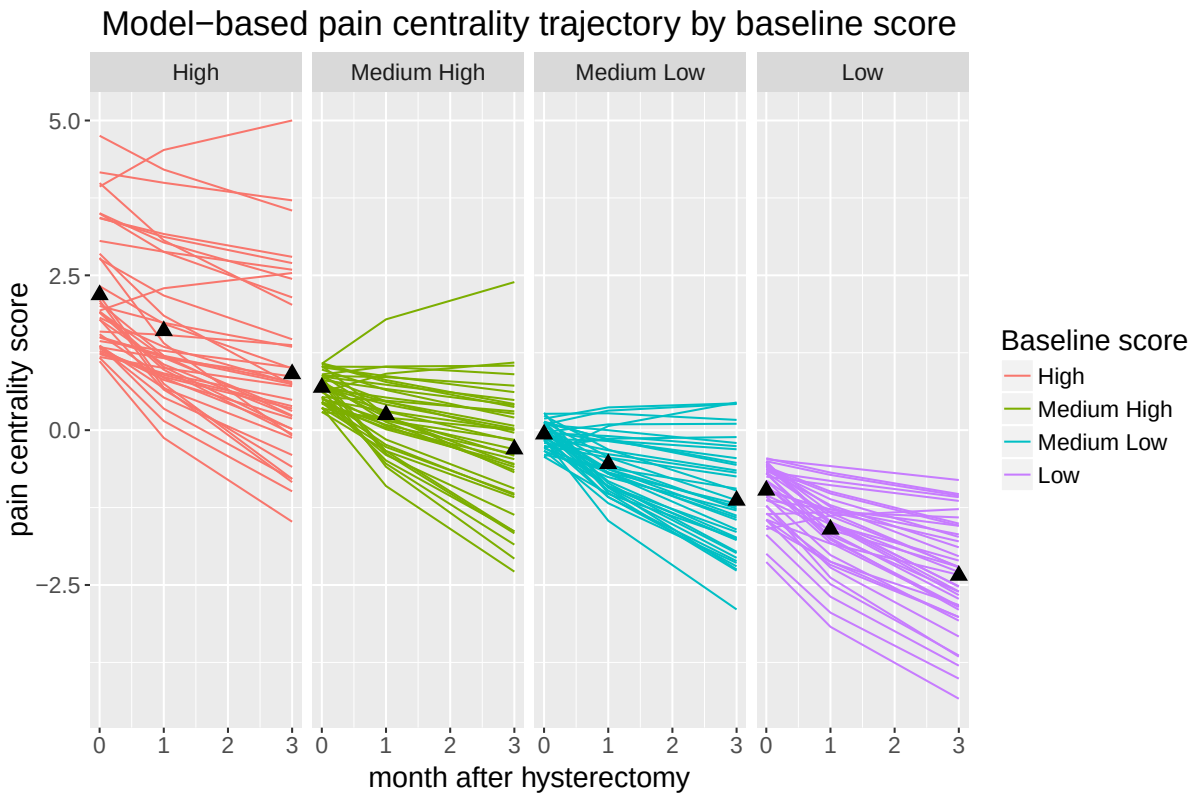

Figure 4.4: Model-based pain centrality trajectory $-\log(U_i(t))$ by baseline score $-a_i$. Patients were grouped into four equal sized clusters based on the 1st quartile (Q_1), median (Q_2) and the 3rd quartile (Q_2) of baseline scores $-a_i$. Groups are labeled based on baseline score as follows: "High": $-a_i \geq Q_3$; "Medium High": $Q2 \leq -a_i < Q_3$; "Medium Low": $Q1 \leq -a_i < Q_2$; "Low": $-a_i < Q_1$. The black triangle points represent the sample mean at each time point.

4.7 Discussion

In this article, we proposed a flexible joint model framework for multiple longitudinal outcomes of mixed scales and time-to-event data, incorporating shared latent trait trajectory. A logistic link is used to accommodate nonparametrically transformed continuous, ordinal, count, multinomial outcomes. A proportional odds survival model is used to model the survival data. We provided an example of using the random effects to link between the survival outcome and the longitudinal responses of mixed types. However, our joint modeling framework is flexible in the association structure between longitudinal and

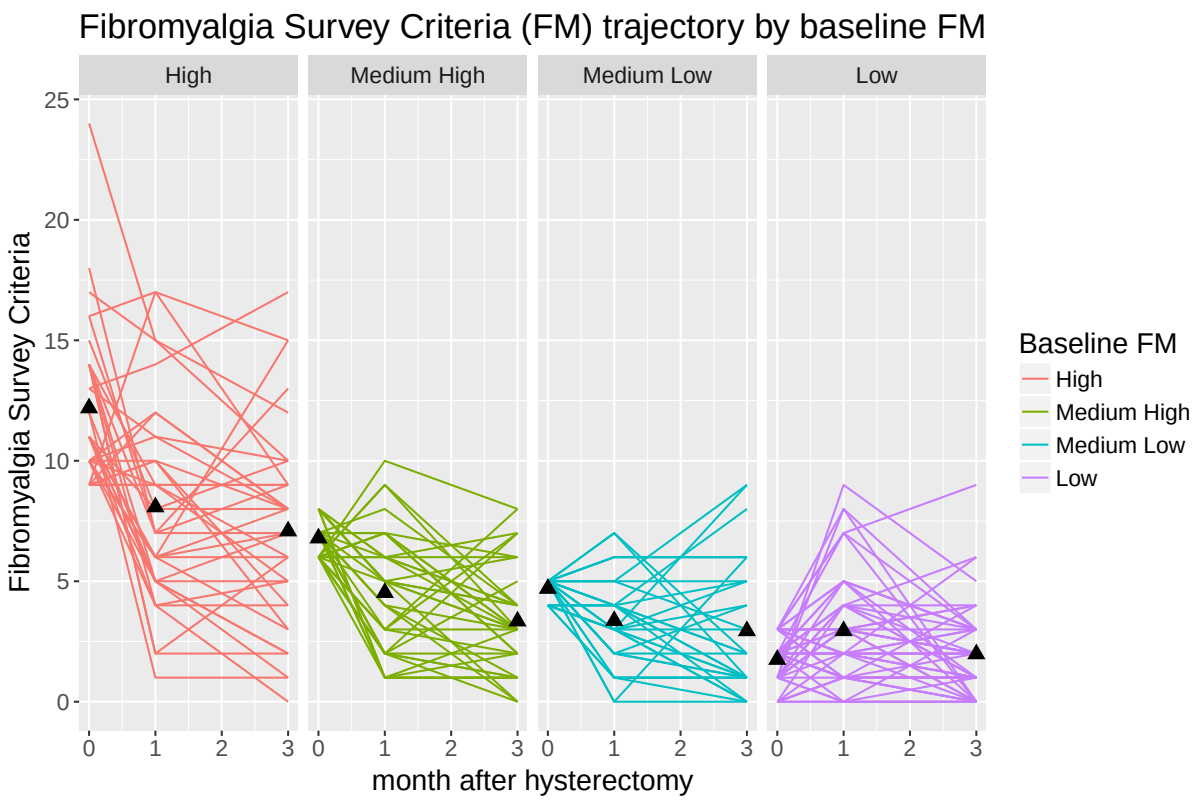

Figure 4.5: Fibromyalgia Survey Criteria (FM) trajectory by baseline FM score. Patients were grouped into four equal sized clusters based on the 1st quartile (Q_1), median (Q_2) and the 3rd quartile (Q_2) of baseline FM. Groups are labeled based on baseline FM score (FM_0) as follows: "High": $FM_0 \geq Q_3$; "Medium High": $Q2 \leq FM_0 < Q_3$; "Medium Low": $Q1 \leq FM_0 < Q_2$; "Low": $FM_0 < Q_1$. The black triangle points represent the sample mean at each time point.

survival outcomes. *Hickey et al.* (2016) and *Rizopoulos* (2012) provided a comprehensive discussion of different association structures that can be potentially incorporated into our model. In addition, the proposed joint model has the flexibility to extend to multiple correlated event times data or a competing risk data.

The resulting parametric and nonparametric estimators are $n^{-1/2}$ consistent and asymptotically normal. The form of the subject-specific latent trait trajectory $U_i(t)$ enjoys a flexible nonparametric specification. We provided an example of using random effects of time to describe the trajectory. However, the latent trait trajectory can be modeled as a stochastic process. One advantage of a joint model is that our model is robust to

unbalanced longitudinal data and we can estimate subject-specific latent trait trajectory even when some of the outcomes are missing.

We have devised numerically efficient and stable estimation and inference procedures based on the maximum likelihood, EM algorithm, DCA algorithm, and classical optimization procedures applied to a finite dimensional profile likelihood. These methods allowed us to conduct a simulation study and to study the dynamic latent trait of pain centrality in patients undergoing hysterectomy. We developed a model-based measurement of the latent trait as a subject-specific function predicted using the model, given the multivariate longitudinal phenotype observed on the subject. Compared to the ad-hoc Fibromyalgia score proposed earlier to characterize pain centrality, our model-based measure of the latent trait is better correlated with pain responses, and is more sensitive in less central patients who experience less pain.

Acknowledgments

This work is supported by grants 1R01HD088712-01A1 "Peripheral and central nervous system correlates of persistent post-hysterectomy pain" and 5P50AR070600-02 "University of Michigan Fibromialgia Center of Research Translation (CORT)" from the National Institutes of Health.

CHAPTER V

Conclusion

Latent variable models have been extensively employed to study the unobservable subject-specific trait and to account for the unobserved heterogeneity between subjects. Existing cure models are limited to modeling static cure status. Most of the latent trait models that accommodate multivariate responses of different scales either were developed under Gaussian framework or require explicitly specified link within exponential family. In this dissertation, we constructed a dynamic cure model that allowed the cure status to change over time. We also proposed a flexible shared latent variable model to accommodate nonparametrically transformed continuous, ordinal, count, multinomial and time-to-event outcomes, under cross-sectional and longitudinal settings. The proposed share latent variable models do not rely on Gaussian assumption and is generic regarding the distribution of the latent variable. The methods proposed in this dissertation represents a contribution to statistical methodology useful for latent variable models, specifically the relaxation of parametric assumptions in statistical models.

Regarding the dynamic cure modeling framework, we modeled the conditional hazard function for terminal event as a change-point function driven by the latent event of cure as an illustrative example. In general, any stochastic hazard process $\mathcal{U}(t)$ that has an absorbing boundary of 0 leads to a cure model. The proposed model framework has the flexibility to incorporate a wide variety of dynamic cure models. We model the time to

cure and time to failure with a proportional hazard model. Other link functions can be naturally incorporated. Notice that if the time to cure model is itself a cure model, the baseline hazard for cure H_2 is bounded, and the cure event may not happen. Predicting something that is fully unobserved, such as the time to cure, should be treated with caution because the model typically has less power for parameters associated with latent components. However, the model-based predictions represent a useful tool to generate hypotheses on the latent effects and to guide further confirmatory studies pursuing more explicit measurements.

The shared latent variable models developed in this dissertation provide a flexible statistical framework to joint modeling nonparametrically transformed continuous, ordinal, count, multinomial and time-to-event responses. As an example, we consider only one latent factor in our model. However, our model can be extended to allow for multiple factors. Extra care should be exercised regarding model identifiability and factor selections. Under the longitudinal setting, we provided an example of using the random effects to link between the survival outcome and the longitudinal responses of mixed types. However, our joint modeling framework is flexible in the association structure between longitudinal and survival outcomes. In addition, the proposed joint model has the flexibility to extend to multiple correlated event times data or a competing risk data. In addition, the form of the subject-specific latent trait trajectory $U_i(t)$ enjoys a flexible nonparametric specification. We provided an example of using random effects of time to describe the trajectory. However, the latent trait trajectory can be modeled as a stochastic process. One advantage of a joint model is that our model is robust to unbalanced longitudinal data and we can estimate subject-specific latent trait trajectory even when some of the outcomes are missing. Application to hysterectomy patient data indicated that the proposed method can offer an improved model-based measurement of the latent trait that better utilizes the information encoded in the multivariate multi-scale observed phenotype. Compared to the ad-hoc Fibromyalgia score proposed earlier to characterize pain centrality, our model-

based measure of the latent trait is better correlated with pain responses, and is more sensitive in less central patients who experience less pain.

Although we illustrated the proposed models using examples in cancer and pain researches, these models can potentially be adapted to a wide spectrum of problems. Precise measurement of subject-specific latent trait that account for population heterogeneity certainly play an important role in precision medicine. The methods in this dissertation enables us to characterize the effect of dynamic factor on latent cure process, to extract useful information from a variety of observable responses, to explain the unobserved heterogeneity, to make prediction and react to potential future trait progression. We hope that this dissertation contributes to statistical methods for cure models and has broader relevance in the statistical literature on latent variable models.

APPENDICES

APPENDIX A

A Semiparametric Joint Survival Model with A Time-Dependent Cure Process

A.1 Joint and Marginal Distributions of Proposed Model

Based on the proposed model hazard functions (2.3) and (2.4), we can show the following model quantities:

1. The marginal density function and survival function of time to cure T_u:

$$f_{T_u}(t_u) = \eta(t_u)h_2(t_u)e^{-\int_0^{t_u}\eta(x)dH_2(x)}$$
$$S_{T_u}(t_u) = e^{-\int_0^{t_u}\eta(x)dH_2(x)}$$

2. The conditional density function and conditional survival function of time to failure T given T_u:

$$f(t \mid T_u) = \mathbb{1}(T_u > t)\theta(t)h_1(t)e^{-\int_0^t \theta(x)dH_1(x)}$$
$$S(t \mid T_u) = e^{-\int_0^t \mathbb{1}(T_u>x)\theta(x)dH_1(x)}$$

3. The joint density function of time to failure and time to cure (T, T_u):

$$\begin{aligned} f_{T,T_u}(t, t_u) &= f(t \mid T_u) f_{T_u}(t_u) \\ &= \mathbb{1}(T_u > t)\theta(t)\eta(t_u)h_1(t)h_2(t_u)e^{-\int_0^t \theta(x)dH_1(x)}e^{-\int_0^{t_u} \eta(x)dH_2(x)} \end{aligned}$$

4. The marginal density and survival function of time to failure T:

$$\begin{aligned} f(t) &= \mathbb{E}\left\{f(t \mid T_u)\right\} \\ &= \int_t^\infty \theta(t)h_1(t)e^{-\int_0^t \theta(x)dH_1(x)}\eta(t_u)e^{-\int_0^{t_u} \eta(x)dH_2(x)}dH_2(t_u) \\ &= \theta(t)h_1(t)e^{-\int_0^t \theta(x)dH_1(x)}e^{-\int_0^t \eta(x)dH_2(x)} \\ S(t) &= \mathbb{E}\left\{S(t \mid T_u)\right\} \\ &= \int_t^\infty \eta(t_u)e^{-\int_0^t \theta(x)dH_1(x)}e^{-\int_0^{t_u} \eta(x)dH_2(x)}dH_2(t_u) + \\ &\int_0^t \eta(t_u)e^{-\int_0^{t_u} \theta(x)dH_1(x)}e^{-\int_0^{t_u} \eta(x)dH_2(x)}dH_2(t_u) \\ &= e^{-\int_0^t \theta(x)dH_1(x)}e^{-\int_0^t \eta(x)dH_2(x)} + \\ &\int_0^t \eta(t_u)e^{-\int_0^{t_u} \theta(x)dH_1(x)}e^{-\int_0^{t_u} \eta(x)dH_2(x)}dH_2(t_u) \end{aligned}$$

A.2 Prediction of Survival Function for The Onset of Cure

From prediction perspective, we are interested in the conditional survival function of the onset of cure given observed information. The prediction is given to those who do not experience the failure event ($\delta = 0$). Given observed data $(X, \delta = 0)$ and estimates

of η, θ, $\{dH_1\}$, and $\{dH2\}$, the survival function of time to cure T_u can be derived as

$$S_{T_u}(t_u|X, \delta = 0) = \frac{\int_{t_u}^{\infty} L_0(s)\eta(s)e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)}{\int_0^{\infty} L_0(s)\eta(s)e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)}, \tag{S.1}$$

where L_0 is the complete data likelihood (2.9) assuming T_u is known.

Consider a subject who is censored at time X. Note the denominator of (S.1) is the expectation of the conditional survival function (2.5) with respect to T_u. Therefore, the denominator of (S.1) is the marginal survival function (2.7). The numerator of (S.1) can be derived under two conditions below:

1. As $t_u \leq X$,

$$\begin{aligned}
&\int_{t_u}^{\infty} L_0(s)\eta(s)e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s) \\
=&\int_{X}^{\infty} \eta(s)e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s) + \int_{t_u}^{X} \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s) \\
=&e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^X \eta(y)dH_2(y)} + \int_{t_u}^{X} \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)
\end{aligned}$$

2. As $t_u > X$,

$$\int_{t_u}^{\infty} L_0(s)\eta(s)e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s) = e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^{t_u} \eta(y)dH_2(y)}$$

Combing the results in (a) and (b), we obtain the survival function for the onset of cure T_u given a censored observation at time X as

$$S_{T_u}(t_u|X, \delta = 0) = \frac{e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^{X\vee t_u} \eta(y)dH_2(y)} + \int_{t_u}^{X\vee t_u} \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)}{e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^X \eta(y)dH_2(y)} + \int_0^X \eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s)}$$

A.3 Functional Derivatives

Consider a functional $J(f)$, where $f = f(x)$ is a function from a certain class. Define a Fréchet differential δ_s (variation of the functional) taken in the direction of a specific function $g(x \mid s) = \mathbb{1}(x > s)$, a unit-jump function at time $x = s$, where s is treated as a parameter. Consider a one-dimensional ϵsubmodel $f + \epsilon g$ that perturbs f in the direction g by the amount of a real number ϵ. A necessary condition for J to be maximized at f is that $\delta_s J = 0$ uniformly over s, where δ_s is defined as

$$\delta_s J(f, g) = \left. \frac{\partial J(f + \epsilon g)}{\partial \epsilon} \right|_{\epsilon=0,\, g=\mathbb{1}(\cdot > s)} . \tag{S.2}$$

Suppose now that $J\big(\bar{f}(t)\big) = \int_0^t \varphi\left(f(x)\right) df(x)$ is a functional that depends on the past trajectory of the function f and $\varphi(y)$ is a differentiable function. The operator δ_s differentiates over the local behavior of f at the point s, and is zero if the perturbation of f does not occur at this location. When differentiating over the vector $\Omega = (\beta, H_1, H_2)^T$, we will use the differential operator $\Delta_s J(\Omega) = (\frac{\partial J}{\partial \beta}, \delta_s J\big(\bar{H}_1\big), \delta_s J\big(\bar{H}_2\big))$, whose last element is a function of s, and δ_s is with respect to H_1 or H_2. For the specific models of this paper, differentiating functions and expectations of a linear functional of the form $J\big(\bar{f}(t)\big) = \int_0^t \varphi\left(f(x)\right) df(x)$ is of interest. Applying (S.2), we have

$$\begin{aligned}
\delta_s J\big(\bar{f}(t)\big) &= \frac{\partial J\big(\bar{f}(t)\big)}{\partial f(s)} \\
&= \int_0^t \varphi\prime\left(f(x)\right) df(x) \left. \frac{\partial f + \epsilon g}{\partial \epsilon} \right|_{\epsilon=0, g=\mathbb{1}(x>s)} + \int_0^t \varphi\left(f(x)\right) d \left. \frac{\partial f + \epsilon g}{\partial \epsilon} \right|_{\epsilon=0, g=\mathbb{1}(x>s)} \\
&= \int_0^t \varphi\prime\left(f(x)\right) df(x)\, \mathbb{1}(x > s) + \int_0^t \varphi\left(f(x)\right) d\, \mathbb{1}(x > s) \\
&= \mathbb{1}(x > s) \left[\int_0^t \varphi\prime\left(f(x)\right) df(x) + \varphi\left(f(s)\right) \right], s > 0
\end{aligned}$$

This definition of the functional derivative corresponds to taking the derivative with respect to a jump in H at timepoint t when H is a step function. Throughout the paper we assume that integration and differentiation are exchangeable, and this can be verified directly for all relevant functionals of the paper. Expression (S.2) is valid whether or not f is a continuous or step function. In the case of step-function, s is restricted to jump points of f to make g in the same class. Then the local functional derivative $\delta_s J = \dfrac{\partial J}{\partial dH(s)}$ for any point of jump s, which is the traditional differentiation over the jump sizes typically used to derive Nonparametric Maximum Likelihood estimates from the log-likelihood. This observation endows the differentiation with the meaning of $\delta_s = \dfrac{\partial}{\partial dH(s)}$ in the continuous case as well.

A.4 EM Algorithm

We use the methods of EM algorithm in *Tsodikov* (2003) to estimate the infinite dimensional baseline hazard function $(\{dH_1(\beta)\}, \{dH_2(\beta)\})$. Consider a single observation data (X, δ), it's contribution to the joint likelihood of time to failure and time to cure can be expressed as

$$L_0\left(\{dH_1\}, \{dH_2\} \mid X, \delta, T_u\right) f(T_u) = \\ \left[\mathbb{1}(T_u > X)\theta(X)dH_1(X)\right]^{\delta} e^{-\int_0^X \mathbb{1}(T_u > s)\theta(s)dH_1(s)} \eta(T_u) dH_2(T_u) e^{-\int_0^{T_u} \eta(x)dH_2(x)} \quad \text{(S.3)}$$

The rest of this section is organized as follows. For baseline hazard function of failure $\{dH_1\}$, we first drive the E step for the censored and failed cases respectively, then we derive the M step to maximize the likelihood with respect to $\{dH_1\}$, which results in a closed-form expression similar to the weighted Breslow-type estimators (*Chen* (2009)). The same steps were applied to the derivation of EM algorithm for baseline hazard func-

tion of cure $\{dH_2\}$. We introduce the following notations

$$G_{1i}^{(k)}(u,v) = e^{-\int_0^u \theta_i(y)dH_1^{(k)}(y)} e^{-\int_0^v \eta_i(y)dH_2^{(k)}(y)}$$

$$G_{2i}^{(k)}(u,v) = \int_u^v \eta_i(x) e^{-\int_0^x \theta_i(y)dH_1^{(k)}(y)} e^{-\int_0^x \eta_i(y)dH_2^{(k)}(y)} dH_2^{(k)}(x)$$

$$p_i^{(k)}(s) = \left[\frac{G_{1i}^{(k)}(X_i,X_i) + G_{2i}^{(k)}(s^+,X_i)}{G_{1i}^{(k)}(X_i,X_i) + G_{2i}^{(k)}(0,X_i)} \right]^{1-\delta_i},$$

$$\Psi_i^{(k)}(s) = Y_i(s) \left[\frac{G_{1i}^{(k)}(X_i,X_i) - G_{1i}^{(k)}(s,s) + G_{2i}^{(k)}(s,X_i)}{G_{1i}^{(k)}(X_i,X_i) + G_{2i}^{(k)}(0,X_i)} \right]^{1-\delta_i},$$

$$\mu_i^{(k)}(s) = \left[\frac{Y_i(s)G_{1i}^{(k)}(s,s) + (1-Y_i(s))G_{1i}^{(k)}(X_i,s)}{G_{1i}^{(k)}(X_i,X_i) + G_{2i}^{(k)}(0,X_i)} \right]^{1-\delta_i} \left[(1-Y_i(s)) \frac{G_{1i}^{(k)}(X_i,s)}{G_{1i}^{(k)}(X_i,X_i)} \right]^{\delta_i}$$

A.4.1 EM algorithm for $\{dH_1\}$

Applying the functional derivative definition (S.2) to the joint log likelihood (S.3) with respect to $dH_1(s)$, we obtain the conditional score function for $dH_1(s)$ as

$$\begin{aligned} U_{0,dH_1}(s) &= \delta_s \log L_0(\{dH_1\},\{dH_2\} \mid T_u) + \delta_s \log f(T_u) \\ &= \frac{\delta 1(X=s)}{dH_1(s)} - \theta(s)\, \mathbb{1}(T_u > s)\, \mathbb{1}(X \geq s) \\ &= \frac{dN(s)}{dH_1(s)} - Y(s)\theta(s)\, \mathbb{1}(T_u > s) \end{aligned}$$

1. **E step**

   Consider an observation censored at time X and $\delta = 0$. The likelihood contribution conditional on T_u is $L_0 = e^{-\int_0^X \mathbb{1}(T_u > y)\theta(y)dH_1(y)}$. The unconditional score is

$$U_{dH_1}(s) = \mathbb{E}\left[U_{0,dH_1}(s) \middle\| L_0^{(k)} \right] = -Y(s)\theta(s) \frac{\mathbb{E}\left[\mathbb{1}(T_u > s) L_0^{(k)} \right]}{\mathbb{E}\left[L_0^{(k)} \right]}$$

Since E operator only involve parameters from k^{th} iteration but not from $(k+1)^{th}$ iteration, we keep the iteration index for $(k+1)^{th}$ iteration as needed, and drop iteration index (k) for brevity. Therefore, in the rest of EM algorithm section, any dH_1 and dH_2 without an iteration index implies the k^{th} iteration. Note the denominator $\mathbb{E}\left[L_0^{(k)}\right]$ is the marginal survival function (2.7). The numerator of the unconditional score is

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(T_u > s)L_0^{(k)}\right] = \mathbb{E}\left[\mathbb{1}(T_u > s)e^{-\int_0^X \mathbb{1}(T_u>y)\theta(y)dH_1(y)}\right] \\
&= \int_{s+}^{\infty} e^{-\mathbb{1}(X<t_u)\int_0^X \theta(y)dH_1(y) - \mathbb{1}(X\geq t_u)\int_0^{t_u}\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= \int_{X^+}^{\infty} \eta(t_u)e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) + \\
&\int_{s^+}^{X} \eta(t_u)e^{-\int_0^{t_u}\theta(y)dH_1(y)}e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^X \eta(y)dH_2(y)} + \\
&\int_{s^+}^{X} \eta(t_u)e^{-\int_0^{t_u}\theta(y)dH_1(y)}e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= G_1^{(k)}(X,X) + G_2^{(k)}(s^+,X)
\end{aligned}
$$

Therefore, the contribution of an observation censored at time X to the unconditional score is

$$
U_{dH_1}(s) = -Y(s)\theta(s)\frac{G_1^{(k)}(X,X) + G_2^{(k)}(s^+,X)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)} \tag{S.4}
$$

Now we consider an observation failed at time X and $\delta = 1$. The likelihood contribution conditional on T_u is $L_0 = \mathbb{1}(T_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}$. The

unconditional score is

$$U_{dH_1}(s) = \mathbb{E}\left[U_{0,dH_1}(s)\middle\| L_0^{(k)}\right] = \frac{dN(s)}{dH_1^{(k+1)}(s)} - Y(s)\theta(s)\frac{\mathbb{E}\left[\mathbb{1}(T_u > s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]}$$

Note the denominator $\mathbb{E}\left[L_0^{(k)}\right]$ is the marginal density function (2.8). The numerator of the unconditional score is

$$\begin{aligned}
&Y(s)\mathbb{E}\left[\mathbb{1}(T_u > s)L_0^{(k)}\right] = Y(s)\mathbb{E}\left[\mathbb{1}(T_u > s)\,\mathbb{1}(T_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\right] \\
&= \int_{s+}^{\infty} Y(s)\,\mathbb{1}(t_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= Y(s)\int_{X^+}^{\infty}\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= Y(s)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^X \eta(y)dH_2(y)}
\end{aligned}$$

The contribution of an observation failed at time X to the unconditional score is

$$U_{dH_1}(s) = \frac{dN(s)}{dH_1^{(k+1)}(s)} - Y(s)\theta(s) \tag{S.5}$$

Combing the above results (S.4) and (S.5), we obtain the marginal score function for $dH_1(s)$ as

$$\begin{aligned}
U_{dH_1}(s) &= \frac{dN(s)}{dH_1^{(k+1)}(s)} - Y(s)\theta(s)\left[\frac{G_1^{(k)}(X,X) + G_2^{(k)}(s^+,X)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)}\right]^{1-\delta} \\
&= \frac{dN(s)}{dH_1^{(k+1)}(s)} - Y(s)\theta(s)p_i^{(k)}(s)
\end{aligned}$$

2. **M step**

Suppose there are n independent observations with data (X_i, δ_i) for $i = 1\cdots n$. The estimator for $dH_1^{(k+1)}(s)$ can be obtained by solving $\sum_{i=1}^n U_{dH_1}(s) = 0$. The solution

results in a Breslow-type estimator

$$dH_1^{(k+1)}(s) = \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\theta_i(s)p_i^{(k)}(s)}$$

This constitutes a self-consistent equation that can be solved iteratively (*Tsodikov* (2003)).

A.4.2 EM algorithm for $\{dH_2\}$

Again, we consider a single observation with data (X, δ). To derive the EM algorithm for dH_2, we first apply the functional derivative definition (S.2) to the joint log likelihood (S.3) with respect to $dH_2(s)$ to obtain the conditional score function for $dH_2(s)$ as

$$\begin{aligned} U_{0,dH_2}(s) &= \delta_s \log L_0(\{dH_1\}, \{dH_2\} \mid T_u) + \delta_s \log f(T_u) \\ &= \frac{\mathbb{1}(T_u = s)}{dH_2(s)} - \eta(s)\, \mathbb{1}(T_u \geq s) \end{aligned}$$

1. **E step**

   The unconditional score for dH_2 is the expectation of the conditional score function with respect to T_u:

$$U_{dH_2}(s) = \mathbb{E}\left[U_{0,dH_2}(s)\middle\| L_0^{(k)}\right] = \frac{1}{dH_2^{(k+1)}} \frac{\mathbb{E}\left[\mathbb{1}(T_u = s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} - \eta(s)\frac{\mathbb{E}\left[\mathbb{1}(T_u \geq s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} \tag{S.6}$$

   Consider an observation censored at time X and $\delta = 0$. The likelihood contribution conditional on T_u is $L_0 = e^{-\int_0^X \mathbb{1}(T_u > y)\theta(y)dH_1(y)}$. The denominator $\mathbb{E}\left[L_0^{(k)}\right]$ is the marginal survival function (2.7). One of the numerator $\mathbb{E}\left[\mathbb{1}(T_u = s)L_0^{(k)}\right]$ can be

derived as

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}(T_u = s)L_0^{(k)}\right] &= \mathbb{E}\left[\mathbb{1}(T_u = s)e^{-\int_0^X \mathbb{1}(T_u > y)\theta(y)dH_1(y)}\right] \\
&= \int_0^\infty \mathbb{1}(t_u = s)e^{-\mathbb{1}(X<t_u)\int_0^X \theta(y)dH_1(y) - \mathbb{1}(X\geq t_u)\int_0^{t_u}\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= \int_0^X \mathbb{1}(t_u = s)\eta(t_u)e^{-\int_0^{t_u}\theta(y)dH_1(y)}e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&\qquad + \int_X^\infty \mathbb{1}(t_u = s)\eta(t_u)e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\
&= \mathbb{1}(X \geq s)\eta(s)e^{-\int_0^s \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2^{(k)}(s) \\
&\qquad + \mathbb{1}(X < s)\eta(s)e^{-\int_0^X \theta(y)dH_1(y)}e^{-\int_0^s \eta(y)dH_2(y)}dH_2^{(k)}(s) \\
&= \eta(s)dH_2^{(k)}(s)\left[Y(s)G_1^{(k)}(s,s) + (1-Y(s))\,G_1^{(k)}(X,s)\right]
\end{aligned}
$$

Therefore, the first term in (S.6) can be written as

$$
\frac{1}{dH_2^{(k+1)}}\frac{E\left[\mathbb{1}(T_u = s)L_0^{(k)}\right]}{E\left[L_0^{(k)}\right]} = \frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)}\frac{\eta(s)\left[Y(s)G_1^{(k)}(s,s) + (1-Y(s))\,G_1^{(k)}(X,s)\right]}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)}
$$

Next, we calculate the numerator expression of the second term $\mathbb{E}\left[\mathbb{1}(T_u \geq s)L_0^{(k)}\right]$ under two cases, $X \geq s$ and $X < s$, respectively.

(a) $X \geq s; Y(s) = 1$

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(T_u \geq s) L_0^{(k)}\right] \\
&= \int_s^{\infty} e^{-\mathbb{1}(X<t_u)\int_0^X \theta(y)dH_1(y) - \mathbb{1}(X\geq t_u)\int_0^{t_u}\theta(y)dH_1(y)} \eta(t_u) e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&= \int_{X^+}^{\infty} e^{-\int_0^X \theta(y)dH_1(y)} \eta(t_u) e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&+ \int_s^{X} \eta(t_u) e^{-\int_0^{t_u}\theta(y)dH_1(y)} e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&= e^{-\int_0^X \theta(y)dH_1(y)} e^{-\int_0^X \eta(y)dH_2(y)} + \int_s^X \eta(t_u) e^{-\int_0^{t_u}\theta(y)dH_1(y)} e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&= G_1^{(k)}(X, X) + G_2^{(k)}(s, X)
\end{aligned}
$$

(b) $X < s; Y(s) = 0$

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(T_u \geq s) L_0^{(k)}\right] = \\
&= \int_s^{\infty} e^{-\mathbb{1}(X<t_u)\int_0^X \theta(y)dH_1(y) - \mathbb{1}(X\geq t_u)\int_0^{t_u}\theta(y)dH_1(y)} \eta(t_u) e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&= \int_s^{\infty} e^{-\int_0^X \theta(y)dH_1(y)} \eta(t_u) e^{-\int_0^{t_u}\eta(y)dH_2(y)} dH_2(t_u) \\
&= e^{-\int_0^X \theta(y)dH_1(y)} e^{-\int_0^s \eta(y)dH_2(y)} \\
&= G_1^{(k)}(X, s)
\end{aligned}
$$

Combine the results (i) and (ii), we obtain the expression for the second term in

(S.6) as

$$\frac{\mathbb{E}\left[\mathbb{1}(T_u \geq s) L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} = \frac{Y(s)\left[G_1^{(k)}(X,X) + G_2^{(k)}(s,X)\right] + (1-Y(s))\, G_1^{(k)}(X,s)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)}$$

Therefore, the contribution of a censored observation at time X to the unconditional score for dH_2 can be written as

$$\begin{aligned} U_{dH_2}(s) = \frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} & \frac{\eta(s)\left[Y(s)G_1^{(k)}(s,s) + (1-Y(s))\, G_1^{(k)}(X,s)\right]}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)} \\ & - \eta(s)\frac{Y(s)\left[G_1^{(k)}(X,X) + G_2^{(k)}(s,X)\right] + (1-Y(s))\, G_1^{(k)}(X,s)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)} \\ = -\eta(s)Y(s) & \frac{G_1^{(k)}(X,X) - G_1^{(k)}(s,s) + G_2^{(k)}(s,X)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)} \\ & + \left(\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} - 1\right)\eta(s)\frac{Y(s)G_1^{(k)}(s,s) + (1-Y(s))G_1^{(k)}(X,s)}{G_1^{(k)}(X,X) + G_2^{(k)}(0,X)} \end{aligned} \tag{S.7}$$

Now we consider an observation failed at time X and $\delta = 1$. The likelihood contribution conditional on T_u is $L_0 = \mathbb{1}(T_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}$, so the denominator of (S.6), $\mathbb{E}\left[L_0^{(k)}\right]$, is the marginal density function (2.8). The term $\mathbb{E}\left[\mathbb{1}(T_u = s)L_0^{(k)}\right]$ can be derived as

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}(T_u = s)L_0^{(k)}\right] &= \mathbb{E}\left[\mathbb{1}(T_u = s)\,\mathbb{1}(T_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\right] \\ &= \int_0^\infty \mathbb{1}(t_u = s)\,\mathbb{1}(t_u > X)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u) \\ &= \mathbb{1}(X < s)\theta(X)dH_1(X)e^{-\int_0^X \theta(y)dH_1(y)}\eta(s)e^{-\int_0^s \eta(y)dH_2(y)}dH_2(s) \\ &= (1-Y(s))\theta(X)dH_1(X)\eta(s)dH_2^{(k)}(s)G_1^{(k)}(X,s) \end{aligned}$$

The first term in (S.6) can be written as

$$\frac{1}{dH_2^{(k+1)}}\frac{\mathbb{E}\left[\mathbb{1}(T_u=s)L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]}=(1-Y(s))\eta(s)\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)}\frac{G_1^{(k)}(X,s)}{G_1^{(k)}(X,X)}$$

Next, we calculate the expression of the term $\mathbb{E}\left[\mathbb{1}(T_u\geq s)L_0^{(k)}\right]$ under two cases, $X\geq s$ and $X<s$, respectively.

(a) $X\geq s; Y(s)=1$

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(T_u\geq s)L_0^{(k)}\right]=\\
&=\int_s^\infty \mathbb{1}(t_u\geq s)\,\mathbb{1}(t_u>X)\theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u)\\
&=\int_{X^+}^\infty \theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u)\\
&=\theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}e^{-\int_0^X\eta(y)dH_2(y)}\\
&=\theta(X)dH_1(X)G_1^{(k)}(X,X)
\end{aligned}$$

(b) $X<s; Y(s)=0$

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}(T_u\geq s)L_0^{(k)}\right]=\\
&=\int_s^\infty \mathbb{1}(t_u\geq s)\,\mathbb{1}(t_u>X)\theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u)\\
&=\int_s^\infty \theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}\eta(t_u)e^{-\int_0^{t_u}\eta(y)dH_2(y)}dH_2(t_u)\\
&=\theta(X)dH_1(X)e^{-\int_0^X\theta(y)dH_1(y)}e^{-\int_0^s\eta(y)dH_2(y)}\\
&=\theta(X)dH_1(X)G_1^{(k)}(X,s)
\end{aligned}$$

Combine the above results (i) and (ii), we obtain the expression for the second term

in (S.6) as

$$\frac{\mathbb{E}\left[\mathbb{1}(T_u \geq s) L_0^{(k)}\right]}{\mathbb{E}\left[L_0^{(k)}\right]} = Y(s) + (1 - Y(s)) \frac{G_1^{(k)}(X, s)}{G_1^{(k)}(X, X)}$$

Therefore, the contribution of an observation failed at time X to the unconditional score for dH_2 can be written as

$$\begin{aligned} U_{dH_2}(s) &= \frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} \frac{\eta(s)(1 - Y(s)) G_1^{(k)}(X, s)}{G_1^{(k)}(X, X)} - \eta(s) Y(s) - \eta(s) \frac{(1 - Y(s)) G_1^{(k)}(X, s)}{G_1^{(k)}(X, X)} \\ &= -\eta(s) Y(s) + \left(\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} - 1 \right) \eta(s) \frac{(1 - Y(s)) G_1^{(k)}(X, s)}{G_1^{(k)}(X, X)} \end{aligned} \tag{S.8}$$

Combing expressions (S.7) and (S.8), we obtain the unconditional score function for $dH_2(s)$ as

$$\begin{aligned} U_{dH_2}(s) &= -Y(s)\eta(s) \left[\frac{G_1^{(k)}(X, X) - G_1^{(k)}(s, s) + G_2^{(k)}(s, X)}{G_1^{(k)}(X, X) + G_2^{(k)}(0, X)} \right]^{1-\delta} + \\ &\left(\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} - 1 \right) \eta(s) \left[\frac{Y(s) G_1^{(k)}(s, s) + (1 - Y(s)) G_1^{(k)}(X, s)}{G_1^{(k)}(X, X) + G_2^{(k)}(0, X)} \right]^{1-\delta} \left[\frac{(1 - Y(s)) G_1^{(k)}(X, s)}{G_1^{(k)}(X, X)} \right]^{\delta} \\ &= -\eta(s) \Psi^{(k)}(s) + \left(\frac{dH_2^{(k)}(s)}{dH_2^{(k+1)}(s)} - 1 \right) \eta(s) \mu^{(k)}(s) \end{aligned}$$

2. **M step**

Suppose there are n independent observations with data (X_i, δ_i) for $i = 1 \cdots n$. The

estimator for $dH_2^{(k+1)}(s)$ can be obtained by solving $\sum_{i=1}^{n} U_{dH_2}(s) = 0$. We have

$$dH_2^{(k+1)}(s) = \frac{\left[\sum_{i=1}^{n} \eta_i(s)\mu_i^{(k)}(s)\right] dH_2^{(k)}(s)}{\sum_{i=1}^{n} \eta_i(s)\left[\mu_i^{(k)}(s) + \Psi_i^{(k)}(s)\right]}$$

The above equation solves for $dH_2(s)$ iteratively until convergence. We obtain a consistent estimate of dH_2 at convergence (*Tsodikov* (2003)).

A.5 Property of Martingale Transform

Let $V(s) = \int_0^\tau \varepsilon(t, s; \beta, H_1, H_2)dM(t)$, where ε is a predictable function such that it does not depend on s when $t < s$. Consider the increment of $V(s)$ over s

$$dV(t) = \int_0^\tau \varepsilon(t, s + ds)dM(t) - \varepsilon(t, s)dM(t) = \int_0^\tau d_s\varepsilon(t, s)dM(t), \tag{S.9}$$

where $d_s\varepsilon(t, s)$ is the partial derivative of $\varepsilon(t, s)$ with respect to s. Since $M(s)$ is a martingale process adapted to filtration $\mathcal{F}_{s-}$, we know $E\{dM(t)|\mathcal{F}_{s-}\} = 1(t < s)dM(t)$.

Taking an expectation conditional of (S.9) on filtration,

$$\begin{aligned} E\{dV(s)|\mathcal{F}_{s-}\} &= \int_0^\tau E\{d_s\varepsilon(t, s)dtdM(t)|\mathcal{F}_{s-}\} \\ &= \int_0^\tau d_s\varepsilon(t, s)E\{dM(t)|\mathcal{F}_{s-}\} = \int_0^s d_s\varepsilon(t, s)dM(t) \end{aligned}$$

Since $\varepsilon(t, s)$ by definition is a predictable function such that it does not depend on s for $t < s$, so $d_s\varepsilon(t, s) = 0$ when $t < s$. We have $E\{dV(t)|\mathcal{F}_{s-}\} = 0$. Therefore, $V(t)$ is a martingale.

A.6 Asymptotic Properties

In this section we provide the technical details to present the consistency and weak convergence properties for the proposed NPMLE weak convergence for the proposed NPMLE $\hat{\Omega} = \left(\hat{\beta}, \{d\hat{H}_1\}, \{d\hat{H}_2\}\right)$.

Let $\|\cdot\|_{l^\infty[0,\tau]}$ denote the supremum norm in $[0,\tau]$, and $\|w\|_{BV[0,\tau]}$ the total variation of $w(t)$ in $[0,\tau]$. Define $\mathcal{Q} = \{w(t) : \|w\|_{BV[0,\tau]} \leq 1\}$. $\hat{H}_(t) = \left(\hat{H}_1(t), \hat{H}_2(t)\right)$ may be regarded as bounded linear functional in $l^\infty[\mathcal{Q}] \times l^\infty[\mathcal{Q}]$, and $\left\{\hat{\beta} - \beta^0, \hat{H}_(t) - H^0(t)\right\}$ a random element in the metric space $\mathcal{R}^p \times l^\infty(\mathcal{Q}) \times l^\infty(\mathcal{Q})$, where p is the dimension of β^0. We denote $\mathcal{H}$ as the compact convex set in the metric space $\mathcal{R}^p \times l^\infty(\mathcal{Q}) \times l^\infty(\mathcal{Q})$ where $\Omega^0 \in \mathcal{H}$. By *Fleming and Harrington* (1991) (p289-p290), the following regularity conditions are required to establish asymptotic properties of NPMLE.

(1) The true $H_k, k = 1, 2$, is strictly increasing and continuously differentiable. The true value of parameter set $\Omega = (\beta, H_1, H_2)$ falls in the interior of the compact convex set $\mathcal{H}$.

(2) The covariate process $\mathbf{z}_k(t), k = 1, 2$, are left continuous with total bounded variation (BV) within $[0,\tau]$, with probability one. Also, $\mathbf{z}_k(t), k = 1, 2$ are linearly independent in the sense that, if there exist $a(t)$ and c such that $a(t) + c^T\mathbf{z}(t) = 0$ with probability one, then $a(t) = 0$ and $c = 0$.

(3) With probability one, $\mathbb{E}(Y(\tau)|\mathbf{z}_1(t), \mathbf{z}_2(t)) > 0$, $P(\delta = 0, T = \tau|\mathbf{z}_1(t), \mathbf{z}_2(t)) > 0$. The at risk set $Y(t)$ will not shrink to empty.

(4) The score operator for Ω is Fréchet differentiable at Ω^0 with a continuously invertible derivative $-\mathcal{I}^0$. The hessian matrix $\mathcal{I}_n$ evaluated at the true values of H and β is positive definite, and converges in probability to a deterministic and invertible operator $\mathcal{I}^0$.

A.6.1 Proof of Proposition II.1

To establish consistency of NPMLE of Ω, i.e., $\|\hat{H}_1(t) - H_1(t)^0\|_{l^\infty(\mathcal{Q})} \xrightarrow{p} 0$, $\|\hat{H}_2(t) - H_2(t)^0\|_{l^\infty(\mathcal{Q})} \xrightarrow{p} 0$ and $|\hat{\beta} - \beta^0| \xrightarrow{p} 0$, we assume the above regularity conditions (1)-(4) hold. In addition, we need to verify the following two conditions hold:

(a) *identifiability condition*:The model is identifiable in the sense that $\Lambda = \Lambda^0$ uniformly over Ω implies $\Omega = \Omega^0$. Then for any sequence $\Omega_n \in \mathcal{H}$, the compact convex set in the metric space $\mathcal{R}^p \times l^\infty(\mathcal{Q}) \times l^\infty(\mathcal{Q})$, $\liminf_{n\to\infty} \ell(\Omega_n) \geq \ell(\Omega^0)$ implies $\|\Omega_n - \Omega^0\| \xrightarrow{p} 0$.

(b) *uniform convergence condition*: for any $\Omega \in \mathcal{H}$ we have uniform convergence, i.e.,

$$\sup_{\Omega\in\mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.$$

Since $\ell_n(\hat{\Omega}) = \sup_{\Omega\in\mathcal{H}} \ell_n(\Omega) + o_p(1)$, based on Theorem 2.12 in *Kosorok* (2008), given that the regularity conditions (1)-(4), identifiability condition and uniform convergence condition hold, we have the consistency of NPMLE: $\|\hat{\Omega} - \Omega^0\| \xrightarrow{p} 0$.

We verify these conditions in the following steps:

To verify the identifiability condition, we need to make use of the large sample limit of the likelihood. Let $F(t)$ be the crude cumulative density function of failure in the presence of censoring, and $R(t)$ be the crude survival function in the presence of censoring. The true function of $F(t)$ and $R(t)$ are denoted as $F_0(t)$ and $R_0(t)$. The model hazard function is a function of Ω and can be denoted as $d\Lambda(t) = \gamma\left(t; \beta, \overline{H}_1(t), \overline{H}_2(t)\right) dH_1(t) = \gamma(t;\Omega)dH_1(t)$. Notice the fact that $dF(t) = R(t)d\Lambda(t)$. Denote the true value of Ω as Ω^0. The true likelihood as a function of Ω can be written as

$$\ell(\Omega, \Omega^0) = \mathbb{E} \int_0^\tau \left[\log d\Lambda(t) dF^0(t) - R^0(t) d\Lambda(t)\right],$$

where the expectation is taken with respect to the covariate process $\mathbf{z}(t) = (\mathbf{z}_1(t), \mathbf{z}_2(t))$.

Consider the negative "true" Kullback-Leibler distance, i.e.,

$$D = \ell(\Omega, \Omega^0) - \ell(\Omega^0, \Omega^0),$$

The distance can be written as

$$D = \mathbb{E} \int_0^\tau \left(\log \frac{d\Lambda(t)}{d\Lambda^0(t)} - \frac{d\Lambda(t)}{d\Lambda^0(t)} + 1 \right) dF^0(t)$$

Consider a non-positive convex function $\phi(x) = \log x - x + 1$. The function $\phi(x)$ has a unique maximizer at $x = 1$, and $\max_x \phi(x) = \phi(1) = 0$. Notice D can be written as

$$D = \mathbb{E} \int_0^\tau \phi \left(\frac{d\Lambda(t)}{d\Lambda^0(t)} \right) dF^0(t)$$

Therefore, D has a unique maximum when $d\Lambda(t) = d\Lambda^0(t)$ uniformly over Ω. Under an identifiable model this means D has a unique maximum at Ω^0.

Since maximizing D is equivalent to maximizing likelihood $\ell(\Omega, \Omega^0)$, and D has a unique maximum, therefore, $\Omega^0 = \text{argmax}_{\Omega \in \mathcal{H}} \ell(\Omega)$ is unique. We assume the model $\ell(\Omega, \Omega^0)$ is identifiable in the sense that $\Lambda = \Lambda^0$ uniformly over Ω implies $\Omega = \Omega^0$ uniformly. Furthermore, since Λ is assumed to be a continuous and differentiable functional of H, so is the likelihood function $\ell(\Omega)$. Based on Lemma 14.3 of *Kosorok* (2008), we have $\liminf_{n \to \infty} \ell(\Omega_n) \geq \ell(\Omega^0)$, i.e., the identifiability condition holds.

To verify the uniform convergence condition, we need to make use of the uniform law of large numbers for the empirical process. If the regularity condition (1) holds, Ω is in the class of functions of bounded variation with integrable envelope, and so the hazard function $H_1(t)$ and $H_2(t)$ are bounded. Therefore, $\mathcal{H}$ is in a Glivenko-Cantelli class whose ε-entropy with bracketing number is bounded by A/ε, where A is a constant. Since the

functionals Λ and $\ell(\Omega)$ are assumed continuous and the envelope of Ω is integrable, the integrand in $\ell(\Omega)$ is also Glivenko-Cantelli by the preservation theorems. Therefore, we may apply the uniform law of large numbers for the empirical process to the sequence D_n, the limited sample counterpart of D as

$$D_n = \ell_n(\Omega, \Omega^0) - \ell_n(\Omega^0, \Omega^0),$$

where

$$\ell_n(\Omega, \Omega^0) = n^{-1} \sum_{i=1}^{n} \int_0^\tau \{[\log \gamma_i(t; \Omega) + \log dH_t] dN_i(t) - Y_i(t)\gamma_i(t; \Omega) dH_1(t)\}$$

such that

$$\sup_{\Omega \in \mathcal{H}} |D_n(\Omega) - D(\Omega)| \xrightarrow{p} 0$$

$$\sup_{\Omega \in \mathcal{H}} |\ell_n(\Omega) - \ell(\Omega)| \xrightarrow{p} 0.$$

Therefore the uniform convergence condition holds.

A.6.2 Proof of Proposition II.2

Theorem II.2 can be proved by the martingale theory applied to the score functions (2.12), (2.13) and (2.14). We know the proposed NPMLE $\hat{\Omega}$ solves the score equation $U(\Omega) = 0$ where $U(\Omega) = (U_\beta, U_{H_1(s)}, U_{H_2(s)})^T$ is the score functions for parameter set Ω. Let Ω^0 be the set of true parameters. Since $U(\Omega^0)$ are martingales, by the martingale central limit theory $n^{-1/2}U(\Omega^0)$ converges weakly to $V(\Omega) = (V_\beta, V_{H_1(s)}, V_{H_2(s)})^T$, where V_β is a zero-mean normal random variable and $V_{H_1(s)}, V_{H_2(s)}$ are zero-mean Gaussian processes. The variance-covariance function of $(V_\beta, V_{H_1(s)}, V_{H_2(s)})$ can be derived below.

The predictable variation process for the score process $n^{-1/2}U_{H_k(s)}, k = 1, 2$, (equations

(2.12) and (2.13)) is

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\varepsilon_{ki}^2(t,s;\beta,H_1,H_2)Y_i(t)\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)dH_1(t),$$

for $k=1,2$, respectively.

An $n\to\infty$, the martingale score process $n^{-1/2}U_{H_k(s)}, k=1,2$, converges weakly to a zero-mean Gaussian process $V_{H_k(s)}, k=1,2$ with covariance function

$$\sigma^2_{H_k}(s,u;\beta^0,H_1^0,H_2^0)=$$
$$\int_0^{\xi}\varepsilon_k(t,s;\beta,H_1^0,H_2^0)\varepsilon_k(t,u;\beta,H_1^0,H_2^0)P(T\geq t)\gamma_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)dH_1^0(t),$$

for $s,u\in[0,\tau]$.

Similarly, as $n\to\infty$, the martingale score process $n^{-1/2}U_\beta$ converges weakly to a zero-mean Gaussian process V_β with covariance

$$\sigma^2_\beta(\beta^0)=\int_0^{\tau}\frac{\dot{\gamma}^2_{i,\beta}\left(t;\beta^0,\overline{H}_1(t),\overline{H}_2(t)\right)}{\gamma_i\left(t;\beta^0,\overline{H}_1(t),\overline{H}_2(t)\right)}P(T\geq t)dH_1(t)$$

As $n\to\infty$, $n^{-1/2}U(H_k^0(t),\beta^0)$, for some t and $k=1,2$, is a martingale and converges to a zero-mean Gaussian process with deterministic covariance function

$$\sigma^2_{H_k,\beta}(s;\beta_0,H_1^0,H_2^0)=\int_0^{\tau}\varepsilon_k(t,s;\beta_0,H_1^0,H_2^0)\dot{\gamma}_i\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)P(T\geq t)dH_1^0(t).$$

Let the normalized likelihood ℓ/n converges in probability to ℓ^0 and $U^0=\left(\frac{\partial\ell^0}{\partial\beta},\frac{\partial\ell^0}{\partial dH(t)}\right)^T$, $dH(t)=(dH_1(t),dH_2(t))$. Introduce an integral equation operator with respect to $(\hat{\Omega}-\Omega^0)$

as follow

$$\mathcal{I}_0(t,s) = \frac{\partial U^0}{\partial \Omega} = -\left(\begin{array}{cc} \frac{\partial^2 \ell^0}{\partial\beta\partial\beta^T} & \frac{\partial^2 \ell^0}{\partial\beta\partial dH(s)} \\ \frac{\partial^2 \ell^0}{\partial dH(t)\partial\beta^T} & \frac{\partial^2 \ell^0}{\partial dH(t)\partial dH(s)} \end{array} \right)_{\Omega=\Omega^0},$$

where $\Omega^0 = (\beta^0, H_t^0)$. The operator $\mathcal{I}_0$ acts on an arbitrary vector-function element $\Omega_s = (\beta, dH(s))^T$ as follow

$$\mathcal{I}_0(t,s)\Omega_s = -\left(\begin{array}{c} \frac{\partial^2 \ell^0}{\partial\beta\partial\beta^T}\beta + \frac{\partial^2 \ell^0}{\partial\beta\partial dH(s)} dH(s) \\ \frac{\partial^2 \ell^0}{\partial dH(t)\partial\beta^T}\beta + \frac{\partial^2 \ell^0}{\partial dH(t)\partial dH(s)} dH(s) \end{array} \right)_{\Omega=\Omega^0}$$

Expand the score $U_t(\hat{\Omega})$ at the true value of parameter set Ω^0, we have

$$n^{1/2} U_t(\hat{\Omega}) = V(t) - n^{-1/2}\mathcal{I}^0(t,s)(\hat{\Omega} - \Omega) + o_p(1)$$

Since $U_t(\hat{\Omega}) = 0$, we have

$$n^{-1/2}\mathcal{I}^0(t,s)(\hat{\Omega} - \Omega) = V(t) + o_p(1) \tag{S.10}$$

Assume that the Fredholm operator expressed by the kernel $\mathcal{I}_0$ of the Fredholm integral equation (S.10) of the first kind is square integrable, and that the equation $\mathcal{I}_0\Omega = 0$ has only the trivial solution $\Omega = 0$. By Theorem 3.3.1 of *Van Der Vaart and Wellner* (1996), equation (S.10) has the unique solution, and there exists the inverse information operator $\mathcal{I}_0^{-1}(t,s)$ such that

$$n^{1/2}(\hat{\Omega} - \Omega^0) = n^{-1/2}(\mathcal{I}^0)^{-1} V(t) + o_p(1)$$

By differentiating the equation $\mathbb{E}[U(\Omega^0)] = 0$ with respect to Ω at the true parameter

value Ω^0 we obtain the variance of the normalized score Gaussian process $V(t)$

$$\mathcal{I}_0(t,s) = -\begin{pmatrix} \frac{\partial \ell^0}{\partial \beta}\frac{\partial \ell^0}{\partial \beta^T} & \frac{\partial \ell^0}{\partial \beta}\frac{\partial \ell^0}{\partial dH(s)} \\ \frac{\partial \ell^0}{\partial dH(t)}\frac{\partial \ell^0}{\partial \beta^T} & \frac{\partial \ell^0}{\partial dH(t)}\frac{\partial \ell^0}{\partial dH(s)} \end{pmatrix}_{\Omega^0},$$

which is equivalent to the likelihood second derivatives $\mathcal{I}^0$. *Andersen et al.* (1993) showed that for a differentiable functional $F(\Omega)$, by functional delta method, $n^{1/2}\{F(\hat{\Omega}) - F(\Omega)\}$ converges weakly to a zero-mean Gaussian process with variance-covariance function $\dot{F}(\Omega)^T(\mathcal{I}^0)^{-1}\dot{F}(\Omega)$, where $\dot{F}(\Omega) = \frac{\partial F}{\partial \Omega}$. Apply the above functional delta method to (2.21), and replacing $\mathcal{I}_0$ by its consistent estimate $n^{-1}\mathcal{I}_n$, we obtain the asymptotic properties stated in Theorem II.2.

A.6.3 Proof of Proposition II.3

In this section, we show that the covariance matrix for β obtained from the profile likelihood $\ell_{pr}(\beta) = \ell\left(\beta, d\hat{H}(\beta) = (d\hat{H}_1(\beta), d\hat{H}_2(\beta))\right)$ converges to the β submatrix of the covariance matrix obtained from the full likelihood at the true model.

Denote the elements of information matrix from the full likelihood at the true model as

$$\mathcal{I}_0(t,s) = \frac{\partial U^0}{\partial \Omega} = -\begin{pmatrix} \frac{\partial^2 \ell^0}{\partial \beta \partial \beta^T} & \frac{\partial^2 \ell^0}{\partial \beta \partial dH(s)} \\ \frac{\partial^2 \ell^0}{\partial dH(t)\partial \beta^T} & \frac{\partial^2 \ell^0}{\partial dH(t)\partial dH(s)} \end{pmatrix}_{\Omega=\Omega^0} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta H} \\ \mathcal{I}_{H\beta} & \mathcal{I}_{HH} \end{pmatrix}$$

Apply the general four blocks matrix inverse formula to the above information matrix, we have the variance covariance matrix for Ω at the true model as

$$\mathcal{I}_0^{-1}(t,s) = \begin{pmatrix} Q^{-1} & -Q^{-1}\mathcal{I}_{\beta H}\mathcal{I}_{HH}^{-1} \\ -\mathcal{I}_{HH}^{-1}\mathcal{I}_{H\beta}Q^{-1} & \mathcal{I}_{HH}^{-1} + \mathcal{I}_{HH}^{-1}\mathcal{I}_{H\beta}Q^{-1}\mathcal{I}_{\beta H}\mathcal{I}_{HH}^{-1} \end{pmatrix},$$

where $Q = \mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta H}\mathcal{I}_{HH}^{-1}\mathcal{I}_{H\beta}$.

We first show that the second derivative of the likelihood with respect to dH is a consistent estimator of the true information submatrix $\mathcal{I}_{HH}$.

Denote $d\Lambda(t) = \gamma\left(t;\beta,\overline{H}_1(t),\overline{H}_2(t)\right)dH_1(t)$, the second partial derivative of the likelihood with respect to dH can be written as

$$\begin{aligned}
\frac{\partial \ell^2}{\partial dH(s)dH(y)} &= \frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \left\{\frac{\partial^2 \log d\Lambda_i(x)}{\partial dH(s)dH(y)}dN_i(x) - \frac{\partial^2 d\Lambda_i(x)}{\partial dH(s)dH(y)}Y_i(x)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \left\{\frac{\partial^2 \log d\Lambda_i(x)}{\partial dH(s)dH(y)}dN_i(x) - \frac{\partial^2 \log d\Lambda_i(x)}{\partial dH(s)dH(y)}d\Lambda_i(x)Y_i(x) - \frac{1}{\partial d\Lambda_i(x)}\frac{\partial d\Lambda_i(x)}{\partial dH(s)}\frac{\partial d\Lambda_i(x)}{\partial dH(y)}Y_i(x)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \left\{\frac{\partial^2 \log d\Lambda_i(x)}{\partial dH(s)dH(y)}dM_i(x) - \frac{\partial \log d\Lambda_i(x)}{\partial dH(s)}\frac{\partial \log d\Lambda_i(x)}{\partial dH(y)}Y_i(x)d\Lambda_i(x)\right\}
\end{aligned}$$

Note that by martingale central limit theory, the process $\frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \frac{\partial^2 \log d\Lambda_i(x)}{\partial dH(s)dH(y)}dM_i(x)$ converges to a zero-mean Gaussian process. Therefore,

$$-\frac{\partial \ell^2}{\partial dH(s)dH(y)} = \frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \frac{\partial \log d\Lambda_i(x)}{\partial dH(s)}\frac{\partial \log d\Lambda_i(x)}{\partial dH(y)}Y_i(x)d\Lambda_i(x) + o_p(1)$$

Notice that by the weak law of large number, $\frac{1}{n}\sum_{i=1}^{n}\int_s^\tau \frac{\partial \log d\Lambda_i(x)}{\partial dH(s)}\frac{\partial \log d\Lambda_i(x)}{\partial dH(y)}Y_i(x)d\Lambda_i(x)$ is the consistent estimator of predictable variation process for the score function $\mathcal{U}_{H(s)}$. Therefore, $-\frac{\partial \ell^2}{\partial dH(s)dH(y)}$ is in fact a consistent estimator of the true information submatrix $\mathcal{I}_{HH}$. From now on, we denote $-\frac{\partial \ell^2}{\partial dH(s)dH(y)}$ as $\hat{\mathcal{I}}_{HH}$.

Next, we will show that the covariance matrix for β obtained from the profile likelihood converges to $\mathcal{Q}^{-1}$, the true covariance matrix of β obtained from the full likelihood at the true model. The first-order partial derivatives of the baseline hazard functions with respect to β is a Jacobian matrix $J_{H\beta} = \frac{\partial d\hat{H}(\beta)}{\partial \beta}$. The profile score function for β can be derived as

$$\mathcal{U}_\beta^{pr} = \frac{d\ell_{pr}}{d\beta} = \left.\frac{\partial \ell}{\partial dH(s)}\right|_{d\hat{H}}\frac{\partial d\hat{H}(s)}{\partial \beta} + \frac{\partial \ell_{pr}}{\partial \beta} = \frac{\partial \ell_{pr}}{\partial \beta}$$

Since $d\hat{H}$ solves the profile score equation $\frac{\partial \ell}{\partial dH(s)} = 0$, so we have $\frac{\partial \ell}{\partial dH(s)}\Big|_{d\hat{H}} = 0$.

The profile Hessian matrix can be derived as

$$
\begin{aligned}
\mathcal{I}^{pr}_{\beta\beta} =& -\frac{d^2\ell_{pr}}{d\beta d\beta^T} \\
=& -\frac{\partial}{\partial dH(y)}\left(\frac{\partial \ell_{pr}}{\partial dH(s)}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(s)}{\partial \beta}\right)\frac{\partial d\hat{H}(y)}{\partial \beta^T} - \frac{\partial}{\partial \beta^T}\left(\frac{\partial \ell}{\partial dH(s)}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(s)}{\partial \beta}\right) \\
& -\frac{\partial}{\partial dH(y)}\left(\frac{\partial \ell}{\partial \beta}\right)\Big|_{d\hat{H}}\frac{\partial d\hat{H}(y)}{\partial \beta^T} - \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} \\
=& -\frac{\partial^2 \ell}{\partial dH(s)\partial dH(y)}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(s)}{\partial \beta}\frac{\partial d\hat{H}(y)}{\partial \beta^T} - \frac{\partial \ell}{\partial dH(s)}\Big|_{d\hat{H}}\frac{\partial^2 d\hat{H}(s)}{\partial \beta \partial dH(y)}\frac{\partial d\hat{H}(y)}{\partial \beta^T} \\
& -\frac{\partial^2 \ell}{\partial dH(s)\partial \beta^T}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(s)}{\partial \beta} - \frac{\partial \ell}{\partial dH(s)}\Big|_{d\hat{H}}\frac{\partial^2 d\hat{H}(s)}{\partial \beta \partial \beta^T} - \frac{\partial^2 \ell}{\partial \beta \partial dH(y)}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(y)}{\partial \beta^T} - \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} \\
=& J_{\beta H}\hat{\mathcal{I}}_{HH}J_{H\beta} - 0 + J_{\beta H}\hat{\mathcal{I}}_{H\beta} - 0 + \hat{\mathcal{I}}_{\beta H}J_{H\beta} + \hat{\mathcal{I}}_{\beta\beta} \\
=& J_{\beta H}\hat{\mathcal{I}}_{HH}J_{H\beta} + J_{\beta H}\hat{\mathcal{I}}_{H\beta} + \hat{\mathcal{I}}_{\beta H}J_{H\beta} + \hat{\mathcal{I}}_{\beta\beta} \qquad \text{(S.11)}
\end{aligned}
$$

To express the Jacobian in terms of $\mathcal{I}$, we make use of the fact that $\frac{\partial \ell_{pr}}{\partial dH(s)} = 0$. We have

$$
\begin{aligned}
0 = \frac{d}{d\beta}\left(\frac{\partial \ell_{pr}}{\partial dH(s)}\right) &= \frac{\partial^2 \ell}{\partial dH(s)\partial \beta^T}\Big|_{d\hat{H}} + \frac{\partial^2 \ell}{\partial dH(s)\partial dH(y)}\Big|_{d\hat{H}}\frac{\partial d\hat{H}(y)}{\partial \beta} = -\hat{\mathcal{I}}_{H\beta} - \hat{\mathcal{I}}_{HH}J_{H\beta} \\
J_{H\beta} &= -\hat{\mathcal{I}}^{-1}_{HH}\hat{\mathcal{I}}_{H\beta} \qquad \text{(S.12)}
\end{aligned}
$$

Replace the Jacobian matrix in (S.11) with the expression (S.12), we obtain the profile Hessian matrix for β as

$$
\mathcal{I}^{pr}_{\beta\beta} = \hat{\mathcal{I}}_{\beta\beta} - \hat{\mathcal{I}}_{\beta H}\hat{\mathcal{I}}^{-1}_{HH}\hat{\mathcal{I}}_{H\beta}
$$

Assuming the regularity conditions hold, by the weak law of large number,

$$
\begin{aligned}
&\hat{\mathcal{I}}_{\beta\beta} - \hat{\mathcal{I}}_{\beta H}\hat{\mathcal{I}}^{-1}_{HH}\hat{\mathcal{I}}_{H\beta} \xrightarrow{p} \mathcal{I}_{\beta\beta} + \mathcal{I}_{\beta H}\mathcal{I}^{-1}_{HH}\mathcal{I}_{H\beta} = Q \\
&\mathcal{I}^{pr^{-1}}_{\beta\beta} \xrightarrow{p} Q^{-1}
\end{aligned}
$$

The covariance matrix for β obtained from the full likelihood at true model, Q^{-1}, can be consistently estimated by the profile likelihood covariance operator $\mathcal{I}_{\beta\beta}^{pr^{-1}}$.

A.7 Observed Information Matrix

The observed information matrix $\mathcal{I}_n$ for the parameter set $\hat{\Omega} = (\hat{\beta}, \{d\hat{H}_1\}, \{d\hat{H}_2\})$ can be obtained explicitly by taking derivative of the negative score functions. For brevity, we denote $dH_k(t_j) = dH_{kj}$ for $k = 1, 2$ and $\gamma_i\left(t; \beta, \overline{H}_1(t), \overline{H}_2(t)\right) = \gamma_i\left(t; \Omega_t\right)$. We introduce the following notation for $k = 1, 2, l = 1, 2$ and for the failure time $t_* \in (t_1, \ldots, t_J)$:

$$\begin{aligned}
\gamma_{i,H_k,H_l}(t;\Omega_t) &= Y_i(t_*)\frac{\partial\dot{\gamma}_{i,H_k}\left(t;\Omega_t\right)}{\partial dH_{l*}} \\
\gamma_{i,\beta,\beta}(t;\Omega_t) &= \frac{\partial\dot{\gamma}_{i,\beta}\left(t;\Omega_t\right)}{\partial\beta} \\
\gamma_{i,\beta,H_l}(t;\Omega_t) &= Y_i(t_*)\frac{\partial\dot{\gamma}_{i,\beta}\left(t;\Omega_t\right)}{\partial dH_{l*}}
\end{aligned}$$

The explicit form of observed information matrix can be expressed as follow, for $j \neq l$:

$$\begin{aligned}
\mathcal{I}_{dH_{1j}dH_{1j}} = &\sum_{i=1}^{n}\left\{\frac{dN_i(t_j)}{dH_{1j}^2} + Y_i(t_j)\left[\dot{\gamma}_{i,H_1}\left(t_j;\Omega_{tj}\right) + \int_{t_j^+}^{\tau}\gamma_{i,H_1,H_1}\left(t;\Omega_t\right)dH_1(t)\right]\right\} - \\
&\sum_{i=1}^{n}Y_i(t_j)\int_{t_j^+}^{\tau}\left[\frac{\gamma_{i,H_1,H_1}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,H_1}^{\otimes 2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)}\right]dN_i(t)
\end{aligned}$$

$$\mathcal{I}_{dH_{1j}dH_{1l}} = \sum_{i=1}^{n} Y_i(t_j) \left[\dot{\gamma}_{i,H_1}\left(t_j;\Omega_{tj}\right) \mathbb{1}(t_l < t_j) + \int\limits_{t_j^+ \vee t_l^+}^{\tau} \gamma_{i,H_1,H_1}\left(t;\Omega_t\right) dH_1(t) \right] -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+ \vee t_l^+}^{\tau} \left[\frac{\gamma_{i,H_1,H_1}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,H_1}^{\otimes 2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{dH_{2j}dH_{2j}} = \sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+}^{\tau} \gamma_{i,H_2,H_2}\left(t;\Omega_t\right) dH_1(t) -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+}^{\tau} \left[\frac{\gamma_{i,H_2,H_2}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,H_2}^{\otimes 2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{dH_{2j}dH_{2l}} = \sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+ \vee t_l^+}^{\tau} \gamma_{i,H_2,H_2}\left(t;\Omega_t\right) dH_1(t) -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+ \vee t_l^+}^{\tau} \left[\frac{\gamma_{i,H_2,H_2}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,H_2}^{\otimes 2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{dH_{1j}dH_{2l}} = \sum_{i=1}^{n} Y_i(t_j) \left[\dot{\gamma}_{i,H_2}\left(t_j;\Omega_{tj}\right) \mathbb{1}(t_l < t_j) + \int\limits_{t_j^+ \vee t_l^+}^{\tau} \gamma_{i,H_1,H_2}\left(t;\Omega_t\right) dH_1(t) \right] -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+ \vee t_l^+}^{\tau} \left[\frac{\gamma_{i,H_1,H_2}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,H_1}\left(t;\Omega_t\right)\gamma_{i,H_2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{\beta dH_{1j}} = \sum_{i=1}^{n} Y_i(t_j) \left[\gamma_{i,\beta}(t_j;\Omega_{tj}) + \int\limits_{t_j^+}^{\tau} \gamma_{i,\beta,H_1}\left(t;\Omega_t\right) dH_1(t) \right] -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+}^{\tau} \left[\frac{\gamma_{i,\beta,H_1}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,\beta}\left(t;\Omega_t\right)\gamma_{i,H_1}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{\beta dH_{2j}} = \sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+}^{\tau} \gamma_{i,\beta,H_2}\left(t;\Omega_t\right) dH_1(t) -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int\limits_{t_j^+}^{\tau} \left[\frac{\gamma_{i,\beta,H_2}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,\beta}\left(t;\Omega_t\right)\gamma_{i,H_2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

$$\mathcal{I}_{\beta\beta} = \sum_{i=1}^{n} Y_i(t_j) \int_{t_j^+}^{\tau} \gamma_{i,\beta,\beta}\left(t;\Omega_t\right) dH_1(t) -$$
$$\sum_{i=1}^{n} Y_i(t_j) \int_{t_j^+}^{\tau} \left[\frac{\gamma_{i,\beta,\beta}\left(t;\Omega_t\right)}{\gamma_i\left(t;\Omega_t\right)} - \frac{\gamma_{i,\beta}^{\otimes 2}\left(t;\Omega_t\right)}{\gamma_i^2\left(t;\Omega_t\right)} \right] dN_i(t)$$

A.8 Gompertz Survival Model for Death from Other Causes

In Section 2.8, we consider the practical concept of cure. Modeling practical concept of cure requires modeling residual survival for other causes from the age of prostate cancer diagnosis. Denote T_{oc} as the time to death from other causes from birth and $S_{oc}(t|Z)$ as the survival function of T_{oc}. Then the residual survival for other causes is

$$P(T_{oc} > t + a | T_{oc} > a, Z) = \frac{P(T_{oc} > t + a|Z)}{P(T_{oc} > a|Z)} = \frac{S_{oc}(t + a|Z)}{S_{oc}(a|Z)} = S_{oc}(t|a, Z)$$

where a is the age at prostate cancer diagnosis. Without loss of generality, assume Z is a null set. Assume age at prostate cancer diagnosis is proportional hazard covariate for $S_{oc}(t|a)$.

- Under cox model framework,

$$\frac{S_{oc}(t + a)}{S_{oc}(a)} = e^{-[H(t+a)-H(a)]}, \text{and } S_{oc}(t|a) = e^{-H^*(t)e^{\beta a}}$$
$$H(t + a) - H(a) = H^*(t)e^{\beta a} \quad \forall a, t$$

- Take $a = 0$, the above implies $H(t) = H^*(t) \quad \forall t$. Therefore we have

$$H(t + a) - H(a) = H(t)e^{\beta a} \tag{A.1}$$

- Solve equation (A.1) by differential equation:

$$H'(a) = H'(0)e^{\beta a}$$

$$H(t) = C \cdot (e^{\beta t} - 1)$$

where C is a positive constant. Recognize that $H(t)$ is a Gompertz cumulative hazard function (characterization of Gompertz distribution).

Table A.1: Characteristics of studied patients

Characteristic		One primary only (N=189,264)	Secondary cancer patients (N=11,730)
		mean (SD) or n (%)	
First primary cancer (prostate)			
Age at diagnosis (yr)		66 (9.4)	69 (8.3)
Stage	*Localized/Regional*	183,338 (96.9)	11,489 (97.9)
	Distant	5,926 (3.1)	241 (2.1)
Grade*	*I/II*	100,173 (52.9)	7,149 (60.9)
	III/IV	89,091 (47.1)	4,581 (39.1)
Secondary cancer			
Age at diagnosis (yr)			72 (8.3)
Stage	*Localized*		5,626 (48.0)
	Regional		2,783 (23.7)
	Distant		3,321 (28.3)
Median follow-up time (months)		56	72
No. patients died of prostate cancer		8,028 (4.2)	488 (4.2)
	Secondary cancer stage		
	Localized		180
	Regional		113
	Distant		195
Median follow-up time to secondary cancer (months)			
	Secondary cancer stage		
	Localized		34
	Regional		36
	Distant		38

*Grade I: Well differentiated
Grade II: Moderately differentiated
Grade III: Poorly differentiated
Grade IV: Undifferentiated; anaplastic

Table A.2: Characteristics of secondary cancer patients

Characteristic	n (%)
Median follow-up time to secondary cancer (months)	
Localized	34
Regional	36
Distant	38
Median follow-up time from secondary cancer diagnosis to death (months)	
Localized	33
Regional	19
Distant	8
No. patients died of prostate cancer	
Localized	180
Regional	113
Distant	195
No. patients died of other causes	
Localized	1,375 (24.4*)
Regional	1,330 (48.8)
Distant	2,256 (67.9)

*Percentage denominator: total number of death among localized secondary cancer patients

Table A.3: Frequency table of causes of death for secondary cancer patients n (%*)

Cause of death	Alive	Prostate cancer	Other causes	Total
Secondary cancer stage				
Localized	4,071 (72.4)	180 (3.2)	1,375 (24.4)	5,626 (100)
Regional	1,340 (48.1)	113 (4.1)	1,330 (48.8)	2,783 (100)
Distant	870 (26.2)	195 (6.9)	2,256 (67.9)	3,321 (100)

*row percentage

APPENDIX B

A Semiparametric Latent Trait Model for Multiple Mixed Continuous, Categorical, and Time-to-event Outcomes

B.1 Proportional Odds Ratio

From the survival function (3.5), we can see that β_j is the proportional odds ratio for one unit increase in Z_j on Y_j, given U.

$$\frac{\frac{P(Y_j \leq y_j|U,Z_j=z_j+1)}{P(Y_j>y_j|U,Z_j=z_j+1)}}{\frac{P(Y_j \leq y_j|U,Z_j=z_j)}{P(Y_j>y_j|U,Z_j=z_j)}} = \frac{U^{\alpha_j} e^{(z_j+1)\beta_j} H_j(y_j)}{U^{\alpha_j} e^{z_j\beta_j} H_j(y_j)} = e^{\beta_j}$$

B.2 EM Algorithm

The EM algorithm is an efficient iterative procedure to compute the maximum likelihood estimate in the presence of missing data. Each iteration of the EM algorithm consists of the E-step and the M-step. In the E-step, the missing data are estimated by the conditional expectation given the observed data and the model parameters at the current iteration. In the M-step, the likelihood function is maximized by substituting the missing data with conditional expectation obtained in the E-step. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

Let Y be the observed data and U be the missing data. The complete data likelihood is $L_0(\theta) = f(Y, U|\theta)$ where θ is the model parameter. Denote the complete data log-likelihood as $\ell_0(\theta) = \log L_0(\theta)$. The conditional score function can be obtained by taking derivative of the complete data log-likelihood $\mathcal{U}_0 = \frac{\partial \ell_0(\theta)}{\partial \theta}$. Introduce the conditional expectation operator $\mathbb{E}[f\|g] = \frac{\mathbb{E}[f \cdot g]}{\mathbb{E}[g]}$. The E-M algorithm at the kth iteration consists of the following E-step and M-step:

1. E-step:

   Determine the conditional expectation given observed data Y,

$$Q(\theta|\theta^{(k)}) = \mathbb{E}_{U|Y,\theta^{(k)}}[\ell_0(\theta)] = \int \ell_0(\theta) f(u|Y, \theta^{(k)}) du = \int \ell_0(\theta) \frac{f(u, Y|\theta^{(k)})}{f(Y|\theta^{(k)})} du$$
$$= \frac{\int \ell_0(\theta) f(Y, u|\theta^{(k)}) du}{f(Y|\theta^{(k)})} = \frac{\int \ell_0(\theta) f(Y, u|\theta^{(k)}) du}{\int f(Y, u|\theta^{(k)}) du} = \frac{\mathbb{E}[\ell_0(\theta) \cdot L_0(\theta^{(k)})]}{\mathbb{E}[L_0(\theta^{(k)})]} = \mathbb{E}\left[\ell_0(\theta) \middle\| L_0^{(k)}\right] \tag{B.1}$$

   where $L_0^{(k)} = L_0(\theta^{(k)})$.

2. M-step:

   Update θ by the value that maximizes (B.1) with respect to θ.

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)})$$

   This can be achieved by taking derivative of (B.1) and setting the resulting equation to zero,

$$\frac{\partial Q(\theta|\theta^{(k)})}{\partial \theta} = \mathbb{E}\left[\frac{\partial \ell_0(\theta)}{\partial \theta} \middle\| L_0^{(k)}\right] = \mathbb{E}\left[\mathcal{U}_0(\theta) \middle\| L_0^{(k)}\right] = 0 \tag{B.2}$$

   And $\mathbb{E}\left[\mathcal{U}_0(\theta^{(k+1)}) \middle\| L_0^{(k)}\right] = 0$.

   Notice that $\mathbb{E}[\mathcal{U}_0(\theta)\|L_0]$ is the marginal score function for θ. We can see this by

writing out the marginal score function,

$$\mathcal{U}(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial \log \mathbb{E}[L_0(\theta)]}{\partial \theta} = \frac{\frac{\partial \mathbb{E}[L_0(\theta)]}{\partial \theta}}{\mathbb{E}[L_0(\theta)]} = \frac{\mathbb{E}\left[\frac{\partial L_0(\theta)}{\partial \theta}\right]}{\mathbb{E}[L_0(\theta)]} = \frac{\mathbb{E}\left[\frac{\partial \ell_0(\theta)}{\partial \theta} \cdot L_0(\theta)\right]}{\mathbb{E}[L_0(\theta)]} = \mathbb{E}\left[\ell_0(\theta) \| L_0\right]$$

Therefore, the M-step maximizes the conditional expectation of the conditional score function given the observed data and the model parameter at the current iteration. For the following derivation of the EM-DCA algorithm, we adopt the expression (B.2) to represent the E-step and M-step in the algorithm.

B.3 EM-DCA Algorithm

We use the methods of EM algorithm in *Tsodikov* (2003) and the difference of convex functions algorithm (DCA) to iteratively estimate the infinite dimensional transformation functions $(H_1(\Omega), \ldots, H_m(\Omega))$. Consider n independent subjects with m distinct outcomes. For subject $i = 1, \ldots, n$, we observe the covariate vectors $(Z_{i1}, \ldots, Z_{im})$ corresponding to a vector of outcomes $(Y_{i1}, \ldots, Y_{im})$ and censoring indicators $(\delta_{i1}, \cdots, \delta_{im})$. As an example, we assume the shared latent variable for the i^{th} subject $U_i \sim Gamma(a_i, b_i)$, where $a_i = \exp(\eta_1 Z_i)$ and $b_i = \exp(\eta_2 Z_i)$. The full parameter sets consist of parameter vectors $\Omega = (\alpha, \beta, \eta)$ and infinite-dimensional $H = (H_1(\cdot), \ldots, H_m(\cdot))$. The joint likelihood of the observed data (Y, Z) and the shared latent variable U can be expressed as

$$\begin{aligned} & L_0(\Omega, H|Y, Z, U) f(U; \eta) \\ & = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\frac{1}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij}^-)} - \frac{1}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})} \right]^{\delta_{ij}} \left[\frac{1}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})} \right]^{1-\delta_{ij}} \frac{b_i^{a_i}}{\Gamma(a_i)} U_i^{a_i - 1} e^{-b_i U_i} \end{aligned} \tag{B.3}$$

Taking the log of (B.3),we obtain the joint log likelihood

$$
\begin{aligned}
&\ell_0(\Omega, H|Y, Z, U) + \log f(U;\eta) \\
&= \sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\left[\delta_{ij}\log\left(\frac{U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij})}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]}\right) + (1-\delta_{ij})\log\left(\frac{1}{1+U_i^{\alpha_j}\gamma_i H_j(y_{ij})}\right)\right]\right. \\
&\left. + \frac{b_i^{a_i}}{\Gamma(a_i)} + (a_i-1)\log(U_i) - b_i U_i\right\} \\
&= \sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\delta_{ij}\log U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij}) - \sum_{j=1}^{m}\delta_{ij}\log[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)] - \sum_{j=1}^{m}\log[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]\right. \\
&\left. + \log\frac{b_i^{a_i}}{\Gamma(a_i)} + (a_i-1)\log(U_i) - b_i U_i\right\}. \qquad \text{(B.4)}
\end{aligned}
$$

1. **Difference of convex algorithm (DCA)**

   DCA is a version of the Minorize-Maximization(MM) algorithm that iteratively optimizes an objective function that can be expressed as the difference of concave functions. Consider two concave functions $f(x)$ and $g(x)$. The objective function to be maximized is $f(x) - g(x)$. By the subgradient inequality of concave function,

$$
\begin{aligned}
g(x) &\leq g(x^*) + \nabla g(x^*)(x - x^*) \\
f(x) - g(x) &\geq f(x) - (g(x^*) + \nabla g(x^*)(x - x^*))
\end{aligned}
$$

   Notice that the objective function (B.4) is a difference between two concave functions

of H. We construct the surrogate function as

$$
\begin{aligned}
&S(dH_j, dH_j^{(k)}) \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} \delta_{ij} \log U_i^{\alpha_j} \gamma_{ij} dH_j(y_{ij}) - \sum_{j=1}^{m} \delta_{ij} \log[1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij}^-)] - \sum_{j=1}^{m} \log[1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})] \right. \\
&\left. + \log \frac{b_i^{a_i}}{\Gamma(a_i)} + (a_i - 1) \log(U_i) - b_i U_i \right\} \qquad \text{(B.5)} \\
&- (dH_j(y_{ij}) - dH_j^{(k)}(y_{ij})) \left[\sum_{i=1}^{n} \frac{\delta_{ij} \, \mathbb{1}(y_{ij}^- \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij}^-)} - \sum_{i=1}^{n} \frac{\mathbb{1}(y_{ij} \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij})} \right] \qquad \text{(B.6)}
\end{aligned}
$$

The surrogate function (B.5) satisfies $S(dH^{(k)}, dH^{(k)}) = \ell_0(dH^{(k)}) + \log f(U)$ and $S(dH, dH^{(k)}) \leq \ell_0(dH) + \log f(U)$. By MM algorithm theory, the next iteration $dH^{(k+1)}$ that maximizes $S(dH, dH^{(k)})$ in M-step will improve the likelihood.

Instead of maximizing imputed joint log likelihood, we employ DCA and maximize imputed surrogate function (B.5). Applying the functional derivative to the surrogate function (B.5) with respect to $dH_j(s)$, we obtain the conditional score function for $dH_j(s)$ as

$$
\begin{aligned}
\mathcal{U}_{0,dH_j}(s) &= \delta_s S(dH, dH^{(k)}) \\
&= \frac{\sum_{i=1}^{n} \delta_{ij} \, \mathbb{1}(y_{ij} = s)}{dH_j(y_{ij})} - \sum_{i=1}^{n} \frac{\delta_{ij} \, \mathbb{1}(y_{ij}^- \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij}^-)} - \sum_{i=1}^{n} \frac{\mathbb{1}(y_{ij} \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij})} \\
&= \frac{\sum_{i=1}^{n} dN_{ij}(s)}{dH_j(y_{ij})} - \sum_{i=1}^{n} \frac{\delta_{ij} \, \mathbb{1}(y_{ij}^- \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij}^-)} - \sum_{i=1}^{n} \frac{\mathbb{1}(y_{ij} \geq s) U_i^{\alpha_j} \gamma_{ij}}{1 + U_i^{\alpha_j} \gamma_{ij} H_j^{(k)}(y_{ij})} \qquad \text{(B.7)}
\end{aligned}
$$

where $\sum_{i=1}^{n} dN_{ij}(s) = \sum_{i=1}^{n} \delta_{ij} \, \mathbb{1}(y_{ij} = s)$ records the number of observations in outcome Y_j are of value s.

2. **E step**

Based on Appendix B.2, the marginal score function can be represented by the

conditional expectation operator as below

$$\begin{aligned}\mathcal{U}_{dH_j}(s) &= \mathbb{E}\left[\mathcal{U}_{0,dH_j}(s)\middle\| L_0^{(k)}\right] \\ &= \frac{\sum_{i=1}^n dN_{ij}(s)}{dH_j(s)} - \sum_{i=1}^n \delta_{ij}\,\mathbb{1}(y_{ij}^- \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right] \\ &- \sum_{i=1}^n \mathbb{1}(y_{ij} \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right].\end{aligned}$$

The conditional expectation $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right]$ and $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right]$ are obtained by Laplace approximation (*Laplace* (1986)).

Let $u = e^v$. The conditional expectation $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right]$ can be written as

$$\begin{aligned}&\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\middle\| L_0\right] = \mathbb{E}\left[\frac{e^{v_i\alpha_j}}{1+e^{v_i\alpha_j}\gamma_{ij}H_j(y_{ij})}\middle\| L_0\right] \\ &= \frac{\int_{-\infty}^{\infty}\frac{e^{v\alpha_j}}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\prod_{j=1}^m\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij}^-)}-\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)^{\delta_{ij}}\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)^{1-\delta_{ij}}\frac{b_i^{a_i}}{\Gamma(a_i)}e^{va_i-b_ie^v}dv}{\int_{-\infty}^{\infty}\prod_{j=1}^m\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij}^-)}-\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)^{\delta_{ij}}\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)^{1-\delta_{ij}}\frac{b_i^{a_i}}{\Gamma(a_i)}e^{va_i-b_ie^v}dv}\end{aligned} \tag{B.8}$$

The numerator and the denominator of (B.8) can be approximated by Laplace's method. We obtain

$$\begin{aligned}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\middle\| L_0\right] &\approx \frac{\frac{e^{\hat{v}_i\alpha_j}}{1+e^{\hat{v}_i\alpha_j}\gamma_{ij}H_j(y_{ij})}\frac{b_i^{a_i}}{\Gamma(a_i)}e^{mf(\hat{v}_i)}\sqrt{\frac{2\pi}{m|f''(\hat{v}_i)|}}}{\frac{b_i^{a_i}}{\Gamma(a_i)}e^{mf(\hat{v}_i)}\sqrt{\frac{2\pi}{m|f''(\hat{v}_i)|}}} \\ &= \frac{e^{\hat{v}_i\alpha_j}}{1+e^{\hat{v}_i\alpha_j}\gamma_{ij}H_j(y_{ij})} = \frac{\hat{U}_i^{\alpha_j}}{1+\hat{U}_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\end{aligned}$$

where

$$\begin{aligned}f(v) = \frac{1}{m}\Bigg\{&va_i - b_ie^v \\ &+\sum_{j=1}^m\left[\delta_{ij}\log\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij}^-)}-\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)+(1-\delta_{ij})\left(\frac{1}{1+e^{v\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)\right]\Bigg\}\end{aligned}$$

Here $\hat{v_i}$ is assumed not an endpoint of the interval of integration and is the global maximum of $f(v)$ for the i^{th} subject. That is, $f'(\hat{v_i}) = 0$.

Similar approximation method is applied to the term $\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right] \approx \frac{\hat{U_i}^{\alpha_j}}{1+\hat{U_i}^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)}$.

3. **M step**

The estimator for $dH_j^{(k+1)}(s)$ that maximize $\mathbb{E}\left(S(dH, dH^{(k)})\middle\| L_0^{(k)}\right)$ can be obtained by solving $\mathcal{U}_{dH_j}(s) = 0$. The solution results in a Breslow-type estimator

$$dH_j^{(k+1)}(s) = \frac{\sum_{i=1}^n dN_{ij}(s)}{\sum_{i=1}^n \delta_{ij}\,\mathbb{1}(y_{ij}^- \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_j^{(k)}(y_{ij}^-)}\middle\| L_0^{(k)}\right] + \sum_{i=1}^n \mathbb{1}(y_{ij} \geq s)\gamma_{ij}\mathbb{E}\left[\frac{U_i^{\alpha_j}}{1+U_i^{\alpha_j}\gamma_{ij}H_k^{(k)}(y_{ij})}\middle\| L_0^{(k)}\right]} \tag{B.9}$$

This constitutes a self-consistent equation that can be solved iteratively (*Tsodikov* (2003)).

Note that the conditional expectation operator $\mathbb{E}\left[\quad\cdot\quad\middle\| L_0^{(k)}\right]$ is linear and does not alter convexity properties. Hence, iterations based on (B.9) will also constitute an MM algorithm monotonically improves the likelihood.

B.4 Asymptotic Properties

B.4.1 Proof of Theorem III.1

For each of $i = 1, \cdots, n$th independent subject, we observed $\{Y_{i1}, \ldots, Y_{im}, \delta_{i1}, \ldots, \delta_{im}, Z_{i1}, \ldots, Z_{im}\}$. Note that for non-survival outcomes, all event indicator δ_{ij} are one. The contribution to the marginal likelihood can be obtained by taking the expectation of complete data likelihood. We can write the observed data likelihood function as:

$$L_n(\Omega, H|Y, Z) = \prod_{i=1}^m \mathbb{E}\left\{\prod_{j=1}^m \left(\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})}\right)^{1-\delta_{ij}} \left(\frac{U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij})}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]}\right)^{\delta_{ij}}\right\}.$$

And the observed data log-likelihood:

$$\ell_n(\Omega, H|Y, Z)$$
$$= \sum_{i=1}^{n} \log \mathbb{E}\left\{ \prod_{j=1}^{m} \left(\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})} \right)^{1-\delta_{ij}} \left(\frac{U_i^{\alpha_j}\gamma_{ij}dH_j(y_{ij})}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]} \right)^{\delta_{ij}} \right\} \tag{B.10}$$

To prove theorem 1, we show that any convergent sub-sequence of $(\hat{\Omega}_n, \hat{H}_n)$ must converge to (Ω_0, H_0). Since $\hat{\Omega}_n$ and $\hat{H}_n$ belong to a compact set, we can assume that $\hat{\Omega}_n \to \Omega^*$ and $\hat{H}_n(\cdot)$ converges point-wise to a monotone function $H^*(\cdot)$ within its domain D. We will show that $\Omega^* = \Omega_0$ and $H^*(y) = H_0(y)$ for all y within the domain.

The observed data log-likelihood (B.10) can be written as

$$\ell_n(\Omega, H|Y, Z)$$
$$= \sum_{i=1}^{n} \log \mathbb{E}\left\{ \prod_{j=1}^{m} \left(\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})} \right)^{1-\delta_{ij}} \left(\frac{U_i^{\alpha_j}\gamma_{ij}}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]} \right)^{\delta_{ij}} \right\} \prod_{j=1}^{m} dH_j(y_{ij})^{\delta_{ij}}$$
$$= \sum_{i=1}^{n} \log \mathbb{E}\left\{ \prod_{j=1}^{m} \left(\frac{1}{1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})} \right)^{1-\delta_{ij}} \left(\frac{U_i^{\alpha_j}\gamma_{ij}}{[1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij}^-)][1+U_i^{\alpha_j}\gamma_{ij}H_j(y_{ij})]} \right)^{\delta_{ij}} \right\}$$
$$+ \sum_{i=1}^{n}\sum_{j=1}^{m} \delta_{ij} \log(dH_j(y_{ij}))$$

By differentiating $\ell(\Omega, H)$ with respect to $dH_k(s)$,

$$\frac{\partial \ell_n}{\partial dH_k(s)} = \frac{\sum_{i=1}^{n} \delta_{ik}\, \mathbb{1}(y_{ik} = s)}{dH_k(s)} - \sum_{i=1}^{n} \frac{\mathbb{E}[R_{1i}(\Omega, H)R_{2i}(s, \Omega, H)]}{\mathbb{E}[R_{1i}(\Omega, H)]}$$
$$= \frac{\sum_{i=1}^{n} dN_{ik}(s)}{dH_k(s)} - \sum_{i=1}^{n} \frac{\int R_{1i}(\Omega, H, u)R_{2i}(s, \Omega, H, u)f(u;\eta)du}{\int R_{1i}(\Omega, H, u)f(u;\eta)du}$$

where

$$R_{1i}(\Omega, H, u) = \prod_{j=1}^{m} \left(\frac{1}{1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})} \right)^{1-\delta_{ij}} \left(\frac{U_i^{\alpha_j} \gamma_{ij}}{[1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij}^-)][1 + U_i^{\alpha_j} \gamma_{ij} H_j(y_{ij})]} \right)^{\delta_{ij}}$$

$$R_{2i}(s, \Omega, H, u) = \frac{(1 + \delta_{ik})\, \mathbb{1}(y_{ik} \geq s) U_i^{\alpha_k} \gamma_{ik}}{1 + U_i^{\alpha_k} \gamma_{ik} H_k(y_{ik})}$$

Setting the derivative to zero, we obtain the equation

$$\frac{\sum_{i=1}^{n} dN_{ik}(s)}{dH_k(s)} = \sum_{i=1}^{n} \frac{\int R_{1i}(\Omega, H, u) R_{2i}(s, \Omega, H, u) f(u; \eta) du}{\int R_{1i}(\Omega, H, u) f(u; \eta) du}$$

Therefore, we see that $d\hat{H}_k$ satisfies the equation

$$\frac{\sum_{i=1}^{n} dN_{ik}(s)}{d\hat{H}_k(s)} = \sum_{i=1}^{n} \frac{\int R_{1i}(\hat{\Omega}, \hat{H}, u) R_{2i}(s, \hat{\Omega}, \hat{H}, u) f(u; \hat{\eta}) du}{\int R_{1i}(\hat{\Omega}, \hat{H}, u) f(u; \hat{\eta}) du} \tag{B.11}$$

Construct a function $\widetilde{H}$ by imitating $\hat{H}$ and we will show $\widetilde{H}$ uniformly converges to H_0. Define $\widetilde{H}$ as a step function with jumps only at the y_{ij} for which $\delta_{ij} = 1$ and $d\widetilde{H}_k$ satisfies the equation

$$\frac{\sum_{i=1}^{n} dN_{ik}(s)}{d\widetilde{H}_k(s)} = \sum_{i=1}^{n} \frac{\int R_{1i}(\Omega_0, H_0, u) R_{2i}(s, \Omega_0, H_0, u) f(u; \eta_0) du}{\int R_{1i}(\Omega_0, H_0, u) f(u; \eta_0) du} \tag{B.12}$$

By definition, $\widetilde{H}_k(s) = \sum_{i=1}^{n} \mathbb{1}(y_{ik} \leq s) d\widetilde{H}_k(y_{ik})$. By Glivenko-Cantelli theorem, $\widetilde{H}_k(s)$ converges almost surely to $\mathbb{E}\{\mathbb{1}(y_{ik} \leq s) f_{Y_k}(s)/\mu(y_{ik})\}$, where

$$\begin{aligned}
\mu(s) &= \mathbb{E}\left\{ \frac{\int R_{1i}(\Omega_0, H_0, u) R_{2i}(s, \Omega_0, H_0, u) f(u; \eta_0) du}{\int R_{1i}(\Omega_0, H_0, u) f(u; \eta_0) du} \right\} \\
&= \mathbb{E}\left\{ \mathbb{E}\left[\frac{(1 + \delta_{ik})\, \mathbb{1}(y_{ik} \geq s) U_i^{\alpha_{0k}} \gamma_{ik}^0}{1 + U_i^{\alpha_{0k}} \gamma_{ik}^0 H_{0k}(y_{ik})} \middle| R_{1i}(\Omega_0, H_0, u) \right] \right\}, \\
f_{Y_k}(s) &= \mathbb{E}[dN_{ik}(s)] = \mathbb{E}[\delta_{ik}\, \mathbb{1}(y_{ik} = s)]
\end{aligned}$$

Denote $S_c(\cdot|Z)$ the survival function of censoring time C given Z.

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{(1+\delta_{ik})\,\mathbb{1}(y_{ik}\geq s)U_i^{\alpha_{0k}}\gamma_{ik}^0}{1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y_{ik})}\middle| R_{1i}(\Omega_0,H_0,u)\right\}\\
=&\mathbb{E}\left\{2\int_s^\infty \frac{U_i^{\alpha_{0k}}\gamma_{ik}^0}{1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y)}\frac{U_i^{\alpha_{0k}}\gamma_{ik}^0 h_{0k}(y)}{[1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y)]^2}S_c(y|Z_{ik})dy\middle| R_{1i}(\Omega_0,H_0,u)\right\}\\
-&\mathbb{E}\left\{\int_s^\infty \frac{U_i^{\alpha_{0k}}\gamma_{ik}^0}{1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y)}\frac{1}{1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y)}dS_c(y|Z_{ik})\middle| R_{1i}(\Omega_0,H_0,u)\right\}\\
=&\mathbb{E}\left\{\frac{S_c(s|Z_{ik})U_i^{\alpha_{0k}}\gamma_{ik}^0}{[1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(s)]^2}\middle| R_{1i}(\Omega_0,H_0,u)\right\}
\end{aligned}
$$

where the second equality follows from integration by part. Therefore we have

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{\mathbb{1}(y_{ik}\leq s)f_{Y_{ik}}(s)}{\mu(y_{ik})}\right\}=\mathbb{E}\left\{\mathbb{E}\left[\int_0^s \frac{S_c(y|Z_{ik})U_i^{\alpha_{0k}}\gamma_{ik}^0 h_{0k}(y)}{\mu(y)[1+U_i^{\alpha_{0k}}\gamma_{ik}^0 H_{0k}(y)]^2}dy\middle| R_{1i}(\Omega_0,H_0,u)\right]\right\}\\
&=\int_0^s h_{0k}(y)dy=H_{0k}(s)
\end{aligned}
$$

Therefore, $\widetilde{H}_k(s)$ converges uniformly to $H_{0k}(s)$ in its specific domain D_k. By plugging in (B.11) into $\ell_n(\hat{\Omega},\hat{H})$, we obtain

$$
\begin{aligned}
&\ell_n(\hat{\Omega},\hat{H})\\
=&\sum_{i=1}^n \log\mathbb{E}\left\{\prod_{j=1}^m\left(\frac{1}{1+U_i^{\hat{\alpha}_j}\hat{\gamma}_{ij}\hat{H}_j(y_{ij})}\right)^{1-\delta_{ij}}\left(\frac{U_i^{\hat{\alpha}_j}\hat{\gamma}_{ij}}{[1+U_i^{\hat{\alpha}_j}\hat{\gamma}_{ij}\hat{H}_j(y_{ij}^-)][1+U_i^{\hat{\alpha}_j}\hat{\gamma}_{ij}\hat{H}_j(y_{ij})]}\right)^{\delta_{ij}}\right\}\\
&+\sum_{i=1}^n\sum_{j=1}^m \delta_{ij}\log(d\hat{H}_j(y_{ij}))\\
=&\sum_{i=1}^n \log\int R_{1i}(\hat{\Omega},\hat{H},u)f(u;\hat{\eta})du+\sum_{i=1}^n\sum_{j=1}^m\delta_{ij}\log\left(\sum_{k=1}^n dN_{kj}(y_{ij})\right)\\
&-\sum_{i=1}^n\sum_{j=1}^m\delta_{ij}\log\left(\sum_{k=1}^n\frac{\int R_{1k}(\hat{\Omega},\hat{H},u)R_{2k}(y_{ij},\hat{\Omega},\hat{H},u)f(u;\hat{\eta})du}{\int R_{1k}(\hat{\Omega},\hat{H},u)f(u;\hat{\eta})du}\right).
\end{aligned}
$$

Likewise, plug in (B.12) into $\ell_n(\Omega_0, \widetilde{H})$, we obtain

$$\ell_n(\Omega_0, \widetilde{H})) = \sum_{i=1}^{n} \log \int R_{1i}(\Omega_0, \widetilde{H}, u) f(u; \eta_0) du + \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} \log \left(\sum_{k=1}^{n} dN_{kj}(y_{ij}) \right) - \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} \log \left(\sum_{k=1}^{n} \frac{\int R_{1k}(\Omega_0, \widetilde{H}, u) R_{2k}(y_{ij}, \Omega_0, \widetilde{H}, u) f(u; \eta_0) du}{\int R_{1k}(\Omega_0, \widetilde{H}, u) f(u; \eta_0) du} \right).$$

Define

$$R_{3i}(s, \Omega, H) = \frac{\int R_{1i}(\Omega, H, u) R_{2i}(s, \Omega, H, u) f(u; \eta) du}{\int R_{1i}(\Omega, H, u) f(u; \eta) du}.$$

We see that $\hat{H}_k(y)$ is continuous with respect to $\widetilde{H}_k(y)$ and

$$\hat{H}_k(y) = \int \frac{\sum_{i=1}^{n} R_{3i}(s, \Omega_0, H_0)}{\sum_{i=1}^{n} R_{3i}(s, \hat{\Omega}, \hat{H})} d\widetilde{H}_k(s)$$

for $k = 1, \cdots, m$.

If the kth outcome is continuous or time-to-event, by taking limits on both sides of the above equation, we see that $H_k^*(y)$ is continuous with respect to $H_{0k}(y)$ so that $H_k^*(y)$ is differentiable with respect to y. In addition, $d\hat{H}(y)/d\widetilde{H}(y)$ converges to $dH^*(y)/dH_0(y)$ uniformly in y.

Since $(\hat{\Omega}, \hat{H})$ are NPMLEs for $\ell_n(\Omega, H)$, we know $\ell_n(\hat{\Omega}, \hat{H}) - \ell_n(\Omega, H) \geq 0$ for any Ω, H. We have

$$\begin{aligned} 0 \leq\ & n^{-1}\{\ell_n(\hat{\Omega}, \hat{H}) - \ell_n(\Omega_0, \widetilde{H})\} \\ =& n^{-1} \sum_{i=1}^{n} \log \int R_{1i}(\hat{\Omega}, \hat{H}, u) f(u; \hat{\eta}) du - n^{-1} \sum_{i=1}^{n} \log \int R_{1i}(\Omega_0, \widetilde{H}, u) f(u; \eta_0) du \\ &+ n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} \log \frac{d\hat{H}_j(y_{ij})}{d\widetilde{H}_j(y_{ij})} \end{aligned} \tag{B.13}$$

As $n \to \infty$ in (B.13), we have

$$\begin{aligned} 0 &\leq \ell(\Omega^*, H^*) - \ell(\Omega_0, H_0) \\ &= \mathbb{E}\left[\log \frac{\int R_{1i}(\Omega^*, H^*, u) f(u;\eta^*) du \prod_{j=1}^m dH_j^*(y_{ij})^{\delta_{ij}}}{\int R_{1i}(\Omega_0, H_0, u) f(u;\eta_0) du \prod_{j=1}^m dH_{0j}(y_{ij})^{\delta_{ij}}}\right] \end{aligned} \tag{B.14}$$

is the negative Kullback-Leibler information. By definition, (Ω_0, H_0) maximizes $\ell(\Omega, H)$, therefore, (B.14) has a unique maximum when

$$\prod_{j=1}^m dH_j^*(y_{ij})^{\delta_{ij}} \int R_{1i}(\Omega^*, H^*, u) f(u;\eta^*) du = \prod_{j=1}^m dH_{0j}(y_{ij})^{\delta_{ij}} \int R_{1i}(\Omega_0, H_0, u) f(u;\eta_0) du \tag{B.15}$$

uniformly over (Ω, H). Under an identifiable model this means (B.14) has a unique maximum at (Ω_0, H_0). Since maximizing (B.14) is equivalent to maximizing likelihood $\ell(\Omega^*, H^*)$, and (B.14) has a unique maximum, therefore, $(\Omega_0, H_0) = \text{argmax}_{(\Omega,H)\in\mathcal{H}}\ell(\Omega, H)$ is unique. Write out (B.15)

$$\int \prod_{j=1}^m \frac{[U_i^{\alpha_j^*} \gamma_{ij}^* dH_j^*(y_{ij})]^{\delta_{ij}}}{[1 + U_i^{\alpha_j^*} \gamma_{ij}^* H_j^*(y_{ij})]^{1+\delta_{ij}}} f(u;\eta^*) du = \int \prod_{j=1}^m \frac{[U_i^{\alpha_{0j}} \gamma_{ij}^0 dH_{0j}(y_{ij})]^{\delta_{ij}}}{[1 + U_i^{\alpha_{0j}} \gamma_{ij}^0 H_{0j}(y_{ij})]^{1+\delta_{ij}}} f(u;\eta_0) du \tag{B.16}$$

We will show (B.16) implies that $\Omega^* = \Omega_0$ and $H^* = H_0$. For an integer q such that $1 \leq q \leq m$. Let $\delta_{ij} = 1, y_{ij} = 0$ in (B.16) for $j = 1, \cdots, q$. For $j = q+1, \cdots, m$ we perform the following: if $\delta_{ij} = 0$, replace y_{ij} with τ_j (upper bound of D_j); if $\delta_{ij} = 1$, we

integrate y_{ij} in its domain. Applying the above actions to (B.16), we obtain

$$\begin{aligned}
&\int \prod_{j=1}^{q} \left\{ U_i^{\alpha_j^*} \gamma_{ij}^* dH_j^*(0) \right\} \prod_{j=q+1}^{m} \left\{ \frac{U_i^{\alpha_j^*} \gamma_{ij}^* H_j^*(\tau_j)}{1 + U_i^{\alpha_j^*} \gamma_{ij}^* H_j^*(\tau_j)} \right\}^{\delta_{ij}} \left\{ \frac{1}{1 + U_i^{\alpha_j^*} \gamma_{ij}^* H_j^*(\tau_j)} \right\}^{1-\delta_{ij}} f(u; \eta^*) du \\
= &\int \prod_{j=1}^{q} \left\{ U_i^{\alpha_{0j}} \gamma_{ij}^0 dH_{0j}(0) \right\} \prod_{j=q+1}^{m} \left\{ \frac{U_i^{\alpha_{0j}} \gamma_{ij}^0 H_{0j}(\tau_j)}{1 + U_i^{\alpha_{0j}} \gamma_{ij}^0 H_{0j}(\tau_j)} \right\}^{\delta_{ij}} \left\{ \frac{1}{1 + U_i^{\alpha_{0j}} \gamma_{ij}^0 H_{0j}(\tau_j)} \right\}^{1-\delta_{ij}} f(u; \eta_0) du
\end{aligned} \tag{B.17}$$

Since $\{\delta_{ij} : j = q+1, \cdots, m\}$ are arbitrary, we sum the two sides of (B.17) over all possible combinations of $\{\delta_{ij} : j = q+1, \cdots, m\}$ and we get

$$\int \prod_{j=1}^{q} \left\{ u^{\alpha_j^*} \gamma_{ij}^* dH_j^*(0) \right\} f(u; \eta^*) du = \int \prod_{j=1}^{q} \left\{ u^{\alpha_{0j}} \gamma_{ij}^0 dH_{0j}(0) \right\} f(u; \eta_0) du$$

Thus,

$$\prod_{j=1}^{q} dH_j^*(0) \gamma_{ij}^* \int \left\{ \prod_{j=1}^{q} u^{\alpha_j^*} \right\} f(u; \eta^*) du = \prod_{j=1}^{q} dH_{0j}(0) \gamma_{ij}^0 \int \left\{ \prod_{j=1}^{q} u^{\alpha_{0j}} \right\} f(u; \eta_0) du$$

Regularity condition (1) implies that $dH_j^*(0) > 0$ for any j. Take $q = 1$, we have

$$\begin{aligned}
&dH_1^*(0) \gamma_{i1}^* \int u^{\alpha_1^*} f(u; \eta^*) du = dH_{01}(0) \gamma_{i1}^0 \int u^{\alpha_{01}} f(u; \eta_0) du \\
&\log dH_1^*(0) + Z_{i1}^T \beta^* + \log \int u^{\alpha_1^*} f(u; \eta^*) du = \log dH_{01}(0) + Z_{i1}^T \beta_0 + \log \int u^{\alpha_{01}} f(u; \eta_0) du \\
&\log \frac{dH_1^*(0)}{dH_{01}(0)} + Z_{i1}^T (\beta^* - \beta_0) + \log \frac{\int u^{\alpha_1^*} f(u; \eta^*) du}{\int u^{\alpha_{01}} f(u; \eta_0) du} = 0
\end{aligned} \tag{B.18}$$

Since outcome index q is interchangeable between outcomes, so equation (B.18) applies to any q. According to condition (4) and (5), equation (B.18) implies $\alpha^* = \alpha_0$, $\beta^* = \beta_0$, $\eta^* = \eta_0$ and $dH^*(0) = dH_0(0)$.

Next we show that $H_j^* = H_{0j}$ for all $= 1, \cdots, m$. We let $\delta_{i1} = 1$ in (B.16) and integrate y_{i1} from 0 to s. In addition, for $j = 2, \cdots, m$, if $\delta_{ij} = 0$, we replace Y_{ij} with τ_j; if $\delta_{ij} = 1$,

we integrate y_{ij} from 0 to τ_j. Then we sum the resulting equations over all possible $\{\delta_{ij} : j = 2, \cdots, m\}$ to obtain

$$\int \left\{ \frac{u^{\alpha_1^*} \gamma_{i1}^* H_1^*(s)}{1 + u^{\alpha_1^*} \gamma_{i1}^* H_1^*(s)} \right\} f(u; \eta^*) du = \int \left\{ \frac{u^{\alpha_{01}} \gamma_{i1}^0 H_{01}(s)}{1 + u^{\alpha_{01}} \gamma_{i1}^0 H_{01}(s)} \right\} f(u; \eta^0) du$$

The two sides of the above equation are strictly monotone in $H_1^*(s)$ and $H_{01}(s)$, respectively. Therefore, we have $H_1^*(s) = H_{01}(s)$. Since the outcome index is arbitrary, the above result also applies to $j = 2, \cdots, m$. We have $H^*(s) = H_0(s)$.

We conclude that $||\hat{\Omega} - \Omega_0|| \to 0$, and $||\hat{H}(s) - H_0(s)|| \to 0$ for all $s \in D$. Thus, we established uniform convergence $\sum_{k=1}^m \sup_{y_k \in D_k} |\hat{H}_k(y_k) - H_{0k}(y_k)| \to 0$.

B.4.2 Proof of Theorem III.2

Consider the set

$$\begin{aligned} \mathcal{H} = \{(v, w_1, \cdots, w_m) : v \in R^{d_\Omega}, w_k(\cdot) \text{ is a function on } D_k; \\ |v| \leq 1, ||w_k||_{BV[D_k]} \leq 1, k = 1, \cdots, m\} \end{aligned}$$

where $||w_k||_{BV[D_k]}$ denotes the total variation of $w_k(\cdot)$ in D_k, and d_Ω is the dimension of Ω. Define a sequence $S_n(\Omega, H)[v, w_1, \cdots, w_m]$ mapping a neighborhood of (Ω_0, H_0) into $l^\infty(\mathcal{H})$ as follows:

$$\begin{aligned} S_n(\Omega, H)[v, w_1, \cdots, w_m] &= \frac{d}{d\epsilon} n^{-1} \ell_n \left(\Omega + \epsilon v, H_k(y) + \epsilon \int_{-\infty}^{y} w_k(s) dH_k(s), k = 1, \cdots, m \right) \Bigg|_{\epsilon=0} \\ &= A_{n0}[v] + \sum_{k=1}^{m} A_{nk}[w_k] \end{aligned}$$

where $A_{np}, p = 0, \cdots, m$, are linear functionals on R^{d_Ω} and $BV[D_k]$, respectively. Let $\dot{\ell}_\Omega$ and $\dot{\ell}_{H_k}(w_k)$ be the score function for Ω and the score for H_k along the path $H_k(y) +$

$\epsilon \int w_k(s) dH_k(s)$, then

$$A_{n0}[v] = \mathcal{P}_n[v^T \dot{\ell}_\Omega], A_{nk}[w_k] = \mathcal{P}_n[\dot{\ell}_{H_k}(w_k)], k = 1, \cdots, m$$

where $\mathcal{P}_n$ denotes the empirical measure based on n independent subjects.

Correspondingly we define the limit map $S : (\Omega, H) \to l^\infty(\mathcal{H})$ as

$$S(\Omega, H)[v, w_1, \cdots, w_m] = A_0[v] + \sum_{k=1}^{m} A_k[w_k],$$

where the linear functionals $A_p, p = 0, \cdots, m$, are the expectation of the empirical average of $A_{np}, p = 0, \cdots, m$. By definition, $S_n(\hat{\Omega}, \hat{H}) = 0$ and $S(\Omega_0, H_0) = 0$.

Since $\mathcal{H}$ is a Donsker class and the functionals $A_{np}, p = 0, \cdots, m$, are bounded Lipschitz functionals with respect to $\mathcal{H}$, $\sqrt{n}(S_n(\Omega_0, H_0) - S(\Omega_0, H_0))$ converges to a tight Gaussian process on $l^\infty(\mathcal{H})$. The first condition in Theorem 2 of *Murphy* (1995) holds.

By regularity condition (6), the score operator $S(\Omega, H)$ is Fréchet differentiable at Ω^0 with a continuously invertible derivative $-\mathcal{I}_0$. The hessian matrix $\mathcal{I}_n$ evaluated at the true values of H and Ω is positive definite, and converges in probability to a deterministic and invertible operator $\mathcal{I}_0$. Thus the second condition in Theorem 2 of *Murphy* (1995) holds. The derivative of $S(\Omega, H)$ at (Ω_0, H_0), denoted as $-\mathcal{I}_0$, is a map from the space $(\Omega - \Omega_0, H - H_0)$ to $l^\infty(\mathcal{H})$. The fourth condition in Theorem 2 of *Murphy* (1995), the approximation condition below can be verified along the lines of appendix in *Murphy* (1995)

$$\begin{aligned}
&\sup |(S_n - S)(\hat{\Omega}, \hat{H}) - (S_n - S)(\Omega_0, H_0)| \\
&= o_p\left(n^{-1/2} \vee \left\{ ||\hat{\Omega} - \Omega_0|| + \sum_{k=1}^{m} \sup |\hat{H}_k(y) - H_{0k}(y)| \right\}\right).
\end{aligned}$$

In order to verify the third condition in Theorem 2 of *Murphy* (1995), we want to show that $-\mathcal{I}_0$, denoted $\dot{S}_0$, is continuously invertible. $\dot{S}_0$ maps $(\Omega - \Omega_0, H - H_0)$ to a

bounded functional on $\mathcal{H}$. By *Zeng et al.* (2005), we will prove the invertibility of $\dot{S}_0$ by verifying that $\dot{S}_0(\Omega - \Omega_0, H - H_0)[v, w_1, \cdots, w_m] = 0$ implies $v = 0$ and $w_k(y) = 0$ uniformly, $k = 1, \cdots, m$.

For a small constant ϵ, choose $\Omega = \Omega_0 + \epsilon v, H_k(y) = H_{0k}(y) + \epsilon \int_\infty^y w_k(s) dH_{0k}(s)$. Then,

$$0 = \dot{S}_0(\Omega - \Omega_0, H - H_0)[v, w_k, k = 1, \cdots, m] = \epsilon \mathbb{E}\left[\left(\dot{\ell_\Omega}[v] + \sum_{k=1}^{m} \dot{\ell_{H_k}}[w_k]\right)^2\right]$$

This means

$$\dot{\ell_\Omega}[v] + \sum_{k=1}^{m} \dot{\ell_{H_k}}[w_k] = 0 \tag{B.19}$$

Closely following the lines in Appendix of *Zeng et al.* (2005), we can see that the equation (B.19) entails $v = 0$ and $w_k(\cdot) = 0$ uniformly. Therefore, the derivative of the score operator at (Ω_0, H_0), denoted as $-\mathcal{I}_0$ is continuously invertible. By Theorem 2 of *Murphy* (1995), $\sqrt{n}\left\{\hat{\Omega} - \Omega, \hat{H}(\cdot) - \hat{H}_0(\cdot)\right\}$ converges weakly to a zero-mean Gaussian process. Furthermore,

$$\sqrt{n}\dot{S}_0(\Omega - \Omega_0, H - H_0)[v, w_k, k = 1, \cdots, m] = \sqrt{n}(\mathcal{P}_n - \mathcal{P})\left[v^T \dot{\ell}_\Omega + \sum_{k=1}^{m} \dot{\ell}_{H_k}[w_k]\right] + o_p(1) \tag{B.20}$$

Thus, $\hat{\Omega}$ is semiparametrically efficient since $\hat{\Omega}$ is asymptotically linear estimator for Ω_0, and its influence function belong to the space spanned by the score function (*Zeng et al.* (2005)).

B.4.3 Proof of Theorem III.3

Observe that $\sqrt{n}\dot{S}_0(\Omega - \Omega_0, H - H_0)[v, w_k, k = 1, \cdots, m]$ is the expectation of the second derivative of the normalized log-likelihood along the direction $(\hat{\Omega} - \Omega_0, \hat{H} - H_0)$

and the direction $\left(v, \int w dH_0\right)$. Therefore, $\sqrt{n}\dot{S}_0(\Omega-\Omega_0, H-H_0)[v, w_k, k=1,\cdots,m]$ can be approximated by

$$\sqrt{n}(v^T, \bar{w}^T)(\mathcal{I}_n/n)\begin{pmatrix}\hat{\Omega}-\Omega_0\\ d\hat{H}(s)-dH_0(s)\end{pmatrix},$$

where $\bar{w}$ denotes the set of vectors $\{w_k(s) : dN_{ij}(s)=1\}$. On the other hand,

$$\sqrt{n}(\mathcal{P}_n-\mathcal{P})\left[v^T\dot{\ell}_\Omega+\sum_{k=1}^m \dot{\ell}_{H_k}[w_k]\right] \to^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{1/2}\mathbf{G}$$

where G is standard multivariate Gaussian. Therefore, equation (B.20) implies that

$$\sqrt{n}(v^T, \bar{w}^T)(\mathcal{I}_n/n)\begin{pmatrix}\hat{\Omega}-\Omega_0\\ d\hat{H}(s)-dH_0(s)\end{pmatrix} \to^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{1/2}\mathbf{G}$$

$$\sqrt{n}(v^T, \bar{w}^T)\begin{pmatrix}\hat{\Omega}-\Omega_0\\ d\hat{H}(s)-dH_0(s)\end{pmatrix} \to^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{-1/2}\mathbf{G}$$

Thus, $\sqrt{n}\left\{v^T(\hat{\Omega}-\Omega_0)+\sum_{k=1}^m \int w_k d(\hat{H}_k-H_{0k})\right\}$ converges to a zero mean Gaussian process with with variance-covariance matrix $n(v^T, \bar{w}^T)\mathcal{I}_0^{-1}(v^T, \bar{w}^T)^T$.

Table B.1: Pearson correlation between age and the six pain responses

	Age
Fibromyalgia survey criteria	−0.22
Opioid use	−0.05
BPI pain severity	−0.21
BPI surgical pain	−0.29
HADS depression	−0.16
HADS anxiety	−0.24

APPENDIX C

A Semiparametric Joint Latent Trait Model for Multiple Mixed Longitudinal Continuous, Categorical Outcomes and Time-to-event Data

C.1 EM-DCA Algorithm

We use the methods of EM algorithm in *Tsodikov* (2003) and the difference of convex functions algorithm (DCA) to iteratively estimate the infinite dimensional transformation functions $(H_1(\Omega), \ldots, H_J(\Omega), H_s(\Omega))$ and $U(\Omega)$. Consider n independent subjects with J distinct longitudinal outcomes. As an example, we assume the latent trait function takes the form $U_i(t) = U(t)e^{a_i+b_it}$. The shared random variables (a_i, b_i) follow a joint distribution $f(a, b|\theta_i) = f(a|\theta_{1i}, \theta_{2i})f(b|\eta_3)$ where $f(a|\theta_{1i}, \theta_{2i})$ is a log-Gamma density function with shape $\theta_{1i} = \exp(\eta_1 Z_i)$, and rate $\theta_{2i} = \exp(\eta_2 Z_i)$. $f(b|\eta_3)$ is a log-Gamma density function with both shape and rate being η_3. The full joint likelihood of the

observed data (Y, X) and the shared latent variable $U_i(t)$ can be expressed as

$$
\begin{aligned}
&L_0(\Omega, \mathbf{H}|\mathbf{Y}, \mathbf{Z}, U(t), a_i, b_i) \prod_{i=1}^{n} f(a_i|\theta_{1i}, \theta_{2i}) f(b_i|\eta_3) \\
&= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} \left[\frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)} - \frac{1}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})} \right] \right. \\
&\left. \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right) \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a_i \theta_{1i} - \theta_{2i} e^{a_i}} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b_i \eta_3 - \eta_3 e^{b_i}} \right\}
\end{aligned} \tag{C.1}
$$

Taking the log of (C.1),we obtain the full joint log-likelihood

$$
\begin{aligned}
&\ell_0(\Omega, \mathbf{H}|\mathbf{Y}, \mathbf{Z}, U(t), a_i, b_i) + \sum_{i=1}^{n} \log f(a_i|\theta_{1i}, \theta_{2i}) f(b_i|\eta_3) \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K} \log \left(\frac{U_i(t_k)^{\alpha_j} \gamma_{ij} dH_j(y_{ijk})}{[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})]} \right) - \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] \right. \\
&\left. + \delta_i \log \left(\frac{e^{v_0 a_i + v_1 b_i} \gamma_{is} dH_s(T_i)}{1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)} \right) + \log \left(\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a_i \theta_{1i} - \theta_{2i} e^{a_i}} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b_i \eta_3 - \eta_3 e^{b_i}} \right) \right\} \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \sum_{k=1}^{K} \{ \log U_i(t_k)^{\alpha_j} \gamma_{ij} dH_j(y_{ijk}) - \log[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})] \} \right. \\
&+ \delta_i (v_0 a_i + v_1 b_i + \log \gamma_{is} dH_s(T_i)) - (1 + \delta_i) \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] + \theta_{1i} \log \theta_{2i} - \log \Gamma(\theta_{1i}) \\
&\left. + a_i \theta_{1i} - \theta_{2i} e^{a_i} + \eta_3 \log \eta_3 - \log \Gamma(\eta_3) + b_i \eta_3 - \eta_3 e^{b_i} \right\}.
\end{aligned} \tag{C.2}
$$

1. **Difference of convex algorithm (DCA)**

   DCA is a version of the Minorize-Maximization(MM) algorithm that iteratively optimizes an objective function that can be expressed as the difference of concave functions. Consider two concave functions $f(x)$ and $g(x)$. The objective function

to be maximized is $f(x) - g(x)$. By the subgradient inequality of concave function,

$$g(x) \leq g(x^*) + \nabla g(x^*)(x - x^*)$$
$$f(x) - g(x) \geq f(x) - (g(x^*) + \nabla g(x^*)(x - x^*))$$

2. **EM-DCA for $\{dH_j\}_{j=1,\cdots,J}$**
For each outcome $j = 1, \cdots, J$, notice that the objective function (C.2) is a difference between two concave functions of dH_j. We construct the surrogate function as

$$\begin{aligned}
&S(dH_j, dH_j^{(m)}) \\
&= \sum_{i=1}^{n} \Bigg\{ \sum_{j=1}^{J} \sum_{k=1}^{K} \{\log U_i(t_k)^{\alpha_j} \gamma_{ij} dH_j(y_{ijk}) - \log[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})]\} \\
&+ \delta_i(v_0 a_i + v_1 b_i + \log \gamma_{is} dH_s(T_i)) - (1 + \delta_i) \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] + \theta_{1i} \log \theta_{2i} - \log \Gamma(\theta_{1i}) \\
&+ a_i \theta_{1i} - \theta_{2i} e^{a_i} + \eta_3 \log \eta_3 - \log \Gamma(\eta_3) + b_i \eta_3 - \eta_3 e^{b_i} \Bigg\} \\
&- (dH_j(x) - dH_j^{(m)}(x)) \left[\sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\mathbb{1}(y_{ijk}^- \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk}^-)} + \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\mathbb{1}(y_{ijk} \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})} \right]
\end{aligned} \tag{C.3}$$

The surrogate function (C.3) satisfies $S(dH_j^{(m)}, dH_j^{(m)}) = \ell_0(dH_j^{(m)}) + \log f(a, b)$ and
$S(dH_j, dH_j^{(m)}) \leq \ell_0(dH_j) + \log f(a, b)$. By MM algorithm theory, the next iteration $dH_j^{(m+1)}$ that maximizes $S(dH_j, dH_j^{(m)})$ in M-step will improve the likelihood.

Instead of maximizing imputed joint log-likelihood, we employ DCA and maximize imputed surrogate function (C.3). Applying the functional derivative to the surrogate function (C.3) with respect to $dH_j(x)$, we obtain the conditional score function

for $dH_j(x)$ as

$$\begin{aligned}
\mathcal{U}_{0,dH_j(x)} &= \delta_x S(dH_j, dH_j^{(m)}) \\
&= \frac{\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x)}{dH_j(x)} - \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk}^-)} \\
&- \sum_{i=1}^n \sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) U_i(t_k)^{\alpha_j} \gamma_{ij}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})}
\end{aligned} \tag{C.4}$$

where $\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x) = \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)$ records the number of observations in the jth outcome are of value x.

- E step

  Taking expectation of (C.4), we obtain the marginal score function

$$\begin{aligned}
\mathcal{U}_{dH_j(x)} &= \mathbb{E}\left[\mathcal{U}_{0,dH_j(x)} \middle\| L_0^{(k)}\right] \\
&= \frac{\sum_{i=1}^n \sum_{k=1}^K dN_{ijk}(x)}{dH_j(x)} - \sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a_i + b_i t_k)}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk}^-)} \middle\| L_0^{(m)}\right] \\
&- \sum_{i=1}^n \mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a_i + b_i t_k)}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})} \middle\| L_0^{(m)}\right].
\end{aligned}$$

  The imputation terms $\mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk}^- \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a_i + b_i t_k)}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk}^-)} \middle\| L_0^{(m)}\right]$ and $\mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a_i + b_i t_k)}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})} \middle\| L_0^{(m)}\right]$ are obtained by multivariate Laplace approximation (*Laplace* (1986)).

  Specifically,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a_i + b_i t_k)}}{1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j^{(m)}(y_{ijk})} \middle\| L_0\right] \\
&= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\sum_{k=1}^K \frac{\mathbb{1}(y_{ijk} \geq x) \gamma_{ij} U(t_k) e^{\alpha_j(a + b t_k)}}{1 + U(t_k) e^{\alpha_j(a + b t_k)} \gamma_{ij} H_j^{(m)}(y_{ijk})}\right) L_0 \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a\theta_{1i} - \theta_{2i} e^a} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b\eta_3 - \eta_3 e^b} da db}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} L_0 \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a\theta_{1i} - \theta_{2i} e^a} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b\eta_3 - \eta_3 e^b} da db}
\end{aligned} \tag{C.5}$$

where

$$L_0 = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} \left[\frac{1}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)} - \frac{1}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})} \right] \right.$$
$$\left. \left(\frac{e^{v_0 a_i + v_1 b_i}\gamma_{is} dH_s(T_i)}{1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)} \right) \right\}$$

The numerator and the denominator of (C.5) can be approximated by multivariate Laplace's method. We obtain

$$\mathbb{E}\left[\sum_{k=1}^{K} \frac{\mathbb{1}(y_{ijk} \geq x)\gamma_{ij}U(t_k)e^{\alpha_j(a_i+b_i t_k)}}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijk})} \middle\| L_0 \right]$$
$$\approx \frac{\left(\sum_{k=1}^{K} \frac{\mathbb{1}(y_{ijk}\geq x)\gamma_{ij}U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_i t_k)}}{1+U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_i t_k)}\gamma_{ij}H_j^{(m)}(y_{ijk})} \right) \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{f(\hat{a}_i,\hat{b}_i)} 2\pi |-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}}{\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{f(\hat{a}_i,\hat{b}_i)} 2\pi |-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}}$$
$$= \sum_{k=1}^{K} \frac{\mathbb{1}(y_{ijk} \geq x)\gamma_{ij}U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_i t_k)}}{1+U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_i t_k)}\gamma_{ij}H_j^{(m)}(y_{ijk})}$$

where

$$f(a,b) = \left\{ a\theta_{1i} - \theta_{2i}e^a + b\eta_3 - \eta_3 e^b \right.$$
$$+ \sum_{j=1}^{J}\sum_{k=1}^{K} \log\left[\frac{1}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)} - \frac{1}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})} \right]$$
$$\left. + \log\left(\frac{e^{v_0 a_i+v_1 b_i}\gamma_{is}dH_s(T_i)}{1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)} \right)^{\delta_i} \left(\frac{1}{1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)} \right) \right\} \quad \text{(C.6)}$$

and $H(f)(\hat{a}_i,\hat{b}_i)$ is the hessian matrix of $f(a,b)$ evaluated at $(\hat{a}_i,\hat{b}_i)$; $|\cdot|$ denotes matrix determinant.

Here $(\hat{a}_i,\hat{b}_i)$ are assumed not at the boundary of the interval of integration and are the global maximum of $f(a,b)$ for the i^{th} subject. That is, $\nabla f(\hat{a}_i,\hat{b}_i) = 0$. Similar approximation method is applied to the term

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{\mathbb{1}(y_{ijk}^{-}\geq x)\gamma_{ij}U(t_k)e^{\alpha_j(a_i+b_it_k)}}{1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijk}^{-})}\middle\| L_0^{(m)}\right]\approx\sum_{k=1}^{K}\frac{\mathbb{1}(y_{ijk}\geq x)\gamma_{ij}U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_it_k)}}{1+U(t_k)e^{\alpha_j(\hat{a}_i+\hat{b}_it_k)}\gamma_{ij}H_j^{(m)}(y_{ijk}^{-})}.$$

- M step

  The estimator for $dH_j^{(m+1)}(x)$ that maximizes $\mathbb{E}\left(S(dH_j,dH_j^{(m)})\middle\| L_0^{(m)}\right)$ can be obtained by solving $\mathcal{U}_{dH_j(x)}=0$. The solution results in a Breslow-type estimator

$$dH_j^{(m+1)}(x)$$
$$=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}dN_{ijk}(x)}{\sum_{i=1}^{n}\mathbb{E}\left[\sum_{k=1}^{K}\frac{\mathbb{1}(y_{ijk}^{-}\geq x)\gamma_{ij}U_i^{(m)}(t_k)^{\alpha_j}}{1+U_i^{(m)}(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ij}^{-})}\middle\| L_0^{(m)}\right]+\sum_{i=1}^{n}\mathbb{E}\left[\sum_{k=1}^{K}\frac{\mathbb{1}(y_{ijk}\geq x)\gamma_{ij}U_i^{(m)}(t_k)^{\alpha_j}}{1+U_i^{(m)}(t_k)^{\alpha_j}\gamma_{ij}H_j^{(m)}(y_{ijk})}\middle\| L_0^{(m)}\right]} \tag{C.7}$$

  This constitutes a self-consistent equation that can be solved iteratively (*Tsodikov* (2003)). Note that the imputation operator $\mathbb{E}\left[\quad\cdot\quad\middle\| L_0^{(m)}\right]$ is linear and does not alter convexity properties. Hence, iterations based on (C.7) will also constitute an MM algorithm monotonically improves the likelihood.

3. **EM-DCA for $\{dH_s\}$**

   For the time-to-event outcome, the objective function (C.2) is a difference between two concave functions of dH_s. We construct the surrogate function as

$$S(dH_s,dH_s^{(m)})$$
$$=\sum_{i=1}^{n}\left\{\sum_{j=1}^{J}\sum_{k=1}^{K}\{\log U_i(t_k)^{\alpha_j}\gamma_{ij}dH_j(y_{ijk})-\log[1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^{-})][1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})]\}\right.$$
$$+\delta_i(v_0a_i+v_1b_i+\log\gamma_{is}dH_s(T_i))-(1+\delta_i)\log[1+e^{v_0a_i+v_1b_i}\gamma_{is}H_s^{(m)}(T_i)]+\theta_{1i}\log\theta_{2i}-\log\Gamma(\theta_{1i})$$
$$\left.+a_i\theta_{1i}-\theta_{2i}e^{a_i}+\eta_3\log\eta_3-\log\Gamma(\eta_3)+b_i\eta_3-\eta_3e^{b_i}\right\}$$
$$-\left(dH_s(x)-dH_s^{(m)}(x)\right)\sum_{i=1}^{n}\frac{Y_{is}(x)(1+\delta_i)e^{v_0a_i+v_1b_i}\gamma_{is}}{1+e^{v_0a_i+v_1b_i}\gamma_{is}H_s^{(m)}(T_i)} \tag{C.8}$$

   The surrogate function (C.8) satisfies $S(dH_s^{(m)},dH_s^{(m)})=\ell_0(dH_s^{(m)})+\log f(a,b)$ and

   $S(dH_s,dH_s^{(m)})\leq\ell_0(dH_s)+\log f(a,b)$. By MM algorithm theory, the next iteration

$dH_s^{(m+1)}$ that maximizes $S(dH_s, dH_s^{(m)})$ in M-step will improve the likelihood.

We employ DCA and maximize imputed surrogate function (C.8). Applying the functional derivative to the surrogate function (C.8) with respect to $dH_s(x)$, we obtain the conditional score function for $dH_s(x)$ as

$$\mathcal{U}_{0,dH_s(x)} = \delta_x S(dH_s, dH_s^{(m)}) = \frac{\sum_{i=1}^n dN_{is}(x)}{dH_s(x)} - \sum_{i=1}^n \frac{Y_{is}(x)(1+\delta_i)e^{v_0 a_i + v_1 b_i}\gamma_{is}}{1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s^{(m)}(T_i)} \tag{C.9}$$

- E step

  Taking expectation of (C.9), we obtain the marginal score function for $dH_s(x)$

  $$\mathcal{U}_{dH_s(x)} = \mathbb{E}\left[\mathcal{U}_{0,dH_s(x)} \middle\| L_0^{(m)}\right] = \frac{\sum_{i=1}^n dN_{is}(x)}{dH_s(x)} - \sum_{i=1}^n \mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0 a_i + v_1 b_i}}{1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s^{(m)}(T_i)} \middle\| L_0^{(m)}\right].$$

  We use multivariate Laplace approximation for the conditional expectation $\mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0 a_i + v_1 b_i}}{1+e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s^{(m)}(T_i)} \middle\| L_0^{(m)}\right]$. Specifically,

  $$\begin{aligned}
  &\mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0 a_i + v_1 b_i}}{1 + e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s^{(m)}(T_i)} \middle\| L_0\right] \\
  &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0 a_i + v_1 b_i}}{1+e^{v_0 a_i + v_1 b_i}\gamma_{is} H_s^{(m)}(T_i)} L_0 \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a\theta_{1i} - \theta_{2i}e^a} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b\eta_3 - \eta_3 e^b} da db}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_0 \frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})} e^{a\theta_{1i} - \theta_{2i}e^a} \frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)} e^{b\eta_3 - \eta_3 e^b} da db}
  \end{aligned} \tag{C.10}$$

  where L_0 is shown as function (2).

  Applying multivariate Laplace's approximation to the numerator and the de-

nominator of (C.10), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0a_i+v_1b_i}}{1+e^{v_0a_i+v_1b_i}\gamma_{is}H_s^{(m)}(T_i)}\middle\|L_0\right] \\
\approx&\frac{\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0\hat{a}_i+v_1\hat{b}_i}}{1+e^{v_0\hat{a}_i+v_1\hat{b}_i}\gamma_{is}H_s^{(m)}(T_i)}\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{f(\hat{a}_i,\hat{b}_i)}2\pi|-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}}{\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{f(\hat{a}_i,\hat{b}_i)}2\pi|-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}} \\
=&\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0\hat{a}_i+v_1\hat{b}_i}}{1+e^{v_0\hat{a}_i+v_1\hat{b}_i}\gamma_{is}H_s^{(m)}(T_i)}
\end{aligned}$$

where $f(a,b)$ is the function (C.6) and $\nabla f(\hat{a}_i,\hat{b}_i)=0$.

- M step

  The estimator for $dH_s^{(m+1)}(x)$ that maximizes $\mathbb{E}\left(S(dH_s,dH_s^{(m)})\middle\|L_0^{(m)}\right)$ can be obtained by solving $\mathcal{U}_{dH_s(x)}=0$. The solution results in a Breslow-type estimator

$$dH_s^{(m+1)}(x)=\frac{\sum_{i=1}^n dN_{is}(x)}{\sum_{i=1}^n\mathbb{E}\left[\frac{Y_{is}(x)(1+\delta_i)\gamma_{is}e^{v_0a_i+v_1b_i}}{1+e^{v_0a_i+v_1b_i}\gamma_{is}H_s^{(m)}(T_i)}\middle\|L_0^{(m)}\right]}$$

  which is a self-consistent equation that can be solved iteratively.

4. **EM-DCA for** $\{U\}$

   Notice that the objective function (C.2) is a difference between two concave functions

of $U(t)$. Construct the surrogate function

$$
\begin{aligned}
&S(U, U^{(m)}) \\
&= \sum_{i=1}^{n} \Bigg\{ \sum_{j=1}^{J} \sum_{k=1}^{K} \{ \log U(t_k) e^{\alpha_j (a_i + b_i t_k)} \gamma_{ij} dH_j(y_{ijk}) \\
&- \log[1 + U^{(m)}(t_k) e^{\alpha_j (a_i + b_i t_k)} \gamma_{ij} H_j(y_{ijk}^-)][1 + U^{(m)}(t_k) e^{\alpha_j (a_i + b_i t_k)} \gamma_{ij} H_j(y_{ijk})] \} \\
&+ \delta_i (v_0 a_i + v_1 b_i + \log \gamma_{is} dH_s(T_i)) - (1 + \delta_i) \log[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)] + \theta_{1i} \log \theta_{2i} - \log \Gamma(\theta_{1i}) \\
&+ a_i \theta_{1i} - \theta_{2i} e^{a_i} + \eta_3 \log \eta_3 - \log \Gamma(\eta_3) + b_i \eta_3 - \eta_3 e^{b_i} \Bigg\} \\
&- \Big(U(x) - U^{(m)}(x) \Big) \left(\sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx}^-)}{1 + U^{(m)}(x) e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx}^-)} + \right. \\
&\left. \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\mathbb{1}(H_j(y_{ijx}) \neq \infty) \alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx})}{1 + U^{(m)}(x) e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx})} \right)
\end{aligned}
\tag{C.11}
$$

The surrogate function (C.11) satisfies $S(U^{(m)}, U^{(m)}) = \ell_0(U^{(m)}) + \log f(a, b)$ and $S(U, U^{(m)}) \leq \ell_0(U) + \log f(a, b)$. By MM algorithm theory, the next iteration $U^{(m+1)}$ that maximizes $S(U, U^{(m)})$ in M-step will improve the likelihood.

We employ DCA and maximize imputed surrogate function (C.11). Applying the functional derivative to the surrogate function (C.11) with respect to $U(x)$, we obtain the conditional score function for $U(x)$ as

$$
\begin{aligned}
\mathcal{U}_{0,U(x)} = \frac{\partial \ell_0}{\partial U(x)} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{J} \alpha_j \, \mathbb{1}(H_j(y_{ijx}) \neq \infty)}{U(x)} \\
&- \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx}^-)}{1 + U^{(m)}(x) e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx}^-)} \\
&- \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{\mathbb{1}(H_j(y_{ijx}) \neq \infty) \alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx})}{1 + U^{(m)}(x) e^{\alpha_j (a_i + b_i x)} \gamma_{ij} H_j(y_{ijx})}
\end{aligned}
\tag{C.12}
$$

- E step

Taking expectation of (C.12), we obtain the marginal score function for $U(x)$:

$$\begin{aligned}\mathcal{U}_{U(x)} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\alpha_j\,\mathbb{1}(H_j(y_{ijx})\neq\infty)}{U(x)} \\ &- \sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{E}\left[\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx}^-)}\middle\| L_0^{(m)}\right] \\ &- \sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{E}\left[\frac{\mathbb{1}(H_j(y_{ijx})\neq\infty)\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx})}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx})}\middle\| L_0^{(m)}\right].\end{aligned}$$

We use multivariate Laplace approximation for the imputed terms. Specifically,

$$\begin{aligned}&\mathbb{E}\left[\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx}^-)}\middle\| L_0\right] \\ &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx}^-)}L_0\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}e^{a\theta_{1i}-\theta_{2i}e^a}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{b\eta_3-\eta_3e^b}dadb}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}L_0\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}e^{a\theta_{1i}-\theta_{2i}e^a}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{b\eta_3-\eta_3e^b}dadb}\end{aligned} \tag{C.13}$$

where L_0 is the function (2).

Applying multivariate Laplace's approximation to the numerator and the denominator of (C.13), we obtain

$$\begin{aligned}&\mathbb{E}\left[\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(a_i+b_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U_i^{(m)}(x)^{\alpha_j}\gamma_{ij}H_j(y_{ijx}^-)}\middle\| L_0\right] \\ &\approx \frac{\frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(\hat{a}_i+\hat{b}_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U^{(m)}(x)^{\alpha_j}e^{\alpha_j(\hat{a}_i+\hat{b}_ix)}\gamma_{ij}H_j(y_{ijx}^-)}\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{f(\hat{a}_i,\hat{b}_i)}2\pi|-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}}{\frac{\theta_{2i}^{\theta_{1i}}}{\Gamma(\theta_{1i})}\frac{\eta_3^{\eta_3}}{\Gamma(\eta_3)}e^{f(\hat{a}_i,\hat{b}_i)}2\pi|-H(f)(\hat{a}_i,\hat{b}_i)|^{-1/2}} \\ &= \frac{\alpha_j U^{(m)}(x)^{\alpha_j-1}e^{\alpha_j(\hat{a}_i+\hat{b}_ix)}\gamma_{ij}H_j(y_{ijx}^-)}{1+U^{(m)}(x)^{\alpha_j}e^{\alpha_j(\hat{a}_i+\hat{b}_ix)}\gamma_{ij}H_j(y_{ijx}^-)}\end{aligned}$$

where $f(a,b)$ is the function (C.6) and $\nabla f(\hat{a}_i,\hat{b}_i)=0$.

- M step

  The estimator for $U(x)$ that maximizes $\mathbb{E}\left(S(U,U^{(m)})\middle\| L_0^{(m)}\right)$ can be obtained

by solving $\mathcal{U}_{U(x)} = 0$. The solution results in a Breslow-type estimator

$$U^{(m+1)}(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\alpha_j\,\mathbb{1}(H_j(y_{ijx}) \neq \infty)}{\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{E}\left[\Theta_{1ij}^{(m)}\middle\|L_0^{(m)}\right] - \mathbb{E}\left[\Theta_{2ij}^{(m)}\middle\|L_0^{(m)}\right]}$$

where

$$\Theta_{1ij}^{(m)} = \frac{\alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j(a_i + b_i x)}\gamma_{ij} H_j^{(m)}(y_{ijx}^-)}{1 + U_i^{(m)}(x)^{\alpha_j}\gamma_{ij} H_j^{(m)}(y_{ijx}^-)}$$
$$\Theta_{2ij}^{(m)} = \frac{\mathbb{1}(H_j(y_{ijx}) \neq \infty)\alpha_j U^{(m)}(x)^{\alpha_j - 1} e^{\alpha_j(a_i + b_i x)}\gamma_{ij} H_j^{(m)}(y_{ijx})}{1 + U_i^{(m)}(x)^{\alpha_j}\gamma_{ij} H_j^{(m)}(y_{ijx})}$$

The above updating equation for $U(x)$ is a self-consistent equation that can be solved iteratively.

C.2 Asymptotic Properties

C.2.1 Proof of Theorem IV.1

To prove theorem IV.1, we show that any convergent sub-sequence of $(\hat{\Omega}_n, \hat{U}_n, \hat{\mathbf{H}}_n)$ must converge to $(\Omega_0, U_0, \mathbf{H}_0)$. Since $\hat{\Omega}_n, \hat{U}_n$ and $\hat{\mathbf{H}}_n$ belong to a compact set, we can assume that $\hat{\Omega}_n \to \Omega^*$, $\hat{U}_n(\mathbf{t}) \to U^*(\mathbf{t})$ and $\hat{\mathbf{H}}_n(\cdot)$ converges point-wise to a monotone function $\mathbf{H}^*(\cdot)$ within its domain $\mathbf{D}$. We will show that $\Omega^* = \Omega_0, U^*(\mathbf{t}) = U_0(\mathbf{t})$ and $\mathbf{H}^*(y) = \mathbf{H}_0(y)$ for all y within the domain.

The marginal loglikelihood (4.3) can be written as

$$
\begin{aligned}
&\ell_n(\Omega, \mathbf{H}|\mathbf{Y}, \mathbf{Z}) \\
&= \sum_{i=1}^{n} \log \mathbb{E}\left\{ \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{e^{\alpha_j(a_i+b_i t_k)}\gamma_{ij}}{[1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)][1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})]} \frac{(e^{v_0 a_i+v_1 b_i}\gamma_{is})^{\delta_i}}{[1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)]^{1+\delta_i}} \right\}. \\
&\prod_{j=1}^{J}\prod_{k=1}^{K} U(t_k)^{\alpha_j} dH_j(y_{ijk}) dH_s(T_i)^{\delta_i} \\
&= \sum_{i=1}^{n} \log \mathbb{E}\left\{ \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{e^{\alpha_j(a_i+b_i t_k)}\gamma_{ij}}{[1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)][1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})]} \frac{(e^{v_0 a_i+v_1 b_i}\gamma_{is})^{\delta_i}}{[1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)]^{1+\delta_i}} \right\} \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} \alpha_j \log U(t_k) + \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} \log dH_j(y_{ijk}) + \sum_{i=1}^{n} \delta_i \log dH_s(T_i).
\end{aligned}
$$

By differentiating ℓ_n with respect to $dH_s(x)$,

$$
\begin{aligned}
\frac{\partial \ell_n}{\partial dH_s(x)} &= \frac{\sum_{i=1}^{n} \delta_i \, \mathbb{1}(T_i = x)}{dH_s(x)} - \sum_{i=1}^{n} \frac{\mathbb{E}[R_{1i}(\Omega, U, H) R_{si}(x, \Omega, U, H)]}{\mathbb{E}[R_{1i}(\Omega, U, H)]} \\
&= \frac{\sum_{i=1}^{n} dN_i(x)}{dH_s(x)} - \sum_{i=1}^{n} \frac{\int\int R_{1i}(\Omega, U, H, a, b) R_{si}(x, \Omega, H, a, b) f_{ab}(a, b|\eta) da db}{\int\int R_{1i}(\Omega, H, U, a, b) f_{ab}(a, b|\eta) da db}
\end{aligned}
$$

where

$$
\begin{aligned}
&R_{1i}(\Omega, U, H, a, b) \\
&= \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{e^{\alpha_j(a_i+b_i t_k)}\gamma_{ij}}{[1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk}^-)][1+U_i(t_k)^{\alpha_j}\gamma_{ij}H_j(y_{ijk})]} \frac{(e^{v_0 a_i+v_1 b_i}\gamma_{is})^{\delta_i}}{[1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)]^{1+\delta_i}} \\
&R_{si}(x, \Omega, U, H, a, b) = \frac{(1+\delta_i)\, \mathbb{1}(T_i \geq x) e^{v_0 a_i+v_1 b_i}\gamma_{is}}{1+e^{v_0 a_i+v_1 b_i}\gamma_{is}H_s(T_i)}
\end{aligned}
$$

Setting the derivative to zero, we obtain the equation

$$
\frac{\sum_{i=1}^{n} dN_i(x)}{dH_s(x)} = \sum_{i=1}^{n} \frac{\int\int R_{1i}(\Omega, U, H, a, b) R_{si}(x, \Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}{\int\int R_{1i}(\Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}
$$

Therefore, we see that $d\hat{H}_s$ satisfies the equation

$$\frac{\sum_{i=1}^n dN_i(x)}{d\hat{H}_s(x)} = \sum_{i=1}^n \frac{\int\int R_{1i}(\hat{\Omega},\hat{U},\hat{H},a,b)R_{si}(x,\hat{\Omega},\hat{U},\hat{H},a,b)f(a,b|\hat{\eta})dadb}{\int\int R_{1i}(\hat{\Omega},\hat{U},\hat{H},a,b)f_{ab}(a,b|\hat{\eta})dadb} \tag{C.14}$$

Construct a function $\widetilde{H}_s$ by imitating $\hat{H}_s$ and we will show $\widetilde{H}_s$ uniformly converges to H_{0s}. Define $\widetilde{H}_0$ as a step function with jumps only at the T_i for which $\delta_i = 1$ and $d\widetilde{H}_s$ satisfies the equation

$$\frac{\sum_{i=1}^n dN_i(x)}{d\widetilde{H}_s(x)} = \sum_{i=1}^n \frac{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)R_{si}(x,\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f(a,b|\eta_0)dadb} \tag{C.15}$$

By definition, $\widetilde{H}_s(x) = \sum_{i=1}^n \mathbb{1}(T_i \le x)d\widetilde{H}_s(T_i)$. By Glivenko-Cantelli theorem, $\widetilde{H}_s(x)$ converges almost surely to $\mathbb{E}\{\mathbb{1}(T_i \le x)f_T(x)/\mu(T_i)\}$, where

$$\begin{aligned}
\mu(x) &= \mathbb{E}\left\{\frac{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)R_{si}(s,\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{(1+\delta_i)\,\mathbb{1}(T_i \ge x)e^{v_0a_i+v_1b_i}\gamma_{is}^0}{1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(T_i)}\middle| R_{1i}(\Omega_0,U_0,H_0,a,b)\right]\right\}, \\
f_T(x) &= \mathbb{E}[dN_i(x)] = \mathbb{E}[\delta_i\,\mathbb{1}(T_i = x)]
\end{aligned}$$

Denote $S_c(\cdot|Z)$ the survival function of censoring time C given Z.

$$\begin{aligned}
&\mathbb{E}\left\{\frac{(1+\delta_i)\,\mathbb{1}(T_i \ge x)e^{v_0a_i+v_1b_i}\gamma_{is}^0}{1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(T_i)}\middle| R_{1i}(\Omega_0,U_0,H_0,a,b)\right\} \\
=&\mathbb{E}\left\{2\int_x^\infty \frac{e^{v_0a_i+v_1b_i}\gamma_{is}^0}{1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_s(t)}\frac{e^{v_0a_i+v_1b_i}\gamma_{is}^0 h_{0s}(t)}{[1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(t)]^2}S_c(t|Z_{is})dt\middle| R_{1i}(\Omega_0,U_0,H_0,a,b)\right\} \\
-&\mathbb{E}\left\{\int_x^\infty \frac{e^{v_0a_i+v_1b_i}\gamma_{is}^0}{1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(t)}\frac{1}{1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(t)}dS_c(t|Z_{is})\middle| R_{1i}(\Omega_0,U_0,H_0,a,b)\right\} \\
=&\mathbb{E}\left\{\frac{S_c(x|Z_{is})e^{v_0a_i+v_1b_i}\gamma_{is}^0}{[1+e^{v_0a_i+v_1b_i}\gamma_{is}^0 H_{0s}(x)]^2}\middle| R_{1i}(\Omega_0,U_0,H_0,a,b)\right\}
\end{aligned}$$

where the second equality follows from integration by part. Therefore we have

$$\mathbb{E}\left\{\frac{\mathbb{1}(T_i \le x) f_T(x)}{\mu(T_i)}\right\} = \mathbb{E}\left\{\mathbb{E}\left[\int_0^x \frac{S_c(t|Z_{ik}) e^{v_0 a_i + v_1 b_i}\gamma_{is}^0 h_{0k}(t)}{\mu(t)[1+e^{v_0 a_i + v_1 b_i}\gamma_{is}^0 H_{0k}(t)]^2} dt \middle| R_{1i}(\Omega_0, U_0, H_0, a, b)\right]\right\}$$
$$= \int_0^x h_{0s}(t)dt = H_{0s}(x)$$

Therefore, $\widetilde{H}_s(x)$ converges uniformly to $H_{0s}(x)$ in $[0, \tau]$.

Next, differentiating ℓ_n with respect to $dH_j(x)$,

$$\frac{\partial \ell_n}{\partial dH_j(x)} = \frac{\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)}{dH_j(x)} - \sum_{i=1}^n \frac{\mathbb{E}[R_{1i}(\Omega, U, H) R_{2i}(x, \Omega, U, H)]}{\mathbb{E}[R_{1i}(\Omega, U, H)]}$$
$$= \frac{\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)}{dH_j(x)} - \sum_{i=1}^n \frac{\int\int R_{1i}(\Omega, U, H, a, b) R_{2i}(x, \Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}{\int\int R_{1i}(\Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}$$

where

$$R_{2i}(x, \Omega, U, H, a, b) = \sum_{k=1}^K \frac{2\,\mathbb{1}(y_{ijk} \ge x) U_i^{\alpha_j}(t_k)\gamma_{ij}}{1 + U_i^{\alpha_j}(t_k)\gamma_{ij} H_j(y_{ijk})}$$

Setting the derivative to zero, we obtain the equation

$$\frac{\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)}{dH_j(x)} = \sum_{i=1}^n \frac{\int\int R_{1i}(\Omega, U, H, a, b) R_{2i}(x, \Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}{\int\int R_{1i}(\Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}$$

Therefore, we see that $d\hat{H}_j$ satisfies the equation

$$\frac{\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}(y_{ijk} = x)}{d\hat{H}_j(x)} = \sum_{i=1}^n \frac{\int\int R_{1i}(\Omega, U, H, a, b) R_{2i}(x, \Omega, U, H, a, b) f_{ab}(a, b|\eta) da db}{\int\int R_{1i}(\Omega, U, H, a, b) f_{ab}(a, b|\eta) da db} \tag{C.16}$$

Construct a function $\widetilde{H}_j$ by imitating $\hat{H}_j$ and we will show $\widetilde{H}_j$ uniformly converges to H_{0j}. Define $\widetilde{H}_j$ as a step function with jumps at the $y_{ijk}, i = 1, \cdots, n, k = 1, \cdots, K$ and

$d\widetilde{H}_j$ satisfies the equation

$$\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(y_{ijk}=x)}{d\widetilde{H}_j(x)}=\sum_{i=1}^{n}\frac{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)R_{2i}(x,\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb} \tag{C.17}$$

By definition, $\widetilde{H}_j(x)=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(y_{ijk}\le x)d\widetilde{H}_j(y_{ijk})$. By Glivenko-Cantelli theorem, $\widetilde{H}_j(x)$ converges almost surely to $\mathbb{E}\left\{\sum_{k=1}^{K}\mathbb{1}(y_{ijk}\le x)f_{Y_j}(x)/\mu(y_{ijk})\right\}$, where

$$\begin{aligned}\mu(x)&=\mathbb{E}\left\{\frac{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)R_{2i}(x,\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb}\right\}\\&=\mathbb{E}\left\{\sum_{k=1}^{K}\mathbb{E}\left[\frac{2\,\mathbb{1}(y_{ijk}\ge x)U_i^{\alpha_{0j}}(t_k)\gamma_{ik}^0}{1+U_i^{\alpha_{0k}}(t_k)\gamma_{ik}^0H_{0k}(y_{ik})}\middle|R_{1i}(\Omega_0,U_0,H_0,a,b)\right]\right\},\\f_{Y_j}(x)&=\mathbb{E}[\mathbb{1}(y_{ijk}=x)]\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{E}\left\{\frac{2\,\mathbb{1}(y_{ijk}\ge x)U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0}{1+U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0H_{0j}(y_{ijk})}\middle|R_{1i}(\Omega_0,U_0,H_0,a,b)\right\}\\=&\mathbb{E}\left\{2\int_x^{\infty}\frac{U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0}{1+U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0H_{0j}(y)}\frac{U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0h_{0j}(y)}{[1+U_i^{\alpha_{0j}}\gamma_{ij}^0H_{0j}(y)]^2}dy\middle|R_{1i}(\Omega_0,U_0,H_0,a,b)\right\}\\=&\mathbb{E}\left\{\frac{U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0}{[1+U_i^{\alpha_{0j}}(t_k)\gamma_{ij}^0H_{0j}(x)]^2}\middle|R_{1i}(\Omega_0,U_0,H_0,a,b)\right\}\end{aligned}$$

Therefore we have

$$\begin{aligned}&\mathbb{E}\left\{\sum_{k=1}^{K}\frac{\mathbb{1}(y_{ijk}\le x)f_{Y_j}(x)}{\mu(y_{ijk})}\right\}=\mathbb{E}\left\{\sum_{k=1}^{K}\mathbb{E}\left[\int_0^{x}\frac{U_i^{\alpha_{0j}}\gamma_{ij}^0h_{0j}(y)}{\mu(y)[1+U_i^{\alpha_{0j}}\gamma_{ij}^0H_{0j}(y)]^2}dy\middle|R_{1i}(\Omega_0,U_0,H_0,a,b)\right]\right\}\\&=\int_0^{x}h_{0j}(y)dy=H_{0j}(x)\end{aligned}$$

Therefore, $\widetilde{H}_j(x)$ converges uniformly to $H_{0j}(x)$ in its specific domain D_j.

By plugging in (C.14) and (C.16) into $\ell_n(\hat{\Omega}, \hat{U}, \hat{H})$, we obtain

$$
\begin{aligned}
&\ell_n(\hat{\Omega}, \hat{U}, \hat{H}) \\
&= \sum_{i=1}^{n} \log \mathbb{E} \left\{ \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{e^{\alpha_j (a_i + b_i t_k)} \gamma_{ij}}{[1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk}^-)][1 + U_i(t_k)^{\alpha_j} \gamma_{ij} H_j(y_{ijk})]} \frac{(e^{v_0 a_i + v_1 b_i} \gamma_{is})^{\delta_i}}{[1 + e^{v_0 a_i + v_1 b_i} \gamma_{is} H_s(T_i)]^{1+\delta_i}} \right\} \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \alpha_j \log \hat{U}(t_k) + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log d\hat{H}_j(y_{ijk}) + \sum_{i=1}^{n} \delta_i \log d\hat{H}_s(T_i) \\
&= \sum_{i=1}^{n} \log \int \int R_{1i}(\hat{\Omega}, \hat{U}, \hat{H}, a, b) f_{ab}(a, b|\hat{\eta}) da db + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \hat{U}(t_k)^{\hat{\alpha}_j} \\
&\quad + \sum_{i=1}^{n} \delta_i \log \left(\sum_{l=1}^{n} dN_{ls}(T_i) \right) + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \left(\sum_{l=1}^{n} \sum_{k=1}^{K} \mathbb{1}(y_{ljk} = y_{ijk}) \right) \\
&\quad - \sum_{i=1}^{n} \delta_i \log \left(\sum_{i=1}^{n} \frac{\int \int R_{1i}(\hat{\Omega}, \hat{U}, \hat{H}, a, b) R_{si}(x, \hat{\Omega}, \hat{U}, \hat{H}, a, b) f(a, b|\hat{\eta}) da db}{\int \int R_{1i}(\hat{\Omega}, \hat{U}, \hat{H}, a, b) f_{ab}(a, b|\hat{\eta}) da db} \right) \\
&\quad - \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \left(\sum_{i=1}^{n} \frac{\int \int R_{1i}(\hat{\Omega}, \hat{U}, \hat{H}, a, b) R_{2i}(x, \hat{\Omega}, \hat{U}, \hat{H}, a, b) f_{ab}(a, b|\hat{\eta}) da db}{\int \int R_{1i}(\hat{\Omega}, \hat{U}, \hat{H}, a, b) f_{ab}(a, b|\hat{\eta}) da db} \right).
\end{aligned}
$$

Likewise, plug in (C.15) and (C.17) into $\ell_n(\Omega_0, U_0, \widetilde{H})$, we obtain

$$
\begin{aligned}
&\ell_n(\Omega_0, U_0, \widetilde{H}) \\
&= \sum_{i=1}^{n} \log \int \int R_{1i}(\Omega_0, U_0, \widetilde{H}, a, b) f_{ab}(a, b|\eta_0) da db + \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log U_0(t_k)^{\alpha_{0j}} \\
&\quad + \sum_{i=1}^{n} \delta_i \log \left(\sum_{l=1}^{n} dN_{ls}(T_i) \right) + \sum_{i=1}^{n} \sum_{j=1}^{J} \log \left(\sum_{l=1}^{n} \sum_{k=1}^{K} \mathbb{1}(y_{ljk} = y_{ijk}) \right) \\
&\quad - \sum_{i=1}^{n} \delta_i \log \left(\sum_{i=1}^{n} \frac{\int \int R_{1i}((\Omega_0, U_0, H_0, a, b) R_{si}(x, (\Omega_0, U_0, H_0, a, b, a, b) f(a, b|\eta_0) da db}{\int \int R_{1i}((\Omega_0, U_0, \widetilde{H}, a, b) f_{ab}(a, b|\eta_0) da db} \right) \\
&\quad - \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \left(\sum_{i=1}^{n} \frac{\int \int R_{1i}((\Omega_0, U_0, \widetilde{H}, a, b) R_{2i}(x, (\Omega_0, U_0, H_0, a, b) f_{ab}(a, b|\eta_0) da db}{\int \int R_{1i}((\Omega_0, U_0, H_0, a, b) f_{ab}(a, b|\eta_0) da db} \right).
\end{aligned}
$$

Define

$$R_{ssi}(s,\Omega,U,H) = \frac{\int\int R_{1i}(\Omega,U,H,a,b)R_{si}(s,\Omega,U,H,a,b)f_{ab}(a,b|\eta)dadb}{\int\int R_{1i}(\Omega,U,H,a,b)f_{ab}(a,b|\eta)dadb},$$
$$R_{3i}(s,\Omega,U,H) = \frac{\int\int R_{1i}(\Omega,U,H,a,b)R_{2i}(s,\Omega,U,H,a,b)f_{ab}(a,b|\eta)dadb}{\int\int R_{1i}(\Omega,U,H,a,b)f_{ab}(a,b|\eta)dadb}.$$

We see that $\hat{H}_j(y)$ is continuous with respect to $\widetilde{H}_k(y)$ and $\hat{H}_s(t)$ is continuous with respect to $\widetilde{H}_s(t)$. In addition,

$$\hat{H}_s(t) = \int_0^t \frac{\sum_{i=1}^n R_{ssi}(u,\Omega_0,U_0,H_0)}{\sum_{i=1}^n R_{ssi}(u,\hat{\Omega},\hat{U},\hat{H})} d\widetilde{H}_s(u)$$
$$\hat{H}_j(y) = \int^y \frac{\sum_{i=1}^n R_{3i}(s,\Omega_0,U_0,H_0)}{\sum_{i=1}^n R_{3i}(s,\hat{\Omega},\hat{U},\hat{H})} d\widetilde{H}_j(s)$$

for $j = 1,\cdots,J$.

If the jth outcome is continuous, by taking limits on both sides of the above equations, we conclude that $H_j^*(y)$ is absolutely continuous with respect to $H_{0j}(y)$ so that $H_j^*(y)$ is differentiable with respect to y. In addition, $d\hat{H}_j(y)/d\widetilde{H}_j(y)$ converges to $dH_j^*(y)/dH_{0j}(y)$ uniformly in y. Similarly, we conclude that $H_s^*(t)$ is absolutely continuous with respect to $H_{0s}(t)$ so that $H_s^*(t)$ is differentiable with respect to t. In addition, $d\hat{H}_s(t)/d\widetilde{H}_s(t)$ converges to $dH_s^*(t)/dH_{0s}(t)$ uniformly in t.

Since $(\hat{\Omega},\hat{U},\hat{H})$ are NPMLEs for $\ell_n(\Omega,U,H)$, we know $\ell_n(\hat{\Omega},\hat{U},\hat{H}) - \ell_n(\Omega,U,H) \geq 0$

for any Ω, U, H. We have

$$
\begin{aligned}
&0 \le n^{-1}\{\ell_n(\hat{\Omega},\hat{U},\hat{H}) - \ell_n(\Omega_0,U_0,\widetilde{H})\} \\
=&n^{-1}\sum_{i=1}^{n}\log\int\int R_{1i}(\hat{\Omega},\hat{U},\hat{H},a,b)f_{ab}(a,b|\hat{\eta})dadb \\
&- n^{-1}\sum_{i=1}^{n}\log\int\int R_{1i}(\Omega_0,U_0,\widetilde{H},a,b)f(a,b|\eta_0)dadb \\
&+\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\log\frac{\hat{U}(t_k)^{\hat{\alpha}_j}}{U_0(t_k)^{\alpha_{0j}}} + \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\log\frac{d\hat{H}_j(y_{ijk})}{d\widetilde{H}_j(y_{ijk})} + \sum_{i=1}^{n}\delta_i\log\frac{d\hat{H}_s(T_i)}{d\widetilde{H}_s(T_i)}. \qquad \text{(C.18)}
\end{aligned}
$$

As $n \to \infty$ in (C.18), we have

$$
\begin{aligned}
&0 \le \ell(\Omega^*,U^*,H^*) - \ell(\Omega_0,U_0,H_0) \qquad \text{(C.19)} \\
&= \mathbb{E}\left[\log\frac{\int\int R_{1i}(\Omega^*,U^*,H^*,a,b)f_{ab}(a,b|\eta^*)dadb\prod_{j=1}^{J}\prod_{k=1}^{K}U^*(t_k)^{\alpha_j^*}\prod_{j=1}^{J}\prod_{k=1}^{K}dH_j^*(y_{ijk})dH_s^*(T_i)^{\delta_i}}{\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb\prod_{j=1}^{J}\prod_{k=1}^{K}U_0(t_k)^{\alpha_{0j}}\prod_{j=1}^{J}\prod_{k=1}^{K}dH_{0j}(y_{ijk})dH_{0s}(T_i)^{\delta_i}}\right]
\end{aligned}
$$

is the negative Kullback-Leibler information. By definition, (Ω_0, U_0, H_0) maximizes $\ell(\Omega, U, H)$, therefore, (C.19) has a unique maximum when

$$
\begin{aligned}
&\int\int R_{1i}(\Omega^*,U^*,H^*,a,b)f_{ab}(a,b|\eta^*)dadb\prod_{j=1}^{J}\prod_{k=1}^{K}U^*(t_k)^{\alpha_j^*}\prod_{j=1}^{J}\prod_{k=1}^{K}dH_j^*(y_{ijk})dH_s^*(T_i)^{\delta_i} \\
=&\int\int R_{1i}(\Omega_0,U_0,H_0,a,b)f_{ab}(a,b|\eta_0)dadb\prod_{j=1}^{J}\prod_{k=1}^{K}U_0(t_k)^{\alpha_{0j}}\prod_{j=1}^{J}\prod_{k=1}^{K}dH_{0j}(y_{ijk})dH_{0s}(T_i)^{\delta_i} \qquad \text{(C.20)}
\end{aligned}
$$

uniformly over (Ω, U, H). Under an identifiable model this means (C.19) has a unique maximum at (Ω_0, U_0, H_0). Since maximizing (C.19) is equivalent to maximizing likelihood $\ell(\Omega^*, U^*, H^*)$, and (C.19) has a unique maximum, therefore, $(\Omega_0, U_0, H_0) = \text{argmax}_{(\Omega,U,H)\in\mathcal{H}}\ell(\Omega, U, H)$

is unique. Write out (C.20)

$$\int\int \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* dH_j^*(y_{ijk})}{[1+U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* H_j^*(y_{ijk})]^2} \frac{(e^{v_0^* a_i+v_1^* b_i}\gamma_{is}^* dH_s^*(T_i))^{\delta_i}}{[1+e^{v_0^* a_i+v_1^* b_i}\gamma_{is}^* H_s^*(T_i)]^{1+\delta_i}} f_{ab}(a,b|\eta^*)dadb$$
$$= \int\int \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 dH_{0j}(y_{ijk})}{[1+U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 H_{0j}(y_{ijk})]^2} \frac{(e^{v_0^0 a_i+v_1^0 b_i}\gamma_{is}^0 dH_{0s}(T_i))^{\delta_i}}{[1+e^{v_0^0 a_i+v_1^0 b_i}\gamma_{is}^0 H_{0s}(T_i)]^{1+\delta_i}} f_{ab}(a,b|\eta_0)dadb \tag{C.21}$$

We will show (C.21) implies that $\Omega^* = \Omega_0, U^* = U_0$ and $H^* = H_0$. For an integer q such that $1 \le q \le J$, let $y_{ijk} = 0$ in (C.21) for $j = 1, \cdots, q$. For $j = q+1, \cdots, J$, we integrate y_{ijk} out in its domain. For the survival outcome, let $\delta_i = 1, T_i = 0$. Applying the above actions to (C.21), we obtain

$$\begin{aligned}
&\int\int e^{v_0^* a_i+v_1^* b_i}\gamma_{is}^* dH_s^*(0) \prod_{j=1}^{q}\prod_{k=1}^{K} \left\{U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* dH_j^*(0)\right\} \\
&\quad \prod_{j=1}^{q+1}\prod_{k=1}^{K} \int_{y\in D_j} \frac{U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* dH_j^*(y)}{[1+U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* H_j^*(y)]^2} dy f_{ab}(a,b|\eta^*)dadb \\
=& \int\int e^{v_0^0 a_i+v_1^0 b_i}\gamma_{is}^0 dH_{0s}(0) \prod_{j=1}^{q}\prod_{k=1}^{K} \left\{U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 dH_{0j}(0)\right\} \\
&\quad \prod_{j=1}^{q+1}\prod_{k=1}^{K} \int_{y\in D_j} \frac{U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 dH_{0j}(y)}{[1+U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 H_{0j}(y)]^2} dy f_{ab}(a,b|\eta_0)dadb \\
&\int\int e^{v_0^* a_i+v_1^* b_i}\gamma_{is}^* dH_s^*(0) \prod_{j=1}^{q}\prod_{k=1}^{K} \left\{U^*(t_k)^{\alpha_j^*} e^{\alpha_j^*(a_i+b_i t_k)}\gamma_{ij}^* dH_j^*(0)\right\} f_{ab}(a,b|\eta^*)dadb \\
=& \int\int e^{v_0^0 a_i+v_1^0 b_i}\gamma_{is}^0 dH_{0s}(0) \prod_{j=1}^{q}\prod_{k=1}^{K} \left\{U_0(t_k)^{\alpha_{0j}} e^{\alpha_{0j}(a_i+b_i t_k)}\gamma_{ij}^0 dH_{0j}(0)\right\} f_{ab}(a,b|\eta_0)dadb
\end{aligned} \tag{C.22}$$

Thus,

$$\gamma_{is}^* dH_s^*(0) \prod_{j=1}^{q} \prod_{k=1}^{K} \left\{U^*(t_k)^{\alpha_j^*} \gamma_{ij}^* dH_j^*(0)\right\} \int\int e^{v_0^* a_i + v_1^* b_i + \sum_{j=1}^{q} \sum_{k=1}^{K} \alpha_j^*(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta^*) da db$$

$$= \gamma_{is}^0 dH_{0s}(0) \prod_{j=1}^{q} \prod_{k=1}^{K} \left\{U_0(t_k)^{\alpha_{0j}} \gamma_{ij}^0 dH_{0j}(0)\right\} \int\int e^{v_0^0 a_i + v_1^0 b_i + \sum_{j=1}^{q} \sum_{k=1}^{K} \alpha_{0j}(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta_0) da db$$

Regularity condition (1) implies that $\mathbf{dH}^*(0) > 0$. Take $q = 1$, we have

$$\gamma_{is}^* dH_s^*(0) \prod_{k=1}^{K} \left\{U^*(t_k)^{\alpha_1^*} \gamma_{i1}^* dH_1^*(0)\right\} \int\int e^{v_0^* a_i + v_1^* b_i + \sum_{k=1}^{K} \alpha_1^*(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta^*) da db$$

$$= \gamma_{is}^0 dH_{0s}(0) \prod_{k=1}^{K} \left\{U_0(t_k)^{\alpha_{01}} \gamma_{i1}^0 dH_{01}(0)\right\} \int\int e^{v_0^0 a_i + v_1^0 b_i + \sum_{k=1}^{K} \alpha_{01}(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta_0) da db$$

Take log of both sides,

$$\begin{aligned}
& Z_{is}^T \beta_s^* + \log dH_s^*(0) + \sum_{k=1}^{K} \log U^*(t_k)^{\alpha_1^*} + K Z_{i1}^T \beta_1^* + K \log dH_1^*(0) \\
& + \int\int e^{v_0^* a_i + v_1^* b_i + \sum_{k=1}^{K} \alpha_1^*(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta^*) da db \\
= & Z_{is}^T \beta_{0s} + \log dH_{0s}(0) + \sum_{k=1}^{K} \log U_0(t_k)^{\alpha_{01}} + K Z_{i1}^T \beta_{01} + K \log dH_{01}(0) \\
& + \int\int e^{v_0^0 a_i + v_1^0 b_i + \sum_{k=1}^{K} \alpha_{01}(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta_0) da db \\
& Z_{is}^T(\beta_s^* - \beta_{0s}) + \log \frac{dH_s^*(0)}{dH_{0s}(0)} + \sum_{k=1}^{K} \log \frac{U^*(t_k)^{\alpha_1^*}}{U_0(t_k)^{\alpha_{01}}} + K Z_{i1}^T(\beta_1^* - \beta_{01}) \\
& + K \log \frac{dH_1^*(0)}{dH_{01}(0)} + \log \frac{\int\int e^{v_0^* a_i + v_1^* b_i + \sum_{k=1}^{K} \alpha_1^*(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta^*) da db}{\int\int e^{v_0^0 a_i + v_1^0 b_i + \sum_{k=1}^{K} \alpha_{01}(a_i + b_i t_k)} f_{ab}(a_i, b_i|\eta_0) da db} = 0 \qquad \text{(C.23)}
\end{aligned}$$

Since outcome index q is interchangeable between outcomes, so the equation (C.23) applies to any outcome j. According to condition (4) and (5), equation (C.23) implies $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_0$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\eta}^* = \boldsymbol{\eta}_0, \mathbf{v}^* = \mathbf{v}_0, U^*(\mathbf{t}) = U_0(\mathbf{t})$ and $\mathbf{dH}^*(0) = \mathbf{dH}_0(0)$.

Next we show that $H_j^* = H_{0j}$ for all $j = 1, \cdots, J$. For $j = 1$, we integrate y_{i1k} from 0 to x in (C.21). In addition, for $j = 2, \cdots, J$, we integrate y_{ijk} out in its domain. For the

survival outcome, let $\delta_i = 1$ and we integrate T_i in $[0, \tau]$. We obtain

$$\int\int \frac{U_i^*(t_k)^{\alpha_1^*}\gamma_{i1}^* H_1^*(x)}{1+U_i^*(t_k)^{\alpha_1^*}\gamma_{i1}^* H_1^*(x)} f_{ab}(a,b|\eta^*)dadb = \int\int \frac{U_{0i}(t_k)^{\alpha_{01}}\gamma_{i1}^0 H_{01}(x)}{1+U_{0i}(t_k)^{\alpha_{01}}\gamma_{i1}^0 H_{01}(x)} f_{ab}(a,b|\eta_0)dadb$$

The two sides of the above equation are strictly monotone in $H_1^*(x)$ and $H_{01}(x)$, respectively. Therefore, we have $H_1^*(x) = H_{01}(x)$. Since the outcome index is arbitrary, the above result also applies to $j = 2, \cdots, J$. Apply the similar actions to H_s by integrating out all y_{i1k} in its domain and let $\delta_i = 1$ and we integrate T_i from 0 to t, we get

$$\int\int \frac{e^{v_0^* a + v_1^* b}\gamma_{is}^* H_s^*(t)}{1+e^{v_0^* a + v_1^* b}\gamma_{is}^* H_s^*(t)} f_{ab}(a,b|\eta^*)dadb = \int\int \frac{e^{v_0^0 a + v_1^0 b}\gamma_{is}^0 H_{0s}(t)}{1+e^{v_0^0 a + v_1^0 b}\gamma_{is}^0 H_{0s}(t)} f_{ab}(a,b|\eta_0)dadb$$

The two sides of the above equation are strictly monotone in $H_s^*(t)$ and $H_{0s}(t)$, respectively. Therefore, we have $H_s^*(t) = H_{0s}(t)$. Overall, we have $\mathbf{H}^*(x) = \mathbf{H}_0(x)$ for $x \in \mathbf{D}$.

We conclude that $||\hat{\Omega} - \Omega_0|| \to 0, ||\hat{U}(\mathbf{t}) - U_0(\mathbf{t})|| \to 0$, and $||\hat{\mathbf{H}}(x) - \mathbf{H}_0(x)|| \to 0$ for all $x \in \mathbf{D}$. Thus, we established uniform convergence $\sum_{j=1}^{J} \sup_{y_j \in D_j} |\hat{H}_j(y_j) - H_{0j}(y_j)| \to 0$ and $\sup_{t\in[0,\tau]} |\hat{H}_s(t) - H_{0s}(t)| \to 0$.

C.2.2 Proof of Theorem IV.2

Consider the set

$$\mathcal{H} = \{(v, w_1, \cdots, w_{J+1}) : v \in R^{d+K}, w_j(\cdot) \text{ is a function on } D_j, w_{J+1}(\cdot) \text{ is a function on } [0,\tau];$$
$$|v| \le 1, ||w_j||_{BV[D_j]} \le 1, j = 1, \cdots, J, ||w_{J+1}||_{BV[0,\tau]} \le 1\}$$

where $||w_j||_{BV[D_j]}$ denotes the total variation of $w_j(\cdot)$ in D_j, and d is the dimension of Ω. Define a sequence $S_n(\Omega, U, H)[v, w_1, \cdots, w_{J+1}]$ mapping a neighborhood of (Ω_0, U_0, H_0)

into $l^\infty(\mathcal{H})$ as follows:

$$
\begin{aligned}
&S_n(\Omega, U, H)[v, w_1, \cdots, w_{J+1}] \\
&= \frac{d}{d\epsilon} n^{-1} \ell_n \left((\Omega, U)^T + \epsilon v, H_j(y) + \epsilon \int_{-\infty}^{y} w_j(x) dH_j(x), j = 1, \cdots, J+1 \right) \Bigg|_{\epsilon=0} \\
&= A_{n0}[v] + \sum_{j=1}^{J+1} A_{nj}[w_j]
\end{aligned}
$$

where $A_{np}, p = 0, \cdots, J+1$, are linear functionals on R^{d+K} and $BV[\mathbf{D}]$, respectively. Let $\dot{\ell}_{\Omega,U}$ and $\dot{\ell}_{H_j}(w_j)$ be the score function for (Ω, U) and the score for H_j along the path $H_j(y) + \epsilon \int w_j(s) dH_j(s)$, then

$$
A_{n0}[v] = \mathcal{P}_n[v^T \dot{\ell}_{\Omega,U}], A_{nj}[w_j] = \mathcal{P}_n[\dot{\ell}_{H_j}(w_j)], j = 1, \cdots, J+1
$$

where $\mathcal{P}_n$ denotes the empirical measure based on n independent subjects.

Correspondingly we define the limit map $S : (\Omega, U, H) \to l^\infty(\mathcal{H})$ as

$$
S(\Omega, U, H)[v, w_1, \cdots, w_{J+1}] = A_0[v] + \sum_{j=1}^{J+1} A_j[w_j],
$$

where the linear functionals $A_p, p = 0, \cdots, J+1$, are the expectation of the empirical average of $A_{np}, p = 0, \cdots, J+1$. By definition, $S_n(\hat{\Omega}, \hat{U}, \hat{H}) = 0$ and $S(\Omega_0, U_0, H_0) = 0$.

Since $\mathcal{H}$ is a Donsker class and the functionals $A_{np}, p = 0, \cdots, J+1$, are bounded Lipschitz functionals with respect to $\mathcal{H}$, $\sqrt{n}(S_n(\Omega_0, U_0, H_0) - S(\Omega_0, U_0, H_0))$ converges to a tight Gaussian process on $l^\infty(\mathcal{H})$. The first condition in Theorem 2 of *Murphy* (1995) holds.

By regularity condition (6), the score operator $S(\Omega, U, H)$ is Fréchet differentiable at Ω_0, U_0, H_0 with a continuously invertible derivative $-\mathcal{I}_0$. The hessian matrix $\mathcal{I}_n$ evaluated at the true values of H and Ω, U is positive definite, and converges in probability to a

deterministic and invertible operator $\mathcal{I}_0$. Thus the second condition in Theorem 2 of *Murphy* (1995) holds. The derivative of $S(\Omega, U, H)$ at (Ω_0, U_0, H_0), denoted as $-\mathcal{I}_0$, is a map from the space $(\Omega - \Omega_0, U - U_0, H - H_0)$ to $l^\infty(\mathcal{H})$. The fourth condition in Theorem 2 of *Murphy* (1995), the approximation condition below can be verified along the lines of appendix in *Murphy* (1995)

$$\begin{aligned}&\sup |(S_n - S)(\hat{\Omega}, \hat{U}, \hat{H}) - (S_n - S)(\Omega_0, U_0, H_0)| \\ =&o_p\left(n^{-1/2} \vee \left\{||\hat{\Omega} - \Omega_0|| + ||\hat{U} - U_0|| + \sum_{j=1}^{J+1} \sup |\hat{H}_j(y) - H_{0j}(y)|\right\}\right).\end{aligned}$$

In order to verify the third condition in Theorem 2 of *Murphy* (1995), we want to show that $-\mathcal{I}_0$, denoted $\dot{S}_0$, is continuously invertible. $\dot{S}_0$ maps $(\Omega - \Omega_0, U - U_0, H - H_0)$ to a bounded functional on $\mathcal{H}$. By *Zeng et al.* (2005), we will prove the invertibility of $\dot{S}_0$ by verifying that $\dot{S}_0(\Omega - \Omega_0, U - U_0, H - H_0)[v, w_1, \cdots, w_{J+1}] = 0$ implies $v = 0$ and $w_j(y) = 0$ uniformly, $j = 1, \cdots, J + 1$.

For a small constant ϵ, choose $(\Omega, U)^T = (\Omega_0, U_0)^T + \epsilon v$, $H_j(y) = H_{0j}(y) + \epsilon \int_{-\infty}^{y} w_j(x) dH_{0j}(x)$. Then,

$$0 = \dot{S}_0(\Omega - \Omega_0, U - U_0, H - H_0)[v, w_j, j = 1, \cdots, J + 1] = \epsilon \mathbb{E}\left[\left(\dot{\ell}_{\Omega,U}[v] + \sum_{j=1}^{J+1} \dot{\ell_{H_j}}[w_j]\right)^2\right]$$

This means

$$\dot{\ell}_{\Omega,U}[v] + \sum_{j=1}^{J+1} \dot{\ell_{H_j}}[w_j] = 0 \tag{C.24}$$

Closely following the lines in Appendix of *Zeng et al.* (2005), we can see that the equation (C.24) entails $v = 0$ and $w_j(\cdot) = 0$ uniformly. Therefore, the derivative of the score operator at (Ω_0, U_0, H_0), denoted as $-\mathcal{I}_0$ is continuously invertible. By Theorem 2 of *Murphy* (1995), $\sqrt{n}\left\{\hat{\Omega} - \Omega, \hat{U} - U, \hat{H}(\cdot) - \hat{H}_0(\cdot)\right\}$ converges weakly to a zero-mean Gaussian

process. Furthermore,

$$\begin{aligned}&\sqrt{n}\dot{S}_0(\Omega-\Omega_0, U-U_0, H-H_0)[v, w_j, j=1,\cdots,J+1]\\&=\sqrt{n}(\mathcal{P}_n-\mathcal{P})\left[v^T\dot{\ell}_{\Omega,U}+\sum_{j=1}^{J+1}\dot{\ell}_{H_j}[w_j]\right]+o_p(1)\end{aligned}\tag{C.25}$$

Thus, $\hat{\Omega}, \hat{U}$ is semiparametrically efficient since $\hat{\Omega}, \hat{U}$ is asymptotically linear estimator for Ω_0, U_0, and its influence function belong to the space spanned by the score function (*Zeng et al.* (2005)).

C.2.3 Proof of Theorem IV.3

Observe that $\sqrt{n}\dot{S}_0(\Omega-\Omega_0, U-U_0, H-H_0)[v, w_j, j=1,\cdots,J+1]$ is the expectation of the second derivative of the normalized log-likelihood along the direction $(\hat{\Omega}-\Omega_0, \hat{U}-U_0, \hat{H}-H_0)$ and the direction $\left(v, \int w dH_0\right)$. Therefore, $\sqrt{n}\dot{S}_0(\Omega-\Omega_0, U-U_0, H-H_0)[v, w_j, j=1,\cdots,J+1]$ can be approximated by

$$\sqrt{n}(v^T, \bar{w}^T)(\mathcal{I}_n/n)\begin{pmatrix}(\hat{\Omega}-\Omega_0, \hat{U}-U_0)\\ d\hat{H}(s)-dH_0(s)\end{pmatrix},$$

where $\bar{w}$ denotes the set of vectors $\{w_j(x): dN_{ijk}(x)=1\}$. On the other hand,

$$\sqrt{n}(\mathcal{P}_n-\mathcal{P})\left[v^T\dot{\ell}_{\Omega,U}+\sum_{j=1}^{J+1}\dot{\ell}_{H_j}[w_j]\right]\to^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{1/2}\mathbf{G}$$

where $\mathbf{G}$ is standard multivariate Gaussian. Therefore, equation (C.25) implies that

$$\sqrt{n}(v^T, \bar{w}^T)(\mathcal{I}_n/n) \begin{pmatrix} (\hat{\Omega} - \Omega_0, \hat{U} - U_0) \\ d\hat{H}(s) - dH_0(s) \end{pmatrix} \rightarrow^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{1/2}\mathbf{G}$$

$$\sqrt{n}(v^T, \bar{w}^T) \begin{pmatrix} (\hat{\Omega} - \Omega_0, \hat{U} - U_0) \\ d\hat{H}(s) - dH_0(s) \end{pmatrix} \rightarrow^d (v^T, \bar{w}^T)(\mathcal{I}_n/n)^{-1/2}\mathbf{G}$$

Thus, $\sqrt{n}\left\{v^T(\hat{\Omega} - \Omega_0, \hat{U} - U_0) + \sum_{j=1}^{J+1} \int w_j d(\hat{H}_j - H_{0j})\right\}$ converges to a zero mean Gaussian process with with variance-covariance matrix $n(v^T, \bar{w}^T)\mathcal{I}_0^{-1}(v^T, \bar{w}^T)^T$.

BIBLIOGRAPHY

BIBLIOGRAPHY

An, L. T. H., and P. D. Tao (2005), The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems, *Annals of Operations Research*, *133*(1), 23–46, doi:10.1007/s10479-004-5022-1.

Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993), *Statistical models based on counting processes*, Springer.

Balka, J., A. Desmond, and P. McNicholas (2009), Review and implementation of cure models based on first hitting times for Wiener processes, *Lifetime Data Analysis*, *15*, 147–176.

Berkson, J., and R. Gage (1952), Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, *47*(259), 501–515.

Broet, P., Y. De Rycke, P. Tubert-Bitter, J. Lellouch, B. Asselain, and T. Moreau (2001), A semiparametric approach for the two-sample comparison of survival times with long-term survivors, *Biometrics*, *57*(3), 844–852.

Chen, M.-H., J. Ibrahim, and D. Sinha (1999), A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*, *94*, 909–919.

Chen, Y. (2010), Semiparametric marginal regression analysis for dependent competing risks under an assumed copula, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 235–251.

Chen, Y. (2012), Maximum likelihood analysis of semicompeting risks data with semiparametric regression models, *Lifetime Data Analysis*, *18*(1), 36–57.

Chen, Y.-H. (2009), Weighted breslow-type and maximum likelihood estimation in semiparametric transformation models, *Biometrika*, *96*(3), 591–600.

Cooner, F., S. Banerjee, and B. Carlin (2007), Flexible cure rate modeling under latent activation schemes, *Journal of the American Statistical Association*, *102*(478), 560–572.

Dunson, D. B. (2000), Bayesian latent variable models for clustered mixed outcomes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 355–366.

Dunson, D. B. (2003), Dynamic latent trait models for multidimensional longitudinal data, *Journal of the American Statistical Association*, *98*(463), 555–563.

Dunson, D. B., and A. H. Herring (2005), Bayesian latent variable models for mixed discrete outcomes, *Biostatistics*, *6*(1), 11–25.

Farewell, V. T. (1982), The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, *38*(4), 1041–1046.

Fine, J. P., and R. J. Gray (1999), A proportional hazards model for the subdistribution of a competing risk, *Journal of the American statistical association*, *94*(446), 496–509.

Fleming, T., and D. Harrington (1991), *Counting processes and survival analysis*, Wiley-Interscience.

Ghosh, P., and T. Hanson (2010), A semiparametric bayesian approach to multivariate longitudinal data, *Australian & New Zealand journal of statistics*, *52*(3), 275–288.

Gjessing, H., O. Aalen, and N. Hjort (2003), *Advances in Applied Probability*, *35*(2), 532–550.

Gruhl, J., E. A. Erosheva, P. K. Crane, et al. (2013), A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes, *The Annals of Applied Statistics*, *7*(4), 2361–2383.

Gueorguieva, R., and G. Sanacora (2006), Joint analysis of repeatedly observed continuous and ordinal measures of disease severity, *Statistics in Medicine*, *25*(8), 1307–1322.

Ha, J., and A. Tsodikov (2015), Semiparametric estimation in the proportional hazard model accounting for a misclassified cause of failure, *Biometrics*, *71*(4), 941–949.

He, B., and S. Luo (2016), Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinsons disease, *Statistical methods in medical research*, *25*(4), 1346–1358.

Hickey, G. L., P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona (2016), Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues, *BMC medical research methodology*, *16*(1), 117.

Hoff, P. D. (2007), Extending the rank likelihood for semiparametric copula estimation, *The Annals of Applied Statistics*, pp. 265–283.

Hu, C., and A. Tsodikov (2014a), Semiparametric regression analysis for time-to-event marked endpoints in cancer studies, *Biostatistics*, *15*(3), 513–525.

Hu, C., and A. Tsodikov (2014b), Joint modeling approach for semicompeting risks data with missing nonterminal event status, *Lifetime Data Analysis*, *20*(4), 563–583.

Jaffa, M. A., M. Gebregziabher, D. K. Luttrell, L. M. Luttrell, and A. A. Jaffa (2016), Multivariate generalized linear mixed models with random intercepts to analyze cardiovascular risk markers in type-1 diabetic patients, *Journal of applied statistics*, *43*(8), 1447–1464.

Kosorok, M. (2008), *Introduction to empirical processes and semiparametric inference*, Springer Verlag.

Kuk, A., and C.-H. Chen (1992), A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, *79*(3), 531–541.

Kunihama, T., C. T. Halpern, and A. H. Herring (2016), Nonparametric bayes models for mixed-scale longitudinal surveys, *arXiv preprint arXiv:1606.02381*.

Lange, K., D. R. Hunter, and I. Yang (2000), Optimization transfer using surrogate objective functions, *Journal of computational and graphical statistics*, *9*(1), 1–20.

Laplace, P. S. (1986), Memoir on the probability of the causes of events, *Statistical Science*, *1*(3), 364–378.

Lee, M.-L., and G. Whitmore (2006), Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary, *Statistical Science*, *21*(4), 501–513.

Li, C.-S., and J. M. G. Taylor (2002), A semi-parametric accelerated failure time cure model, *Statistics in Medicine*, *21*(21), 3235–3247.

Lin, H., L. Zhou, R. M. Elashoff, and Y. Li (2014), Semiparametric latent variable transformation models for multiple mixed outcomes, *Statistica Sinica*, pp. 833–854.

Lu, W. (2010), Efficient estimation for an accelerated failure time model with a cure fraction, *Statistica Sinica*, *20*(2), 661–674.

Moustaki, I., and M. Knott (2000), Generalized latent trait models, *Psychometrika*, *65*(3), 391–411.

Murphy, S. A. (1995), Asymptotic theory for the frailty model, *The Annals of Statistics*, *23*(1), 182–198.

Murphy, S. A., and A. W. Van der Vaart (2000), On profile likelihood, *Journal of the American Statistical Association*, *95*(450), 449–465.

Murray, J. S., D. B. Dunson, L. Carin, and J. E. Lucas (2013), Bayesian gaussian copula factor models for mixed data, *Journal of the American Statistical Association*, *108*(502), 656–665.

Muthén, B. (1984), A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika*, *49*(1), 115–132, doi: 10.1007/BF02294210.

Othus, M., Y. Li, and R. Tiwari (2009), A class of semiparametric mixture cure survival models with dependent censoring, *Journal of the American Statistical Association*, *104*(487), 1241–1250.

Othus, M., Y. Li, and R. Tiwari (2012), Change point-cure models with application to estimating the change-point effect of age of diagnosis among prostate cancer patients, *Journal of Applied Statistics*, *39*(4).

Peng, Y. (2003), Estimating baseline distribution in proportional hazards cure models, *Computational Statistics and Data Analysis*, *42*(12), 187 – 201.

Peng, Y., and K. B. G. Dear (2000), A nonparametric mixture model for cure rate estimation, *Biometrics*, *56*(1), 237–243.

Proust-Lima, C., H. Amieva, and H. Jacqmin-Gadda (2013), Analysis of multivariate mixed longitudinal data: a flexible latent process approach, *British Journal of Mathematical and Statistical Psychology*, *66*(3), 470–487.

Proust-Lima, C., J.-F. Dartigues, and H. Jacqmin-Gadda (2016), Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach, *Statistics in medicine*, *35*(3), 382–398.

Putter, H., and H. van Houwelingen (2015), Dynamic frailty models based on compound birth-death processes, *Biostatistics*, *16*(3), 550–564.

Rice, J. D., and A. Tsodikov (2017), Semiparametric time-to-event modeling in the presence of a latent progression event, *Biometrics*, *73*(2), 463–472.

Rizopoulos, D. (2012), *Joint models for longitudinal and time-to-event data: With applications in R*, CRC Press.

Rizopoulos, D., and P. Ghosh (2011), A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event, *Statistics in medicine*, *30*(12), 1366–1380.

Sammel, M. D., L. M. Ryan, and J. M. Legler (1997), Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(3), 667–678.

Shi, J.-Q., and S.-Y. Lee (2000), Latent variable models with mixed continuous and polytomous data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 77–87, doi:10.1111/1467-9868.00220.

Skrondal, A., and S. Rabe-Hesketh (2004), *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, Crc Press.

Snavely, A. C., D. P. Harrington, and Y. Li (2014), A latent variable transformation model approach for exploring dysphagia, *Statistics in medicine*, *33*(25), 4337–4352.

Sy, J., and J. Taylor (2000), Estimation in a cox proportional hazards cure model, *Biometrics*, *56*(1), 227–236.

Tao, P. D., and L. T. H. An (1997), Convex analysis approach to dc programming: theory, algorithms and applications, *Acta Mathematica Vietnamica*, *22*(1), 289–355.

Tao, P. D., and L. T. H. An (1998), A dc optimization algorithm for solving the trust-region subproblem, *SIAM Journal on Optimization*, *8*(2), 476–505.

Taylor, J. (1995), Semi-parametric estimation in failure time mixture models, *Biometrics*, *51*(3), 899–907.

Tsodikov, A. (1998), A proportional hazards model taking account of long-term survivors, *Biometrics*, *54*(4), 1508–1516.

Tsodikov, A. (2002), Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in utah by age and stage, *Statistics in Medicine*, *21*(6), 895–920.

Tsodikov, A. (2003), Semiparametric models: a generalized self-consistency approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(3), 759–774.

Tsodikov, A., and S. Chefo (2008), Generalized self-consistency: Multinomial logit model and poisson likelihood, *Journal of statistical planning and inference*, *138*(8), 2380–2397.

Tsodikov, A., J. G. Ibrahim, and A. Yakovlev (2003), Estimating cure rates from survival data: An alternative to two-component mixture models, *Journal of the American Statistical Association*, *98*(464), 1063–1078.

Van Der Vaart, A., and J. Wellner (1996), *Weak convergence and empirical processes*, Springer Verlag.

Wang, L., P. Du, and H. Liang (2012), Two-component mixture cure rate model with spline estimated nonparametric components, *Biometrics*, *68*(3), 726–735.

Yakovlev, A., and A. Tsodikov (1996), *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific Publ.:Singapore.

Yashin, A., and K. Manton (1997), Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies, *Statistical Science*, *12*, 20–34.

Yin, G., and J. G. Ibrahim (2005), Cure rate models: A unified approach, *Canadian Journal of Statistics*, *33*(4), 559–570.

Zeng, D., and D. Lin (2007), Maximum likelihood estimation in semiparametric regression models with censored data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 507–564.

Zeng, D., and D. Lin (2010), A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data, *Statistica Sinica*, *20*(2), 871.

Zeng, D., D. Y. Lin, and G. Yin (2005), Maximum likelihood estimation for the proportional odds model with random effects, *Journal of the American Statistical Association*, *100*(470), 470–483, doi:10.1198/016214504000001420.

Zhang, J., Y. Peng, and H. Li (2013), A new semiparametric estimation method for accelerated hazards mixture cure model, *Computational Statistics and Data Analysis*, *59*, 95 – 102.