# SEMIPARAMETRIC REGRESSION OF RIGHT- AND INTERVAL-CENSORED COMPETING RISKS DATA

Lu Mao

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Danyu Lin

Donglin Zeng

Michael G. Hudgens

Michael R. Kosorok

Christine U. Oramasionwu

**ABSTRACT**

Lu Mao: Semiparametric Regression of Right- and Interval-Censored Competing Risks Data
(Under the direction of Danyu Lin)

In clinical and epidemiological studies, competing risks data arise when the subject can experience one, and only one, of several mutually exclusive types of events. Competing risks data are often right- or interval-censored. For right-censored data, a semiparametric regression model proposed by Fine and Gray (1999) has become the method of choice for formulating the effects of covariates on the cumulative incidence. Its estimation, however, requires modeling of the censoring distribution and is not statistically efficient. In this project, we present a broad class of semiparametric transformation models which extends the Fine and Gray model, and we derive the nonparametric maximum likelihood estimators (NPMLEs). We develop a simple and fast algorithm for computing the NPMLEs through the profile likelihood. We establish the consistency, asymptotic normality, and semiparametric efficiency of the NPMLEs. In addition, we construct graphical and numerical procedures to evaluate and select models. Then, we demonstrate the advantages of the proposed methods over the existing ones through extensive simulation studies and an application to a major study on bone marrow transplantation.

We extend the same class of transformation models to interval-censored competing risks data. We allow covariates to be time-dependent and accommodate missing event type information. We develop a novel EM algorithm to compute the NPMLEs, and establish the consistency, asymptotic normality, and semiparametric efficiency of the NPMLEs. Extensive numerical studies show that our methods perform well in finite samples. A well-known HIV/AIDS study is provided to illustrate our methods.

Finally, we consider two problems which can be viewed as extensions of the methodologies described. One is for partly interval-censored competing risks data, where some of the risks are interval censored while the rest are right censored. The other concerns interval-censored failure time with a continuous mark. We describe semiparametric regression methods for these data types and analyze data from HIV/AIDS studies using the proposed procedures.

KEY WORDS: EM algorithm; Cumulative incidence; Continuous mark; Interval censoring; Nonparametric maximum likelihood estimation; Time-dependent covariates; Transformation models.

## ACKNOWLEDGMENTS

My principal thanks are due to my advisers Profs. Danyu Lin and Donglin Zeng. The genesis of this dissertation owes much to Prof. Lin's grant proposals on innovative statistical methods for cancer and HIV/AIDS research. Intriguing as the problems appear, their solution entails numerous challenges, many of which are unseen in the literature. Throughout my years as a doctoral student, Profs. Lin and Zeng have lent invaluable assistance to my efforts in solving these problems. Our weekly meetings to discuss the specifics of my work have served as a major source of improvement on my scholastic qualifications as well on the thesis itself. Their mentoring has been both instrumental in problem-solving and inspirational in outlook. In addition, they helped revise the manuscript time and again with minute detail until utter clarity was achieved. I feel truely blessed with their guidance in my crucial years of becoming an independent researcher in biostatistics.

I'd also like to thank my dissertation committee members Profs. Michael G. Hudgens, Michael R. Kosorok, and Christine U. Oramasionwu for reading the manuscript and offering valuable comments and suggestions which have led to significant improvements of the thesis. Furthermore, I am indebted to Prof. Kosorok by benefiting tremendously from the two courses he taught: *Advanced Probability and Statistical Inference* and *Empirical Processes and Semiparametric Inference*. These courses provided the theoretical foundations and requisite tools to tackle the technical problems in this dissertation. In addition, I owe my thanks to Profs. Hudgens and Oramasionwu for providing opportunities to participate in various quantitative studies in HIV/AIDS. Such hands-on experience with real data has allowed me to look at the statistical problems treated in this thesis through the lens of a real-world investigator instead of as mere mathematical constructs to play with.

I feel that my acknowledgments would be unforgivably incomplete if I left out entirely the many scholars whom I have never met but whose work has informed my research beyond measure. Among them I could not overstate the influence of Prof. Aad van der Vaart of Leiden University in the Netherlands, whose magnum opus *Asymptotic Statistics* has become an intimate companion of mine ever since I became a graduate student in Biostatistics. I also derive a great deal of knowledge, skills, and inspirations from the books, lecture notes, and papers of Prof. Jon Wellner of University of Washington and Prof. Sara van de Geer of ETH Zurich in Switzerland. Collectively their works have helped me make this thesis a lot more sound than it otherwise would have been.

Last but certainly not the least, I owe my deepest gratitude to my parents. The tenacity and perseverance much needed for finishing a Ph.D. degree would have eluded me if it weren't for their unswerving love and support. To them I dedicate this crude work of mine.

# TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

In this chapter, we introduce the concepts and ideas that will play a key role in the subsequent development of the thesis.

## 1.1 Competing Risks

Competing risks data arise when each study subject can experience one and only one of several distinct types of events or failures. A classical example of competing risks is death from different causes. In addition, patients who undergo an invasive surgical procedure to treat a particular disease, such as bone marrow transplantation for the treatment of leukemia (Kalbfleisch and Prentice (2002), chapter 8), may experience relapse of that disease or death related to the surgical procedure itself. Another example of competing risks is infection with a pathogen such as HIV-1 (Hudgens et al., 2001), whereby infection with one viral subtype precludes infection with other subtypes.

The competing risks data consist of $(T, D)$, where $T$ denotes the event time and $D = 1, \cdots, K$ indicates the cause of failure. Competing risks may be analyzed through the cause-specific hazard or cumulative incidence function. The cause-specific hazard of the $k$th risk $\Lambda_k^C$ is defined by

$$d\Lambda_k^C(t) = P(t \leq T < t + dt, D = k | T \geq t),$$

i.e., the instantaneous rate of failure from a specific cause at a particular time given that the subject has not experienced a failure of any cause up until that time. In contrast, the cumulative incidence, or sub-distribution function, $F_k$ is defined as

$$F_k(t) = P(T \leq t, D = k),$$

i.e., the marginal (unconditional) probability of occurrence of a specific failure type over a certain time period. The cumulative incidence is deemed more relevant, because the marginal rate is more interpretable and characterizes the subject's ultimate clinical experience (Gray, 1988).

Statistical methods have been developed to make inference about the cumulative incidence. Gray (1988) proposed a nonparametric log-rank-type test for comparing the cumulative incidence functions of a particular failure type among different groups. Specifically, a "sub-distribution hazards" function is defined by $\Lambda_k(t) = \log\{1 - F_k(t)\}$, which is the cumulative hazard function of the (improper) failure time from the $k$th cause in the presence of other risks, namely, $T_k^* \equiv TI(D = k) + \infty I(D \neq k)$, where $I(\cdot)$ is the indicator function. Because of the direct relationship between the cumulative incidence and the sub-distribution hazard function, inference on one quantity can be readily translated to that on the other. To make inference on the sub-distribution function, and hence on the cumulative incidence, a nonparametric estimator for $\Lambda_k$ is given by

$$\widehat{\Lambda}_k(t) = \int_0^t \widehat{S}(u) d\widehat{\Lambda}_k^C(u),$$

where $\widehat{S}(u)$ is the Kaplan-Meier estimator of the overall survival function $P(T > u)$, and $\widehat{\Lambda}_k^C(u)$ is the Nelsen-Aalen estimator of the cause-specific hazard (by treating the other risks as censoring). Then, a log-rank-type test statistic can be constructed by

$$\int_0^\infty W_n(u) d\{\widehat{\Lambda}_{k1}(u) - \widehat{\Lambda}_{k2}(u)\},$$

where $W_n(\cdot)$ is a data-dependent weight function, and $\widehat{\Lambda}_{kj}$ is the nonparametric estimator . The asymptotic distribution of the test statistic is derived by the (functional) delta method.

In the regression setting, Fine and Gray (1999) proposed a semiparametric proportional sub-distribution hazards model for a competing risk. Let $\boldsymbol{Z}(\cdot)$ be the set of (possibly time-dependent external) covariates. They proposed the following model for the conditional sub-distribution hazard function

$$d\Lambda_k(t|\boldsymbol{Z}) = e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)} d\Lambda_k(t),$$

where $\boldsymbol{\beta}_k$ is the regression parameter and $\Lambda_k$ is an arbitrary increasing function satisfying $\Lambda_k(0) = 0$. The authors adapted the familiar partial likelihood estimating equation of the Cox model, and used the inverse probability weighting technique (Robins and Rotnitzky, 1992) to estimate the censoring distribution in order to exclude the censoring effect of the other risks.

## 1.2 Interval Censoring

Interval-censored data arise when the event or failure of interest is not observed at an exact time but is rather known to occur within a time interval. Such data are commonly encountered in disease research, where the ascertainment of an asymptomatic event is costly or invasive and thus can only take place at a small number of monitoring times. For example, in HIV/AIDS studies, blood samples are drawn from at-risk subjects periodically for evidence of HIV sero-conversion. Likewise, biopsies are performed on patients to determine the occurrence or recurrence of cancer.

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

The fact that the failure time is never observed exactly poses tremendous theoretical and computational challenges in semiparametric regression analysis of interval-censored data. Huang

[1] Another direction of generalizing the "case-$k$" scenario lies in the so called independent inspection process (IIP) consider by Lawless (2011, section 2.3.1) which allows future monitoring times to be dependent on all previous data. Treatment of the IIP censoring mechanism is beyond the scope of this thesis.

(1995; 1996) and Huang and Wellner (1997) studied nonparametric maximum likelihood estimation for the proportional hazards and proportional odds models with case-1 and case-2 data. The estimators are obtained by the iterative convex minorant algorithm, which may become unstable for large datasets. Sieve maximum likelihood estimation for the proportional odds model was considered by Rossini & Tsiatis (1996) with case-1 data, and by Huang and Rossini (1997) and Shen (1998) with case-2 data; however, it is difficult to choose an appropriate sieve parameter space, especially the number of knots. For the proportional odds model with case-1 and case-2 data, Rabinowitz et al. (2000) derived an approximate conditional likelihood, which does not perform well in small samples. Gu et al. (2005), Zhang et al. (2005), Sun and Sun (2005), and Zhang and Zhao (2013) constructed rank-based estimators for linear transformation models; such estimators are computationally demanding and statistically inefficient. None of the existing work accommodates time-dependent covariates or handles case-$k$ or mixed-case interval censoring.

### 1.3 Semiparametric Transformation Models

In the univariate setting, let $T$ denote the failure time, and let $\boldsymbol{Z}(\cdot)$ denote the set of potentially time-dependent covariates. Under the semiparametric transformation model, the cumulative hazard function for $T$ conditional on $\boldsymbol{Z}(\cdot)$ takes the form

$$\Lambda(t; \boldsymbol{Z}) = G\left\{\int_0^t e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}(s)} d\Lambda(s)\right\}, \tag{1.1}$$

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

$$G(x) = -\log \int_0^\infty e^{-xt}\phi(t)dt,$$

where $\phi(t)$ is the density function of a frailty with support $[0, \infty)$. The choice of the gamma density with mean 1 and variance $r$ for $\phi(t)$ yields the class of logarithmic transformations: $G(x) = r^{-1}\log(1 + rx)$ $(r \geq 0)$; the choice of the positive stable distribution with parameter $\rho < 1$ yields the class of Box-Cox transformations: $G(x) = \{(1 + x)^{\rho} - 1\}/\rho$. When covariates are all time-independent, model (1.1) can be rewritten as a linear transformation model

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

where $\epsilon$ is an error term with distribution function $1 - \exp\{-G(e^{x})\}$ (Chen et al., 2002). Thus, $\boldsymbol{\beta}$ can be interpreted as the effects of covariates on a transformation of $T$. Semiparametric transformation models have not been considered in the competing risks setting.

# CHAPTER 2: RIGHT-CENSORED COMPETING RISKS

## 2.1 Introduction

Competing risks data arise when each study subject can experience one and only one of several distinct types of events or failures. A classical example of competing risks is death from different causes. In addition, patients who undergo an invasive surgical procedure to treat a particular disease, such as bone marrow transplantation for the treatment of leukemia (Kalbfleisch and Prentice (2002), chapter 8), may experience relapse of that disease or death related to the surgical procedure itself. Another example of competing risks is infection with a pathogen such as HIV-1 (Hudgens et al., 2001), whereby infection with one viral subtype precludes infection with other subtypes.

Competing risks data may be analyzed through the cause-specific hazard or cumulative incidence function. The cause-specific hazard function is the instantaneous rate of failure from a specific cause at a particular time given that the subject has not experienced a failure of any cause up until that point, and the cumulative incidence function measures the probability of occurrence of a specific failure type over a certain time period (Kalbfleisch and Prentice (2002), chapter 8). These two approaches are complementary: the cause-specific hazard is an instantaneous risk function whereas the cumulative incidence characterizes the subject's ultimate clinical experience. Standard survival analysis methods, such as the log-rank test and proportional hazards regression, can be applied to the cause-specific hazard function. The way in which covariates affect the cause-specific hazards may not coincide with the way in which they affect the cumulative incidence (Andersen et al., 2012).

Statistical methods have been developed to make inference about the cumulative incidence. Gray (1988) proposed a nonparametric log-rank-type test for comparing the cumulative incidence functions of a particular failure type among different groups. In the regression setting, Fine and Gray (1999) proposed a semiparametric proportional hazards model for the sub-distribution of a competing risk. This model has become the method of choice with well over 4,000 citations (as of May 2016) and been incorporated into the statistical guidelines of the European Group for Blood and Marrow Transplantation (Iacobelli, 2013).

The Fine and Gray methodology has important limitations, however. First, it requires the modelling of the censoring distribution and may yield invalid inference if the censoring distribution is mis-modeled. Second, the estimation is based on the inverse probability of censoring weighting, such that the estimators are statistically inefficient and numerically unstable. Third, the model is restricted to the proportional sub-distribution hazards structure, which may not hold in practice, and there are no model-checking tools. Fourth, the cause of failure needs to be known for every subject. Finally, joint inference on multiple risks is not provided.

Jeong and Fine (2006; 2007) proposed parametric regression models for the cumulative incidence function and derived maximum likelihood estimators. Their approach does not model the censoring distribution. However, it is difficult to parametrize failure time distributions, especially when there are multiple failure types. Incorrect parametrization can lead to erroneous inference.

In this paper, we develop semiparametric regression methods that avoid the aforementioned limitations of the existing methods. Specifically, we formulate the effects of covariates on the cumulative incidence function using a flexible class of semiparametric transformation models, which encompasses both proportional and non-proportional sub-distribution hazards structures. We allow the cause of failure information to be partially missing. We derive efficient estimators for the proposed models through the NPMLE approach, which does not involve modelling the censoring distribution. We construct simple and fast numerical algorithms based on the profile likelihood (Murphy and van der Vaart, 2000) to obtain the estimators. We establish the asymptotic properties of the estimators through modern empirical process theory (van der Vaart and Wellner, 1996) and semiparametric efficiency theory (Bickel et al., 1993). Our approach allows for joint inference on multiple risks, which is desirable because an increase in the incidence of one risk decreases the incidence of other risks. We also develop numerical and graphical procedures to evaluate and select models.

There is some literature on the NPMLEs for semiparametric models with censored data (e.g., Murphy et al., 1997; Kosorok et al., 2004; Scheike and Martinussen, 2004; Zeng and Lin, 2006; 2007). Our setting is unique in that the regression parameters and infinite-dimensional cumulative hazard functions are all intertwined due to the the constraint that the sum of the cumulative incidence functions must not exceed one. Because of this constraint, existing asymptotic arguments, such as the general theory of Zeng and Lin (2007), do not directly apply. A further complication arises from the missing information on the cause of failure. To tackle these challenges, we use novel techniques to prove the asymptotic properties, especially the consistency. In addition, we develop novel numerical algorithms through the profile likelihood so as to avoid direct maximization over high-dimensional parameters. Finally, we extend the martingale residuals for traditional survival data to competing risks data and study the theoretical properties of the cumulative sums of residuals so as to provide objective model-checking procedures.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the models, describe the estimation procedure, and present the asymptotic results. We also define appropriate residuals and use the cumulative sums of residuals to develop model-checking techniques. In Section 2.3, we conduct simulation studies to assess the performance of the proposed methods in finite samples and to make comparisons with existing methods. We provide an application to a major bone marrow transplantation study in Section 2.4. We make some concluding remarks in Section 2.5. Technical details and additional tables and figures are relegated to the Technical Details in Section 2.6.

## 2.2 Methods

### 2.2.1 Model Specification

We are interested in estimating the effects of a set of covariates $\boldsymbol{Z}$ on a failure time $T$ with $K$ competing causes. We characterize the regression effects through the conditional cumulative incidence functions

$$F_k(t;\boldsymbol{Z}) = \Pr(T \le t, D = k|\boldsymbol{Z}), \quad k = 1,\ldots,K,$$

where $D$ indicates the cause of failure. We formulate each $F_k$ through a class of linear transformation models:

$$g_k\{F_k(t;\boldsymbol{Z})\} = Q_k(t) + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}, \tag{2.2}$$

where $g_k$ is a known increasing function, $Q_k(\cdot)$ is an arbitrary increasing function, and $\boldsymbol{\beta}_k$ is a set of regression parameters.

To allow time-dependent covariates, we consider the conditional hazard function

$$\lambda_k(t;\boldsymbol{Z}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\Pr\{t \le T < t + \Delta t, D = k|T \ge t \cup (T \le t \cap D \ne k), \boldsymbol{Z}\},$$

which pertains to the hazard function of the improper random variable $T_k^* \equiv I(D = k)T + I(D \ne k)\infty$, where $I(\cdot)$ is the indicator function, and $\boldsymbol{Z}$ consists of time-dependent external covariates (Kalbfleisch and Prentice (2002), chapter 6). Clearly, $F_k(t;\boldsymbol{Z}) = 1 - \exp\{-\Lambda_k(t;\boldsymbol{Z})\}$, where $\Lambda_k(t;\boldsymbol{Z}) = \int_0^t \lambda_k(s;\boldsymbol{Z})ds$. We specify that

$$\Lambda_k(t;\boldsymbol{Z}) = G_k\left\{\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)} d\Lambda_k(s)\right\}, \tag{2.3}$$

where $G_k$ is a known increasing function, and $\Lambda_k(\cdot)$ is an arbitrary increasing function. The choices of $G_k(x) = x$ and $G_k(x) = \log(1 + x)$ yield the proportional sub-distribution hazards model (Cox,

1972) and the proportional odds model (Bennett, 1983), respectively. If the covariates are all time-independent, then equation (2.3) can be expressed in the form of (2.2).

### 2.2.2 Parameter Estimation

Suppose that $T$ is subject to right censoring by $C$. Then we observe $\widetilde{T}$ and $\widetilde{D}$ instead of $T$ and $D$, where $\widetilde{T} = \min(T, C)$, and $\widetilde{D} = I(T \le C)D$. Let $\xi$ denote the missing failure cause indicator, that is, $\xi = 1$ if the failure cause is observed, and $\xi = 0$ if otherwise. For better notation, we define $\xi$ to be 1 if $\widetilde{D} = 0$. For a random sample of size $n$, the data consist of $(\widetilde{T}_i, \xi_i, \xi_i\widetilde{D}_i, \boldsymbol{Z}_i)$ $(i = 1, \ldots, n)$. Assuming that $\widetilde{D}$ is missing at random (MAR), the likelihood function for $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Lambda} \equiv (\Lambda_1, \cdots, \Lambda_K)$ takes the form

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \prod_{i=1}^{n} \left[ \prod_{k=1}^{K} F_k'(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})^{I(\xi_i \widetilde{D}_i = k)} \left\{ \sum_{k=1}^{K} F_k'(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \right\}^{I(\xi_i = 0)} \times \left\{ 1 - \sum_{k=1}^{K} F_k(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \right\}^{I(\widetilde{D}_i = 0)} \right], \tag{2.4}$$

where $F_k(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda})$ denotes the conditional cumulative incidence function under model (2.3). Here and in the sequel, $f'(t) = df(t)/dt$ for any function $f$.

To obtain the NPMLEs, we treat $\Lambda_k$ as a right-continuous step function with jump size $\Lambda_k\{t\}$ at time $t$. Then the calculation of the NPMLEs is tantamount to maximizing (2.4) with respect to $\boldsymbol{\beta}$ and $\Lambda_k\{\widetilde{T}_i\}$ for $\widetilde{D}_i = k$ $(k = 1, \ldots, K)$. The maximization can be implemented through optimization algorithms, as described by Zeng and Lin (2007).

For time-independent covariates, we propose an explicit algorithm to compute the NPMLEs. First, we treat the special case with fully observed $\widetilde{D}_i$, i.e., $\xi_i = 1$ for $i = 1, \cdots, n$. Then, the general case will follow with a straightforward EM algorithm treating $\widetilde{D}_i$ as missing data. Let $t_{k1} < \cdots < t_{km_k}$ be the distinct failure times of cause $k$. Denote $d_{kj} = \Lambda_k\{t_{kj}\}$ for $j = 1, \cdots m_k$ and $k = 1, \cdots, K$. Let $t_1 < \cdots < t_m$ be the distinct failure times regardless of cause with $t_{m+1} = \infty$, and let $\delta_1, \cdots, \delta_m$ and $d_1, \cdots, d_m$ be the corresponding causes and jump sizes, respectively.

Write $\Lambda_{kj} = \sum_{l=1}^{j} I(\delta_l = k) d_l$ and $\Lambda_{(kj)} = \sum_{l=1}^{j} d_{kl}$, which pertain to $\Lambda_k$ at times $t_j$ and $t_{kj}$, respectively. With fully observed failure cause, the log-likelihood can be written as

$$\begin{aligned} l_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = & \sum_{k=1}^{K} \sum_{j=1}^{m_k} \log d_{kj} + \sum_{k=1}^{K} \sum_{j=1}^{m_k} H_k\big(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_{(kj)}} \Lambda_{(kj)}\big) \\ & + \sum_{j=1}^{m} \sum_{t_j \le \widetilde{T}_i < t_{j+1}} I(\widetilde{D}_i = 0) \log \left[ \sum_{k=1}^{K} \exp\big\{ - G_k(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i} \Lambda_{kj}) \big\} - K + 1 \right] \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} I(\widetilde{D}_i = k) \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i, \end{aligned} \tag{2.5}$$

where $H_k = \log G_k' - G_k$, and $\boldsymbol{Z}_{(kj)}$ denotes the covariate vector for the subject having the $k$th cause of failure at time $t_{kj}$.

We construct the profile likelihood (Murphy and van der Vaart, 2000) of $\boldsymbol{\beta}$ by "profiling out" $\boldsymbol{\Lambda}$. This task is complicated by the fact that $\partial l_n(\boldsymbol{\beta}, \boldsymbol{\Lambda})/\partial \boldsymbol{d} = \mathbf{0}$ is a system of nonlinear equations of $\boldsymbol{d} \equiv (d_1, \cdots, d_m)^{\mathrm{T}}$. From those equations, however, we can express $d_{kj}$ as a function of the $d_l$'s corresponding to the failure times preceding $t_{kj}$. That is, we can write

$$d_{kj} = q\big(\{d_l : t_l < t_{kj}\}\big), \qquad j = 2, \cdots, m_k, k = 1, \cdots, K, \tag{2.6}$$

where $q$ is some data-dependent function. This equation defines a recursive formula to compute $d_{kj}$ $(j = 2, \cdots, m_k)$ given $\boldsymbol{\beta}$ and $d_{k1}$ $(k = 1, \cdots, K)$; see Technical Details §2.6.

Write $\alpha_k = d_{k1}$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^{\mathrm{T}}$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$. Then $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) \equiv \arg\max_{\boldsymbol{\Lambda}} l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ can be calculated by the recursive formula given in (2.10) in the Technical Details in §2.6. The profile log-likelihood $pl_n(\boldsymbol{\theta}) \equiv l_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\theta}))$ can be computed as well. The first and second derivatives of $pl_n(\boldsymbol{\theta})$ involve the derivatives of $\boldsymbol{\Lambda}$ with respect to $\boldsymbol{\theta}$; they can be obtained through the recursive formula by differentiating both sides of (2.6) with respect to $\boldsymbol{\theta}$. We can then use the Newton-Raphson algorithm to obtain the NPMLE of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}$. We set the initial value of $\boldsymbol{\beta}$ to $\mathbf{0}$ and the initial value of $\alpha_k$ to $1/m_k$.

In the general case, $\widetilde{D}_i$ is missing when $\xi_i = 0$. We propose an EM algorithm assuming that the $\widetilde{D}_i$ are observed in the "full data", whose log-likelihood is precisely (2.5). Thus, we only need to compute $w_{ik}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) \equiv E\{I(\widetilde{D}_i = k) | \widetilde{T}_i, \boldsymbol{Z}_i, \xi_i = 0; \boldsymbol{\beta}, \boldsymbol{\Lambda}\}$ in the E-step. Then, the M-step involves an objective function which is a version of (2.5) weighted by $w_{ik}$ and 1, whose maximization proceeds similarly to that of (2.5) described above. To compute the weights in the E-step, note that

$$w_{ik}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \frac{F_k'(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}{\sum_{l=1}^{K} F_l'(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})}.$$

*Remark 2.1.* In the iterations of the algorithm, the overall survival function

$$1 - \sum_{k=1}^{K} F_k(\widetilde{T}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})$$

may become zero or negative. To improve the convergence of the algorithm, one may impose a small positive number on the survival function as a "buffer" to force it to be strictly positive, along similar lines of Groeneboom and Wellner (1992, page 70). In actual implementations, however, we did not encounter problems arising from survival function becoming non-positive. If one directly maximizes the likelihood, the constraint would automatically be satisfied. Another way to incorporate the constraint is to decompose the cumulative incidence function using the mixture cure model representation of Lu and Peng (2008).

### 2.2.3 Asymptotic Properties

We assume the following regularity conditions:

(C1) The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, lies in the interior of a compact subset of the Euclidean space $\mathbb{R}^p$, where $p$ is the dimension of $\boldsymbol{\beta}$; the true value of $\Lambda_k$, denoted by $\Lambda_{k0}$, is continuously differentiable with $\Lambda_{k0}'(t) > 0$ on $[0, \tau]$ for some constant $\tau > 0$.

(C2) The components of $\boldsymbol{Z}(\cdot)$ are uniformly bounded and have bounded total variation with probability one, and if $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}(t) = d(t)$ almost surely for some constant function $d$ for all $t \in [0, \tau]$, then $\boldsymbol{\beta} = \mathbf{0}$ and $d(t) = 0$.

(C3) With probability one, there exists a constant $\delta_0$ such that $\Pr(T \geq \tau|\boldsymbol{Z}) \geq \delta_0 > 0$ and $\Pr(C \geq \tau|\boldsymbol{Z}) = \Pr(C = \tau|\boldsymbol{Z}) \geq \delta_0 > 0$.

(C4) The function $G_k$ is four-times differentiable with $G_k(0) = 0$ and $G_k'(x) > 0$, and for any $c_0 > 0$,

$$\limsup_{x\to\infty}[\{G_k(c_0x)\}^{-1}\log\{x\sup_{y\leq x}G_k'(y)\}] = 0. \tag{2.7}$$

(C5) With probability one, $\Pr(\xi = 1|\widetilde{D}, \widetilde{T}, T \leq C, \boldsymbol{Z}) = \Pr(\xi = 1|\widetilde{T}, T \leq C, \boldsymbol{Z}) > \delta_0$ for some $\delta_0 > 0$. In particular, the cause of failure is missing at random (MAR).

*Remark 2.2.* Conditions (C1)-(C3) are standard regularity conditions in survival analysis. (C4) is satisfied by the Box-Cox transformations $G_k(x) = \{(1+x)^\gamma - 1\}/\gamma$ $(\gamma \geq 0)$. Equation (2.7) is not satisfied by the logarithmic transformations $G_k(x) = r^{-1}\log(1+rx)$ $(r \geq 0)$. However, this equation, which ensures that $\widehat{\Lambda}_k(\tau)$ stays bounded, is used only in proving the consistency of the NPMLEs, and the proof actually goes through for the logarithmic transformations by the partitioning device described in the technical report of Zeng and Lin (2006). The MAR assumption in (C5) is necessary for the validity of the MLE.

The following theorem on the consistency of the NPMLEs is proved the Technical Details in §2.6.

*Theorem 2.1.* Under Conditions (C1)-(C5), $\widehat{\boldsymbol{\beta}}$ and $\widehat{\Lambda}_k$ $(k = 1, \ldots, K)$ are strongly consistent, i.e.,

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| + \sup_{t\in[0,\tau]}\sum_{k=1}^{K}\left|\widehat{\Lambda}_k(t) - \Lambda_{k0}(t)\right| \longrightarrow 0$$

almost surely, where $||\cdot||$ denotes the Euclidean norm.

*Remark 2.3.* A major challenge in proving this theorem is that the $\Lambda_k$'s are defective (i.e., $\Lambda_k(\tau)$ cannot be arbitrarily large) and constrained by the condition that $1 - \sum_{k=1}^{K} F_k(\tau; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) > 0$. To overcome this technical difficulty, we show that $\liminf_n \left\{1 - \sum_{k=1}^{K} F_k(\tau; \boldsymbol{Z}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})\right\} > 0$.

Let $BV_1$ denote the space of functions on $[0, \tau]$ that are uniformly bounded by 1 and with total variation bounded by 1. Write $\mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}|| \leq 1\}$ and $\mathcal{W} = BV_1^{\otimes K}$, which is the $K$-product space of $BV_1$. Let $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{w} = (h_1, \cdots, h_K) \in \mathcal{W}$. Then we can identify $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ as elements in $l^\infty(\mathcal{V} \times \mathcal{W})$, which is the space of bounded functions on $\mathcal{V} \times \mathcal{W}$, by $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^K \int_0^\tau h_k d\Lambda_k$. Likewise, we identify $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0)$ as random elements in $l^\infty(\mathcal{V} \times \mathcal{W})$ such that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0)[\boldsymbol{v}, \boldsymbol{w}] = \sqrt{n}\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \sqrt{n}\sum_{k=1}^K \int h_k d(\widehat{\Lambda}_k - \Lambda_{k0}).$$

The following theorem on the distribution of the NPMLEs is proved in the Technical Details in §2.6.

*Theorem 2.2.* Under Conditions (C1)-(C5), $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0)$ converges weakly to a zero-mean Gaussian process in $l^\infty(\mathcal{V} \times \mathcal{W})$. In addition, $\widehat{\boldsymbol{\beta}}$ is semiparametric efficient in the sense of Bickel *et al.* (1993).

This theorem implies that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically multivariate zero-mean normal and $\sqrt{n}(\widehat{\Lambda}_1 - \Lambda_{10}, \cdots, \widehat{\Lambda}_K - \Lambda_{K0})$ converges to a multivariate zero-mean Gaussian process on $[0, \tau]^{\otimes K}$, the $K$-product space of $[0, \tau]$. If failure cause is fully observed, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated by the upper left block of $-\left\{\partial^2 pl_n(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}\right\}^{-1}$, which is a natural by-product of the algorithm. In the general case, we can estimate the covariance matrix by the inverse of the whole information matrix for $\boldsymbol{\beta}$ and the nonzero $d_{kj}$'s. This approach also provides variance estimation for the $\widehat{\Lambda}_k$'s; see the Technical Details in §2.6 for justifications.

Since $\Lambda_k(t)$ is positive, we construct its confidence interval by using the log transformation: $\widehat{\Lambda}_k(t)\exp\{\pm z_{1-\alpha/2}\widehat{\xi}_k(t)^{1/2}/\widehat{\Lambda}_k(t)\}$, where $\widehat{\xi}_k(t)$ is the variance estimator of $\widehat{\Lambda}_k(t)$, and $z_{1-\alpha/2}$ is the upper $(1 - \alpha/2)100$th percentile of the standard normal distribution. To estimate $\Lambda_k(t; \boldsymbol{z})$ for covariate value $\boldsymbol{z}$, we subtract $\boldsymbol{z}$ from $\boldsymbol{Z}$; then $\Lambda_k(t)$ corresponds to $\Lambda_k(t; \boldsymbol{z})$. Inference on $F_k(t; \boldsymbol{z})$ follows from the simple relationship $F_k(t; \boldsymbol{z}) = 1 - e^{-\Lambda_k(t;\boldsymbol{z})}$.

### 2.2.4 Model Checking

The class of models given in (2.3) requires specification of the following components: the functional form of each covariate; the link function, i.e., the exponential regression function; the proportionality structure, i.e., the multiplicative effect of the regression function within the transformation; and the transformation function $G_k$. To check these components, we define appropriate residuals and consider cumulative sums of residuals. Specifically, we define $N_k(t) = I(\widetilde{T} \le t, \xi\widetilde{D} = k)$ and $Y(t) = I(\widetilde{T} \ge t)$. For simplicity, we present the results only for the case where the failure cause is fully observed. We provide some remarks on how to construct the procedure for the general case at the end of the subsection.

With $\xi = 1$, the following process is centered at zero

$$\begin{aligned} M_k(t; \boldsymbol{\beta}, \boldsymbol{\Lambda}) &= N_k(t) - \int_0^t Y(u) S^{-1}(u; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) dF_k(u; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \\ &= N_k(t) - \int_0^t Y(u) \Psi_k(u; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) d\Lambda_k(u), \end{aligned} \tag{2.8}$$

where $S(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = 1 - \sum_{k=1}^K F_k(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda})$, and

$$\Psi_k(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = S^{-1}(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \exp\left[ H_k \left\{ \int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(u)} d\Lambda_k(u) \right\} + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) \right].$$

We obtain the residual process $M_k(t; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ by replacing $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ in (2.8) with the NPMLEs.

Let $M_{ki}$ denote the value of $M_k$ for the $i$th subject, and let $Z_{ji}$ denote the $j$th component of $\boldsymbol{Z}_i$. To check the functional form of the $j$th covariate, we consider the cumulative sum of residuals over this covariate:

$$W_{kc}^{(j)}(z, t) = n^{-1/2} \sum_{i=1}^n \int_0^t I\{Z_{ji}(s) \le z\} dM_{ki}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}).$$

To check the link function, we consider the cumulative sum over the linear predictor:

$$W_{kl}(x,t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t I\left\{\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{Z}_i(s) \le x\right\} dM_{ki}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}).$$

To check the transformation function $G_k$, we take the cumulative sum over its argument:

$$W_{ktr}(x,t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t I\left\{\int_0^s e^{\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{Z}_i(u)} d\widehat{\Lambda}_k(u) \le x\right\} dM_{ki}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}).$$

To check the proportionality for the $j$th covariate, we consider the "score" process

$$W_{kp}^{(j)}(t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t \widetilde{Z}_{kji}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) dM_{ki}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}),$$

where $\widetilde{Z}_{kji}(t; \boldsymbol{\beta}, \boldsymbol{\Lambda})$ is the $j$th component of

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}_k} \log \Psi_k(t; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = \boldsymbol{Z}_i(t) + \Bigg[ H_k'\left\{\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(u)} d\Lambda_k(u)\right\} + S^{-1}(t; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \\ \times \exp\left\{H_k\left(\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(u)} d\Lambda_k(u)\right)\right\}\Bigg] \int_0^t \boldsymbol{Z}_i(u) e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(u)} d\Lambda_k(u), \end{aligned}$$

which pertains to the score function of $\boldsymbol{\beta}_k$ based on the data available up to time $t$. Finally, to assess the overall fit of the model, we consider the process

$$W_{ko}(\boldsymbol{z}, t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t I\{\boldsymbol{Z}_i(s) \le \boldsymbol{z}\} dM_{ki}(s; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}).$$

All above processes are special cases of the multi-parameter process

$$W_{kn}(\boldsymbol{x}, t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t f(\boldsymbol{x}, u; \boldsymbol{Z}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) dM_{ki}(u; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}),$$

where $f$ is some function. We use Monte Carlo simulation to evaluate its null distribution. Specifically, we define

$$\widehat{W}_{kn}(\boldsymbol{x},t)=n^{-1/2}\sum_{i=1}^{n}\left\{\int_{0}^{t}f(\boldsymbol{x},u;\boldsymbol{Z}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})dM_{ki}(u;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})+\widehat{S}_i(\boldsymbol{x},t)\right\}Q_i,$$

where $(Q_1,\cdots,Q_n)$ are independent standard normal variables, and $\widehat{S}_1(\boldsymbol{x},t),\cdots,\widehat{S}_n(\boldsymbol{x},t)$ are described in the Technical Details in §2.6. We show in the Technical Details in §2.6 that the conditional distribution of $\widehat{W}_{kn}$ given the observed data $(\widetilde{T}_i,\widetilde{D}_i,\boldsymbol{Z}_i)$ $(i=1,\ldots,n)$ is asymptotically the same as the distribution of $W_{kn}$.

To approximate the null distribution of $W_{kn}$, we simulate the distribution of $\widehat{W}_{kn}$ by repeatedly generating the normal random sample $(Q_1,\cdots,Q_n)$ while holding the observed data fixed. To visually inspect model mis-specification, we compare the observed residual process with a few, say 20, realizations from the simulated process. We can also perform formal goodness-of-fit tests by calculating the $p$-values for the suprema of the residual processes based on a large number, say 1000, realizations. We establish the consistency of the supremum tests in the Technical Details in §2.6.

*Remark 2.4.* In the general case where the failure cause is possibly missing, we need to redefine the mean-zero process $M_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda})$. Specifically, the $N_k(t)$ in (2.8) will be replaced by $\widetilde{N}_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda})\equiv N_k(t)+w_k(\boldsymbol{\beta},\boldsymbol{\Lambda})I(\widetilde{T}\le t,\xi=0)$, where $w_k(\boldsymbol{\beta},\boldsymbol{\Lambda})=\left\{\sum_{l=1}^{K}F_l'(\widetilde{T};\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\right\}^{-1}\times F_k'(\widetilde{T};\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})$. Then, the expansion of the observed cumulative residuals $M_k(t;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ by $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ proceeds similarly as described in the Technical Details in §2.6 for the special case.

### 2.3 Simulation Studies

We conducted extensive simulation studies to evaluate the proposed and existing methods. We set $K=2$ and $\mathbf{Z}=(Z_1,Z_2)^{\mathrm{T}}$, where $Z_1$ is binary with $\Pr(Z_1=-1)=\Pr(Z_1=1)=0.5$, and $Z_2$ is $\mathrm{Un}(-1,1)$. We let $\Lambda_k(t)=\rho_k(1-e^{-t})$, where $\rho_k>0$.

We first simulated data with fully observed failure cause. We compared the NPMLE with the Fine and Gray (1999) (FG) method under proportional sub-distribution hazards models. We let the censoring time be the minimum of a Un$(5, 6)$ variable and an Exp$(0.1)$ variable. We let $\boldsymbol{\beta}_1 = \mathbf{0}$, $\boldsymbol{\beta}_2 = (0.5, 0.5)^{\mathrm{T}}$, $\rho_1 = 0.1$, and $\rho_2 = 0.75$, 1.1, and 1.5, corresponding to $50\%, 40\%$, and $25\%$ censoring, respectively. We focused on the estimation of $\beta_{11}$. The algorithm was deemed convergent when the Euclidean distance between the values of $\boldsymbol{\beta}$ of the current and previous iterations was less than $10^{-4}$ and the number of iterations did not exceed 100. Convergence rates for NPMLE with sample sizes $100, 200$, and $500$ were approximately $99.5\%, 99.8\%$, and $99.9\%$, respectively. For most ($> 98.6\%$) of the simulated datasets, convergence criteria were met with $3$ to $10$ iterations. The simulations were implemented in R. It took about 1 second and 0.5 second on a Dell Inspiron 2000 machine to complete one simulation with NPMLE and with FG for $n = 200$, respectively. The results for the estimation of $\beta_{11}$ are summarized in Table 2.1. For both methods, the estimators are virtually unbiased and the standard error estimators reflect the true variations well. Thus, the confidence intervals have accurate coverage probabilities. However, the standard error of the FG estimator is larger than that of the NPMLE. The difference is more pronounced when there are more events of the second type and lower censoring rate because FG models the censoring distribution while discarding the information in the second type of event.

Then, we evaluated the NPMLE under different transformation models. We considered the family of logarithmic transformations $G_r(x) = r^{-1} \log(1 + rx)$, in which $r = 0$ and $1$ correspond to the proportional sub-distribution hazards and proportional odds models, respectively. We set $\rho_1 = 0.1$ and $\rho_2 = 0.75$, and used the same censoring distributions as in Table 2.1. We focused on the estimation of $\boldsymbol{\beta}_1$ with different true values while fixing $\boldsymbol{\beta}_2$ at $(0.5, 0.5)^{\mathrm{T}}$. The results are summarized in Table 2.2. The NPMLE provides valid inference for all transformation models.

We also evaluated the FG method under mis-specified censoring distributions. We let the censoring time be the minimum of a Un$(3, 6)$ variable and an Exp$(\exp(\eta Z_1))$ variable, where $\eta = 1$ or 2. We let $\boldsymbol{\beta}_1 = \mathbf{0}$, $\boldsymbol{\beta}_2 = (0.5, 0.5)^{\mathrm{T}}$, $\rho_1 = 0.1$, and $\rho_2 = 0.75$. The other conditions were

Table 2.1: Comparison of the NPMLE and Fine and Gray methods under proportional subdistribution hazards models

| | | NPMLE | | | | FG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Censoring | $n$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP | RE |
| 50% | 100 | −0.004 | 0.396 | 0.388 | 0.948 | −0.002 | 0.401 | 0.394 | 0.954 | 1.03 |
| | 200 | 0.000 | 0.270 | 0.267 | 0.951 | −0.008 | 0.272 | 0.276 | 0.946 | 1.01 |
| | 500 | −0.001 | 0.162 | 0.153 | 0.945 | −0.003 | 0.162 | 0.164 | 0.953 | 1.00 |
| 40% | 100 | −0.006 | 0.389 | 0.397 | 0.944 | −0.001 | 0.409 | 0.412 | 0.965 | 1.11 |
| | 200 | −0.003 | 0.263 | 0.267 | 0.942 | −0.008 | 0.278 | 0.273 | 0.954 | 1.12 |
| | 500 | −0.008 | 0.161 | 0.163 | 0.942 | 0.001 | 0.169 | 0.167 | 0.958 | 1.10 |
| 25% | 100 | −0.001 | 0.385 | 0.385 | 0.949 | −0.001 | 0.413 | 0.408 | 0.956 | 1.15 |
| | 200 | 0.001 | 0.257 | 0.264 | 0.957 | 0.006 | 0.270 | 0.272 | 0.949 | 1.10 |
| | 500 | 0.006 | 0.161 | 0.158 | 0.958 | 0.005 | 0.174 | 0.176 | 0.948 | 1.17 |

Bias and SE are the bias and standard error of the parameter estimator; SEE is the empirical average of the standard error estimator; CP is the empirical coverage probability of the 95% confidence interval. Each entry is based on 10,000 replicates. RE is the variance of FG over that of the NPMLE.

the same as those for Table 2.1. Thus, the censoring distribution depends on the covariate in a non-proportional hazards manner. In the FG method, the censoring distribution is estimated by the Kaplan-Meier estimator. Fine and Gray (1999) suggested to use the proportional hazards model for the censoring distribution but did not derive the corresponding variance estimators. Table 2.3 compares the NPMLE, the original FG estimator, and the modification based on the proportional hazards modelling of the censoring distribution, denoted by FG*. FG has considerable bias and the bias becomes greater as the censoring distributions become more uneven. For FG*, the bias is smaller but still appreciable relative to the standard error, especially when the sample size is large. For both FG and FG*, the mean square error is larger than that of the NPMLE.

We then considered estimation of the cumulative hazard functions $\Lambda_k$. We set $\rho_1 = 0.1$, $\rho_2 = 0.75$, and $\boldsymbol{\beta} = \mathbf{0}$. The censoring distributions were the same as in Table 2.1. The results for $\Lambda_1(t)$ are summarized in Table 2.4. The parameter estimators are virtually unbiased, the standard error estimators are accurate, and the confidence intervals have correct coverages.

Next, we compared our NPMLE to the parametric MLE of Jeong and Fine (2007). We let $\rho_1 = 0.1$ and $\rho_2 = 0.75$, and used the set-up of Table 2.1, but set $\boldsymbol{\beta} = \mathbf{0}$ and focused on $\beta_{11}$. Under this setting, $\Lambda_1$ and $\Lambda_2$ are correctly modeled by the parametric method. As shown in Table 2.5,

Table 2.2: Simulation results on the regression parameters under transformation models

| | | | $\beta_{11}$ | | | | $\beta_{12}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $r$ | $(\beta_{11}, \beta_{12})$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 100 | 0 | (0, 0) | 0.003 | 0.395 | 0.391 | 0.945 | 0.008 | 0.608 | 0.608 | 0.950 |
| | | (0, 0.5) | −0.008 | 0.398 | 0.393 | 0.945 | −0.002 | 0.609 | 0.605 | 0.943 |
| | | (0.5, 0.5) | −0.001 | 0.401 | 0.397 | 0.945 | −0.001 | 0.611 | 0.608 | 0.946 |
| | 0.5 | (0, 0) | 0.004 | 0.403 | 0.401 | 0.948 | 0.008 | 0.620 | 0.626 | 0.957 |
| | | (0, 0.5) | 0.005 | 0.407 | 0.405 | 0.946 | 0.003 | 0.621 | 0.619 | 0.947 |
| | | (0.5, 0.5) | 0.000 | 0.412 | 0.417 | 0.956 | −0.001 | 0.623 | 0.623 | 0.953 |
| | 1 | (0, 0) | 0.005 | 0.411 | 0.408 | 0.945 | −0.004 | 0.633 | 0.633 | 0.952 |
| | | (0, 0.5) | −0.007 | 0.414 | 0.408 | 0.944 | −0.007 | 0.636 | 0.641 | 0.958 |
| | | (0.5, 0.5) | −0.007 | 0.417 | 0.418 | 0.951 | 0.004 | 0.639 | 0.633 | 0.942 |
| 200 | 0 | (0, 0) | −0.005 | 0.269 | 0.272 | 0.955 | −0.006 | 0.464 | 0.464 | 0.950 |
| | | (0, 0.5) | −0.001 | 0.273 | 0.278 | 0.960 | −0.007 | 0.468 | 0.466 | 0.948 |
| | | (0.5, 0.5) | 0.005 | 0.274 | 0.274 | 0.951 | −0.001 | 0.469 | 0.467 | 0.946 |
| | 0.5 | (0, 0) | −0.006 | 0.275 | 0.274 | 0.948 | −0.005 | 0.473 | 0.470 | 0.946 |
| | | (0, 0.5) | −0.004 | 0.279 | 0.281 | 0.953 | 0.005 | 0.474 | 0.473 | 0.947 |
| | | (0.5, 0.5) | −0.006 | 0.281 | 0.281 | 0.952 | 0.005 | 0.476 | 0.470 | 0.943 |
| | 1 | (0, 0) | −0.002 | 0.281 | 0.283 | 0.954 | 0.007 | 0.483 | 0.488 | 0.956 |
| | | (0, 0.5) | 0.002 | 0.286 | 0.288 | 0.954 | −0.004 | 0.487 | 0.484 | 0.946 |
| | | (0.5, 0.5) | −0.002 | 0.289 | 0.285 | 0.943 | −0.003 | 0.489 | 0.494 | 0.959 |
| 500 | 0 | (0, 0) | −0.003 | 0.163 | 0.165 | 0.953 | 0.002 | 0.285 | 0.283 | 0.947 |
| | | (0, 0.5) | −0.005 | 0.164 | 0.167 | 0.957 | 0.001 | 0.288 | 0.283 | 0.944 |
| | | (0.5, 0.5) | 0.003 | 0.167 | 0.164 | 0.946 | 0.005 | 0.293 | 0.288 | 0.941 |
| | 0.5 | (0, 0) | −0.004 | 0.165 | 0.163 | 0.945 | −0.002 | 0.291 | 0.294 | 0.954 |
| | | (0, 0.5) | 0.000 | 0.166 | 0.166 | 0.954 | 0.003 | 0.294 | 0.290 | 0.945 |
| | | (0.5, 0.5) | 0.003 | 0.166 | 0.166 | 0.949 | 0.004 | 0.295 | 0.290 | 0.945 |
| | 1 | (0, 0) | −0.002 | 0.169 | 0.169 | 0.951 | −0.002 | 0.297 | 0.301 | 0.955 |
| | | (0, 0.5) | −0.002 | 0.173 | 0.174 | 0.953 | 0.000 | 0.297 | 0.303 | 0.955 |
| | | (0.5, 0.5) | 0.001 | 0.178 | 0.179 | 0.954 | −0.004 | 0.301 | 0.302 | 0.951 |

See the note to Table 2.1.

Table 2.3: Simulation results for the Fine and Gray estimators under mis-specified censoring distributions

| | | NPMLE | | | FG | | | FG* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE |
| 1 | 100 | 0.002 | 0.396 | 0.157 | −0.119 | 0.404 | 0.177 | −0.054 | 0.414 | 0.174 |
| | 200 | 0.003 | 0.270 | 0.073 | −0.116 | 0.275 | 0.089 | −0.049 | 0.275 | 0.078 |
| | 500 | −0.003 | 0.161 | 0.026 | −0.114 | 0.161 | 0.039 | −0.053 | 0.165 | 0.030 |
| 2 | 100 | 0.002 | 0.395 | 0.156 | −0.231 | 0.401 | 0.214 | −0.031 | 0.418 | 0.176 |
| | 200 | 0.002 | 0.266 | 0.071 | −0.223 | 0.282 | 0.129 | −0.028 | 0.281 | 0.080 |
| | 500 | −0.001 | 0.155 | 0.024 | −0.227 | 0.163 | 0.078 | −0.027 | 0.173 | 0.032 |

See the note to Table 2.1. FG and FG* are based on the Kaplan-Meier estimator and proportional hazards model for the censoring distribution, respectively. MSE is the mean square error.

Table 2.4: Simulation results on the estimation of the cumulative hazard function under transformation models

| $n$ | $r$ | $t$ | $\Lambda_1(t)$ | Bias | SE | SEE | CP |
|---|---|---|---|---|---|---|---|
| 100 | 0 | 1 | 0.059 | 0.000 | 0.030 | 0.029 | 0.957 |
| | | 2 | 0.081 | 0.001 | 0.028 | 0.028 | 0.948 |
| | 0.5 | 1 | 0.059 | 0.001 | 0.026 | 0.022 | 0.960 |
| | | 2 | 0.081 | 0.000 | 0.030 | 0.030 | 0.952 |
| | 1 | 1 | 0.059 | 0.000 | 0.024 | 0.026 | 0.943 |
| | | 2 | 0.081 | 0.000 | 0.033 | 0.030 | 0.951 |
| 200 | 0 | 1 | 0.059 | −0.001 | 0.017 | 0.019 | 0.952 |
| | | 2 | 0.081 | 0.000 | 0.022 | 0.021 | 0.950 |
| | 0.5 | 1 | 0.059 | −0.002 | 0.017 | 0.017 | 0.950 |
| | | 2 | 0.081 | 0.001 | 0.025 | 0.025 | 0.955 |
| | 1 | 1 | 0.059 | 0.001 | 0.018 | 0.018 | 0.949 |
| | | 2 | 0.081 | 0.002 | 0.023 | 0.022 | 0.946 |
| 500 | 0 | 1 | 0.059 | 0.000 | 0.010 | 0.011 | 0.956 |
| | | 2 | 0.081 | 0.000 | 0.012 | 0.012 | 0.944 |
| | 0.5 | 1 | 0.059 | 0.000 | 0.011 | 0.012 | 0.954 |
| | | 2 | 0.081 | −0.001 | 0.015 | 0.015 | 0.948 |
| | 1 | 1 | 0.059 | −0.001 | 0.012 | 0.013 | 0.943 |
| | | 2 | 0.081 | 0.000 | 0.013 | 0.013 | 0.954 |

See the note to Table 2.1.

Table 2.5: Comparison of the semiparametric and parametric MLEs of $\beta_{11}$

| | | Semiparametric | | | | Parametric | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $r$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 100 | 0 | 0.004 | 0.391 | 0.388 | 0.943 | −0.007 | 0.365 | 0.365 | 0.956 |
| | 0.5 | 0.004 | 0.400 | 0.399 | 0.953 | −0.002 | 0.372 | 0.371 | 0.957 |
| | 1 | 0.006 | 0.412 | 0.406 | 0.945 | 0.008 | 0.387 | 0.385 | 0.945 |
| 200 | 0 | −0.002 | 0.277 | 0.276 | 0.955 | −0.007 | 0.261 | 0.262 | 0.950 |
| | 0.5 | −0.003 | 0.296 | 0.292 | 0.946 | −0.010 | 0.274 | 0.276 | 0.957 |
| | 1 | −0.003 | 0.309 | 0.305 | 0.943 | −0.002 | 0.287 | 0.288 | 0.944 |
| 500 | 0 | 0.004 | 0.169 | 0.173 | 0.959 | 0.006 | 0.164 | 0.159 | 0.947 |
| | 0.5 | 0.008 | 0.175 | 0.172 | 0.950 | −0.005 | 0.160 | 0.160 | 0.958 |
| | 1 | −0.004 | 0.173 | 0.180 | 0.940 | −0.007 | 0.171 | 0.172 | 0.944 |

See the note to Table 2.1.

the parametric MLE tends to be more efficient than the NPMLE; however, the efficiency gain is rather moderate.

We conducted additional studies to assess the bias of the parametric method under mis-specified failure distributions. We used the proportional sub-distribution hazards models and set $\boldsymbol{\beta}_1 = (0.5, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. We let $\lambda_k(t) = 0.5t\exp(-t^2/2)$. The censoring distributions were the same as in Table 2.1. As shown in Table 2.6, the estimation for the cumulative hazard function is severely biased.

To show that the estimation of $\boldsymbol{\beta}$ can also be biased, we considered a more wiggly hazard function, i.e., $\lambda_k(t) = 0.15\{1 + \cos(\pi t)\}$. We set $\boldsymbol{\beta}_1 = (0.5, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = \mathbf{0}$ and focused on the estimation of $\beta_{11}$. The censoring distributions were the same as in Table 2.1. As shown in Table 2.7, the parametric method underestimates $\beta_{11}$. The bias is considerable relative to the standard error, especially when the sample size is large.

Table 2.6: Comparison of the semiparametric and parametric MLEs under mild mis-specification of the parametric distribution for $\Lambda_k$

| | | | Semiparametric | | | | Parametric | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Parameter | Value | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 100 | $\beta_{11}$ | 0.500 | 0.003 | 0.225 | 0.221 | 0.958 | 0.007 | 0.220 | 0.205 | 0.940 |
| | $\Lambda_1(0.5)$ | 0.047 | 0.009 | 0.024 | 0.024 | 0.944 | 0.057 | 0.020 | 0.018 | 0.141 |
| | $\Lambda_1(1.5)$ | 0.346 | 0.003 | 0.055 | 0.054 | 0.956 | −0.106 | 0.048 | 0.037 | 0.279 |
| 200 | $\beta_{11}$ | 0.500 | 0.006 | 0.146 | 0.147 | 0.956 | 0.001 | 0.135 | 0.131 | 0.958 |
| | $\Lambda_1(0.5)$ | 0.047 | −0.003 | 0.017 | 0.016 | 0.956 | 0.059 | 0.016 | 0.014 | 0.041 |
| | $\Lambda_1(1.5)$ | 0.346 | −0.008 | 0.037 | 0.038 | 0.944 | −0.101 | 0.034 | 0.029 | 0.076 |
| 500 | $\beta_{11}$ | 0.500 | −0.001 | 0.095 | 0.096 | 0.947 | 0.000 | 0.092 | 0.087 | 0.959 |
| | $\Lambda_1(0.5)$ | 0.047 | 0.005 | 0.013 | 0.013 | 0.947 | 0.058 | 0.012 | 0.009 | $< .001$ |
| | $\Lambda_1(1.5)$ | 0.346 | −0.005 | 0.024 | 0.025 | 0.951 | −0.105 | 0.020 | 0.016 | $< .001$ |

See the note to Table 2.1.

Table 2.7: Comparison of the semiparametric and parametric MLEs of $\beta_{11}$ under severe mis-specification of the parametric distribution for $\Lambda_k$

| | Semiparametric | | | | Parametric | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 100 | 0.008 | 0.228 | 0.232 | 0.952 | −0.043 | 0.223 | 0.215 | 0.945 |
| 200 | 0.005 | 0.151 | 0.152 | 0.947 | −0.041 | 0.143 | 0.136 | 0.937 |
| 500 | −0.003 | 0.103 | 0.101 | 0.951 | −0.039 | 0.098 | 0.083 | 0.930 |

See the note to Table 2.1.

We also evaluated the performance of the goodness-of-fit tests in the set-up of Table 2.1 with $\rho_1 = 0.1$ and $\rho_2 = 0.75$. We set $n = 100$, $\boldsymbol{\beta}_1 = (0.5, 0)^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (0, 0)^{\mathrm{T}}$. We evaluated the type I error of the supremum tests for the first risk at the nominal significance level of $0.05$. We simulated 10,000 datasets and used 1,000 normal samples to calculate the $p$-value. For checking the functional form of $Z_1$, the exponential link function, the transformation function, the proportionality on $Z_1$, and the overall fit, the empirical type I error rates were found to be 0.051, 0.059, 0.062, 0.042, and 0.048, respectively. Thus, the asymptotic approximations for the supremum tests are accurate enough for practical use.

Then, we assessed the robustness of the inference on the risk of interest to model mis-specification on other risks. We used the setting in Table 2.1 with $\rho_1 = 0.1$ and $\rho_2 = 0.75$. We generated data

Table 2.8: Simulation results on the regression parameters of one risk under mis-specification of the other risk

| | | $\beta_{11}$ | | | | $\beta_{12}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $(\beta_{11}, \beta_{12})$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 100 | (0, 0) | 0.001 | 0.384 | 0.384 | 0.936 | 0.002 | 0.606 | 0.606 | 0.949 |
| | (0, 0.5) | 0.010 | 0.373 | 0.371 | 0.963 | 0.004 | 0.667 | 0.669 | 0.952 |
| | (0.5, 0.5) | 0.010 | 0.432 | 0.434 | 0.961 | 0.000 | 0.638 | 0.640 | 0.955 |
| 200 | (0, 0) | 0.005 | 0.280 | 0.277 | 0.958 | 0.012 | 0.465 | 0.467 | 0.956 |
| | (0, 0.5) | 0.001 | 0.271 | 0.271 | 0.944 | 0.003 | 0.465 | 0.463 | 0.933 |
| | (0.5, 0.5) | 0.006 | 0.294 | 0.293 | 0.961 | 0.010 | 0.454 | 0.451 | 0.963 |
| 500 | (0, 0) | 0.009 | 0.171 | 0.170 | 0.939 | 0.005 | 0.281 | 0.284 | 0.935 |
| | (0, 0.5) | 0.006 | 0.170 | 0.169 | 0.948 | 0.011 | 0.280 | 0.278 | 0.955 |
| | (0.5, 0.5) | 0.002 | 0.173 | 0.175 | 0.954 | 0.002 | 0.266 | 0.264 | 0.951 |

See the note to Table 2.1.

under the proportional sub-distribution hazards and proportional odds models for the first and second risks, respectively, but fit the proportional sub-distribution hazards models to both risks. As shown in Table 2.8, mis-specification of the second risk has little impact on the inference on the first risk.

Finally, we considered missing causes of failure. We generated the missing indicator for non-right-censored subjects by the logistic model

$$\Pr\big(\xi = 0 \big| \widetilde{D}, \widetilde{T}, T \leq C, \boldsymbol{Z}\big) = \frac{\exp(-0.5 + \gamma Z_1)}{1 + \exp(-0.5 + \gamma Z_1)}.$$

We set $\gamma = 0$ and $-0.2$, which correspond to missing completely at random (MCAR) and MAR, respectively. We used the set-up of Table 2.1 with with $\rho_2 = 0.75$, $\boldsymbol{\beta}_1 = (-0.5, 0)^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (0, 0)^{\mathrm{T}}$ We compared the NPMLE with the FG complete-case analysis (i.e., excluding subjects with missing causes of failure). The results for the estimation of $\beta_{11}$ are summarized in Table 2.9. The NPMLE remains unbiased and the FG is biased under both MCAR and MAR, especially for MAR. The bias from MCAR arises from uneven right-censoring rate. Because $\beta_{11} < 0$, for larger values of $Z_1$, more subjects are right-censored and thus fewer events of $k = 1$ are discarded due to missing cause information, resulting in a positive bias of the regression parameter. The

Table 2.9: Comparison of the NPMLE and Fine and Gray methods in the estimation of $\beta_{11}$ with missing causes of failure

| | $n$ | NPMLE | | | | FG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | SE | SEE | CP | Bias | SE | SEE | CP | RE |
| MCAR | 100 | 0.003 | 0.479 | 0.482 | 0.953 | 0.038 | 0.523 | 0.524 | 0.941 | 1.190 |
| | 200 | −0.001 | 0.291 | 0.295 | 0.955 | 0.050 | 0.319 | 0.321 | 0.929 | 1.197 |
| | 500 | 0.002 | 0.189 | 0.192 | 0.952 | 0.036 | 0.217 | 0.220 | 0.920 | 1.313 |
| MAR | 100 | 0.006 | 0.426 | 0.429 | 0.953 | 0.189 | 0.473 | 0.474 | 0.881 | 1.235 |
| | 200 | −0.006 | 0.306 | 0.305 | 0.949 | 0.203 | 0.336 | 0.337 | 0.825 | 1.202 |
| | 500 | −0.002 | 0.223 | 0.223 | 0.949 | 0.213 | 0.243 | 0.246 | 0.722 | 1.186 |

See the note to Table 2.1.

MAR scenario adds to the bias by making the missing rate smaller among the non-right-censored subjects for larger values of $Z_1$. In addition to the bias issue, the FG is also substantially less efficient for both MCAR and MAR as a result of discarding information contained in observations with unknown causes of failure.

## 2.4 A Bone Marrow Transplantation Study

We present a major study on bone marrow transplantation in patients with multiple myeloma (MM) (Kumar et al., 2011). The standard treatment for MM is autologous hematopoietic stem cell transplantation (auto-HCT). An alternative treatment, allogeneic hematopoietic cell transplantation (allo-HCT), is less commonly used because of its high treatment-related mortality (TRM). However, recent advances in medical care have lowered TRM rates of allo-HCT (Kumar et al., 2011). To evaluate the effects of various risk factors on clinical outcomes after allo-HCT for MM, we consider data collected from years 1995-2005 by the Center for International Blood and Marrow Transplantation Research (CIBMTR). The CIBMTR is comprised of clinical and basic scientists who confidentially share data on their blood and bone marrow transplant patients with the Data Collection Center located at the Medical College of Wisconsin; it provides a repository of information about results of transplants from more than 450 transplant centers worldwide.

The database contains 864 patients, among whom 376 received transplantation in 1995-2000 and 488 received transplantation in 2001-2005. The two competing risks are TRM and relapse of MM. A total of 297 patients experienced TRM, and 348 experienced relapse. Risk factors include cohort indicator (transplantation years 1995-2000 or 2001-2005), type of donor (unrelated or HLA-identical sibling donor), history of a prior auto-HCT (yes or no), and time from diagnosis to transplantation ($\leq$ 24 months or $>$ 24 months).

We first fit the proportional sub-distribution hazards models for both risks and compare our method with the FG method. As shown in Table 2.10, the two methods produce considerably different results. The differences are largely attributed to uneven censoring distributions. By fitting a proportional sub-distribution hazards model to the censoring distribution with the same set of covariates, we find that cohort indicator, prior auto-HCT, and waiting time increase the censoring rate. Thus, the FG estimates of their effects are biased downward for both risks. By contrast, donor type decreases the censoring rate, such that the FG estimates of its effects on the two risks are biased upward. At the significance level of 0.05, waiting time is associated with both risks under the NPMLE method but is not associated with either risk under the FG method. The more recent cohort (years 2001-2005) has a significantly lower incidence of TRM but higher incidence of relapse. Transplantation involving an unrelated donor significantly increases the risk of both TRM and relapse. Prior auto-HCT reduces the risk of TRM but increases the risk of relapse. We test the global null hypothesis that waiting time does not affect TRM or relapse. The $p$-value of the $\chi_2^2$ test is $<$0.001, so that the null hypothesis is strongly rejected. For comparison, we included the results of proportional cause-specific hazards modeld in Table 2.11. The parameter estimates are quite different from those from Table 2.10, but their signs are consistent, indicating that the covariate effects on the cause-specific and the sub-distribution hazards have the same directions.

Next, we consider the family of transformation functions $G_k(x) = r^{-1}\log(1 + rx)$ $(k = 1, 2)$. We fit 4 pairs of models with $r = 0$ or $1$ for the two competing risks. We label the choices of $r = i$

Table 2.10: Proportional sub-distribution hazards analysis of the bone marrow transplantation data

| | | NPMLE | | | FG | | |
|---|---|---|---|---|---|---|---|
| | | Est | SE | $p$-value | Est | SE | $p$-value |
| TRM | | | | | | | |
| | Years 2001-2005 | −0.543 | 0.132 | <0.001 | −0.578 | 0.139 | <0.001 |
| | Unrelated donor | 0.476 | 0.126 | <0.001 | 0.521 | 0.128 | <0.001 |
| | Prior auto-HCT | −0.451 | 0.162 | 0.005 | −0.463 | 0.153 | 0.003 |
| | TX $> 24$ months | 0.296 | 0.123 | 0.017 | 0.248 | 0.135 | 0.065 |
| Relapse | | | | | | | |
| | Years 2001-2005 | 0.518 | 0.129 | <0.001 | 0.401 | 0.122 | 0.001 |
| | Unrelated donor | 0.293 | 0.101 | 0.004 | 0.330 | 0.122 | 0.007 |
| | Prior auto-HCT | 0.399 | 0.116 | 0.001 | 0.351 | 0.125 | 0.005 |
| | TX $> 24$ months | 0.310 | 0.121 | 0.018 | 0.216 | 0.123 | 0.078 |

Table 2.11: Proportional cause-specific hazards analysis of the bone marrow transplantation data

| | | Est | SE | $p$-value |
|---|---|---|---|---|
| TRM | | | | |
| | Years 2001-2005 | −0.461 | 0.133 | <0.001 |
| | Unrelated donor | 0.743 | 0.128 | <0.001 |
| | Prior auto-HCT | −0.362 | 0.148 | 0.014 |
| | TX $> 24$ months | 0.316 | 0.133 | 0.017 |
| Relapse | | | | |
| | Years 2001-2005 | 0.348 | 0.134 | 0.009 |
| | Unrelated donor | 0.770 | 0.120 | <0.001 |
| | Prior auto-HCT | 0.309 | 0.137 | 0.024 |
| | TX $> 24$ months | 0.425 | 0.122 | <0.001 |

$(i = 0, 1)$ for TRM and $r = j$ $(j = 0, 1)$ for relapse as Model $2j + i + 1$. To evaluate these transformations, we use the $\sup_{x,t} |W_{ktr}(x, t)|$ test for $k = 1, 2$. The $p$-values, based on 1000 realizations, for testing the transformation functions for TRM under Models 1-4 are $0.035, 0.543, 0.029$, and $0.382$, respectively, and the corresponding $p$-values for testing the transformation functions for relapse are $0.056, 0.045, 0.282$, and $0.307$. These results suggest that the proportional subdistribution hazards assumption is not appropriate for TRM and relapse. We also fit the models with $r$ ranging from 0 to 2. As shown in Figure 2.1, the log-likelihood is maximized at $r = 0.8$ for TRM and $r = 1.3$ for relapse, which would be the combination selected by the Akaike information criterion. The selected combination is close to Model 4 (i.e., proportional odds models for both risks), which we adopt for ease of interpretation.

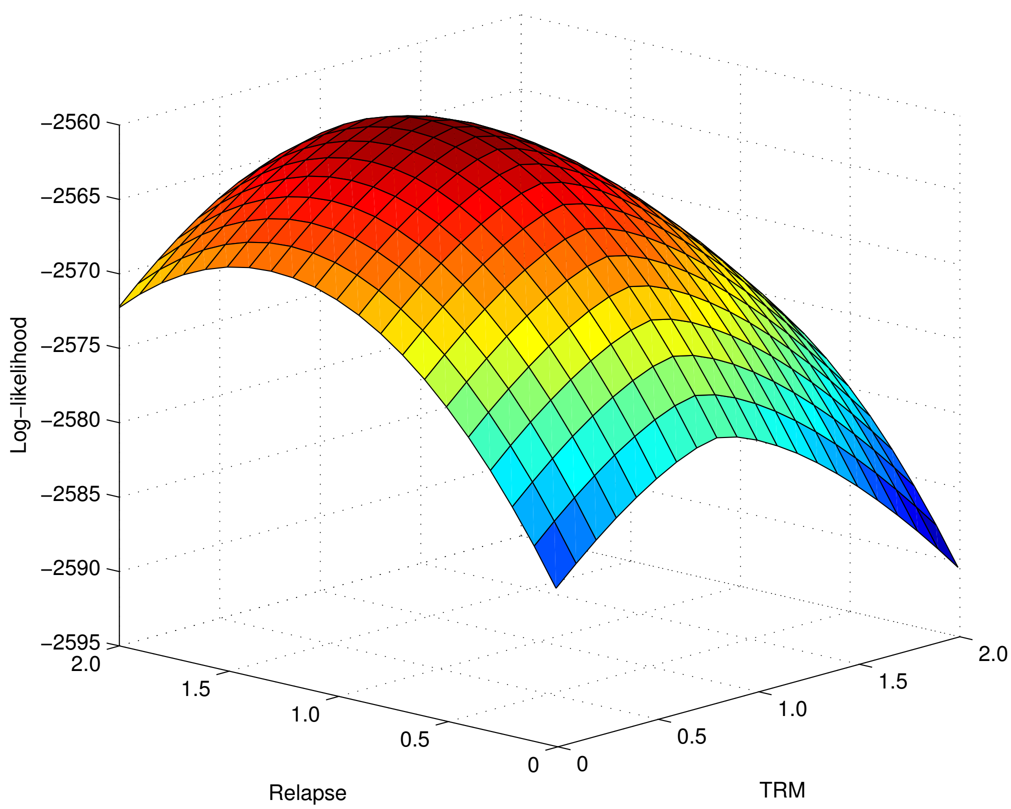


Figure 2.1: Log-likelihood surface for pairs of transformation functions $G_k(x) = r^{-1} \log(1 + rx)$ $(k = 1, 2)$ in the analysis of the bone marrow transplantation data.

The results from Model 4 are summarized in Table 2.12. These results differ markedly from the

Table 2.12: Proportional odds analysis of the bone marrow transplant data

| | | Est | SE | $p$-value |
|---|---|---|---|---|
| TRM | | | | |
| | Years 2001-2005 | $-0.582$ | 0.123 | <0.001 |
| | Unrelated donor | 0.504 | 0.108 | <0.001 |
| | prior auto-HCT | $-0.420$ | 0.143 | 0.003 |
| | TX $>$ 24 months | 0.217 | 0.126 | 0.084 |
| Relapse | | | | |
| | Years 2001-2005 | 0.353 | 0.116 | <0.001 |
| | Unrelated donor | 0.337 | 0.088 | <0.001 |
| | prior auto-HCT | 0.314 | 0.123 | 0.002 |
| | TX $>$ 24 months | 0.343 | 0.138 | 0.013 |

NPMLE results in Table 2.10, and the interpretations of the regression effects are quite different. We assess the proportionality assumption using the $\sup |W_{kp}|$ test. The $p$-values are 0.297, 0.123, 0.687, and 0.673, respectively, for the effects of cohort indicator, donor type, history and waiting time on TRM; the corresponding $p$-values for relapse are 0.818, 0.361, 0.352, and 0.940. Thus, the proportionality assumption holds on all covariates. The $p$-value for the omnibus test is 0.412, indicating overall goodness of fit.

To compare the predictions from the initial proportional sub-distribution hazards models and the chosen proportional odds models, we show in Figure 2.2 the estimated cumulative incidence functions for the two cohorts with HLA-identical sibling donor, no history of prior auto-HCT treatment, and waiting time $\leq 24$ months. The two pairs of models yield rather different estimates, especially for relapse. Thus, proper choice of the transformation function is crucial to accurate prediction of the incidence of competing risks.

### 2.5 Discussion

Our work represents the first likelihood-based approach to semiparametric regression analysis of cumulative incidence functions with competing risks data. It offers major improvements over the pioneer work of Fine and Gray (1999). First, it does not require modelling of the censoring distribution, such that the inference is valid regardless of the censoring patterns. Second, it

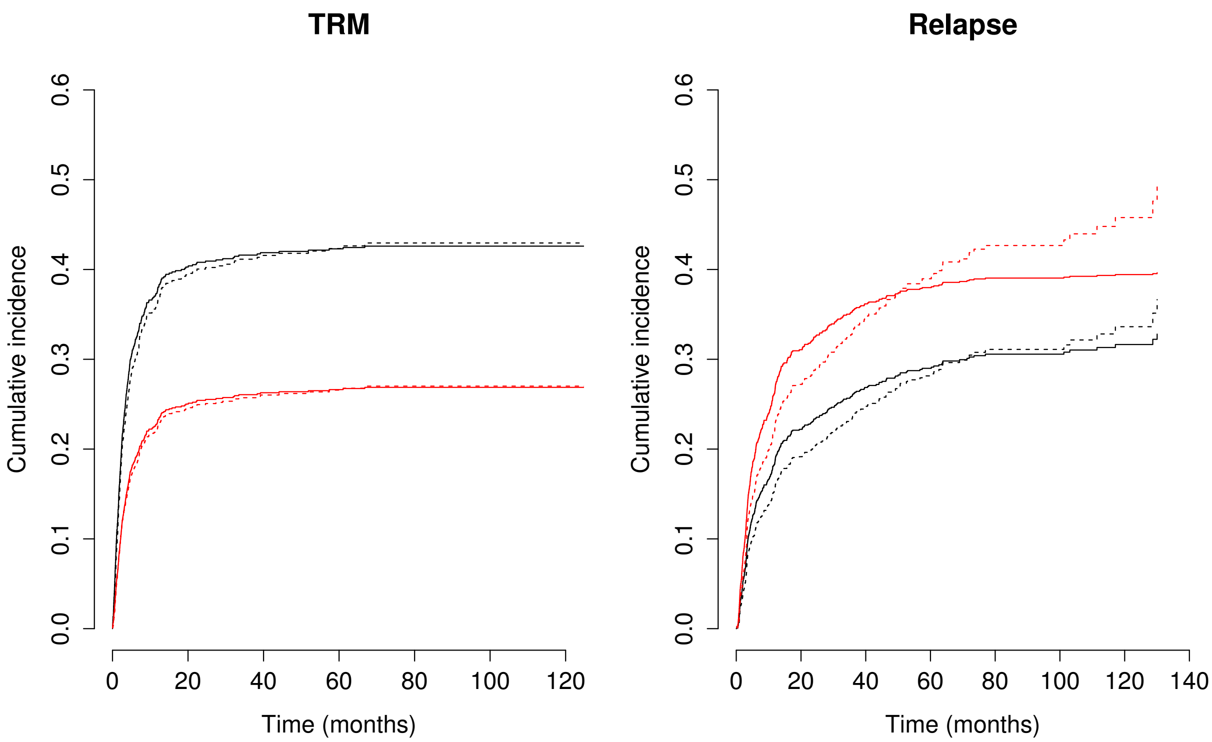


Figure 2.2: Estimated cumulative incidence of TRM and relapse for a subject with HLA-identical sibling donor, no prior auto-HCT treatment, and waiting time $\leq 24$ months. The black and red curves indicate years 1995-2000 and 2001-2005, respectively; the solid and dashed curves pertains to the proportional odds and proportional sub-distribution hazards models, respectively.

provides flexible choices of models for covariate effects, accommodating both proportional and non-proportional sub-distribution hazards structures. Third, it provides efficient parameter estimators. Fourth, it accommodates missing information on the failure cause. Fifth, it allows for simultaneous inference on multiple risks. Finally, it provides graphical and numerical techniques to evaluate and select models. These improvements have important implications in actual data analysis, as demonstrated in the simulation studies and real example.

Because an increase in the incidence of one risk reduces the incidence of other risks, it is necessary to take into account all other risks when interpreting the results on one particular risk. Thus, it is desirable to model all risks even when one is interested in only one of them. Our simulation results show that the inference on one risk is robust to model misspecification on other risks. The FG method only models the risk of interest and thus seems to involve fewer model assumptions. However, it requires modelling the censoring distribution, which is of no scientific interest at all. As shown in our simulation studies, mis-specification of the censoring distribution may bias the inference. In addition, the estimated inverse weights can be quite unstable under heavy censoring.

For maximizing the nonparametric likelihood, many authors have resorted to optimization algorithms. Due to the high dimensionality of the argument, such algorithms are slow and their convergence is not guaranteed. We have developed a recursive formula to compute the profile likelihood in the M-step of an EM algorithm. Our strategy greatly reduces the dimension of the problem and is fast and stable.

For notational simplicity, we have assumed that the covariates are the same for all risks. All theoretical results hold when covariates are risk-specific, i.e., dependent on $k$. Our formulation accommodates time-dependent covariates, but only external time-dependent covariates (Kalbfleisch and Prentice (2002), chapter 6) are allowed. A common example of such covariates is time$\times$covariate interaction; other examples include temperature, particulate levels, and precipitation. For internal time-dependent covariates, the relationship between the cumulative incidence function and the sub-distribution hazard function does not hold, and the likelihood does not conform to (2.4). The FG method is also restricted to external covariates (Latouche et al., 2005).

### 2.6 Technical Details

*Recursive formula for computing the profile likelihood*

We take the derivative of the log-likelihood in (6) with respect to $d_{kj}$ to obtain

$$\frac{\partial l_n(\boldsymbol{\beta}, \boldsymbol{\Lambda})}{\partial d_{kj}} = d_{kj}^{-1} + \sum_{l=j}^{m_k} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_{(kl)}} H_k'\big(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_{(kl)}} \Lambda_{(kl)}\big) - \sum_{t_l \geq t_{kj}} \sum_{t_l \leq \widetilde{T}_i < t_{l+1}} I(\widetilde{D}_i = 0) \frac{\exp\Big\{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i + H_k'\big(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i} \Lambda_{kl}\big)\Big\}}{\sum_{k=1}^K \exp\big\{-G_k(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i} \Lambda_{kl})\big\} - K + 1}. \tag{2.9}$$

Setting (2.9) to 0 for $d_{kj}$ and $d_{k,j+1}$, we obtain

$$d_{k,j+1}^{-1} = d_{kj}^{-1} + e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_{(kj)}} H_k'\big(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_{(kj)}} \Lambda_{(kj)}\big) - \sum_{t_{kj} \leq t_l < t_{k,j+1}} \sum_{t_l \leq \widetilde{T}_i < t_{l+1}} I(\widetilde{D}_i = 0) \frac{\exp\Big\{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i + H_k'\big(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i} \Lambda_{kl}\big)\Big\}}{\sum_{k=1}^K \exp\big\{-G_k(e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i} \Lambda_{kl})\big\} - K + 1}. \tag{2.10}$$

Because the second term on the right side of (2.10) involves only the $d_l$'s with $t_l < t_{k,j+1}$, this equation indeed defines a recursive formula starting with $d_{k1}, k = 1, \cdots, K$.

*Proof of Theorem 2.1*

Let $\mathbb{P}_n$ denote the empirical measure and $P$ the underlying probability measure. Denote $N_k(t) = I(\widetilde{T} \leq t, \xi\widetilde{D} = k), k = 0, 1, \cdots, K$, and $\widetilde{N}(t) = I(\widetilde{T} \leq t, \xi = 0)$. The proof consists of three major steps.

*Step 1.* We show that for large $n$, the NPMLE exists, or equivalently, $\widehat{\Lambda}_k(\tau) < \infty$. The log-likelihood function is

$$\begin{aligned} l_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \mathbb{P}_n \sum_{k=1}^{K} \int_0^\tau \left[ H_k \left\{ \int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda_k(s) \right\} + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) + \log \Lambda_k\{t\} \right] dN_k(t) \\ + \mathbb{P}_n \int_0^\tau \log \left( \sum_{k=1}^{K} \exp \left[ H_k \left\{ \int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda_k(s) \right\} + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) \right] \Lambda_k\{t\} \right) d\widetilde{N}(t) \\ + \mathbb{P}_n \int_0^\tau \log S(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) dN_0(t). \end{aligned} \tag{2.11}$$

Conditions (C2) and (C3) imply that, for large $n$ with probability one, there exists a subject with $\widetilde{T} = \tau$ and $\widetilde{D} = 0$. For this subject, if $\Lambda_k(\tau) = \infty$ for some $k$, then $S(\widetilde{T}; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \leq 0$ and the corresponding term in (2.11) is negative infinity. Thus, $\widehat{\Lambda}_k(\tau) < \infty$.

*Step 2.* Let $\mathcal{Z}$ be the support of $\boldsymbol{Z}$ equipped with the uniform norm. We show that for every $\boldsymbol{z} \in \mathcal{Z}$, with probability one,

$$\liminf_n S(\tau; \boldsymbol{z}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) > 0. \tag{2.12}$$

Note that (2.12) implies that $\limsup_n \widehat{\Lambda}_k(\tau) < \infty$ almost surely for each $k$.

Suppose that $\liminf_n S(\tau; \boldsymbol{z}_0, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \leq 0$ for some $\boldsymbol{z}_0$. By the continuity of $S$ in $\boldsymbol{z}$, there exists a small neighborhood $\mathcal{Z}_0 \subset \mathcal{Z}$ of $\boldsymbol{z}_0$ with $\Pr(\boldsymbol{z} \in \mathcal{Z}_0) > 0$ such that

$$\liminf_n \sup_{\boldsymbol{z} \in \mathcal{Z}_0} S(\tau; \boldsymbol{z}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \leq 0.$$

We take a subsequence, still indexed by $n$, such that $\lim_n \sup_{\boldsymbol{z} \in \mathcal{Z}_0} S(\tau; \boldsymbol{z}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \leq 0$. Choose $M > 0$ such that $\sup_{t \in [0,\tau]} |\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{z}(t)| \leq M$ for every $\boldsymbol{\beta}_k$ and $\boldsymbol{z} \in \mathcal{Z}$, and choose $\epsilon_0 \in (0, K^{-1})$. Define $\overline{\Lambda}_k(t) = (\widehat{\Lambda}_k(t) \wedge \widetilde{M}_k) \vee \widetilde{M}_k/2$, where $\widetilde{M}_k = e^{-M} G_k^{-1}\big\{ -\log(1 - K^{-1} + \epsilon_0)\big\} > 0$. Since $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Lambda}}$ are the NPMLEs,

$$l_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \geq l_n(\widehat{\boldsymbol{\beta}}, \overline{\boldsymbol{\Lambda}}), \tag{2.13}$$

where $\overline{\boldsymbol{\Lambda}} = (\overline{\Lambda}_1. \cdots, \overline{\Lambda}_K)$. To derive a contradiction, we will show that the left side of (2.13) goes to $-\infty$ and the right side is bounded away from $-\infty$. To this end, we will use the following inequalities

$$K^{-1}\sum_{k=1}^{K}\log a_k \le \log\left(\sum_{k=1}^{K} a_k\right) \le \sum_{k=1}^{K}\log a_k + \log K,$$

where $a_1, \cdots, a_K$ are any positive constants. Clearly,

$$\begin{aligned} l_n(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \le & \sum_{k=1}^{K}\left[\log\left\{\widehat{\Lambda}_k(\tau)\sup_{y\le\widehat{\Lambda}_k(\tau)e^M} G_k'(y)\right\} + M\right]\mathbb{P}_n\left\{N_k(\tau)+\widetilde{N}(\tau)\right\} + \log K\mathbb{P}_n\widetilde{N}(\tau) \\ & + \log\sup_{\boldsymbol{z}\in\mathcal{Z}_0} S(\tau;\boldsymbol{z},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})\mathbb{P}_n(\widetilde{T}=\tau,\widetilde{D}=0,\boldsymbol{Z}\in\mathcal{Z}_0). \end{aligned}$$

By (C4), we can show that the right side goes to $-\infty$. On the other hand,

$$\begin{aligned} l_n(\widehat{\boldsymbol{\beta}}, \overline{\boldsymbol{\Lambda}}) \ge & \sum_{k=1}^{K}\left[\log\left\{2^{-1}\widetilde{M}_k\inf_{y\le\widetilde{M}_k e^M} G_k'(y)\right\} - G_k(\widetilde{M}_k e^M) - M\right]\mathbb{P}_n\{N_k(\tau)+K^{-1}\widetilde{N}(\tau)\} \\ & + \log\epsilon_0\mathbb{P}_n(\widetilde{T}=\tau,\widetilde{D}=0), \end{aligned}$$

which is bounded away from $-\infty$. Thus, we obtain a contradiction, so (2.12) holds. It then follows from Helly's selection lemma that, along a subsequence, $\widehat{\Lambda}_k(t)\to\Lambda_k^*(t)$ weakly for some increasing function $\Lambda_k^*(t)$ and $\widehat{\boldsymbol{\beta}}\to\boldsymbol{\beta}^*$ for some vector $\boldsymbol{\beta}^*$.

*Step 3.* We show that $\Lambda_k^*=\Lambda_{k0}$ and $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$. Define $\Lambda_k^\epsilon(t)=\int_0^t\{1+\epsilon h_k(s)\}d\Lambda_k(s)$ as a path through $\Lambda_k$ indexed by $\epsilon$, where $h_k\in BV_1$. By differentiating the log-likelihood of $\Lambda_k^\epsilon$ for a single subject with respect to $\epsilon$ at 0, we obtain the score operator for $\Lambda_k$ as

$$\begin{aligned} \dot{l}_{2k}(\boldsymbol{\beta},\boldsymbol{\Lambda})[h_k] = & \int_0^\tau\left[H_k'\left\{\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}h_k(s)d\Lambda_k(s)+h_k(t)\right]d\widetilde{N}_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) \\ & -\int_0^\tau\left\{\widetilde{\Psi}_k(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\int_0^t h_k(s)e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}dN_0(t), \qquad (2.14) \end{aligned}$$

where $\widetilde{N}_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda})=N_k(t)+w_k(\boldsymbol{\beta},\boldsymbol{\Lambda})\widetilde{N}(t)$, $w_k(\boldsymbol{\beta},\boldsymbol{\Lambda})=\left\{\sum_{l=1}^K F_l'(\widetilde{T};\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\right\}^{-1}F_k'(\widetilde{T};\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})$, and $\widetilde{\Psi}_k(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})=S^{-1}(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}\right]$.

By changing the order of integrations, we re-write (2.14) as

$$\begin{aligned}\dot{l}_{2k}(\boldsymbol{\beta},\boldsymbol{\Lambda})[h_k] = &\int_0^\tau h_k(s)d\widetilde{N}_k(s;\boldsymbol{\beta},\boldsymbol{\Lambda}) \\ &+\int_0^\tau h_k(s)e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}\int_s^\tau H_k'\left\{\int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}d\widetilde{N}_k(s;\boldsymbol{\beta},\boldsymbol{\Lambda})d\Lambda_k(s) \\ &-\int_0^\tau h_k(s)e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}\int_s^\tau \widetilde{\Psi}_k(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})dN_0(t)d\Lambda_k(s).\end{aligned}$$

By definition of the NPMLEs, $\mathbb{P}_n\dot{l}_{2k}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[h_k] = 0$ for all $h_k \in BV_1$. We take $h_k(\cdot) = I(\cdot \le t)$ to obtain

$$\widehat{\Lambda}_k(t) = \int_0^t \frac{\mathbb{P}_n d\widetilde{N}_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})}{\phi_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})},$$

where

$$\begin{aligned}\phi_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) = &\mathbb{P}_n e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}\int_t^\tau \widetilde{\Psi}_k(s;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})dN_0(s) \\ &-\mathbb{P}_n e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}\int_t^\tau H_k'\left\{\int_0^s e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(u)}d\Lambda_k(u)\right\}d\widetilde{N}_k(s;\boldsymbol{\beta},\boldsymbol{\Lambda}).\end{aligned}$$

By Step 2 and the continuity of $S(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})$ in $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$, and with $\boldsymbol{\Lambda}$ equipped with the weak topology, there exist a neighborhood of $\boldsymbol{\beta}^*$, denoted by $\mathcal{B}$, and a neighborhood of $\Lambda_k^*$, denoted by $\mathcal{A}_k$, such that $S(t;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})$ is uniformly bounded away from zero. Therefore, $\{\widetilde{\Psi}_k(\cdot;\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\Lambda}) : \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{\beta} \in \mathcal{B}, \Lambda_k \in \mathcal{A}_k, k = 1,\cdots,K\}$ is a class of functions on $[0,\tau]$ that are uniformly bounded and with total variation uniformly bounded and is thus Donsker (van der Vaart and Wellner, 1996, chapter 2.10). By the permanence of the Donsker property and the uniform law of large numbers,

$$\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B},\Lambda_k\in\mathcal{A}_k,k=1,\cdots,K} |\phi_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) - \phi_k^*(t;\boldsymbol{\beta},\boldsymbol{\Lambda})| \to 0, \tag{2.15}$$

where

$$\phi_k^*(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) = Pe^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}\int_t^\tau \widetilde{\Psi}_k(s;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})dN_0(s) \\ - Pe^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}\int_t^\tau H_k'\left\{\int_0^s e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(u)}d\Lambda_k(u)\right\}d\widetilde{N}_k(s;\boldsymbol{\beta},\boldsymbol{\Lambda}).$$

By (2.15) and the continuity of $\phi_k^*$ in $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$, we have $\phi_k(\cdot;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) \to \phi_k^*(\cdot;\boldsymbol{\beta}^*,\boldsymbol{\Lambda}^*)$ uniformly. We can show that for large $n$, $\phi_k(t;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ is uniformly bounded away from $0$. Furthermore, $\mathbb{P}_n d\widetilde{N}_k(\cdot;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) = \mathbb{P}_n d\widetilde{N}_k(\cdot;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) + o_p(1)$ uniformly by similar Donsker property arguments.

We define

$$\widetilde{\Lambda}_k(t) = \int_0^t \frac{\mathbb{P}_n d\widetilde{N}_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)}{\phi_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)}.$$

It follows from the uniform law of large numbers and the MAR assumption (C5) that uniformly

$$\widetilde{\Lambda}_k(t) \to \int_0^t \frac{Pd\widetilde{N}_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)}{\phi_k^*(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)} = \Lambda_{k0}(t).$$

By the lower bound on $\phi_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$, $\widehat{\Lambda}_k$ is absolutely continuous with respect to $\widetilde{\Lambda}_k$. Furthermore, $d\widehat{\Lambda}_k/d\widetilde{\Lambda}_k$ converges uniformly to a bounded function $\eta(\cdot)$. Thus, $\Lambda_k^*(t) = \int_0^t \eta(s)d\Lambda_{k0}(s)$, and $\Lambda_k^*(t)$ is absolutely continuous with respect to the Lebesgue measure. We denote the derivatives of $\Lambda_k^*(t)$ and $\Lambda_{k0}(t)$ by $\lambda_k^*(t)$ and $\lambda_{k0}(t)$, respectively. By the uniform convergence of the log-likelihood to its expectation and that of the function $d\widehat{\Lambda}_k/d\widetilde{\Lambda}_k$ to $\lambda_k^*/\lambda_{k0}$, together with the

Kullback-Leibler criterion,

$$
\begin{aligned}
&\sum_{k=1}^{K}\int_0^\tau \log\left(\lambda_k^*(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k^*(s)\right\}+\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(t)\right]\right)dN_k(t)\\
&\int_0^\tau \log\left(\sum_{k=1}^{K}\lambda_k^*(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k^*(s)\right\}+\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(t)\right]\right)d\widetilde{N}(t)\\
&\qquad+\int_0^\tau \log S(t;\boldsymbol{Z},\boldsymbol{\beta}^*,\boldsymbol{\Lambda}^*)dN_0(t)\\
&=\sum_{k=1}^{K}\int_0^\tau \log\left(\lambda_{k0}(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s)\right\}+\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)\right]\right)dN_k(t)\\
&\int_0^\tau \log\left(\sum_{k=1}^{K}\lambda_{k0}(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s)\right\}+\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)\right]\right)d\widetilde{N}(t)\\
&\qquad+\int_0^\tau \log S(t;\boldsymbol{Z},\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)dN_0(t)
\end{aligned}
$$

almost surely. In the case that $\widetilde{T}=t$ and $\xi\widetilde{D}=k$,

$$
\begin{aligned}
&\lambda_k^*(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k^*(s)\right\}+\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(t)\right]\\
&\qquad\qquad=\lambda_{k0}(t)\exp\left[H_k\left\{\int_0^t e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s)\right\}+\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)\right].
\end{aligned}
$$

We integrate both sides to obtain $e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(t)}\lambda_k^*(t)=e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)}\lambda_{k0}(t)$. It then follows from (C2) that $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$ and $\Lambda_k^*(t)=\Lambda_{k0}(t)$. Thus, with probability one, $\widehat{\boldsymbol{\beta}}\to\boldsymbol{\beta}_0$ and $\widehat{\Lambda}_k(t)\to\Lambda_{k0}(t)$ pointwise. The latter can be strengthened to uniform convergence since $\Lambda_{k0}$ is continuous.

*Proof of Theorem 2.2*

We denote the empirical process by $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. The score operator for $\Lambda_k$ is given in (2.14). The score function for $\boldsymbol{\beta}$ is $\dot{\boldsymbol{l}}_1 \equiv (\dot{\boldsymbol{l}}_{11}^{\mathrm{T}}, \cdots, \dot{\boldsymbol{l}}_{1K}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\dot{\boldsymbol{l}}_{1k}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \int_0^\tau H_k' \left\{ \int_0^t e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda_k(s) \right\} \int_0^t \boldsymbol{Z}(s) e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda_k(s) d\widetilde{N}_k(t; \boldsymbol{\beta}, \boldsymbol{\Lambda}) + \int_0^\tau \boldsymbol{Z}(t) d\widetilde{N}_k(t; \boldsymbol{\beta}, \boldsymbol{\Lambda}) - \int_0^\tau \widetilde{\Psi}_k(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \int_0^t \boldsymbol{Z}(s) e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda_k(s) dN_0(t).$$

For $\delta > 0$ sufficiently small, the class of functions

$$\left\{ \dot{\boldsymbol{l}}_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}), \dot{l}_{2k}(\boldsymbol{\beta}, \boldsymbol{\Lambda})[h_k] : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| + \sum_{k=1}^K \sup_{t \in [0,\tau]} |\Lambda_k(t) - \Lambda_{k0}(t)| < \delta, h_k \in BV_1, k = 1, \cdots, K \right\}$$

is Donsker. Thus, by the consistency of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$, the continuity of the score functions in the parameters, and the dominated convergence theorem,

$$\mathbb{G}_n \Big\{ \boldsymbol{v}^{\mathrm{T}} \dot{\boldsymbol{l}}_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) + \sum_{k=1}^K \dot{l}_{2k}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})[h_k] \Big\} = \mathbb{G}_n \Big\{ \boldsymbol{v}^{\mathrm{T}} \dot{\boldsymbol{l}}_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) + \sum_{k=1}^K \dot{l}_{2k}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[h_k] \Big\} + o_p(1) \quad (2.16)$$

uniformly in $(\boldsymbol{v}, \boldsymbol{w})$. It remains to show that the map $W : l^\infty(\mathcal{V} \times \mathcal{W}) \to l^\infty(\mathcal{V} \times \mathcal{W})$ given by

$$W(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{v}, \boldsymbol{w}] = P \Big\{ \boldsymbol{v}^{\mathrm{T}} \dot{\boldsymbol{l}}_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) + \sum_{k=1}^K \dot{l}_{2k}(\boldsymbol{\beta}, \boldsymbol{\Lambda})[h_k] \Big\}$$

is Fréchet differentiable at $(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$ with a derivative that is continuously invertible.

It is straightforward to show that

$$\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} W \left( \boldsymbol{\beta}_0 + \epsilon \widetilde{\boldsymbol{v}}, \boldsymbol{\Lambda}_0 + \epsilon \int \widetilde{\boldsymbol{w}} d\boldsymbol{\Lambda}_0 \right) = \widetilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{B}_1[\boldsymbol{v}, \boldsymbol{w}] + \sum_{k=1}^K \int B_{2k}[\boldsymbol{v}, \boldsymbol{w}] \tilde{h}_k d\Lambda_{k0}, \quad (2.17)$$

where the operator $(\boldsymbol{B}_1, B_{21}, \ldots, B_{2K})[\boldsymbol{v}, \boldsymbol{w}]$ can be expressed as

$$
-\begin{pmatrix} \boldsymbol{v} \\ \phi_1^*(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)h_1(t) \\ \vdots \\ \phi_K^*(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)h_K(t) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\zeta}_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\boldsymbol{v} + \sum_{k=1}^K \int h_k(t)\boldsymbol{\nu}_{1k}(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)d\Lambda_{k0}(t) + \boldsymbol{v} \\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\zeta}_{21}(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) + \sum_{j=1}^K \int h_j(s)\nu_{21j}(s,t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)d\Lambda_{j0}(s) \\ \vdots \\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\zeta}_{2K}(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) + \sum_{j=1}^K \int h_j(s)\nu_{2Kj}(s,t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)d\Lambda_{j0}(s) \end{pmatrix}, \tag{2.18}
$$

$\boldsymbol{\zeta}_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) = -P\{\dot{\boldsymbol{l}}_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\dot{\boldsymbol{l}}_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)^{\mathrm{T}}\}$, and $\boldsymbol{\zeta}_{2k}, \boldsymbol{\nu}_{1k}$ and $\nu_{2kj}$ are certain functions. We show that the operator $\boldsymbol{B} \equiv (\boldsymbol{B}_1, B_{21}, \ldots, B_{2K})$ is invertible on its range.

In light of Theorem 2.1,

$$
\phi_k^*(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) = \frac{Pd\widetilde{N}_k(t;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)/dt}{\lambda_{k0}(t)} > 0
$$

under Conditions (C1) and (C3). Thus, the first term in (2.18) is an invertible operator. Because the second term is a compact operator, it suffices to show that the operator $\boldsymbol{B}$ is one-to-one (Rudin, 1973, pages 99-103). Suppose that for some $(\boldsymbol{v},\boldsymbol{w}) \in \mathcal{V}\times\mathcal{W}$, $\boldsymbol{B}(\boldsymbol{v},\boldsymbol{w}) = \boldsymbol{0}$. We then wish to show that $(\boldsymbol{v},\boldsymbol{w}) = \boldsymbol{0}$. By (2.17),

$$
P\left(\boldsymbol{v}^{\mathrm{T}}\dot{\boldsymbol{l}}_1(\boldsymbol{\beta},\boldsymbol{\Lambda}) + \sum_{k=1}^K \dot{l}_{2k}(\boldsymbol{\beta},\boldsymbol{\Lambda})[h_k]\right)^2 = -\left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0} W\left(\boldsymbol{\beta}_0 + \epsilon\boldsymbol{v}, \boldsymbol{\Lambda}_0 + \epsilon\int \boldsymbol{w}d\boldsymbol{\Lambda}_0\right) = 0.
$$

Thus, with probability one,

$$
\boldsymbol{v}^{\mathrm{T}}\dot{\boldsymbol{l}}_1(\boldsymbol{\beta},\boldsymbol{\Lambda}) + \sum_{k=1}^K \dot{l}_{2k}(\boldsymbol{\beta},\boldsymbol{\Lambda})[h_k] = 0. \tag{2.19}
$$

Let $\boldsymbol{v} = (\boldsymbol{v}_1^{\mathrm{T}}, \cdots, \boldsymbol{v}_K^{\mathrm{T}})^{\mathrm{T}}$, and take $dN_k(t) = 1$. It follows from (2.19) that

$$
h_k(t) + \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{Z}(t) = -\left[\int_0^t \left\{h_k(s) + \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{Z}(s)\right\} e^{\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{Z}(s)} d\Lambda_{k0}(s)\right] H_k'\left\{\int_0^t e^{\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{Z}(s)} d\Lambda_{k0}(s)\right\},
$$

which is a homogeneous integral equation of $h_k(t) + \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{Z}(t)$ with $0$ as the only solution. Thus, it follows from (C2) that $\boldsymbol{v}_k = \mathbf{0}$ and $h_k(\cdot) = 0$. Therefore, $\boldsymbol{B}$ is one-to-one and thus invertible. Consequently, the derivative of $W$ is continuously invertible.

For $(\boldsymbol{v}, \boldsymbol{w}) \in \mathcal{V} \times \mathcal{W}$, denote $(\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{w}}) = \boldsymbol{B}^{-1}(\boldsymbol{v}, \boldsymbol{w})$. It follows from (2.16) that

$$\sqrt{n}\Big\{\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \sum_{k=1}^{K}\int_0^\tau h_k d(\widehat{\Lambda}_k - \Lambda_{k0})\Big\} = -\mathbb{G}_n\Big\{\widetilde{\boldsymbol{v}}^{\mathrm{T}}\dot{\boldsymbol{l}}_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) + \sum_{k=1}^{K}\dot{l}_{2k}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\widetilde{h}_k]\Big\} + o_p(1) \tag{2.20}$$

uniformly in $(\boldsymbol{v}, \boldsymbol{w})$. Thus, $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0)$ is asymptotically Gaussian. Take $h_k = 0$ and $\boldsymbol{v}$ to be the unit coordinate vectors in (2.20) to find that the influence function of $\widehat{\boldsymbol{\beta}}$ lies in the tangent space, i.e., the closed linear span of the score functions. By the semiparametric efficiency theory (Bickel et al., 1993), $\widehat{\boldsymbol{\beta}}$ is semiparametric efficient.

*Profile likelihood and information matrix*

We justify the use of the negative inverse of the second derivative of the profile log-likelihood to estimate the covariance matrix of $\widehat{\boldsymbol{\beta}}$ by verifying the conditions in Theorem 1 of Murphy and van der Vaart (2000). From the proof of Theorem 2.2, the invertibility of the whole information operator implies the invertibility of the information operator for $\boldsymbol{\Lambda}$. This ensures that there is a "least favorable direction" $\boldsymbol{h}_k$, which is a vector of dimension $p$ with components in $BV_1$ such that the parametric model $\boldsymbol{\epsilon} \to \big(\boldsymbol{\epsilon}, (\Lambda_{1\boldsymbol{\epsilon}}, \cdots, \Lambda_{K\boldsymbol{\epsilon}})\big)$ with $d\Lambda_{k\boldsymbol{\epsilon}} = \big\{1 + (\boldsymbol{\epsilon} - \boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{h}_k\big\}d\Lambda_{k0}$ is a least favorable submodel. Given $\widetilde{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}_0$, let $\widehat{\boldsymbol{\Lambda}}_{\widetilde{\boldsymbol{\beta}}} \equiv \big(\widehat{\Lambda}_{1\widetilde{\boldsymbol{\beta}}}, \cdots, \widehat{\Lambda}_{K\widetilde{\boldsymbol{\beta}}}\big)$ denote the maximizer of $l_n(\widetilde{\boldsymbol{\beta}}, \cdot)$. We can show that

$$\sup_{t\in[0,\tau]}\sum_{k=1}^{K}\Big|\widehat{\Lambda}_{k\widetilde{\boldsymbol{\beta}}}(t) - \Lambda_{k0}(t)\Big| = O_p(||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| + n^{-1/2})$$

by using the arguments in the proofs of Theorems 1 and 2 and the smoothness of the likelihood in $\boldsymbol{\beta}$. Then the no-bias condition follows from the smoothness property of the score functions in the parameters. Finally, we can verify the Donsker properties of the first and second derivatives of the least favorable submodel log-likelihood since functions of uniformly bounded variation are Donsker.

For the use of the inverse information matrix for $\boldsymbol{\beta}$ and $d_{kj}$'s as an estimator of the covariance matrix of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Lambda}}$, the justification is similar to Theorem 3 of Zeng and Lin (2007). We note that the first and second derivatives of our log-likelihood are smooth on a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$.

*Technical details of model checking procedures*

Using the (functional) delta method, we will show that $W_{kn}(\boldsymbol{x}, t)$ is asymptotically equivalent to

$$\widetilde{W}_{kn}(\boldsymbol{x}, t) = \mathbb{G}_n \Bigg\{ \int_0^t f(\boldsymbol{x}, u; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) dM_k(u; \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) + (\dot{\boldsymbol{l}}_1, \dot{\boldsymbol{l}}_2)\boldsymbol{B}^{-1} P\Big(\boldsymbol{g}_{k11}(\boldsymbol{x}, t), \cdots, \boldsymbol{g}_{k1K}(\boldsymbol{x}, t), g_{k21}(\cdot, \boldsymbol{x}, t), \cdots, g_{k2K}(\cdot, \boldsymbol{x}, t)\Big) \Bigg\} \tag{2.21}$$

uniformly in $\boldsymbol{x}$ and $t$, where $\dot{\boldsymbol{l}}_1$ is the score function for $\boldsymbol{\beta}$, $\dot{\boldsymbol{l}}_2$ is the score operator for $\boldsymbol{\Lambda}$, and $\boldsymbol{B}$ is the information operator, all evaluated at $(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$,

$$\begin{aligned} \boldsymbol{g}_{k1j}(\boldsymbol{x}, t) = &- \int_0^t f(\boldsymbol{x}, s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) Y(s) \widetilde{\Psi}_j(s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) \int_0^s \boldsymbol{Z}(u) e^{\boldsymbol{\beta}_{j0}^{\mathrm{T}} \boldsymbol{Z}(u)} d\Lambda_{j0}(u) \\ &\times \Psi_k(s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) d\Lambda_{k0}(s) - I(j = k) \int_0^t f(\boldsymbol{x}, s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) \Bigg[ \boldsymbol{Z}(s) + \\ &\times H_k' \left\{ \int_0^s e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}(u)} d\Lambda_{k0}(u) \right\} \int_0^s \boldsymbol{Z}(u) e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}(u)} d\Lambda_{k0}(u) \Bigg] Y(s) \Psi_k(s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) d\Lambda_{k0}(s), \end{aligned}$$

and

$$g_{k2j}(u, \boldsymbol{x}, t) = -\int_0^t I(s \geq u) f(\boldsymbol{x}, s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) Y(s) \Psi_k(s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) \exp\{\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}(u)\}$$
$$\times \left[ \widetilde{\Psi}_j(s; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) + I(j = k) H_k' \left\{ \int_0^s e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}(\tilde{u})} d\Lambda_{k0}(\tilde{u}) \right\} \right] d\Lambda_{k0}(s)$$
$$- I(j = k) I(u \leq t) f(\boldsymbol{x}, u; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) Y(u) \Psi_k(u; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0).$$

We replace the unknown quantities in $\widetilde{W}_{kn}$ by their empirical counterparts. Specifically, we estimate the functions $\boldsymbol{g}_{k1j}$ and $g_{k2j}$ by replacing $(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$ with $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$. Denote the resulting expressions by $\widehat{\boldsymbol{g}}_{k1j}$ and $\widehat{g}_{k2j}$. Recall from Section 2.2 that $t_1, \cdots, t_m$ are the distinct failure times and $\delta_1, \cdots, \delta_m$ are the corresponding failure types. We treat the jump sizes of $\boldsymbol{\Lambda}$ at the $t_j$'s as Euclidean parameters, and, along with $\boldsymbol{\beta}$, we calculate the score vector for the $i$th subject, denoted by $\widetilde{\boldsymbol{l}}_i$, and the information matrix $\mathcal{I}_n$. Let $\widetilde{\boldsymbol{g}}_{k1n}(\boldsymbol{x}, t) = \mathbb{P}_n\big(\widehat{\boldsymbol{g}}_{k11}(\boldsymbol{x}, t)^{\mathrm{T}}, \cdots, \widehat{\boldsymbol{g}}_{k1K}(\boldsymbol{x}, t)^{\mathrm{T}}\big)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{g}}_{k2n}(\boldsymbol{x}, t) = \mathbb{P}_n\big(\widehat{g}_{k2\delta_1}(t_1, \boldsymbol{x}, t), \cdots, \widehat{g}_{k2\delta_m}(t_m, \boldsymbol{x}, t)\big)^{\mathrm{T}}$. Also, let $\widehat{S}_i(\boldsymbol{x}, t) = \widetilde{\boldsymbol{l}}_i^{\mathrm{T}} \mathcal{I}_n^{-1} \big(\widetilde{\boldsymbol{g}}_{k1n}(\boldsymbol{x}, t)^{\mathrm{T}}, \widetilde{\boldsymbol{g}}_{k2n}(\boldsymbol{x}, t)^{\mathrm{T}}\big)^{\mathrm{T}}$. Then we obtain

$$\widehat{W}_{kn}(\boldsymbol{x}, t) = n^{-1/2} \sum_{i=1}^n \left\{ \int_0^t f(\boldsymbol{x}, u; \boldsymbol{Z}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) dM_{ki}(u; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) + \widehat{S}_i(\boldsymbol{x}, t) \right\} Q_i,$$

as given in Section 2.2.4.

Let $\mathcal{X}$ denote the space of $\boldsymbol{x}$, and consider $W_{kn}$ and $\widehat{W}_{kn}$ as random elements in $l^\infty(\mathcal{X} \times [0, \tau])$. In addition, let $BL_1$ be the space of Lipschitz functions on $l^\infty(\mathcal{X} \times [0, \tau])$ that are uniformly bounded by 1 and with Lipschitz norm bounded by 1. It is convenient to metrize the laws on $l^\infty(\mathcal{X} \times [0, \tau])$ by $\rho(\mathbb{Z}_1, \mathbb{Z}_2) = \sup_{h \in BL_1} |Eh(\mathbb{Z}_1) - Eh(\mathbb{Z}_2)|$, where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are random elements in $l^\infty(\mathcal{X} \times [0, \tau])$ (van der Vaart and Wellner, 1996). We impose the following regularity conditions on the function $f(\boldsymbol{x}, t; \boldsymbol{\beta}, \boldsymbol{\Lambda})$, whose dependence on $\boldsymbol{Z}$ is suppressed for notational simplicity.

(D1) For some $\delta > 0$, the class of functions

$$\left\{ f(\boldsymbol{x}, t; \boldsymbol{\beta}, \boldsymbol{\Lambda}) : \boldsymbol{x} \in \mathcal{X}, t \in [0, \tau], ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| + \sum_{k=1}^{K} \sup_{s \in [0,\tau]} |\Lambda_k(s) - \Lambda_{k0}(s)| < \delta \right\}$$

is a uniformly bounded $P$-Donsker class.

(D2) There exists a constant $M > 0$ such that, with probability one, the total variation of $f(\boldsymbol{x}, \cdot; \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$ is bounded by $M$ for all $\boldsymbol{x} \in \mathcal{X}$.

(D3) For all $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ such that $\boldsymbol{\beta} \to \boldsymbol{\beta}_0$ and $\boldsymbol{\Lambda} \to \boldsymbol{\Lambda}_0$,

$$\sup_{\boldsymbol{x},t} E|f(\boldsymbol{x}, t; \boldsymbol{\beta}, \boldsymbol{\Lambda}) - f(\boldsymbol{x}, t; \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)| \to 0.$$

*Remark 3.5.* Conditions (D1) and (D2) are satisfied by all processes considered in the main text. Condition (D3) is satisfied by $W_{kc}$, $W_{kp}$, and $W_{ko}$ and is also satisfied by $W_{kl}$ and $W_{ktr}$ if there is at least one continuous covariate.

*Theorem 3.3.* Under Conditions (C1)-(C4) and (D1)-(D3),

$$\sup_{h \in BL_1} |E_{\boldsymbol{Q}} h(\widehat{W}_{kn}) - E h(W_{kn})| \longrightarrow 0$$

almost surely, where $E_{\boldsymbol{Q}}$ denotes expectation with respect to $\boldsymbol{Q}$.

*Proof.* Our main task is to show that

$$W_{kn} = \widetilde{W}_{kn} + o_p(1) \qquad \text{in } l^{\infty}(\mathcal{X} \times [0, \tau]). \tag{2.22}$$

Then the conditional distribution of $\widehat{W}_{kn}$ can be shown to be asymptotically the same as the distribution of $\widetilde{W}_{kn}$ by using the uniform central limit theorem (van der Vaart and Wellner, 1996, Thm 2.11.1).

To show (2.22), we define

$$\Lambda_{kc}(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) = \int_0^t \Psi_k(s;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})d\Lambda_k(s).$$

Then $M_k(t;\boldsymbol{\beta},\boldsymbol{\Lambda}) = N_k(t) - \int_0^t Y(s)d\Lambda_{kc}(s;\boldsymbol{\beta},\boldsymbol{\Lambda})$. Clearly,

$$\begin{aligned} W_{kn} &= \mathbb{G}_n \int_0^t f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})dM_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) \\ &\quad + \sqrt{n}P\left[\int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)Y(s)d\{\Lambda_{kc}(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \Lambda_{kc}(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}\right] \\ &\quad + \sqrt{n}P\left[\int_0^t \{f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) - f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})\}Y(s)d\{\Lambda_{kc}(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \Lambda_{kc}(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}\right]. \end{aligned} \tag{2.23}$$

Because $\Lambda_{kc}(\cdot;\boldsymbol{\beta},\boldsymbol{\Lambda})$ is a Hadamard differentiable function of $(\boldsymbol{\beta},\boldsymbol{\Lambda})$, for almost every $\boldsymbol{Z}$, $\sqrt{n}\{\Lambda_{kc}(\cdot;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \Lambda_{kc}(\cdot;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}$ converges to a zero-mean Gaussian process on $[0,\tau]$. Then, by Conditions (D2) and (D3), the third term on the right side of (2.23) is $o_p(1)$. By the delta method and the linearization result on $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ given in the proof of Theorem 2, the second term is asymptotically linear in the second term on the right side of (2.21). The proof is complete if we can show that

$$\mathbb{G}_n \int_0^t f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})dM_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) = \mathbb{G}_n \int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)dM_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) + o_p(1)$$

uniformly in $\boldsymbol{x}$ and $t$. By Conditions (D1) and (D2), together with the permanence of the Donsker property, the class of functions

$$\left\{\int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta},\boldsymbol{\Lambda})dM_k(s;\boldsymbol{\beta},\boldsymbol{\Lambda}) : \boldsymbol{x}\in\mathcal{X}, t\in[0,\tau], ||\boldsymbol{\beta}-\boldsymbol{\beta}_0|| + \sum_{k=1}^K \sup_{s\in[0,\tau]} |\Lambda_k(s) - \Lambda_{k0}(s)| < \delta\right\}$$

is Donsker. Thus, it suffices to show that

$$\sup_{\boldsymbol{x},t} P\left(\int_0^t f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})dM_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)dM_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right)^2 \to 0 \qquad (2.24)$$

in probability. Note that

$$\begin{aligned}
&P\left|\int_0^t f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})dM_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)dM_k(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right| \\
&\leq \int_0^t P|f(\boldsymbol{x},s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)||dM_k(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})| \\
&\quad + P\left|\int_0^t f(\boldsymbol{x},s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)d\{\Lambda_{kc}(s;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - \Lambda_{kc}(s;\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}\right|. \qquad (2.25)
\end{aligned}$$

The second term on the right side of (2.25) is uniformly $o_p(1)$. For the first term, note that from Step 2 in the proof of Theorem 1, for every $\boldsymbol{Z}$, $\Psi_k(s;\boldsymbol{Z},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ is uniformly bounded as $n\to\infty$. Thus, the total variation of $M_k(\cdot;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ is uniformly bounded in $n$ for every $\boldsymbol{Z}$. Therefore, Condition (D3) implies that the first term is uniformly $o_p(1)$. Hence, (2.24) follows from the uniform boundedness of $f$. This completes the proof of (2.22).

To show that the conditional distribution of $\widehat{W}_{kn}$ is asymptotically equivalent to the distribution of $\widetilde{W}_{kn}$, we appeal to Theorem 2.11.1 of van der Vaart and Wellner (1996). With $\widehat{W}_{kn}$ playing the role of $\sum_{i=1}^{m_n} Z_{ni}$ in that theorem, where the $Q_i$ are the random quantities, it suffices to show that $\widehat{W}_{kn}$ converges to a Gaussian process indexed by $(\boldsymbol{x},t,\boldsymbol{\beta},\boldsymbol{\Lambda})$. The $\rho$ function in that theorem can be chosen to be the sum of the distances of the four components of the index. Then, the first and second displays of that theorem follow from the continuity of the influence function of $\widetilde{W}_{kn}$ in the index. The entropy integral condition (2.11.2) and the (almost sure) convergence of the covariance function are direct consequences of $\widehat{W}_{kn}$ being a multiplier Gaussian process.

We state below the consistency results on the supremum tests. We omit the proofs, which can be obtained by extending the arguments of Chen et al. (2012). As in Chen et al. (2012), we assume that covariates are time-independent.

1. *Omnibus tests.* The results in (i) and (ii) pertain to the goodness-of-fit tests for a particular risk and all risks, respectively.

   (a) The test $\sup_{\boldsymbol{z},t} |W_{ko}(\boldsymbol{z}, t)|$ is consistent against any alternative hypothesis such that there do not exist $\boldsymbol{\beta}_k$ and $c > 0$ such that $\Lambda_k'(0; \boldsymbol{Z}) = c \exp(\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z})$.

   (b) The test $\max_{1 \leq k \leq K} \sup_{\boldsymbol{z},t} |W_{ko}(\boldsymbol{z}, t)|$ is consistent against any alternative hypothesis such that there do not exist $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ such that $\Lambda_k(t; \boldsymbol{Z}) = G_k(\exp(\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z})\, \Lambda_k(t))$ for all $k$, $t$, and $\boldsymbol{Z}$.

2. *Functional form of covariates.* Assume that the components of $\boldsymbol{Z}$ are independent. The test $\sup_{z,t} |W_{kc}^{(j)}(z, t)|$ is consistent against any alternative such that

$$\Lambda_k(t; \boldsymbol{Z}) = G_k(\exp(\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}^{(-j)}) g(Z_j) \Lambda_{k0}(t))$$

   for some $\boldsymbol{\beta}_{k0}$ and $\Lambda_{k0}$, where $\boldsymbol{Z}^{(-j)}$ is the covariate vector with $Z_j$ removed, and $g$ is not an exponential function.

3. *Link function.* Assume that for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, $E\{g(\exp(\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{Z}))| \exp(\boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{Z})\} = c_0 \exp(\boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{Z})$ for some $c_0 > 0$ implies that $g(x) = cx^a$ for some constants $a$ and $c$. Then the test $\sup_{x,t} |W_{kl}(x, t)|$ is consistent against any alternative that $\Lambda_k(t; \boldsymbol{Z}) = G_k(g(\exp(\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}))\, \Lambda_{k0}(t))$ for some $\boldsymbol{\beta}_{k0}$ and $\Lambda_{k0}$, where $g(x)$ is not a mononomial function in the form of $cx^a$.

4. *Proportionality.* Assume that $Z$ is binary and that $xG_k''(x)/G_k'(x) \neq -1$. Then the test $\sup_t |W_{kp}(t)|$ is consistent against any alternative such that $\Lambda_k(t; Z) = G_k(\exp(\beta_k(t) Z)\, \Lambda_k(t))$ with $\beta_k'(0) \neq 0$.

5. *Transformation function.* Assume that for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, $E\{g(\exp(\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{Z}))| \exp(\boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{Z})\}$ $= \exp(\boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{Z})$ implies that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Then the supremum test $\sup_{x,t} |W_{ktr}(x, t)|$ is consistent against any alternative such that $\Lambda_k(t; \boldsymbol{Z}) = G_{k0}(\exp(\boldsymbol{\beta}_{k0}^{\mathrm{T}} \boldsymbol{Z}) \Lambda_{k0}(t))$ for some $\boldsymbol{\beta}_{k0}$, $\Lambda_{k0}$, and $G_{k0}$, where $G_{k0}$ is different from the adopted transformation $G_k$.

# CHAPTER 3: INTERVAL-CENSORED COMPETING RISKS

## 3.1 Introduction

In clinical and epidemiological studies, the event of interest is often asymptomatic such that the event time or failure time cannot be exactly observed but is rather known to lie in an interval between two examination times. An additional complication arises when there are several distinct causes or types of failure. The resulting data are referred to as interval-censored competing risks. Such data are commonly encountered in HIV/AIDS research, where sero-conversion with different HIV-1 viral subtypes is determined through periodic blood tests (Hudgens et al., 2001). The data are also encountered in cancer clinical trials, where different types of adverse events may occur during reporting periods and the therapy and follow-up on each patient is terminated upon occurrence of any adverse event.

Because none of the failure times are observed exactly under interval censoring, it is much more challenging, both theoretically and computationally, to deal with interval-censored than right-censored data. In the nonparametric setting, Hudgens et al. (2001) adapted the self-consistency algorithm of Turnbull (1976) to compute the nonparametric maximum likelihood estimator (NPMLE). Jewell et al. (2003) studied the NPMLE and other estimators with current status data, where each subject is examined only once. Groeneboom et al. (2008a; 2008b) established rigorous asymptotic theory for the NPMLE with current status data. Li and Fine (2012) considered kernel-smoothed estimation in current status data. In the regression setting, Delord and Génin (2015) used multiple imputation based on sets of complete or right-censored data; the method is computationally intensive, especially for large datasets. Li (2016) proposed spline models to fit interval-censored data under the Fine and Gray (1999) model; however, the number and locations of knots are hard to choose in practice.

In this article, we consider a general class of semiparametric regression models for competing risks with potentially time-varying covariates. This class of models encompasses both proportional and non-proportional sub-distribution hazards structures. We study the NPMLEs for these models when there is a random sequence of examination times for each subject and the cause of failure information may be partially missing. We develop a fast and stable EM-type algorithm by extending the self-consistency formula of Turnbull (1976). We establish that, under mild conditions, the proposed estimators for the regression parameters are consistent and asymptotically normal. In addition, the estimators attain the semiparametric efficiency bound with a covariance matrix that can be consistently estimated by the profile likelihood method (Murphy and van der Vaart, 2000). The proofs involve careful use of modern empirical processes theory (van der Vaart and Wellner, 1996) and semiparametric efficiency theory (Bickel et al., 1993) to address unique challenges posed by the combination of interval censoring and competing risks. We evaluate the operating characteristics of the proposed numerical and inferential procedures through extensive simulation studies. Finally, we describe an application to a clinical study of HIV/AIDS in which a cohort of injecting drug users were followed for detection of sero-conversion with HIV-1 viral subtypes B and E.

### 3.2 Theory and Methods

#### 3.2.1 Data and Models

As in Chapter 2, denote $T$ and $D$ as the time and cause of failure, respectively, and let $\boldsymbol{Z}(\cdot)$ be the $p$-vector of (possibly time-dependent) covariates. We model the conditional sub-distribution hazard function for each risk using the same class of semiparametric transformation models specified in (2.3). We consider a general interval-censoring scheme under which each subject has an arbitrary sequence of examination times and the information on the cause of failure is possibly missing. Specifically, let $U_1 < U_2 < \ldots < U_J$ denote a random sequence of examination times, where $J$ is a random integer. Define $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_J)^{\mathrm{T}}$, where $\Delta_j = I(U_{j-1} < T \leq U_j)$

$(j = 1, \ldots, J)$, and $U_0 = 0$. Also, write $\widetilde{D} = DI(\boldsymbol{\Delta} \neq \mathbf{0})$. In addition, let $\xi$ indicate, by the values 1 versus 0, whether or not the cause of failure is observed. We set $\xi = 1$ if $\widetilde{D} = 0$. Then, the observed data for a random sample of $n$ subjects consist of $(J_i, \boldsymbol{U}_i, \boldsymbol{\Delta}_i, \xi_i, \xi_i\widetilde{D}_i, \boldsymbol{Z}_i)$ $(i = 1, \ldots, n)$, where $\boldsymbol{U}_i = (U_{i0}, U_{i1}, \ldots, U_{i,J_i})^{\mathrm{T}}$, and $\boldsymbol{\Delta}_i = (\Delta_{i1}, \ldots, \Delta_{i,J_i})^{\mathrm{T}}$.

### 3.2.2 Nonparametric Maximum Likelihood Estimation

Write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_K)$. We estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ by the nonparametric maximum likelihood approach. Suppose that $(T, D)$ is independent of $(\boldsymbol{U}, J)$ conditional on $\boldsymbol{Z}$ and that the cause of failure is missing at random (MAR). Then, the likelihood for $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ can be written as

$$
\begin{aligned}
L_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \prod_{i=1}^{n} \Bigg[ \prod_{j=1}^{J_i} \prod_{k=1}^{K} \Bigg( & \exp\left[-G_k\left\{\int_0^{U_{i,j-1}} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(t)} d\Lambda_k(t)\right\}\right] \\
& - \exp\left[-G_k\left\{\int_0^{U_{ij}} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(t)} d\Lambda_k(t)\right\}\right] \Bigg)^{I(\xi_i \widetilde{D}_i = k, \Delta_{ij} = 1)} \\
\times \Bigg\{ \sum_{k=1}^{K} \Bigg( & \exp\left[-G_k\left\{\int_0^{U_{i,j-1}} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(t)} d\Lambda_k(t)\right\}\right] \\
& - \exp\left[-G_k\left\{\int_0^{U_{ij}} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(t)} d\Lambda_k(t)\right\}\right] \Bigg) \Bigg\}^{I(\xi_i = 0, \Delta_{ij} = 1)} \\
\times \Bigg( \sum_{k=1}^{K} \exp & \left[-G_k\left\{\int_0^{U_{iJ_i}} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}_i(t)} d\Lambda_k(t)\right\}\right] - K + 1 \Bigg)^{I(\boldsymbol{\Delta}_i = \mathbf{0})} \Bigg],
\end{aligned}
$$

where the last term corresponds to the overall survival function $1 - \sum_{k=1}^{K} F_k(U_{i,J_i}; \boldsymbol{Z}_i)$. Let $(L_i, R_i]$ denote the interval among $(U_{i0}, U_{i1}], \ldots, (U_{i,J_i}, \infty]$ that contains $T_i$. Then, the above likelihood can be written as

$$\prod_{k=1}^{K}\prod_{i:\xi_i\widetilde{D}_i=k}\left(\exp\left[-G_k\left\{\int_0^{L_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]-\exp\left[-G_k\left\{\int_0^{R_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]\right)$$
$$\times\prod_{i:\xi_i=0}\left\{\sum_{k=1}^{K}\left(\exp\left[-G_k\left\{\int_0^{L_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]-\exp\left[-G_k\left\{\int_0^{R_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]\right)\right\}$$
$$\times\prod_{i:R_i=\infty}\left(\sum_{k=1}^{K}\exp\left[-G_k\left\{\int_0^{L_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]-K+1\right). \tag{3.26}$$

To maximize (3.26), we treat $\Lambda_k$ as a right-continuous step function that jumps at the right ends of the intervals. Specifically, let $t_{k1}<\ldots<t_{k,m_k}$ denote the distinct values of the $R_i$ with $\xi_i\widetilde{D}_i=k$ or $\xi_i=0$. In addition, let $\lambda_{kj}$ denote the jump size of $\Lambda_k$ at $t_{kj}$, and let $\boldsymbol{Z}_{ikj}=\boldsymbol{Z}_i(t_{kj})$. Then, the likelihood given in (3.26) becomes

$$\prod_{k=1}^{K}\prod_{i:\xi_i\widetilde{D}_i=k}\left[\exp\left\{-G_k\left(\sum_{t_{kj}\le L_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}-\exp\left\{-G_k\left(\sum_{t_{kj}\le R_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}\right]$$
$$\times\prod_{i:\xi_i=0}\left(\sum_{k=1}^{K}\left[\exp\left\{-G_k\left(\sum_{t_{kj}\le L_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}-\exp\left\{-G_k\left(\sum_{t_{kj}\le R_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}\right]\right)$$
$$\times\prod_{i:R_i=\infty}\left[\sum_{k=1}^{K}\exp\left\{-G_k\left(\sum_{t_{kj}\le L_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}-K+1\right]. \tag{3.27}$$

### 3.2.3 Numerical Algorithm

Direct maximization of (3.27) is difficult due to the high dimensionality of the $\lambda_{kj}$. This task is further complicated by the fact that, unlike the right-censoring case, the maximizers for some of the $\lambda_{kj}$ are zero and thus lie at the boundary of the parameter space. Herein, we propose a novel EM algorithm that extends the "self-consistency" formula of Turnbull (1972) for one-sample estimation with a single risk to regression analysis with competing risks.

Let $N_k(u,v]$ denote the number of event of the $k$th type that occurs in the interval $(u,v]$. For the $i$th subject with $R_i < \infty$, let $s_{ik1} < s_{ik2} < \ldots < s_{ik,j_{ik}}$ denote the distinct values of the $t_{kj}$ in the interval $(L_i, R_i]$, such that the interval $(L_i, R_i]$ is partitioned into a sequence of sub-intervals $(s_{ik0}, s_{ik1}], \ldots, (s_{ik,j_{ik}-1}, s_{ik,j_{ik}}]$, where $s_{ik0} = L_i$. In the "complete data", we know which sub-interval the failure time lies in, along with the cause of failure. Then, the complete-data log-likelihood takes the form

$$\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\sum_{j=1}^{j_{ik}} I(R_i < \infty) N_{ki}(s_{ik,j-1}, s_{ikj}] \log \Delta F_k(s_{ikj}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) + I(R_i = \infty) \log S(L_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})\right\},$$

where $S(t; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = 1 - \sum_{k=1}^{K} F_k(t; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k)$, $F_k(t; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) = 1 - \exp\{-\Lambda_k(t; \boldsymbol{Z})\}$, and $\Delta F_k(t; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)$ is the jump size of $F_k(\cdot; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k)$ at $t$.

In the M-step, we maximize

$$\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\sum_{j=1}^{j_{ik}} \widehat{w}_{ikj} \log \Delta F_k(s_{ikj}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) + I(R_i = \infty) \log S(L_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})\right\}, \tag{3.28}$$

where $\widehat{w}_{ikj}$ is the conditional probability that the $i$th subject experiences a failure of the $k$th cause in $(s_{ik,j-1}, s_{ikj}]$ given the subject's failure information. If $\xi_i \widetilde{D}_i = k'$, then

$$\widehat{w}_{ikj} = E\left\{N_{ki}(s_{ik,j-1}, s_{ikj}]\,\middle|\,N_{k'i}(L_i, R_i] = 1\right\} = I(k = k') \frac{\Delta F_k(s_{ikj}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}{\sum_{l=1}^{j_{ik}} \Delta F_k(s_{ikl}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}.$$

If $\xi_i = 0$, then

$$\widehat{w}_{ikj} = E\left\{N_{ki}(s_{ik,j-1}, s_{ikj}]\,\middle|\,\sum_{k'=1}^{K} N_{k'i}(L_i, R_i] = 1\right\} = \frac{\Delta F_k(s_{ikj}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)}{\sum_{k'=1}^{K}\sum_{l=1}^{j_{ik'}} \Delta F_{k'}(s_{ik'l}; \boldsymbol{Z}_i, \boldsymbol{\beta}_{k'}, \Lambda_{k'})}.$$

If $R_i = \infty$, then $\widehat{w}_{ikj} = 0$.

In the maximization of (3.28), we update the parameters using a one-step "self-consistency" type formula. Specifically, observe that the jump size of the sub-distribution function can be expressed as

$$\Delta F_k(t_{kj}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) = \widetilde{G}_k\left(\sum_{j'=1}^{j} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj'}}\lambda_{kj'}\right) e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\lambda_{kj},$$

where $\widetilde{G}_k(x) = \exp\{-G_k(x)\}G_k'(x)$. Here and in the sequel, $f'(t) = df(t)/dt$ for any function $f$. Then, the objective function in (3.28) can be written as

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m_k}\widehat{w}_{ijk}\left\{\log\lambda_{kj} + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj} + \log\widetilde{G}_k\left(\sum_{j'=1}^{j} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj'}}\lambda_{kj'}\right)\right\} + \sum_{i:R_i=\infty}\log S(L_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}). \tag{3.29}$$

We set the derivative of (3.29) with respect to $\lambda_{kj}$ to zero to obtain an updating formula for $\lambda_{kj}$

$$\lambda_{kj} = \left(\sum_{i=1}^{n}\widehat{w}_{ikj}\right)\Bigg[-\sum_{i=1}^{n}\sum_{j'=j}^{m_k}\widehat{w}_{ikj'}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj'}}\frac{\widetilde{G}_k'}{\widetilde{G}_k}\left(\sum_{j''=1}^{j'} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj''}}\lambda_{kj''}\right) + \sum_{i:R_i=\infty, L_i\geq t_{kj}} S(L_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Lambda})^{-1}\widetilde{G}_k\left(\sum_{t_{kj'}\leq L_i} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj'}}\lambda_{kj'}\right)e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\Bigg]^{-1}.$$

To update $\boldsymbol{\beta}$, we use a one-step Newton-Raphson algorithm based on (3.29).

We set the initial values of $\boldsymbol{\beta}$ to $\mathbf{0}$ and the initial value of $\lambda_{kj}$ to $n^{-1}$. We iterate between the E- and M-steps until the sum of the absolute differences of the parameter estimates between two successive iterations is less than a small number, say $10^{-3}$. To increase the chance of reaching the global maximum, we suggest to use a range of initial values for $\boldsymbol{\beta}$. The resulting estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ are denoted as $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Lambda}}$, respectively.

The proposed algorithm has several desirable features. First, the conditional expectations in the E-step have simple analytic forms. Second, the M-step involves only a single analytic update for $\lambda_{kj}$ and thus avoids large-scale optimization over high-dimensional parameters. Finally, the algorithm is applicable to any transformation functions.

### 3.2.4 Asymptotic Properties

We impose the following regularity conditions:

(C1) The true value for $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}_0$, is contained in a known compact set of the Euclidean space $\mathbb{R}^d$, where $d = Kp$. The true value for $\Lambda_k(\cdot)$, denoted as $\Lambda_{k0}(\cdot)$, is continuously differentiable with positive derivatives on $[\zeta, \tau]$ with $\Lambda_{k0}(\zeta) > 0$, where $[\zeta, \tau]$ is the union of the support of $(U_1, \ldots, U_J)$, and $0 < \zeta < \tau$. In addition, $\Pr(T \geq \tau | \boldsymbol{Z}) > c$ with probability one for some positive constant $c$.

(C2) The vector of covariates $\boldsymbol{Z}(t)$ is uniformly bounded with uniformly bounded total variation over $[\zeta, \tau]$, and its left-limits exist for all $t$. In addition, $E[g\{\boldsymbol{Z}_{(l)}(t)\}]$ $(l = 1, 2)$ is continuously differentiable in $[\zeta, \tau]$, where $g(\cdot)$ is any continuously differentiable function, and $\boldsymbol{Z}_{(1)}(t)$ and $\boldsymbol{Z}_{(2)}(t)$ are vectors of increasing functions whose difference is $\boldsymbol{Z}(t)$.

(C3) If $h(t) + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) = 0$ for all $t \in [\zeta, \tau]$ with probability one, then $h(t) = 0$ for $t \in [\zeta, \tau]$ and $\boldsymbol{\beta}_k = \mathbf{0}$.

(C4) The number of examination times, $J$, is positive, and $E(J) < \infty$. In addition, $\Pr(U_J = \tau | J, \boldsymbol{Z}) > \delta$ for some positive constant $\delta$, and there exists some positive constant $\eta$ such that $\Pr\{\min_{1 \leq j \leq J}(U_j - U_{j-1}) \geq \eta | J, \boldsymbol{Z}\} = 1$. Furthermore, the sum of the marginal density functions of $U_1, \ldots, U_J$ is strictly positive on $[\zeta, \tau]$. Finally, the conditional density function of $\boldsymbol{U}$ given $\boldsymbol{Z}$ and $J = j$, denoted as $g_j(u_1, \ldots, u_j | \boldsymbol{Z}, J = j)$, is strictly positive on $[\zeta, \tau]$ with continuous second-order partial derivatives with respect to $u_1, \ldots, u_j$ when $u_{j'} - u_{j'-1} > \eta$ $(j' = 2, \ldots, j)$ and is continuously differentiable with respect to $\boldsymbol{Z}$.

(C5) The transformation function $G_k$ is twice-continuously differentiable on $[0, \infty)$ with $G_k(0) = 0$, $G_k'(x) > 0$, and $G_k(\infty) = \infty$.

(C6) With probability one, $\Pr(\xi = 1 | T, D, \boldsymbol{Z}, J, \boldsymbol{U}, \boldsymbol{\Delta} \neq 0) = \Pr(\xi = 1 | \boldsymbol{Z}, J, \boldsymbol{U}, \boldsymbol{\Delta} \neq 0) > c_0$ for some positive constant $c_0$. Moreover, the conditional probability of $\xi = 1$ given $\boldsymbol{Z}$, $J = j$,

$\boldsymbol{U} = (u_1, \ldots, u_j)^{\mathrm{T}}$, and $\boldsymbol{\Delta} \neq 0$, denoted as $\rho_j(u_1, \ldots, u_j, \boldsymbol{Z})$, has continuous second-order partial derivatives with respect to $u_1, \ldots, u_j$ when $u_{j'} - u_{j'-1} > \eta$ $(j' = 2, \ldots, j)$ and is continuously differentiable with respect to $\boldsymbol{Z}$.

*Remark 3.1.* Condition (C1) pertains to standard assumptions on the parameter space. Condition (C2) places some smoothness structures on the covariate process $\boldsymbol{Z}(t)$. Condition (C3) is a standard assumption on the linear independence of covariates. Condition (C4) requires the distributions of the examination times be continuous. Condition (C5) is satisfied by the class of logarithmic transformations: $G_k(x) = r^{-1} \log(1 + rx)$ $(r \geq 0)$ and the class of Box-Cox transformations: $G_k(x) = \rho^{-1}\{(1 + x)^{\rho} - 1\}$ $(\rho \geq 0)$. Condition (C6) ensures that the MAR assumption holds for the cause of failure.

The following two theorems state the consistency and weak convergence of the proposed estimators.

*Theorem 3.1.* Under Conditions (C1)-(C6), $\widehat{\boldsymbol{\beta}}$ and $\widehat{\Lambda}_k$ $(k = 1, \ldots, K)$ are strongly consistent, i.e.,

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0| + \sup_{t \in [\zeta, \tau]} \sum_{k=1}^{K} \left| \widehat{\Lambda}_k(t) - \Lambda_{k0}(t) \right| \longrightarrow 0$$

almost surely, where $| \cdot |$ denotes the Euclidean norm.

*Theorem 3.2.* Under Conditions (C1)-(C6), $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges in distribution to a normal random vector with mean zero and with a covariance matrix that attains the semiparametric efficiency bound.

The proof of Theorem 3.1 relies on establishing a preliminary bound for $\widehat{\Lambda}_1(\tau), \ldots, \widehat{\Lambda}_K(\tau)$, which are constrained together by the overall survival function. The key to the proof of Theorem 3.2 is to show the existence of the least favorable sub-model for $\boldsymbol{\beta}$ by careful use of modern empirical process theory. An important intermediate step is to show the rate of convergence for $\widehat{\boldsymbol{\Lambda}}$ as $n^{-1/3}$.

The profile likelihood for $\boldsymbol{\beta}$ is defined as

$$pl_n(\boldsymbol{\beta}) = \sup_{\boldsymbol{\Lambda} \in \mathcal{C}} \log L_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}),$$

where $\mathcal{C}$ is the space of $\boldsymbol{\Lambda}$ in which $\Lambda_k$ is a step function with non-negative jumps at $t_{kj}$. We propose to estimate the covariance matrix of $\widehat{\boldsymbol{\beta}}$ by the negative inverse of the matrix whose $(j,k)$th element is

$$\frac{pl_n(\widehat{\boldsymbol{\beta}}) - pl_n(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_k) - pl_n(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_j) + pl_n(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_k + h_n \boldsymbol{e}_j)}{h_n^2},$$

where $\boldsymbol{e}_j$ is the $j$th canonical vector in $\mathbb{R}^d$, and $h_n$ is a constant of the order $n^{-1/2}$. The consistency of this estimator follows from the profile likelihood theory of Murphy and van der Vaart (2000). To evaluate $pl_n(\boldsymbol{\beta})$, we adopt the EM algorithm described in Section 2.2 by using $\widehat{\boldsymbol{\Lambda}}$ as the initial value for $\boldsymbol{\Lambda}$ and holding $\boldsymbol{\beta}$ fixed.

### 3.2.5 Reduced-Data Likelihood

We can estimate the parameters separately for each risk by using the reduced-data likelihood along the lines of Jewell et al. (2003) and Hudgens et al. (2014). Assume that there are no missing values on the causes of failure. In the reduced data, the subjects who experience the risk of interest and those who are right censored remain intact, whereas those who experience the other risks are treated as right censored at the last examination time $U_J$ (Hudgens et al. 2014). For the $k$th risk, the reduced data consist of $\{J_i, \boldsymbol{U}_i, I(\widetilde{D}_i = k)\boldsymbol{\Delta}_i, \boldsymbol{Z}_i\}$ $(i = 1, \ldots, n)$, and the corresponding likelihood for $(\boldsymbol{\beta}_k, \Lambda_k)$ is

$$\prod_{i=1}^{n}\Bigg\{\prod_{j=1}^{J_i}\Bigg(\exp\left[-G_k\left\{\int_0^{U_{i,j-1}} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right] - \exp\left[-G_k\left\{\int_0^{U_{ij}} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]\Bigg)^{I(\widetilde{D}_i=k)\Delta_{ij}} \times \exp\left[-G_k\left\{\int_0^{U_{iJ_i}} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_i(t)}d\Lambda_k(t)\right\}\right]^{I(\widetilde{D}_i\neq k)}\Bigg\},$$

which can be written as

$$\prod_{i:\widetilde{D}_i=k}\left[\exp\left\{-G_k\left(\sum_{t_{kj}\leq L_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\} - \exp\left\{-G_k\left(\sum_{t_{kj}\leq R_i}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}\right] \times \prod_{i:\widetilde{D}_i\neq k}\exp\left\{-G_k\left(\sum_{t_{kj}\leq U_{i,J_i}}\lambda_{kj}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}_{ikj}}\right)\right\}, \tag{3.30}$$

We can maximize (3.30) by adapting the algorithm of Sections 2.2. The resulting estimators are referred to as naive estimators.

With the exception of semiparametric efficiency, the asymptotic properties of the full-data NPMLEs carry over to the naive estimators by setting $K = 1$. The reason that efficiency is not attained by the naive estimators is because the reduced data do not contain all relevant information about the risk of interest. For example, the overall survival function, which combines the risk of interest with other risks in the full-data likelihood, is not included in the reduced-data likelihood.

In the reduced data for a particular risk, a subject who experiences a different risk is treated as right censored at the last potential examination time $U_J$. This examination time may be unknown if the available data consist only of $(L_i, R_i)$ $(i = 1, \ldots, n)$. It might be tempting to replace $U_{i,J_i}$ in (3.30) by $R_i$ if the $i$th subject experiences a different risk and $U_{i,J_i}$ is unavailable. However, the corresponding likelihood would involve parameters for other risks, such that the resulting naive estimators would be biased.

We have assumed that $\xi_i = 1$ for all $i = 1, \ldots, n$. When there are missing values on the causes of failure, it is natural to consider only complete cases, i.e., subjects with non-missing values. (Note that right-censored observations are complete cases because their causes of failure are naturally unknown.) The complete-case analysis, however, is generally biased. Suppose, for instance, that missingness is completely random (among subjects who are not right censored). Since shorter failure times are less likely to be right censored (than longer failure times) and thus more likely to be associated with missing causes and be discarded, the cumulative incidence will be underestimated. If the probability of right-censoring depends on covariates, then the naive estimator for the regression parameter will also be biased.

### 3.3 Simulation Studies

We carried out simulation studies to assess the performance of the NPMLE and naive methods in realistic settings. We let $Z_1(t) = B_1 I(t \leq V) + B_2 I(t > V)$ and $Z_2 \sim$ Unif$[0, 1]$, where $B_1$ and $B_2$ are independent Bernoulli$(0.5)$, and $V \sim$ Unif$[0, 3]$. In addition, we let $U_1$ and $U_2 - U_1$ be two independent random variables distributed as the minimum of 1.5 and an exponential random variable with hazard $e^{0.5Z_2 - 0.5}$. We considered $K = 2$ and $G_k(x) = r^{-1}\log(1 + rx)$ with $r = 0$, 1, and $0.5$. We set $\Lambda_1(t) = \Lambda_2(t) = 0.2(1 - e^{-t})$, $\boldsymbol{\beta}_1 = (0.25, -0.25)^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (-0.25, 0.25)^{\mathrm{T}}$. We first assumed that the causes of failure are completely observed. Under these conditions, the event rate for each cause was roughly $15\%$. We set the convergence threshold to $10^{-3}$ in the EM algorithm and $h_n = n^{-1/2}$ in the variance estimation. For both the NPMLE and the naive estimator, the algorithms converged in about 100 iterations.

The results for $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12})^{\mathrm{T}}$ are summarized in Table 3.13. For both the NPMLE and naive methods, the parameter estimators are virtually unbiased, and the standard error estimators reflect the true variability well. As a result, the empirical coverage probabilities of the confidence intervals are close to the nominal level. The NPMLE has smaller variance than the naive estimator. As shown in Figure 3.3, the NPMLE has little bias in estimating the cumulative hazard function,

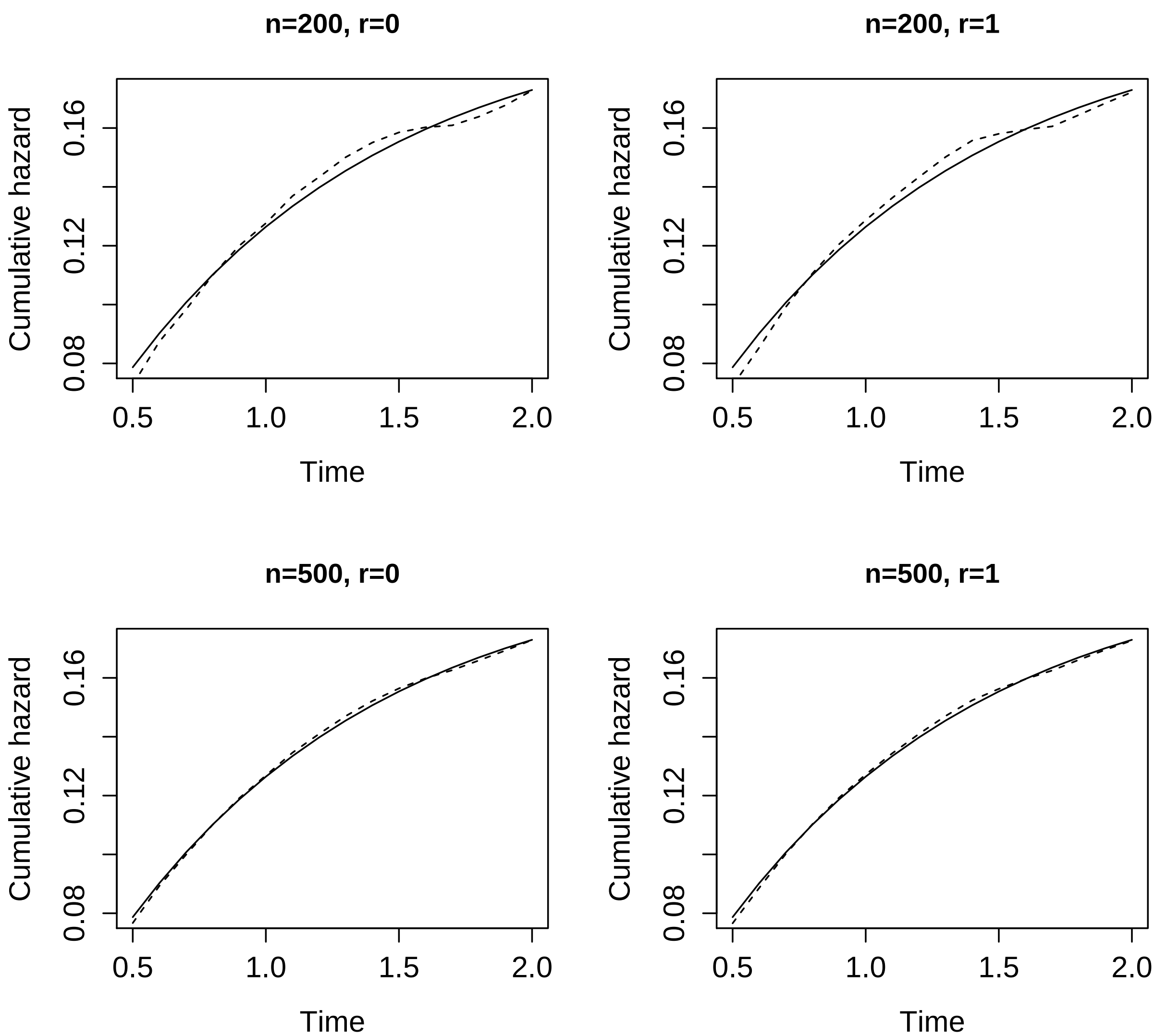


Figure 3.3: Estimation of the cumulative hazard function $\Lambda_1(\cdot)$ by the NPMLE. The true values and the mean estimates (based on 10,000 replicates) are shown by the solid and dashed curves, respectively.

Table 3.13: Simulation results on the estimation of $\boldsymbol{\beta}_1$ with complete data.

| | | | NPMLE | | | | Naive | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $r$ | | Bias | SE | SEE | CP | Bias | SE | SEE | CP | RE |
| 200 | 0 | $\beta_{11}$ | 0.012 | 0.297 | 0.299 | 0.954 | 0.009 | 0.308 | 0.311 | 0.957 | 1.076 |
| | | $\beta_{12}$ | −0.005 | 0.664 | 0.666 | 0.952 | −0.002 | 0.697 | 0.701 | 0.960 | 1.101 |
| | 0.5 | $\beta_{11}$ | 0.001 | 0.304 | 0.312 | 0.965 | −0.001 | 0.316 | 0.323 | 0.958 | 1.082 |
| | | $\beta_{12}$ | 0.006 | 0.671 | 0.678 | 0.962 | 0.005 | 0.698 | 0.701 | 0.956 | 1.081 |
| | 1 | $\beta_{11}$ | −0.009 | 0.307 | 0.313 | 0.959 | −0.009 | 0.318 | 0.322 | 0.954 | 1.076 |
| | | $\beta_{12}$ | 0.010 | 0.680 | 0.686 | 0.962 | 0.012 | 0.712 | 0.717 | 0.953 | 1.096 |
| 500 | 0 | $\beta_{11}$ | −0.005 | 0.184 | 0.186 | 0.953 | −0.011 | 0.197 | 0.198 | 0.953 | 1.136 |
| | | $\beta_{12}$ | −0.006 | 0.419 | 0.420 | 0.952 | 0.002 | 0.439 | 0.442 | 0.958 | 1.100 |
| | 0.5 | $\beta_{11}$ | 0.001 | 0.197 | 0.195 | 0.947 | −0.005 | 0.203 | 0.203 | 0.953 | 1.062 |
| | | $\beta_{12}$ | −0.005 | 0.426 | 0.427 | 0.950 | 0.001 | 0.441 | 0.440 | 0.948 | 1.070 |
| | 1 | $\beta_{11}$ | −0.002 | 0.192 | 0.190 | 0.946 | 0.003 | 0.202 | 0.206 | 0.954 | 1.107 |
| | | $\beta_{12}$ | −0.001 | 0.434 | 0.435 | 0.955 | 0.006 | 0.451 | 0.453 | 0.951 | 1.077 |

Bias and SE are the bias and standard error of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the variance of the naive estimator over that of the NPMLE. Each entry is based on 10,000 replicates.

especially when $n = 500$.

Next, we simulated missing values on the causes of failure by assuming that the probability of $\xi = 0$ is 0.3 given that the subject is not right censored. The results for the estimation of $\boldsymbol{\beta}_1$ by the NPMLE and the complete-case naive estimator are summarized in Table 3.14. The NPMLE continues to perform well. The naive estimator is substantially less efficient and is severely biased for $\beta_{12}$ as a result of the dependence of the right censoring time on $Z_2$.

## 3.4 An HIV/AIDS Study

The Bangkok Metropolitan Administration (BMA) conducted a prospective study on a cohort of 1,209 initially HIV-seronegative injecting drug users (IDUs) (Hudgens et al., 2001). This study was designed to investigate risk factors for HIV incidence and develop better prevention strategies. The subjects were followed from 1995 to 1998 at 15 BMA drug treatment clinics. Blood tests were conducted on each participant in approximately every 4 months post recruitment for evidence of HIV-1 sero-conversion (i.e., detection of HIV-1 antibodies in the serum). As of December 1998,

Table 3.14: Simulation results on the estimation of $\boldsymbol{\beta}_1$ with missing data.

| $n$ | $r$ | | NPMLE Bias | NPMLE SE | NPMLE SEE | NPMLE CP | Naive Bias | Naive SE | Naive SEE | Naive CP | RE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 200 | 0 | $\beta_{11}$ | 0.004 | 0.320 | 0.325 | 0.956 | 0.005 | 0.348 | 0.352 | 0.958 | 1.178 |
| | | $\beta_{12}$ | 0.013 | 0.704 | 0.707 | 0.955 | 0.116 | 0.770 | 0.777 | 0.926 | 1.197 |
| | 0.5 | $\beta_{11}$ | −0.003 | 0.329 | 0.330 | 0.953 | 0.007 | 0.357 | 0.358 | 0.952 | 1.178 |
| | | $\beta_{12}$ | 0.000 | 0.710 | 0.715 | 0.960 | 0.131 | 0.772 | 0.777 | 0.921 | 1.182 |
| | 1 | $\beta_{11}$ | −0.011 | 0.332 | 0.333 | 0.948 | −0.012 | 0.361 | 0.366 | 0.952 | 1.186 |
| | | $\beta_{12}$ | 0.000 | 0.716 | 0.722 | 0.956 | 0.123 | 0.773 | 0.775 | 0.929 | 1.167 |
| 500 | 0 | $\beta_{11}$ | 0.000 | 0.206 | 0.210 | 0.951 | 0.014 | 0.226 | 0.228 | 0.952 | 1.197 |
| | | $\beta_{12}$ | −0.001 | 0.448 | 0.448 | 0.950 | 0.135 | 0.485 | 0.488 | 0.821 | 1.172 |
| | 0.5 | $\beta_{11}$ | 0.004 | 0.212 | 0.211 | 0.948 | 0.006 | 0.224 | 0.226 | 0.954 | 1.117 |
| | | $\beta_{12}$ | −0.005 | 0.449 | 0.453 | 0.954 | 0.124 | 0.489 | 0.488 | 0.826 | 1.186 |
| | 1 | $\beta_{11}$ | 0.002 | 0.209 | 0.212 | 0.954 | −0.013 | 0.225 | 0.223 | 0.948 | 1.157 |
| | | $\beta_{12}$ | 0.002 | 0.457 | 0.461 | 0.957 | 0.141 | 0.495 | 0.498 | 0.813 | 1.173 |

See the note to Table 3.13.

there were 133 HIV-1 sero-conversions and approximately 2,300 person-years of follow-up. Out of the 133 sero-conversions, 27 were of viral subtype B, and 99 of subtype E. The subtypes for the remaining 7 were unknown but assumed to be either B or E.

We apply the proposed methods to the data derived from this study by treating the two HIV-1 subtypes as competing risks with partially missing information. We investigate the influence of potential risk factors on the cumulative incidence of HIV-1 sero-conversion. The potential risk factors include age at baseline (in years), gender (male vs female), whether the subject had a history of needle sharing (yes vs no), whether s/he had drug injection at jail (yes vs no), and whether s/he had an imprisonment history before recruitment. We consider logarithmic transformation functions: $G_1(x) = r_1^{-1}\log(1 + r_1 x)$, and $G_2(x) = r_2^{-1}\log(1 + r_2 x)$. The log-likelihood is maximized at $r_1 = 1.6$ and $r_2 = 0.2$, which is the combination of transformation functions that would be selected by the AIC criterion. Table 3.15 shows the results for this combination, as well as $r_1 = r_2 = 0$ (proportional hazards) and $r_1 = r_2 = 1$ (proportional odds).

There are considerable differences in the parameter estimates between the NPMLE and naive methods, and the latter tends to produce larger standard errors than the former. The effects of risk factors on the two competing risks are quite different. By the NPMLE, needle sharing significantly increases the incidence of sero-conversion with HIV-1 subtype B, but its effect on subtype E is minimal. In contrast to previous regression analysis focusing on the cause-specific hazard (Hudgens et al., 2002), its estimated effects on the two risks have opposite directions as a result of the trade-off between the cumulative incidences. In addition, younger age and imprisonment history significantly increase the incidence of sero-conversion with subtype E; and drug injection has a marginally significant effect on the incidence of subtype E. None of these factors, however, are significantly associated with the incidence of subtype B. To illustrate the joint inference by the NPMLE, we conduct a joint test for the effects of imprisonment history on the two viral subtypes. The $\chi_2^2$ test statistic is 12.8, which is highly significant.

Table 3.15: Analysis of the BMA HIV-1 study.

| | NPMLE | | | | | | Naive | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Subtype B | | | Subtype E | | | Subtype B | | | Subtype E | | |
| | Est | SE | $p$-value | Est | SE | $p$-value | Est | SE | $p$-value | Est | SE | $p$-value |
| Proportional hazards | | | | | | | | | | | | |
| Age | −0.027 | 0.182 | 0.883 | −0.268 | 0.099 | 0.007 | 0.009 | 0.199 | 0.963 | −0.285 | 0.097 | 0.003 |
| Gender | −0.116 | 0.444 | 0.793 | 0.618 | 0.277 | 0.026 | 0.020 | 0.551 | 0.970 | 0.628 | 0.377 | 0.096 |
| Needle | 1.038 | 0.240 | <0.001 | −0.031 | 0.196 | 0.875 | 1.033 | 0.301 | 0.001 | −0.010 | 0.265 | 0.969 |
| Drug | 0.098 | 0.139 | 0.479 | 0.374 | 0.194 | 0.054 | 0.018 | 0.186 | 0.924 | 0.334 | 0.198 | 0.091 |
| Prison | −0.557 | 0.424 | 0.189 | 0.727 | 0.216 | 0.001 | −0.599 | 0.500 | 0.231 | 0.737 | 0.227 | 0.001 |
| Proportional odds | | | | | | | | | | | | |
| Age | 0.085 | 0.184 | 0.646 | −0.305 | 0.104 | 0.003 | −0.010 | 0.194 | 0.959 | −0.272 | 0.132 | 0.039 |
| Gender | −0.143 | 0.454 | 0.752 | 0.677 | 0.355 | 0.057 | 0.022 | 0.531 | 0.967 | 0.729 | 0.431 | 0.091 |
| Needle | 0.975 | 0.238 | <0.001 | −0.064 | 0.222 | 0.775 | 1.071 | 0.282 | <0.001 | −0.055 | 0.277 | 0.841 |
| Drug | 0.118 | 0.113 | 0.300 | 0.425 | 0.194 | 0.028 | −0.019 | 0.203 | 0.926 | 0.334 | 0.199 | 0.093 |
| Prison | −0.588 | 0.450 | 0.191 | 0.757 | 0.193 | <0.001 | −0.513 | 0.543 | 0.344 | 0.717 | 0.244 | 0.003 |
| Selected model | | | | | | | | | | | | |
| Age | 0.085 | 0.182 | 0.639 | −0.264 | 0.090 | 0.003 | −0.053 | 0.203 | 0.794 | −0.256 | 0.119 | 0.032 |
| Gender | −0.107 | 0.507 | 0.832 | 0.521 | 0.350 | 0.136 | 0.018 | 0.585 | 0.976 | 0.591 | 0.430 | 0.169 |
| Needle | 0.998 | 0.235 | <0.001 | −0.075 | 0.244 | 0.760 | 0.996 | 0.274 | <0.001 | −0.025 | 0.281 | 0.930 |
| Drug | 0.123 | 0.133 | 0.356 | 0.362 | 0.195 | 0.064 | 0.030 | 0.190 | 0.876 | 0.325 | 0.240 | 0.176 |
| Prison | −0.552 | 0.421 | 0.190 | 0.769 | 0.188 | <0.001 | −0.508 | 0.520 | 0.329 | 0.716 | 0.241 | 0.003 |

Est and SE denote the parameter estimate and (estimated) standard error.

To illustrate prediction, we display in Figure 3.4 the NPMLE and naive estimates of the cumulative incidence function for a 32-year old female with a history of needle sharing, drug injection, and imprisonment before recruitment. The cumulative incidence of sero-conversion with HIV-1 subtype E is much higher than that of subtype B. The naive method yields appreciably different estimates than the NPMLE, especially for subtype E. The discrepancies are likely due to omission of failures with unknown causes, which are mostly assigned to subtype E by the NPMLE.

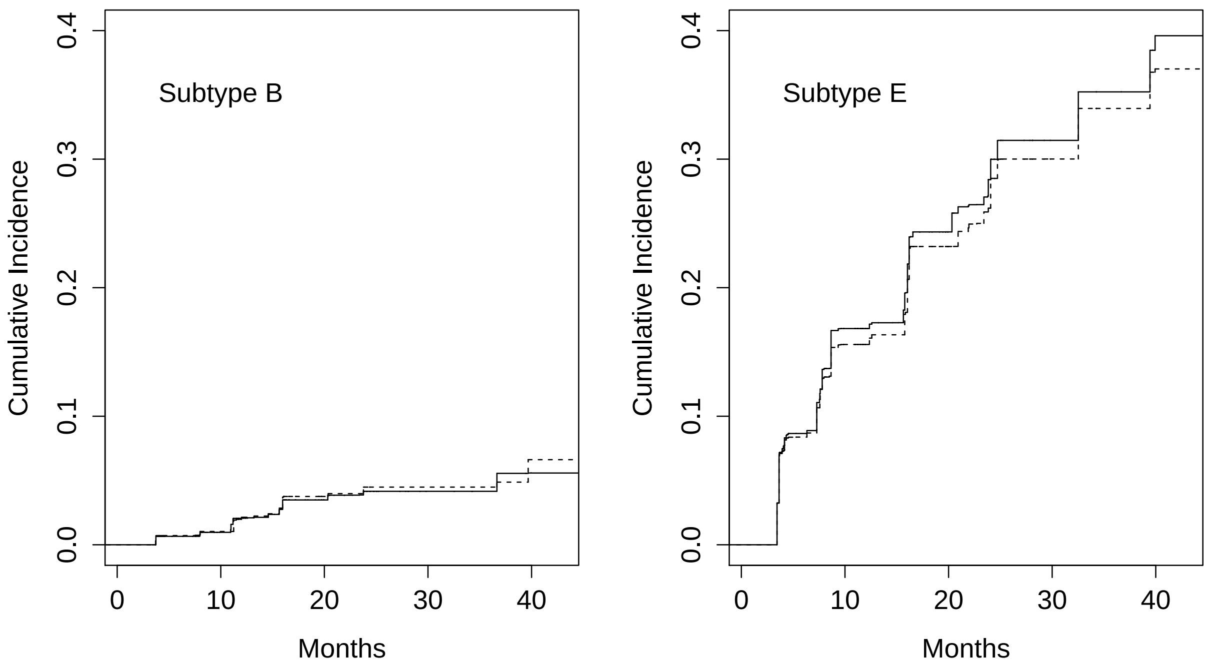


Figure 3.4: Estimated cumulative incidence of sero-conversion in the BMA HIV-1 study for a 32-year old woman with a history of needle sharing, drug injection, and imprisonment before recruitment. The solid and dashed curves pertain to the NPMLE and naive methods, respectively.

## 5. DISCUSSION

There is some literature on the NPMLE for interval-censored univariate failure time data; see Huang (1996) and Huang and Wellner (1997). It is theoretically more difficult to study the NPMLE for interval-censored competing risks data with partially missing causes of failure. First, the multiplicity of failure types requires that the least favorable directions hold simultaneously for all nuisance tangent spaces. Second, as the probability of missing cause of failure information may depend on the examination times and the covariates, showing the existence of a solution to the normal equations for the least favorable directions requires careful arguments for the smoothness of the score and information operators.

The computation of the NPMLE with interval-censored competing risks data also poses new challenges. The iterative convex minorant algorithm (Huang and Wellner, 1997) works well for univariate failure time data with time-invariant covariates but is not applicable to competing risks or time-varying covariates. We have devised a novel EM algorithm by treating failure times as missing data and allowing for unknown causes of failure. Our algorithm is very stable at each iteration and its convergence is fast. We have not encountered any non-convergence in our extensive numerical studies.

In practice, it may be difficult to determine the causes of failure for all study subjects, especially when the ascertainment requires an extra (and possibly costly) step. In the BMA study, for instance, the HIV-1 subtype information required genotyping the viral DNA (Hudgens et al., 2001). The NPMLE approach enables one to make valid and efficient inference in the presence of missing information on the cause of failure.

We have studied both the NPMLE and naive estimators. The naive estimator may be preferable if the interest lies in only a subset of risks. However, the naive estimator is less satisfactory than the NPMLE for several reasons. First, the naive estimator is not statistically efficient. Secondly, it cannot properly handle unknown causes of failure. Finally, it does not provide simultaneous inference, which is often desirable because an increase in the incidence of one risk naturally reduces the incidence of other risks.

In some applications, a subset of risks are interval-censored while the rest are right-censored. For example, in the Breastfeeding, Antiretrovirals, and Nutrition (BAN) study (Hudgens et al., 2014), there were three competing risks: infant HIV-infection, weening, and infant death prior to infection or weening. While the first two risks were interval-censored, the time to death was known exactly or right censored. We plan to extend our work to this type of competing risks data.

### 3.5 Technical Details

Let $\mathbb{P}_n$ denote the empirical measure from $n$ i.i.d. observations, $\mathbb{P}$ denote the underlying probability measure, and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ denote the corresponding empirical process. Write $\boldsymbol{\Lambda}_0 = (\Lambda_{10}, \ldots, \Lambda_{K0})$. In addition, let $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ denote the density function with parameter $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ for a single observation, i.e.,

$$\begin{aligned} p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}(J, \boldsymbol{U}, \boldsymbol{\Delta}, \xi, \xi\widetilde{D}, \boldsymbol{Z}) = \prod_{j=1}^{J} \Bigg( & \prod_{k=1}^{K} \{F_k(U_j; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k)\}^{I(\xi\widetilde{D}=k, \Delta_j=1)} \\ & \times \left[ \sum_{k=1}^{K} \{F_k(U_j; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k)\} \right]^{I(\xi=0, \Delta_j=1)} \Bigg) \\ & \times S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda})^{I(\widetilde{D}=0)}. \end{aligned}$$

*Proof of Theorem 3.1*

We first show that, with probability one, $\widehat{\Lambda}_k(\tau)$ is asymptotically bounded for each $k$, i.e., $\limsup_n \widehat{\Lambda}_k(\tau) < \infty$. To this end, we define

$$\widetilde{N}_k(t) = \sum_{j=1}^{J} \{I(U_j \leq t, \xi\widetilde{D} = k) + w_k I(U_j \leq t, \xi = 0)\},$$

where $w_k = \Pr(\widetilde{D} = k | \xi = 0)$. By conditions (C4) and (C6), the measure $E[d\widetilde{N}_k(\cdot)]$ is absolutely continuous with respect to the Lebesgue measure on $[\zeta, \tau)$ with a density that is uniformly bounded from above and bounded away from zero. Define

$$\widetilde{\Lambda}_k(t) = \int_0^t \frac{\Lambda'_{k0}(s)}{E[d\widetilde{N}_k(s)]/ds} \mathbb{P}_n[d\widetilde{N}_k(s)].$$

By the Glivenko-Cantelli theorem, $\widetilde{\Lambda}_k(t)$ converges uniformly to $\Lambda_{k0}(t)$ with probability one under condition (C1). Since $\widetilde{\Lambda}_k(\cdot)$ is a right-continuous step function taking jumps at the $U_j$'s,

$$\mathbb{P}_n \log p_{\widehat{\boldsymbol{\beta}},\widehat{\Lambda}} \geq \mathbb{P}_n \log p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}, \tag{3.31}$$

where $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\Lambda}_1, \ldots, \widetilde{\Lambda}_K)$.

We consider the class of functions $\left\{\int_0^{U_j} e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)} d\Lambda_k(t) : \boldsymbol{\beta}_k \in \mathcal{B}_k, \Lambda_k \in BV_c[0,\tau]\right\}$, where $\mathcal{B}_k$ is the space for $\boldsymbol{\beta}_k$, and $BV_c[0,\tau]$ is the space of functions on $[0,\tau]$ with total variation bounded by $c$. This class is contained in the convex hull of the class $\mathcal{F}_1 \equiv \{I(U_j \geq t)e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}\}$. Because $\{\boldsymbol{\beta}_k^T \boldsymbol{Z}(t) : t \in [0,\tau]\}$ and $\{I(U_j \geq t) : t \in [0,\tau]\}$ are VC-classes, the convex hull of $\mathcal{F}_1$ is Donsker according to Theorem 2.6.9 of van der Vaart and Wellner (1996). Thus, $p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}$ (indexed by the possible values of $\widetilde{\boldsymbol{\Lambda}}$) belongs to a Glivenko-Cantelli class. By condition (C1), $S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}) \geq S(\tau; \boldsymbol{Z}, \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}})$, and the latter converges uniformly in $\boldsymbol{Z}$ to $S(\tau; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$, which is bounded by some positive constant from below. Thus, $S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}})$ is bounded away from zero when $n$ is large enough. Similarly, $F_k(U_j; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \widetilde{\Lambda}_k) - F_k(U_{j-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \widetilde{\Lambda}_k)$ converges uniformly in $\boldsymbol{Z}$ to $F_k(U_j; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) - F_k(U_{j-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0})$, which is bounded away from 0 by condition (C4). It then follows from the preservation property of the log-transformation that $\log p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}$ belongs to a Donsker class and is therefore Glivenko-Cantelli. As a result, $|\mathbb{P}_n \log p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}} - \mathbb{P}_n \log p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}| \to 0$ with probability one, and $\mathbb{P}_n \log p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}$ is bounded away from $-\infty$. Hence, inequality (3.31) entails that $\liminf_n \mathbb{P}_n \log p_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}} > -\infty$ with probability one.

Clearly,

$$\begin{aligned}
\mathbb{P}_n \log p_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}} &\leq \mathbb{P}_n I(\widetilde{D} = 0) \log S(U_j; \boldsymbol{Z}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \\
&\leq \mathbb{P}_n I(\boldsymbol{\Delta} = \boldsymbol{0}, U_J = \tau) \log S(\tau; \boldsymbol{Z}, \widehat{\boldsymbol{\beta}}, \widehat{\Lambda}) \\
&\leq \log \left\{ \sum_{k=1}^K e^{-G_k(\widehat{\Lambda}_k(\tau)c_0} \right\} \mathbb{P}_n I(\boldsymbol{\Delta} = \boldsymbol{0}, U_J = \tau),
\end{aligned}$$

where $c_0$ is a positive constant bounding $e^{\widehat{\boldsymbol{\beta}}_k^T \boldsymbol{Z}(\tau)}$ from below. By condition (C4), $\mathbb{P}_n I(\boldsymbol{\Delta} = \mathbf{0}, U_J = \tau) \to \mathbb{P} I(\boldsymbol{\Delta} = \mathbf{0}, U_J = \tau)$, which is a strictly positive constant. Therefore,

$$\limsup_n \widehat{\Lambda}_k(\tau) < \infty, \quad k = 1, \ldots, K.$$

We can restrict $\widehat{\Lambda}_k$ to a uniformly bounded class of functions equipped with the weak topology on $[\zeta, \tau]$. By Helly's selection lemma, for every subsequence, there exists a further subsequence such that $\widehat{\boldsymbol{\beta}}_k \to \boldsymbol{\beta}_k^*$ for some $\boldsymbol{\beta}_k^*$ and $\widehat{\Lambda}_k$ converges weakly to some $\Lambda_k^*$. Our remaining task is to show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}_0$.

Write $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\beta}_K^{*\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Lambda}^* = (\Lambda_1^*, \ldots, \Lambda_K^*)$. By the arguments for verifying the Glivenko-Cantelli property of $\{\log p_{\boldsymbol{\beta}, \widetilde{\boldsymbol{\Lambda}}}\}$, we can show that the class $\{\log[p_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}} + p_{\boldsymbol{\beta}, \widetilde{\boldsymbol{\Lambda}}}]/2\}$ is also Glivenko-Cantelli. Thus,

$$\mathbb{P}_n \left[ \log \frac{p_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}} + p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}}}{2} \right] = \mathbb{P} \left[ \log \frac{p_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}} + p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}}}{2} \right] + o(1)$$

with probability one. Because

$$\mathbb{P}_n \left[ \log \frac{p_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}} + p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}}}{2} \right] \geq \mathbb{P}_n \left[ \log p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}} \right],$$

and

$$\mathbb{P}_n \left[ \log p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}} \right] = \mathbb{P} \left[ \log p_{\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0} \right] + o(1),$$

we conclude that

$$\mathbb{P} \left[ \log \frac{p_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}} + p_{\boldsymbol{\beta}_0, \widetilde{\boldsymbol{\Lambda}}}}{2} \right] \geq \mathbb{P} \left[ \log p_{\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0} \right] + o(1).$$

By the Taylor series expansion,

$$
\begin{aligned}
& \Big|\mathbb{P}\left[\log\frac{p_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}}+p_{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Lambda}}}}{2}\right]-\mathbb{P}\left[\log\frac{p_{\boldsymbol{\beta}^*,\boldsymbol{\Lambda}^*}+p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}}{2}\right]\Big| \\
\lesssim\ & |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|+\mathbb{P}\left\{\sum_{k=1}^{K}\sum_{j=1}^{J}\Big|\int_0^{U_j}e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\widehat{\Lambda}_k(s)-\int_0^{U_j}e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k^*(s)\Big|\right\} \\
& +\sum_{k=1}^{K}\sup_{t\in\tau}\left|\widetilde{\Lambda}_k(t)-\Lambda_{k0}(t)\right| \\
\lesssim\ & |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|+\mathbb{P}\left[\sum_{k=1}^{K}\sum_{j=1}^{J}\left\{\left|\widehat{\Lambda}_k(U_j)-\Lambda_k^*(U_j)\right|+\int|\widehat{\Lambda}_k(s)-\Lambda_k^*(s)|I(U_j\geq s)|d\boldsymbol{Z}(s)|\right\}\right] \\
& +o(1) \\
=\ & |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|+\sum_{k=1}^{K}\int|\widehat{\Lambda}_k(s)-\Lambda_k^*(s)|d\nu(s)+o(1),
\end{aligned}
$$

where $A\lesssim B$ means that $A\leq cB$ for a positive constant $c$, and $d\nu(s)=d\mathbb{P}[\sum_{j=1}^{J}I(U_j\geq s)]+\mathbb{P}[\sum_{j=1}^{J}I(U_j\geq s)|d\boldsymbol{Z}(s)|]$. Consequently,

$$
\mathbb{P}\left[\log\frac{p_{\boldsymbol{\beta}^*,\boldsymbol{\Lambda}^*}+p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}}{2}\right]\geq\mathbb{P}\left[\log p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}\right].
$$

By the non-negativeness of the Kullback-Leibler information, the above inequality implies $p_{\boldsymbol{\beta}^*,\boldsymbol{\Lambda}^*}=p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}$ almost surely. Letting $\xi\widetilde{D}=k$ and $\Delta_j=1$ $(j=1,\ldots,J)$, we obtain

$$
\begin{gathered}
F_k(U_1;\boldsymbol{Z},\boldsymbol{\beta}_k^*,\Lambda_k^*)=F_k(U_1;\boldsymbol{Z},\boldsymbol{\beta}_{k0},\Lambda_{k0}) \\
F_k(U_2;\boldsymbol{Z},\boldsymbol{\beta}_k^*,\Lambda_k^*)-F_k(U_1;\boldsymbol{Z},\boldsymbol{\beta}_k^*,\Lambda_k^*)=F_k(U_2;\boldsymbol{Z},\boldsymbol{\beta}_{k0},\Lambda_{k0})-F_k(U_1;\boldsymbol{Z},\boldsymbol{\beta}_{k0},\Lambda_{k0}) \\
\vdots \\
F_k(U_J;\boldsymbol{Z},\boldsymbol{\beta}_k^*,\Lambda_k^*)-F_k(U_{J-1};\boldsymbol{Z},\boldsymbol{\beta}_k^*,\Lambda_k^*)=F_k(U_J;\boldsymbol{Z},\boldsymbol{\beta}_{k0},\Lambda_{k0})-F_k(U_{J-1};\boldsymbol{Z},\boldsymbol{\beta}_{k0},\Lambda_{k0}).
\end{gathered}
$$

By condition (C5), $\int_0^{U_j} e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)}d\Lambda_{k0}(t) = \int_0^{U_j} e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(t)}d\Lambda_k^*(t)$ $(j = 1, \dots, J)$, which, under condition (C4), yields

$$\int_0^t e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s) = \int_0^t e^{\boldsymbol{\beta}_k^{*\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k^*(s), \tag{3.32}$$

for all $t \in [\zeta, \tau]$. Assume that $\mathbf{0}$ is in the support of $\boldsymbol{Z}$ (by subtracting $\boldsymbol{Z}$ by any value within its support). We take $\boldsymbol{Z}(\cdot) = \mathbf{0}$ on $[\zeta, \tau]$ to find $\Lambda_k^*(t) = \Lambda_{k0}(t)$. After taking derivatives on both sides of (3.32) and using condition (C3), we have $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_{k0}$ and $\Lambda_k^* = \Lambda_{k0}$. Thus, $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$, and $\widehat{\Lambda}_k$ converges weakly to $\Lambda_{k0}$. The latter convergence can be strengthened to uniform convergence in $t \in [\zeta, \tau]$ since $\Lambda_{k0}$ is continuous.

*Proof of Theorem 3.2*

In light of Theorem 3.1, we restrict parameter $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ to the following space:

$$\mathcal{A} \equiv \{(\boldsymbol{\beta}, \boldsymbol{\Lambda}) : 1/M \le \Lambda_k(\zeta) \le \Lambda_k(\tau) \le M, k = 1, \dots, K, |\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \delta_0\}$$

for some $M > 0$ and $\delta_0 > 0$. The proof consists of four major steps.

*Step 1 (Establishing the convergence rate for the NPMLE).* We define a distance function $\rho$ on the parameter space as

$$\begin{aligned}\rho\{(\boldsymbol{\beta}, \boldsymbol{\Lambda}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\} &= |\boldsymbol{\beta} - \boldsymbol{\beta}_0| \\ &\quad + \left[\mathbb{P}\left\{\sum_{k=1}^K \sum_{j=1}^J \left\{\int_0^{U_j} e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s) - \int_0^{U_j} e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s)\right\}^2\right\}\right]^{1/2}.\end{aligned}$$

We wish to show that

$$\rho\{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\} = O_P(n^{-1/3}). \tag{3.33}$$

The achieve this goal, we verify the conditions in Theorem 3.4.4 of van der Vaart and Wellner (1996). First, we calculate the entropy of the class $\mathcal{M} \equiv \{\log p_{\boldsymbol{\beta},\boldsymbol{\Lambda}} : (\boldsymbol{\beta}, \boldsymbol{\Lambda}) \in \mathcal{A}\}$. Let

$N_{[]}(\epsilon, \mathcal{F}, L_2(\mathbb{P}))$ denote the $\epsilon$-bracketing number of any class $\mathcal{F}$ with the $L_2(\mathbb{P})$ distance. Since the $\boldsymbol{\Lambda}$ space in $\mathcal{A}$ consists of increasing and uniformly bounded functions on $[\zeta, \tau]$, Lemma 2.2 of van de Geer (2000) implies that $\log N_{[]}(\epsilon, \mathcal{A}_k, ||\cdot||_2) \lesssim \epsilon^{-1}$, where $\mathcal{A}_k$ is the space for $\Lambda_k$ in $\mathcal{A}$, and $||\cdot||_2$ denotes the $L_2$ distance with respect to the Lebesgue measure on $[0, \tau]$. Given $\epsilon > 0$, we can find $\exp\{O(\epsilon^{-1})\}$ brackets $[\Lambda_{kj}^l, \Lambda_{kj}^u]$, indexed by $j$, with $||\Lambda_{kj}^u - \Lambda_{kj}^l||_2 \leq \epsilon$, to cover $\mathcal{A}_k$. In addition, there are $O(\epsilon^{-d})$ number of brackets covering $\mathcal{B}$, the space of $\boldsymbol{\beta}$ in $\mathcal{A}$. Thus, there are a total of $O(1/\epsilon^d) \times \exp\{O(K\epsilon^{-1})\}$ brackets covering $\mathcal{B} \times \mathcal{A}_1 \times \ldots \mathcal{A}_K \supset \mathcal{A}$. Since $E(J) < \infty$, we have $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}} \in L_2(\mathbb{P})$ for all $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$. Furthermore, by conditions (C1), (C2), (C4), and (C5),

$$
\begin{aligned}
\mathbb{P}\left(p_{\boldsymbol{\beta},\Lambda} - p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}\right)^2 \lesssim \sum_{j=1}^{\infty} \mathbb{P}\pi_j(\boldsymbol{Z}) \Bigg[ & \sum_{j'=1}^{j} \sum_{k=1}^{K} \Big\{ F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) \\
& - F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) + F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \Big\}^2 \\
& + \sum_{k=1}^{k} \Big\{ F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) \\
& - F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) + F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \Big\}^2 \\
& \Big\{ S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) - S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) \Big\}^2 \Bigg] \\
\lesssim \sum_{k=1}^{K} \sum_{j=1}^{\infty} \mathbb{P}\pi_j(\boldsymbol{Z}) & \left\{ \int_0^{U_j} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}(t)} d\Lambda_k(t) - \int_0^{U_j} e^{\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{Z}(t)} d\Lambda_{k0}(t) \right\}^2 \\
\lesssim |\boldsymbol{\beta} - \boldsymbol{\beta}_0|^2 + \sum_{k=1}^{K} & \left( \sum_{j=1}^{\infty} \mathbb{P}\pi_j(\boldsymbol{Z}) \sum_{j'=1}^{j} \{\Lambda_k(U_{j'}) - \Lambda_{k0}(U_{j'})\}^2 \right. \\
& \left. + \int_0^{\tau} \{\Lambda_k(t) - \Lambda_{k0}(t)\}^2 \sum_{j=1}^{\infty} j \mathbb{P}\pi_j(\boldsymbol{Z}) |d\boldsymbol{Z}(t)| \Big] \right).
\end{aligned}
$$

The measures $\sum_{j=1}^{\infty} \sum_{j'=1}^{j} \mathbb{P}[\pi_j(\boldsymbol{Z}) dI(t \geq U_{j'})]$ and $\sum_{j=1}^{\infty} j\mathbb{P}[\pi_j(\boldsymbol{Z})|d\boldsymbol{Z}(t)|]$ have bounded densities with respect to the Lebesgue measure in $[\zeta, \tau)$, and the former has a finite point mass at $\tau$. Thus, any $\epsilon$-bracket for $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ is also an $\epsilon$-bracket (up to a constant) for $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ in $L_2(P)$-space. Hence, the $\epsilon$-bracket number for $\{p_{\boldsymbol{\beta},\boldsymbol{\Lambda}} : (\boldsymbol{\beta}, \boldsymbol{\Lambda}) \in \mathcal{A}\}$ under the $L_2(P)$-norm is of the order

$O(1/\epsilon^d) \times \exp\{O(K\epsilon^{-1})\}$. Because of the restriction of the parameters in $\mathcal{A}$, $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ is uniformly bounded away from 0, such that $\log p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ is a Lipschitz transformation of the density $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$. Thus, the bracketing number for $\mathcal{M}$ is of the same order as the $L_2(\mathbb{P})$ bracketing number of the class $\{p_{\boldsymbol{\beta},\boldsymbol{\Lambda}} : (\boldsymbol{\beta}, \boldsymbol{\Lambda}) \in \mathcal{A}\}$. As a result,

$$\log N_{[]}(\epsilon, \mathcal{M}, L_2(\mathbb{P})) \lesssim d \log(1/\epsilon) + K\epsilon^{-1}. \tag{3.34}$$

The bracketing entropy integral satisfies

$$J_{[]}(\delta, \mathcal{M}, L_2(\mathbb{P})) = \int_0^\delta \sqrt{1 + \log N(\epsilon, \mathcal{M}, L_2(\mathbb{P}))} d\epsilon \lesssim \sqrt{\delta}.$$

By Theorem 3.4.1 of van der Vaart and Wellner (1996), we let $r_n^2 \phi_n(r_n^{-1}) \lesssim \sqrt{n}$, where $\phi_n(x) = \sqrt{x}$, so that $r_n \lesssim n^{-1/3}$ and $H(p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}, p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}) = O_p(n^{-1/3})$, where $H(p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}, p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0})$ is the Hellinger distance between two densities $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ and $p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}$ defined as

$$H^2(p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}, p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}) = \int \left(\sqrt{p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}} - \sqrt{p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0}}\right)^2 d\mu$$

for a dominating measure $\mu$.

Since $p_{\boldsymbol{\beta},\boldsymbol{\Lambda}}$ is uniformly bounded and bounded away from zero, the mean-value theorem implies that

$$
\begin{aligned}
H(p_{\boldsymbol{\beta},\Lambda}, p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0})^2 &\gtrsim \int (p_{\boldsymbol{\beta},\Lambda} - p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0})^2 p_{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0} d\mu \\
&\gtrsim \sum_{j=1}^{\infty} \mathbb{P}\pi_j(\boldsymbol{Z}) \Bigg[ \sum_{j'=1}^{j} \sum_{k=1}^{K} \Big\{ F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) \\
&\qquad - F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) + F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \Big\}^2 \\
&\qquad + \sum_{k=1}^{K} \Big\{ F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) \\
&\qquad - F_k(U_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) + F_k(U_{j'-1}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \Big\}^2 \\
&\qquad \Big\{ S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) - S(U_J; \boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) \Big\}^2 \Bigg] \\
&\gtrsim \rho^2\{(\boldsymbol{\beta}, \boldsymbol{\Lambda}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\},
\end{aligned}
$$

where $A \gtrsim B$ means that $A \geq cB$ for a positive constant $c$. Hence, we have proved (3.33).

*Step 2 (Deriving the least favorable direction for $\boldsymbol{\beta}_0$).* The score function for $\boldsymbol{\beta}$ is

$$
\begin{aligned}
l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \sum_{k=1}^{K} \int_0^{\tau} \Big\{ &\sum_{j=1}^{J} I(\Delta_j = 1, \xi = 1, \widetilde{D} = k) B_k(t, U_{j-1}, U_j; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) \\
&+ \sum_{j=1}^{J} I(\Delta_j = 1, \xi = 0) B_{0k}(t, U_{j-1}, U_j; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \\
&+ I(\boldsymbol{\Delta} = \boldsymbol{0}) \widetilde{B}_k(t, U_J; \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \Big\} e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t)} \boldsymbol{X}_k(t) d\Lambda_k(t),
\end{aligned}
$$

where

$$B_k(t,u,v;\boldsymbol{Z},\boldsymbol{\beta}_k,\Lambda_k)=\frac{\widetilde{G}_k\left\{\int_0^v e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}I(t\le v)-\widetilde{G}_k\left\{\int_0^u e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}I(t\le u)}{F(v;\boldsymbol{Z},\boldsymbol{\beta}_k,\Lambda_k)-F(u;\boldsymbol{Z},\boldsymbol{\beta}_k,\Lambda_k)},$$

$$B_{0k}(t,u,v;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})=\frac{\widetilde{G}_k\left\{\int_0^v e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}I(t\le v)-\widetilde{G}_k\left\{\int_0^u e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}I(t\le u)}{\sum_{k'=1}^{k}\left\{F_{k'}(v;\boldsymbol{Z},\boldsymbol{\beta}_{k'},\Lambda_{k'})-F_{k'}(u;\boldsymbol{Z},\boldsymbol{\beta}_{k'},\Lambda_{k'})\right\}},$$

$$\widetilde{B}_k(t,u;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})=-\frac{\widetilde{G}_k\left\{\int_0^u e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_k(s)\right\}I(t\le u)}{S(u;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})},$$

and $\boldsymbol{X}_k(t)$ is a $d$-dimensional vector whose $\{(k-1)p+1\}$th to $(kp)$th components are $\boldsymbol{Z}(t)$ and whose other components are 0. To obtain the score operator for $\boldsymbol{\Lambda}$, we consider the one-dimensional parametric model $d\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}}=\{(1+\epsilon h_1)d\Lambda_1,\ldots,(1+\epsilon h_K)d\Lambda_K\}$, where

$$\boldsymbol{h}=(h_1,\ldots,h_K)\in L_2(\boldsymbol{\Lambda}_0)\equiv L_2(\Lambda_{10})\otimes\ldots\otimes L_2(\Lambda_{K0}).$$

After differentiating the log-likelihood along this submodel, we obtain the score operator for $\boldsymbol{\Lambda}$

$$\begin{aligned}l_{\boldsymbol{\Lambda}}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}]=\sum_{k=1}^{K}\int_0^\tau\Big\{&\sum_{j=1}^{J}I(\Delta_j=1,\xi=1,\widetilde{D}=k)B_k(t,U_{j-1},U_j;\boldsymbol{Z},\boldsymbol{\beta}_k,\Lambda_k)\\&+\sum_{j=1}^{J}I(\Delta_j=1,\xi=0)B_{0k}(t,U_{j-1},U_j;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\\&+I(\boldsymbol{\Delta}=\boldsymbol{0})\widetilde{B}_k(t,U_J;\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Lambda})\Big\}e^{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{Z}(t)}h_k(t)d\Lambda_k(t).\end{aligned}$$

The least favorable direction for $\boldsymbol{\beta}_0$, denoted by $\boldsymbol{h}^*=(\boldsymbol{h}_1^*,\ldots,\boldsymbol{h}_K^*)$, is a $d$-dimensional vector with components in $L_2(\Lambda_{k0})$ satisfying $\mathbb{P}(l_{\boldsymbol{\beta}}-l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*])l_{\boldsymbol{\Lambda}}[\boldsymbol{h}]=0$ for all $\boldsymbol{h}\in L_2(\boldsymbol{\Lambda}_0)$, where $l_{\boldsymbol{\beta}}=l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$ and $l_{\boldsymbol{\Lambda}}=l_{\boldsymbol{\Lambda}}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$. Equivalently, it solves the norm equation

$$l_{\boldsymbol{\Lambda}}^*l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*]=l_{\boldsymbol{\Lambda}}^*l_{\boldsymbol{\beta}} \tag{3.35}$$

where $l_{\boldsymbol{\Lambda}}^*:L_2(\mathbb{P})\to L_2(\boldsymbol{\Lambda}_0)$ is the adjoint operator of $l_{\boldsymbol{\Lambda}}$. The operator $l_{\boldsymbol{\Lambda}}^*l_{\boldsymbol{\Lambda}}$ is also the information operator for $\mathbf{\Lambda}_0$.

To show the existence of $\boldsymbol{h}^*$, we define an inner product of the Hilbert space $L_2(\mathbf{\Lambda}_0)$ as

$$< \boldsymbol{h}_1, \boldsymbol{h}_2 >_{L_2(\mathbf{\Lambda}_0)} = \sum_{k=1}^{K} \int_{\zeta}^{\tau} h_{1k} h_{2k} d\Lambda_{k0}.$$

Given $\boldsymbol{h}_1, \boldsymbol{h}_2 \in L_2(\mathbf{\Lambda}_0)$,

$$\mathbb{P}(l_{\mathbf{\Lambda}}[\boldsymbol{h}_1] l_{\mathbf{\Lambda}}[\boldsymbol{h}_2]) = \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{\zeta}^{\tau} \int_{\zeta}^{\tau} \mathbb{P} C_{kl}(t, s, \boldsymbol{Z}) h_{2k}(t) d\Lambda_{k0}(t) h_{1l}(s) d\Lambda_{l0}(s),$$

where

$$\begin{aligned}
C_{kl}(t, s, \boldsymbol{Z}) = \sum_{j=1}^{\infty} \pi_j(\boldsymbol{Z}) \int_{\zeta}^{\tau} \cdots \int_{\zeta}^{\tau} \Bigg[ & I(k = l) \rho_j(u_1, \ldots, u_j, \boldsymbol{Z}) \sum_{j'=1}^{j} \widetilde{F}_k(u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \\
& \times B_k(t, u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) B_k(s, u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) \\
& + \{1 - \rho_j(u_1, \ldots, u_j, \boldsymbol{Z})\} \sum_{j'=1}^{j} \widetilde{F}(u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_0, \mathbf{\Lambda}_0) \\
& \times B_{0k}(t, u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{k0}, \Lambda_{k0}) B_{0l}(s, u_{j'-1}, u_{j'}; \boldsymbol{Z}, \boldsymbol{\beta}_{l0}, \Lambda_{l0}) \\
& + S(u_j; \boldsymbol{Z}, \boldsymbol{\beta}_0, \mathbf{\Lambda}_0) \widetilde{B}_k(t, u_j; \boldsymbol{Z}, \boldsymbol{\beta}_0, \mathbf{\Lambda}_0) \widetilde{B}_l(s, u_j; \boldsymbol{Z}, \boldsymbol{\beta}_0, \mathbf{\Lambda}_0) g_j(u_1, \ldots, u_j, \boldsymbol{Z}) \Bigg] \\
& \times e^{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) + \boldsymbol{\beta}_l^{\mathrm{T}} \boldsymbol{Z}(s)} du_1 \ldots du_j,
\end{aligned}$$

$$\widetilde{F}_k(u, v; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) = F_k(v; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(u; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k),$$

$$\widetilde{F}(u, v; \boldsymbol{Z}, \boldsymbol{\beta}, \mathbf{\Lambda}) = \sum_{k=1}^{K} \{F_k(v; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k) - F_k(u; \boldsymbol{Z}, \boldsymbol{\beta}_k, \Lambda_k)\}.$$

Likewise,

$$\mathbb{P}(l_{\mathbf{\Lambda}}[\boldsymbol{h}_1] l_{\boldsymbol{\beta}}) = \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{\zeta}^{\tau} \int_{\zeta}^{\tau} \mathbb{P}\{C_{kl}(t, s, \boldsymbol{Z}) \boldsymbol{X}_k(t)\} d\Lambda_{k0}(t) h_{1l}(s) d\Lambda_{l0}(s).$$

Hence, the normal equation in (3.35) is equivalent to

$$\sum_{k=1}^{K}\int_{\zeta}^{\tau}\mathbb{P}C_{kl}(t,s,\boldsymbol{Z})\boldsymbol{h}_k^*(t)d\Lambda_{k0}(t)=\sum_{k=1}^{K}\int_{\zeta}^{\tau}\mathbb{P}\{C_{kl}(t,s,\boldsymbol{Z})\boldsymbol{X}_k(t)\}d\Lambda_{k0}(t) \tag{3.36}$$

for all $s\in[\zeta,\tau]$ and $l=1,\ldots,K$.

We define a linear operator $\Gamma$ from $L_2(\boldsymbol{\Lambda}_0)$ to itself by

$$\Gamma[\boldsymbol{h}](s)=\left\{\sum_{k=1}^{K}\int_{\zeta}^{\tau}\mathbb{P}C_{kl}(t,s,\boldsymbol{Z})h_k(t)d\Lambda_{k0}(t),\ldots,\sum_{k=1}^{K}\int_{\zeta}^{\tau}\mathbb{P}C_{kl}(t,s,\boldsymbol{Z})h_k(t)d\Lambda_{k0}(t)\right\}$$

and define a new semi-norm $\|\cdot\|$ by $\|\boldsymbol{h}\|=<\Gamma[\boldsymbol{h}],\boldsymbol{h}>_{L_2(\boldsymbol{\Lambda}_0)}$. To show that this semi-norm is a norm, suppose that $\|\boldsymbol{h}\|=0$. Since $\|\boldsymbol{h}\|^2=\mathbb{P}l_{\boldsymbol{\Lambda}}[\boldsymbol{h}]^2$, we see that $l_{\boldsymbol{\Lambda}}[\boldsymbol{h}]=0$ with probability one. Let $\xi\widetilde{D}=k$ and $\Delta_j=1$ $(j=1,\ldots,J)$ successively to obtain that

$$\int_0^{U_1}e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)}h_k(t)d\Lambda_{k0}(t)=0,\ \int_0^{U_2}e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)}h_k(t)d\Lambda_{k0}(t)=0,\ldots.$$

Thus, $\int_0^s e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(t)}h_k(t)d\Lambda_0(t)=0$ for $s\in[\zeta,\tau]$. We take the derivative with respect to $s$ to obtain that $h_k=0$ on $[\zeta,\tau]$. Thus, $\|\boldsymbol{h}\|$ is a norm. Clearly, $\|h\|\le c<h,h>^{1/2}$ for some constant $c$. By the bounded inverse theorem in Banach spaces, we have $<h,h>_{L_2(\boldsymbol{\Lambda}_0)}\le c_1^2\|h\|$ for another constant $c_1$. By the Lax-Milgram theorem (Zeidler, 1995), $\boldsymbol{h}^*$ exists as the solution to the system of equations (3.36). In other words, the least favorable direction exists.

We differentiate the system of equations (3.36) to obtain

$$\sum_{k=1}^{K}\left\{b_{1lk}(t)\boldsymbol{h}_k^*(t)+\int_t^{\tau}b_{2lk}(t,s)\boldsymbol{h}_k^*(s)ds+\int_{\zeta}^{t}b_{3lk}(t,s)\boldsymbol{h}_k^*(s)ds\right\}=\boldsymbol{b}_{4l}(t),$$

where $b_{1lk}(t)>0$, and the columns of the matrix $\boldsymbol{b}_1(t)$, with $b_{1lk}(t)$ as the $lk$th element, are linearly independent. In addition, the functions $b_{2lk},b_{2lk}$, and $\boldsymbol{b}_{4l}$ are continuously differentiable with respect to their arguments. Hence, each component of $\boldsymbol{h}^*$ is continuously differentiable.

*Step 3 (Verifying the positive definiteness of the efficient information matrix $\boldsymbol{\Sigma}$).* If $\boldsymbol{\Sigma}$ is singular, then there exists a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^{\mathrm{T}}E\left\{(l_{\boldsymbol{\beta}}-l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*])(l_{\boldsymbol{\beta}}-l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*])^{\mathrm{T}}\right\}\boldsymbol{v}=0$. It follows that the score function along the sub-model $\{\boldsymbol{\beta}_0+\epsilon\boldsymbol{v},(1+\epsilon\boldsymbol{v}^{\mathrm{T}}\boldsymbol{h}_1^*)d\Lambda_{10},\ldots,(1+\epsilon\boldsymbol{v}^{\mathrm{T}}\boldsymbol{h}_K^*)d\Lambda_{K0}\}$ is zero with probability one. In particular, we let $\xi\widetilde{D}=k$ and $\Delta_j=1$ $(j=1,2,\ldots)$ to obtain $\int_0^{U_j}\widetilde{\boldsymbol{h}}_k(t)d\Lambda_{k0}(t)=0$, where $\widetilde{\boldsymbol{h}}_k(t)=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Z}(t)+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{h}_k^*(t)$. Therefore, with probability one, $\int_0^{t}\widetilde{\boldsymbol{h}}_k(t)d\Lambda_{k0}(t)=0$ for any $t\in[\zeta,\tau]$. This implies that $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Z}(t)+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{h}^*(t)=0$, so $\boldsymbol{v}=\mathbf{0}$ by condition (C3).

*Step 4 (Deriving the asymptotic linear expansion for $\widehat{\boldsymbol{\beta}}$).* Since $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ is the NPMLE, the score along the sub-model $(\boldsymbol{\beta},(1+\epsilon\boldsymbol{h}^*)d\widehat{\boldsymbol{\Lambda}})$ is zero. Thus,

$$\mathbb{P}_n\left\{\boldsymbol{l}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})-l_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\}=0.$$

Equivalently,

$$-\sqrt{n}\mathbb{P}\left\{\boldsymbol{l}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})-l_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\}=\mathbb{G}_n\left\{\boldsymbol{l}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})-l_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\}.$$

Since the components of $\boldsymbol{h}^*$ are continuously differentiable, the (signed) measures $\boldsymbol{h}_k^*d\Lambda_k$ $(k=1,\ldots,K)$ have uniformly bounded total variations. By the preservation of the Donsker property via convex hull, it is easy to show that $\left\{\boldsymbol{l}_{\beta}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})-l_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\}$ is Donsker and converges in $L_2(\mathbb{P})$ to the efficient score $\boldsymbol{l}_{\boldsymbol{\beta}}-l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*]$. Thus,

$$-\sqrt{n}\mathbb{P}\left\{\boldsymbol{l}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})-l_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\}=\mathbb{G}_n\left\{\boldsymbol{l}_{\boldsymbol{\beta}}-l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*]\right\}+o_P(1). \tag{3.37}$$

By the Taylor series expansion at $(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$, together with the property of the least favorable direction $\boldsymbol{h}^*$, the left-hand side of (3.37) is equal to

$$\boldsymbol{\Sigma}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+n^{1/2}O_P\left(|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|^2+\mathbb{P}\left[\sum_{k=1}^{K}\sum_{j=1}^{J}\left\{\int_0^{U_j}e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\widehat{\Lambda}_k(s)-\int_0^{U_j}e^{\boldsymbol{\beta}_{k0}^{\mathrm{T}}\boldsymbol{Z}(s)}d\Lambda_{k0}(s)\right\}^2\right]\right).$$

By (3.33) and the non-singularity of $\boldsymbol{\Sigma}$, the second term is $o_P(1)$, so that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \boldsymbol{\Sigma}^{-1}\mathbb{G}_n \left\{ \boldsymbol{l}_{\boldsymbol{\beta}} - l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*] \right\} + o_P(1).$$

As a result, $\widehat{\boldsymbol{\beta}}$ is an asymptotically linear estimator for $\boldsymbol{\beta}_0$, and its influence function is $\boldsymbol{\Sigma}^{-1} \left\{ \boldsymbol{l}_{\boldsymbol{\beta}} - l_{\boldsymbol{\Lambda}}[\boldsymbol{h}^*] \right\}$, which is the efficient influence function. Hence, $\widehat{\boldsymbol{\beta}}$ is asymptotically normal and semiparametrically efficient.

# CHAPTER 4: EXTENSIONS AND FUTURE RESEARCH

In this chapter, we propose regression procedures for two types of data which are closely akin to right- and interval-censored competing risks. One is the partly interval-censored competing risks and the other is the interval-censored failure time with continuous marks. In the end, we outline several possible directions for future research.

## 4.1 Partly Interval-Censored Competing Risks

### 4.1.1 Introduction

In some studies with competing risks endpoint, some of the risks are interval censored and the rest are right censored. This type of data typically arise if death is among the otherwise asymptomatic risks. For example, in the Women's Interagency HIV Study (WIHS) established in August 1993 to investigate the impact of HIV infection on women in New York, 1,164 HIV-positive women free of clinical AIDS were enrolled and were followed up at 6 month intervals (Lau et al., 2009). The endpoints were first occurrence of treatment initiation, AIDS diagnosis, or death. The first two risks are interval censored while death is observed exactly or right censored. The parametric models proposed by Hudgens et al. (2014) can be used to analyze this type of data. However, no regression methods have been proposed in the literature. In this section, we extend our models and methods to the partly interval-censored competing risks data.

### 4.1.2 Models and Methods

Let $T$ denote the failure time with $K$ competing causes, among which the first $K'$ risks are subject to right censoring, and the remaining to interval censoring. Let $\boldsymbol{Z}(\cdot)$ be a set of $p$-dimensional (possibly time-dependent) covariates. Let $D \in \{1, \cdots, K\}$ denote the cause of failure. We model the conditional sub-distribution hazard for each risk using the semiparametric transformation models specified in (2.3).

Let $U_1 < \ldots < U_J$ be a sequence of random follow-up times, where $J$ is random positive integer. Write $\boldsymbol{U} = (U_0, U_1, \ldots, U_J)^{\mathrm{T}}$, where $U_0 \equiv 0$. Define $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_J)^{\mathrm{T}}$, where $\Delta_j = I(U_{j-1} < T \leq U_j)$ $(j = 1, \ldots, J)$. Write $\widetilde{D} = DI(T \leq U_J)$. We allow $\widetilde{D}$ to be possibly missing. Let $\xi = 1, 0$ indicate whether $\widetilde{D}$ is observed or not. We set $\xi = 1$ if $\widetilde{D} = 0$. However, $\delta \equiv I(1 \leq \widetilde{D} \leq K')$, i.e., the information on whether the (possibly unknown) cause is of the right-censored or interval-censored type, is always observable. For a random sample of $n$ subjects, the observed data consist of

$$\{\delta_i, \delta_i T_i, J_i, \boldsymbol{U}_i, \boldsymbol{\Delta}_i, \xi_i, \xi_i \widetilde{D}_i, \boldsymbol{Z}_i\} \, (i = 1, \ldots, n).$$

Suppose that $(T, D)$ and $\boldsymbol{U}$ are independent given $\boldsymbol{Z}$, and that $\widetilde{D}$ is missing at random (MAR). Then, the likelihood function for $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Lambda} \equiv (\Lambda_1, \cdots, \Lambda_K)$ can be written as

$$\begin{aligned} L_n(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \prod_{i=1}^{n} \Bigg( & \prod_{k=1}^{K'} F_k'(T_i; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k)^{I(\xi_i \widetilde{D}_i = k, \delta_i = 1)} \left\{ \sum_{k=1}^{K'} F_k'(T_i; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) \right\}^{I(\xi_i = 0, \delta_i = 1)} \\ & \times \left[ \prod_{j=1}^{J_i} \prod_{k=K'+1}^{K} \bigg( F_k(U_{ij}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{i,j-1}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) \bigg)^{I(\xi_i \widetilde{D}_i = k, \Delta_{ij} = 1, \delta_i = 0)} \right. \\ & \left. \times \left\{ \sum_{k=K'+1}^{K} \bigg( F_k(U_{ij}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) - F_k(U_{i,j-1}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) \bigg) \right\}^{I(\xi_i = 0, \Delta_{ij} = 1, \delta_i = 0)} \right] \\ & \times \left( 1 - \sum_{k=1}^{K} F_k(U_{i,J_i}; \boldsymbol{Z}_i, \boldsymbol{\beta}_k, \Lambda_k) \right)^{I(\widetilde{D}_i = 0)} \Bigg). \end{aligned} \tag{4.38}$$

We can use an EM algorithm similar to that described in §3.2.3 to compute the NPMLE. Denote the resulting estimators as $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$.

### 4.1.3 Asymptotic Properties

We impose the following regularity conditions:

(C1) The true value for $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}_0$, is contained in a known compact set of the Euclidean space $\mathbb{R}^d$, where $d = Kp$. For $k = 1, \cdots, K$, the true value for $\Lambda_k(\cdot)$, denoted as $\Lambda_{k0}(\cdot)$, is continuously differentiable with positive derivatives on $[\zeta, \tau]$ with $\Lambda_{k0}(\zeta) > 0$, where $[\zeta, \tau]$ is the union of the support of $(U_1, \ldots, U_J)$, and $0 < \zeta < \tau$. In addition, $\Pr(T \geq \tau | \boldsymbol{Z}) > c$ with probability one for some positive constant $c$.

(C2) The vector of covariates $\boldsymbol{Z}(t)$ is uniformly bounded with uniformly bounded total variation over $[\zeta, \tau]$, and its left-limits exist for all $t$. In addition, $E[g\{\boldsymbol{Z}_{(l)}(t)\}]$ $(l = 1, 2)$ is continuously differentiable in $[\zeta, \tau]$, where $g(\cdot)$ is any continuously differentiable function, and $\boldsymbol{Z}_{(1)}(t)$ and $\boldsymbol{Z}_{(2)}(t)$ are vectors of increasing functions whose difference is $\boldsymbol{Z}(t)$.

(C3) For $k = 1, \cdots, K$, if $h(t) + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{Z}(t) = 0$ for all $t \in [\zeta, \tau]$ with probability one, then $h(t) = 0$ for $t \in [\zeta, \tau]$ and $\boldsymbol{\beta}_k = \mathbf{0}$.

(C4) The number of examination times, $J$, is positive, and $E(J) < \infty$. In addition, $\Pr(U_J = \tau | J, \boldsymbol{Z}) > \delta$ for some positive constant $\delta$, and there exists some positive constant $\eta$ such that $\Pr\{\min_{1 \leq j \leq J}(U_j - U_{j-1}) \geq \eta | J, \boldsymbol{Z}\} = 1$. Furthermore, the sum of the marginal density functions of $U_1, \ldots, U_J$ is strictly positive on $[\zeta, \tau]$. Finally, the conditional density function of $\boldsymbol{U}$ given $\boldsymbol{Z}$ and $J = j$, denoted as $g_j(u_1, \ldots, u_j | \boldsymbol{Z}, J = j)$, is strictly positive on $[\zeta, \tau]$ with continuous second-order partial derivatives with respect to $u_1, \ldots, u_j$ when $u_{j'} - u_{j'-1} > \eta$ $(j' = 2, \ldots, j)$ and is continuously differentiable with respect to $\boldsymbol{Z}$.

(C5) For $k = 1, \cdots, K'$, the transformation function $G_k$ is four-times differentiable with $G_k(0) = 0$ and $G_k'(x) > 0$, and for any $c_0 > 0$,

$$\limsup_{x \to \infty} [\{G_k(c_0 x)\}^{-1} \log\{x \sup_{y \leq x} G_k'(y)\}] = 0. \tag{4.39}$$

For $k = K'+1, \cdots, K$, the transformation function $G_k$ is twice-continuously differentiable on $[0, \infty)$ with $G_k(0) = 0$, $G_k'(x) > 0$, and $G_k(\infty) = \infty$.

(C6) With probability one, $\Pr(\xi = 1 | T, D, \boldsymbol{Z}, J, \boldsymbol{U}, \boldsymbol{\Delta} \neq 0) = \Pr(\xi = 1 | \delta, \boldsymbol{Z}, J, \boldsymbol{U}, \boldsymbol{\Delta} \neq 0) > c_0$ for some positive constant $c_0$. Moreover, the conditional probability of $\xi = 1$ given $\delta = 0$, $\boldsymbol{Z}$, $J = j$, $\boldsymbol{U} = (u_1, \ldots, u_j)^{\mathrm{T}}$, and $\boldsymbol{\Delta} \neq 0$, denoted as $\rho_j(u_1, \ldots, u_j, \boldsymbol{Z})$, has continuous second-order partial derivatives with respect to $u_1, \ldots, u_j$ when $u_{j'} - u_{j'-1} > \eta$ $(j' = 2, \ldots, j)$ and is continuously differentiable with respect to $\boldsymbol{Z}$.

*Remark 4.1.* Conditions (C1)-(C4) are identical to (C1)-(C4) in §3.2.4 for interval-censored data. Condition (C5) is a hybrid of the conditions on transformation functions in §2.2.3 for right-censored data and those in §3.2.4 for interval-censored data. Condition (C6) implies the MAR mechanism for the cause of failure and ensures that the smoothness conditions on the missing probabilities, similar to those in (C6) of §3.2.4, hold for the interval-censored risks.

The following theorem establishes the consistency of the NPMLEs.

*Theorem 4.1.* Under Conditions (C1)-(C6), $\widehat{\boldsymbol{\beta}}$ and $\widehat{\Lambda}_k$ $(k = 1, \ldots, K)$ are strongly consistent, i.e.,

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| + \sup_{t \in [\zeta, \tau]} \sum_{k=1}^{K} \left| \widehat{\Lambda}_k(t) - \Lambda_{k0}(t) \right| \longrightarrow 0$$

almost surely, where $|| \cdot ||$ denotes the Euclidean norm.

Denote $BV_1$ as the space of functions on $[\zeta, \tau]$ that are uniformly bounded by 1 and with total variation bounded by 1. Write $\mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}|| \leq 1\}$ and $\mathcal{W} = BV_1^{\otimes K'}$, As in §2.2.3, we identify $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\Lambda}_1 - \Lambda_{10}, \cdots, \widehat{\Lambda}_{K'} - \Lambda_{K'0})$ as random elements in $l^\infty(\mathcal{V} \times \mathcal{W})$. The next theorem on the asymptotic Gaussianity of estimators for the Euclidean parameter and cumulative hazard functions for the right-censored risks.

*Theorem 4.2.* Under Conditions (C1)-(C6), $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \widehat{\Lambda}_1 - \boldsymbol{\Lambda}_{10}, \cdots, \widehat{\Lambda}_{K'} - \boldsymbol{\Lambda}_{K'0})$ converges weakly to a zero-mean Gaussian process in $l^\infty(\mathcal{V} \times \mathcal{W})$. In addition, $\widehat{\boldsymbol{\beta}}$ is semiparametric efficient in the sense of Bickel et al. (1993).

As of consequence, $\widehat{\boldsymbol{\beta}}$ is asymptotically multivariate normal. Its variance can be estimated by the profile likelihood method similar to that described in §3.2.3.

### 4.1.4 Simulation Studies

We conducted simulation studies to assess the finite sample performance of the NPMLEs. We simulated $K = 3$ risks, the first two of which were interval censored and the last right censored. Set $\Lambda_k(t) = 0.2(1 - e^{-t})$ $k = 1, 2, 3$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_3 = (0.25, -0.25)^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (-0.25, 0.25)^{\mathrm{T}}$. For covariates, let $Z_1(t) = B_1 I(t \leq V) + B_2 I(t > V)$ and $Z_2 \sim$ Unif$[0, 1]$, where $B_1$ and $B_2$ are independent Bernoulli$(0.5)$, and $V \sim$ Unif$[0, 3]$. For the examination times $\boldsymbol{U} = (U_1, U_2)$, let $U_1$ and $U_2 - U_1$ be two independent random variables distributed as the minimum of 1.5 and an exponential random variable with hazard $0.5$. We considered $G_k(x) = r^{-1}\log(1 + rx)$ with $r = 0$, $1$, and $0.5$. Under these conditions, each cause had an event rate of roughly $15\%$. We used $10^{-3}$ as threshold for convergence of the EM algorithm and used $h_n = n^{-1/2}$ for variance estimation (see §3.2.3).

The results for $\beta_{11}$ and $\beta_{31}$, the regression coefficients of $Z_1$ for the first and third causes, respectively, are summarized in Table 4.16. For both estimators, the bias is negligible and the standard error estimates reflect the true variations well. In addition, the empirical coverage probability of the 95% confidence interval shows that the normal approximation of distribution is fairly accurate. As shown in Figure 4.5, estimation of the cumulative hazard functions has minimal bias, especially for the right-censored risk ($k = 3$) and for sample size $n = 500$.

## 4.2 Interval-Censored Failure Time with Continuous Marks

### 4.2.1 Introduction

Competing risks data arise when the failure time of interest is associated with a finite set of causes. In many cases, however, the cause of failure may take continuous values, and is thus termed a continuous "mark". For example, in HIV vaccine clinical trials, the time to sero-conversion is

Table 4.16: Simulation results on the estimation of $\beta_{11}$ and $\beta_{31}$.

| $n$ | $r$ | | Bias | SE | SEE | CP |
|---|---|---|---|---|---|---|
| 200 | 0 | $\beta_{11}$ | 0.006 | 0.295 | 0.295 | 0.952 |
| | | $\beta_{31}$ | 0.008 | 0.240 | 0.238 | 0.949 |
| | 0.5 | $\beta_{11}$ | −0.012 | 0.287 | 0.285 | 0.947 |
| | | $\beta_{31}$ | 0.003 | 0.247 | 0.247 | 0.951 |
| | 1 | $\beta_{11}$ | −0.009 | 0.301 | 0.300 | 0.951 |
| | | $\beta_{31}$ | −0.011 | 0.245 | 0.244 | 0.950 |
| 500 | 0 | $\beta_{11}$ | 0.005 | 0.189 | 0.190 | 0.953 |
| | | $\beta_{31}$ | 0.000 | 0.147 | 0.149 | 0.955 |
| | 0.5 | $\beta_{11}$ | −0.001 | 0.171 | 0.177 | 0.962 |
| | | $\beta_{31}$ | 0.006 | 0.159 | 0.162 | 0.958 |
| | 1 | $\beta_{11}$ | −0.003 | 0.191 | 0.197 | 0.963 |
| | | $\beta_{31}$ | 0.006 | 0.150 | 0.149 | 0.950 |

Bias and SE are the bias and standard error of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; Each entry is based on 10,000 replicates.

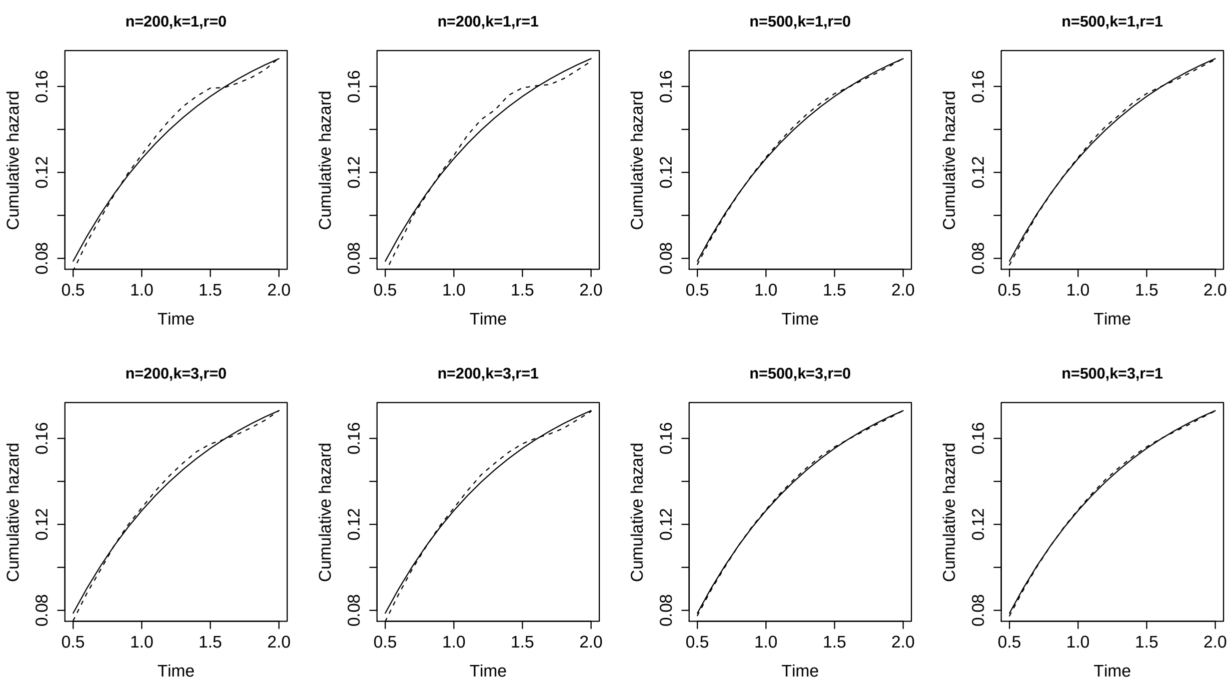


Figure 4.5: Estimation of the cumulative hazard function $\Lambda_k(\cdot)$ by the NPMLE. The true values and the mean estimates (based on 10,000 replicates) are shown by the solid and dashed curves, respectively.

coupled with the genetic distance between the infecting virus and the virus in the vaccine, which carries important information about the specificity of efficacy of the vaccine (Hudgens et al., 2007). Because of the mutational diversity of the HIV, each sero-converted subject may have a unique value for the viral genetic distance, warranting its treatment as a continuous mark. Furthermore, the sero-conversion time is typically not observed exactly but can only be known to lie within a pair of adjacent examination times. In this sense, the event time is said to be interval censored.

Unlike the competing risks data, each failure cause in the continuous mark data cannot be treated individually, and so novel methods are required to analyze such data. For right-censored failure time with a continuous mark, Huang and Louis (1998) studied the nonparametric maximum likelihood estimation (NPMLE) for the joint distribution of the failure time and mark. Gilbert et al. (2008) developed a two-sample testing procedure using a kernel-smoothed log-rank type test. In the regression setting, Sun et al. (2009) proposed the proportional mark-specific hazards model with regression parameters as nonparametric functions of the mark, and used a kernel-smoothed partial likelihood of the Cox model for inference. However, the mark-specific hazard function, in parallel with the cause-specific hazard function in competing risks setting, pertains to the instantaneous risk of failure with a specific mark value conditioning on survival to that point, and does not correspond directly with the more intuitive cumulative incidence (sub-distribution) for that mark value. Moreover, for application to interval-censored failure time in HIV vaccine trials, methods for right-censored data, usually with *ad hoc* right or mid-point imputation, may lead to biased inference.

In the interval-censoring setting, Hudgens et al. (2007) studied the computations of the NPMLE (corresponding to the MLE for right imputed data), the mid-point MLE, and a coarsened-mark MLE, where the mark variable is categorized into discrete groups. Maathuis and Wellner (2008) proved rigorously that the NPMLE for interval-censored continuous mark data is generally inconsistent, but that the coarsened-mark MLE can be consistent, provided that total number of discretizing knots increases with the sample size at appropriate rates. Alternatively, Groeneboom et al. (2012) applied the maximum smoothed likelihood estimation (MSLE) to current-status continuous mark data, where the each subject has only one examination. To our knowledge, no regression methods for interval-censored continuous mark data are available in the literature. In this paper, we extend the approach of coarsened-mark MLE to a class of spline regression models for interval-censored continuous mark data.

### 4.2.2 Models and Methods

Let $T$ denote the failure time and let $X$ denote the continuous mark. Without loss of generality, assume $X \in [0, 1]$. Let $\boldsymbol{Z}$ be a set of possibly time-dependent covariates. We consider the conditional cumulative incidence functions of $T$ associated with specific values of $X$, i.e.,

$$F(t, x; \boldsymbol{Z}) = \frac{\partial}{\partial x} \Pr(T \leq t, X \leq x | \boldsymbol{Z}),$$

which satisfy $\int_0^1 F(\infty, x; \boldsymbol{Z}) dx = 1$.Thus, $F(t, x; \boldsymbol{Z})$ can be interpreted as the "density" of the cumulative incidence function with respect to the mark, in the sense that

$$\Pr(T \leq t, x \leq X < x + \Delta x | \boldsymbol{Z}) \approx \Delta x F(t, x; \boldsymbol{Z}).$$

We propose to model the conditional cumulative incidence density function by

$$F(t, x; \boldsymbol{Z}) = G_x \left\{ \int_0^t e^{\boldsymbol{\beta}(x)^{\mathrm{T}} \boldsymbol{Z}(s)} d\Lambda(s, x) \right\}, \tag{4.40}$$

where $G_x(\cdot)$ is a known family of increasing functions, $\boldsymbol{\beta}(x)$ is vector of nonparametric regression parameters, and $\Lambda(s, x)$ is an arbitrary function increasing in $s$ with $\Lambda(0, x) = 0$ for all $x$.

*Remark 4.2.* We may define the "sub-distribution hazard function" as in competing risks setting

$$d\Lambda(t, x; \boldsymbol{Z}) \Delta x = \frac{dF(t, x; \boldsymbol{Z}) \Delta x}{1 - dF(t, x; \boldsymbol{Z}) \Delta x} \approx dF(t, x; \boldsymbol{Z}) \Delta x$$

So, because of the nature of $F(t, x; \boldsymbol{Z})$ as a density function with respect to $x$, the "sub-distribution hazard function" associated with $x$ is the same as $F(t, x; \boldsymbol{Z})$.

Let $U_1 < \ldots < U_J$ be a sequence of random follow-up times, where $J$ is random positive integer. Write $\boldsymbol{U} = (U_0, U_1, \ldots, U_J)^{\mathrm{T}}$, where $U_0 \equiv 0$. Define $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_J)^{\mathrm{T}}$, where $\Delta_j = I(U_{j-1} < T \leq U_j)$ $(j = 1, \ldots, J)$. Write $\widetilde{X} = XI(T \leq U_J)$. We allow $\widetilde{X}$ to be possibly missing. Let $\xi = 1, 0$ indicate whether $\widetilde{X}$ is observed or not. We set $\xi = 1$ if $\widetilde{X} = 0$. For a random sample of $n$ subjects, the observed data consist of

$$\{J_i, \boldsymbol{U}_i, \boldsymbol{\Delta}_i, \xi_i, \xi_i\widetilde{X}_i, \boldsymbol{Z}_i\}\ (i = 1, \ldots, n).$$

Suppose that $(T, X)$ and $\boldsymbol{U}$ are independent given $\boldsymbol{Z}$, and that $\widetilde{X}$ is missing at random (MAR). Then, the likelihood can be written as

$$\begin{aligned} L_n(\boldsymbol{\beta}, \Lambda) = \prod_{i=1}^{n} \Bigg[ \prod_{j=1}^{J_i} & \Big( F(U_{ij}, \widetilde{X}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) - F(U_{i,j-1}, \widetilde{X}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) \Big)^{I(\xi_i=1, \Delta_{ij}=1)} \\ & \times \left\{ \int_0^1 \Big( F(U_{ij}, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) - F(U_{i,j-1}, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) \Big) dx \right\}^{I(\xi_i=0, \Delta_{ij}=1)} \Bigg] \\ & \times \Bigg( 1 - \int_0^1 F(U_{i,J_i}, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) dx \Bigg)^{I(\boldsymbol{\Delta}=\boldsymbol{0})} \Bigg). \end{aligned} \tag{4.41}$$

Denote $(L_i, R_i]$ as the interval among $(U_{i0}, U_{i1}], \cdots, (U_{i,J_i-1}, U_{i,J_i}], (U_{i,J_i}, \infty)$ that brackets $\widetilde{T}_i$. Then the likelihood (4.41) can be re-written as

$$\begin{aligned} L_n(\boldsymbol{\beta}, \Lambda) = & \prod_{i:\xi_i=1, R_i<\infty} \Big( F(R_i, \widetilde{X}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) - F(L_i, \widetilde{X}_i; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) \Big) \\ & \times \prod_{i:\xi_i=0} \int_0^1 \Big( F(R_i, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) - F(L_i, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) \Big) dx \\ & \times \prod_{i:R_i=\infty} \Big( 1 - \int_0^1 F(L_i, x; \boldsymbol{Z}_i, \boldsymbol{\beta}, \Lambda) dx \Big). \end{aligned} \tag{4.42}$$

Direct maximization of (4.42) is not only difficulty but also likely to lead to biased estimators due to the "representational non-uniqueness" inherent in the continuous mark data structure (Maathuis and Wellner, 2008). Instead, we circumvent this non-uniqueness problem by borrowing strength across neighboring values of $x$. Specifically, we model the nonparametric regression function $\boldsymbol{\beta}(x)$ and baseline function $\Lambda(t, x)$ by splines. Let $\{B_{11}(x), \cdots, B_{1,q_n}(x)\}$ be the set of spline functions, e.g., B-splines, generated from the knots $x_1, \cdots, x_{q_n}$. Then, we model the regression function by $\boldsymbol{\beta}(x, \boldsymbol{\gamma}) \equiv \sum_{j=1}^{q_n} \gamma_j B_j(x)$, where $\boldsymbol{\gamma} \equiv (\gamma_1, \cdots, \gamma_{q_n})^{\mathrm{T}}$ are Euclidean parameters.

The modeling of the baseline function $\Lambda(t, x)$ needs to take into account the fact that $\Lambda(\cdot, x)$ is non-decreasing for every $x$. To this aim, denote $\phi(t, x) = \log\{\partial\Lambda(t, x)/\partial t\}$, and model $\phi(t, x)$ by two-dimensional B-splines. Specifically, let $B_{21}(t), \cdots, B_{2,m_n}(t)$ be the set of splines generated from the knots $t_1, \cdots, t_{m_n}$. Write $C_{ij}(t, x) = B_{1,j}(x)B_{2,i}(t)$, $i = 1, \cdots, m_n$, $j = 1, \cdots, q_n$. Then, we model $\phi(t, x)$ by

$$\phi(t, x; \boldsymbol{\alpha}) = \sum_{i=1}^{m_n} \sum_{j=1}^{q_n} \alpha_{ij} C_{ij}(t, x),$$

where $\boldsymbol{\alpha} = (\alpha_{11}, \cdots, \alpha_{1,q_1}, \alpha_{21}, \cdots, \alpha_{m_n,q_n})^{\mathrm{T}}$. Under these models, the sub-distribution for the $i$th subject can be written as

$$F(t, x; \boldsymbol{Z}) = G_x \left\{ \int_0^t e^{\boldsymbol{\beta}(x,\boldsymbol{\gamma})^{\mathrm{T}} \boldsymbol{Z}(s) + \phi(s,x;\boldsymbol{\alpha})} ds \right\}. \tag{4.43}$$

Inserting (4.43) to the likelihood function, we can compute the MLE by the Newton-Raphson algorithm.

### 4.2.3 Simulation Studies

Let $Z = -1$ and $1$ with equal probability, and let $\beta(x) = 0.25x + (x - 0.5)^2$. For the examination times $\boldsymbol{U} = (U_1, U_2)$, let $U_1$ and $U_2 - U_1$ be two independent random variables distributed as the minimum of 1.5 and an exponential random variable with hazard $0.5$. Since $U_2 \leq 3$, we need only consider $T$ on the interval $[0, 3]$. We simulated $(T, X)$ on $[0, 3] \times [0, 1]$ under the model

$$F(t, x; Z) = 0.3 \exp(\beta(x)Z + x(1 - e^{-t})).$$

Under these settings, the event rate was roughly 50%.

Table 4.17: Simulation results on the estimation of $\beta(x)$.

| | | Two knots | | | | Three knots | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $r$ | Bias | SE | SEE | CP | Bias | SE | SEE | CP |
| 200 | $\beta(0.1) = 0.185$ | $-0.017$ | 0.254 | 0.253 | 0.929 | $-0.009$ | 0.275 | 0.281 | 0.951 |
| | $\beta(0.5) = 0.125$ | 0.008 | 0.218 | 0.222 | 0.949 | 0.001 | 0.234 | 0.234 | 0.947 |
| | $\beta(0.7) = 0.215$ | $-0.023$ | 0.234 | 0.238 | 0.933 | $-0.008$ | 0.251 | 0.252 | 0.942 |
| 500 | $\beta(0.1) = 0.185$ | $-0.012$ | 0.169 | 0.173 | 0.943 | $-0.006$ | 0.186 | 0.187 | 0.947 |
| | $\beta(0.5) = 0.125$ | 0.005 | 0.143 | 0.149 | 0.954 | 0.003 | 0.161 | 0.167 | 0.956 |
| | $\beta(0.7) = 0.215$ | $-0.015$ | 0.149 | 0.157 | 0.949 | $-0.009$ | 0.168 | 0.172 | 0.942 |

See the note to Table 4.16. Each entry is based on 2,000 replicates.

We fitted the data using cubic splines for $\beta(x)$ and tensor cubic splines for $\phi(t, x)$ with sample tertiles (2 knots) and quartiles (3 knots) respectively. The results are summarized in Table 4.17. Overall, the bias is tolerable and the standard error estimators and empirical coverage probabilities are fairly accurate. Specifically, the 3-knot models have smaller bias and larger variance than do the 2-knot ones. In terms of the coverage probability, the 3-knot models have better performance.

### 4.2.4 An HIV Vaccine Study

We considered VAX004, a randomized controlled trial of a HIV-1 vaccine glycoprotein 120 (rgpl20) conducted at 61 sites in North America and the Netherlands (Hudgens et al., 2007). Among the 5403 initially HIV-1 sero-negative subjects, 3598 were assigned to receive vaccine and the remaining 1805 to receive placebo. Each subject was followed with a maximum of three years and examined at roughly 6 month intervals for evidence of sero-conversion. In the treatment group, 241 participants were infected; in the placebo group, 127 were infected. Previous analysis with time to sero-conversion as the primary endpoint did not establish significant efficacy of the vaccine.

An additional objective was to determine whether the efficacy of the vaccine (if any) depends on the genetic sequence of the infecting HIV virus. In particular, for infected subjects the amino acid sequence of the gpl20 region of one HIV-1 isolate was determined, and the set of sequences was aligned together with the gpl20 amino acid sequence of the GNE8 HIV that was represented in the vaccine. Each HIV amino acid sequence was 581 positions long. The distance (mark variable) between each infecting HIV sequence and the GNE8 sequence was computed as a weighted Hamming distance, that is, the percent mismatch in amino acids with the different possible amino acid substitutions (e.g., A versus C) weighted by the estimated probability of mutation.

Hudgens et al. (2007) used the coarsened-mark MLE to estimate the cumulative incidences as a function of the mark variable for the two treatment arms separately. We first fit the data using the proportional hazards model for the sero-conversion time ignoring the genetic distance. The estimated the hazard ratio for treatment is 0.899 with a p-value of 0.172. So, the reduction in risk is not significant.

To obtain interpretable estimates of the treatment effects for different mark values, we used the proposed methods to analyze the data. We used the proportional hazards model adjusting for baseline risk score, which is an integer variable 0-7 summarizing the amount of risky behavior as measured by a baseline questionnaire. We used cubic splines for $\boldsymbol{\beta}(x)$ and tensor cubic splines for $\phi(t, x)$ with sample tertiles of observed mark values and examination times, respectively, as the knots. The estimated regression parameter for treatment is plotted in Figure 4.6. The parameter values for treatment are all negative, suggesting that the vaccine reduces the incidence of sero-conversion for all genetic distances. The effect is more prominent with smaller mark values, e.g., with a weighted Hamming distance between 0.05 to 0.1. The regression parameter for treatment is minimized at $x = 0.082$ with a value of $-0.251$, which suggests that the incidence of sero-conversion in the treatment group is 0.778 times that in the placebo group for genetic distances in a neighborhood of 0.082. The pointwise confidence intervals show that the treatment effect is significant for genetic distance between 0.085 and 0.165.

To test the overall effect of the vaccine, i.e., $\beta_1(x) = 0$, we used a Wald test on all the regression parameters for the splines pertaining to $\boldsymbol{\beta}_1(x)$ and obtained a p-value of $0.125$. Alternatively, we tested the null hypothesis using $\sup_{x\in[0,1]} |\widehat{\beta}_1(x)|$. To evaluate the null hypothesis, suppose we can expand $\widehat{\beta}_1(x)$ around its true value $\beta_{10}(x)$ in terms of its efficient influence function $f$

$$\widehat{\beta}_1(x) = \beta_{10}(x) + n^{-1/2}\sum_{i=1}^{n} f_i(x) + o_P(1).$$

Then, we approximate the null distribution of $\widehat{\beta}_1(x)$ by the conditional distribution of

$$\widetilde{\beta}_1(x) = n^{-1/2}\sum_{i=1}^{n} \widehat{f}_i(x)Q_i,$$

where $\widehat{f}$ is an estimate for $f$ under $H_0$ and $(Q_1, \cdots, Q_n)$ are independent standard normal variables. Specifically, we simulate the distribution of $\widetilde{\beta}_1(x)$ by repeatedly generating the normal random sample $(Q_1, \cdots, Q_n)$ while holding the observed data fixed. We conducted the test by calculating the p-values for the suprema of the processes based on 1000 realizations. The p-value was calculated to be $0.164$. Thus, neither Wald test nor the supremum test shows a significant overall effect of the vaccine.

Finally, we computed the estimated cumulative incidence functions at 3 years for a subject with median risk score in the two treatment groups in Figure 4.7. There are only minor differences between the two arms.

## 4.3 Future Research

### *4.3.1 More on Partly Interval-Censored Competing Risks*

In §4.1 we have proposed regression methods for partly interval-censored competing risks data. It has been suggested therein that the profile likelihood approach (see §3.2.3) be used to

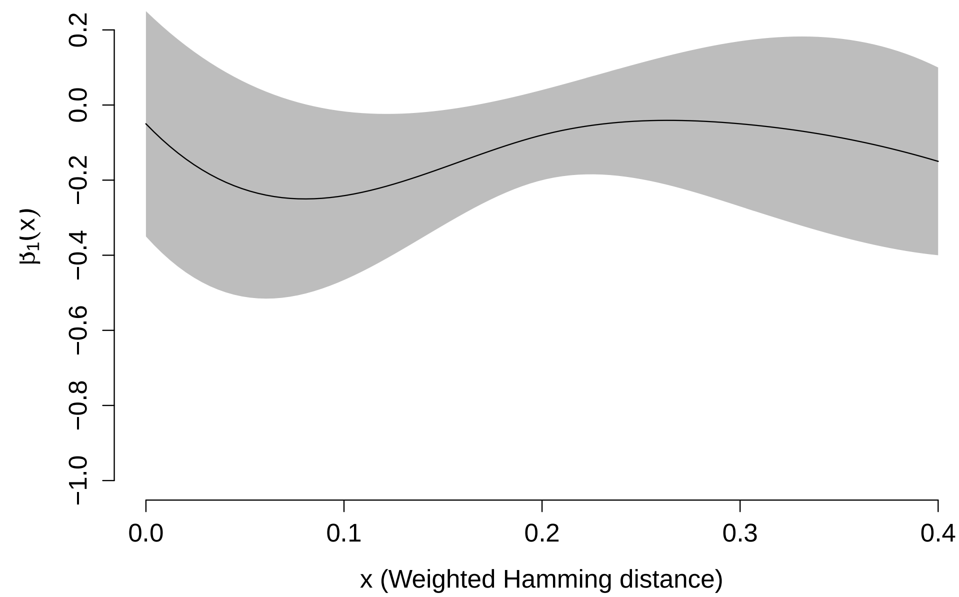


Figure 4.6: Estimated regression parameter for the treatment. The shaded area indicates the pointwise confidence interval.

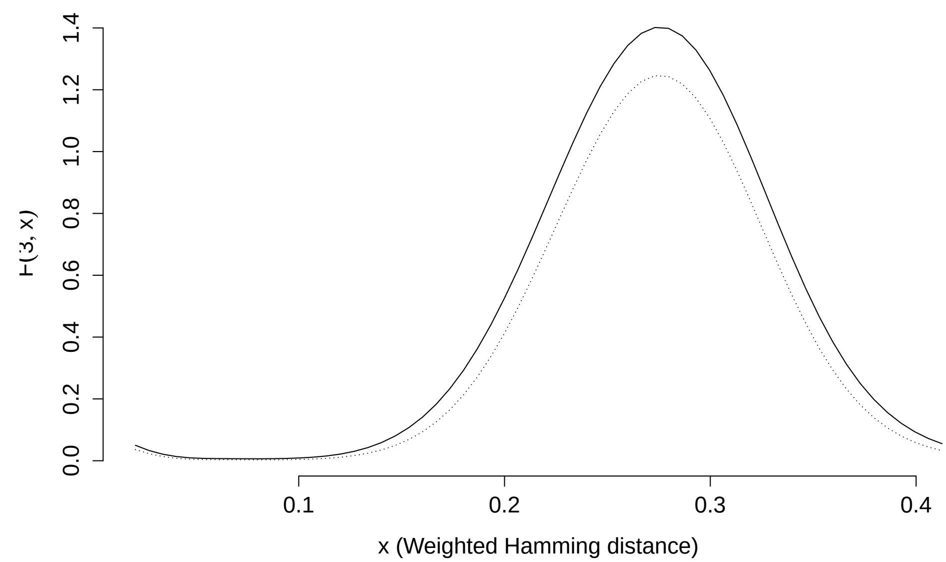


Figure 4.7: Predicted cumulative incidence density at year three for a subject with median baseline risk score. The placebo and treatment groups are shown by the solid and dotted curves, respectively.

estimate the variance of the regression parameter. To estimate the variance for the cumulative hazard functions of the right-censored risks, one could use similar approaches by treating their jump sizes as Euclidean parameters. However, due to the possible high dimension of the jump sizes, numerical differentiation of the profile likelihood is likely to be computationally burdensome and/or unstable. Simple and and efficient estimation of the variance for the cumulative hazard functions remains to be investigated.

In addition, we will collaborate with the Women's Interagency HIV Study (WIHS) to apply our methods to investigate the progression of HIV disease in women. Previous studies treating (interval-censored) AIDS diagnosis as a right-censored endpoint (Lau et al., 2009) were likely to incur bias. Our proposed methods provide a statistically sound and efficient framework for analysis of such data.

#### *4.3.2 Nonparametric Regression of Continuous Marks Data*

We have treated the splines models for the continuous mark data effectively as a (flexible) parametric approach. A nonparametric version is possible by allowing the number of knots to increase with the sample size at proper rates. This entails two challenges. One is concerned with the asymptotic theory, particularly with the growth rate of the number of knots with relation to the sample size. The other is the practical guidelines for choosing the number and locations of knots.

#### *4.3.3 Right- and Interval-Censored Semi-competing Risks Data*

Our methodology for competing risks data is likely to be generalizable to semi-competing risks data, where some risks are not censored by others. For example, (all-cause) mortality and relapse of tumor are common endpoints of semi-competing risks in cancer patients. Unlike the competing risks data, however, frailty or random effects models are probably needed to derive the NPMLE for semi-competing risks data.

### 4.3.4 Variable Selection with Competing Risks Endpoint

Throughout this thesis, we have assumed the number of covariates is fixed and finite (that is, much smaller than the sample size). In many applications with competing risks endpoints, the number of predictors is comparable to or even larger than the sample size, and it thus becomes crucial to identify a relatively small subset of factors that are truly prognostic to the (competing risks) outcome. For example, one would be interested in selecting the genes predictive of different causes of mortality in cancer patients using gene expression data. Recently, Fu et al. (2016) studied various penalization approaches, e.g., LASSO, adaptive LASSO, and SCAD, with the Fine and Gray (1999) objective function. It is of interest to explore such penalized variable selection procedures using the (semiparametric) likelihood functions considered in this thesis.

### 4.3.5 Interval-Censoring under Independent Inspection Process

For interval-censored data, we have considered the mixed-case censoring mechanism, where the whole sequence of monitoring times is assumed to be independent of the failure time and cause conditioning on the covariates. Alternatively, one may be interested in considering the independent inspection process (IIP) mechanism for the monitoring times, in which future monitoring times may depend on all the data observed up to that point. The IIP model seems to be more realistic as in practice the planning of future examinations is often with reference to the available data. In the competing risks setting, the full data likelihoods for the mixed-case and the IIP interval-censorings are identical (Hudgens et al., 2014). However, the reduced-data likelihood (see §3.2.5) under the IIP model involves additional parameters and is different from that under mixed-case censoring. We will study the behavior of the IIP version of "naive estimators" by possibly using ancillary models for monitoring times given observed data.

## REFERENCES


Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012), "Competing Risks in Epidemiology: Possibilities and Pitfalls," *International Journal of Epidemiology*, 41, 861-870.

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273-277.

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

Chen, K., Jin, Z., and Ying, Z. (2002), "Semiparametric Analysis of Transformation Models With Censored Data," *Biometrika*, 89, 659-668.

Chen, L., Lin, D. Y., and Zeng, D. (2012), "Checking Semiparametric Transformation Models With Censored Data," *Biostatistics*, 13, 18-31.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of Transformation Models With Censored Data," *Biometrika*, 82, 835-845.

Cox, D. R. (1972), "Regression Models and Life-tables (With Discussion)," *Journal of the Royal Statistical Society,* Series B, 34, 187-220.

Delord, M. and Génin, E. (2015) "Multiple Imputation for Competing Risks Regression With Interval-Censored Data." *Journal of Statistical Computation and Simulation*, 1-12.

Fine, J. P. and Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496-509.

Fu, Z., Parikh, C. R., and Zhou, B. (2016), "Penalized Variable Selection in Competing Risks Regression," *Lifetime Data Analysis*, 1-24.

Gilbert, P. B., McKeague, I. W., and Sun, Y. (2008), "The 2-sample Problem for Failure Rates Depending on a Continuous Mark: an Application to Vaccine Efficacy," *Biostatistics*, 9, 263-276.

Gray, R. J. (1988), "A Class of $K$-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk," *The Annals of Statistics*, 16, 1141-1154.

Groeneboom, P., Maathuis, M. H., and Wellner, J. A. (2008a), "Current Status Data With Competing Risks: Consistency and Rates of Convergence of the MLE," *The Annals of Statistics*, 36, 1031-1063.

Groeneboom, P., Maathuis, M. H., and Wellner, J. A. (2008b), "Current Status Data With Competing Risks: Limiting Distribution of the MLE," *The Annals of Statistics*, 36, 1064-1089.

Groeneboom, P. and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, New York, NY:Springer.

Groeneboom, P., Jongbloed, G., and Witte, B. I. (2012), "A Maximum Smoothed Likelihood Estimator in the Current Status Continuous Mark Model," *Journal of Nonparametric Statistics*, 24, 85-101.

Gu, M. G., Sun, L., and Zuo, G. (2005), "A Baseline-Free Procedure for Transformation Models under Interval Censorship," *Lifetime Data Analysis*, 11, 473-488.

Huang, J. (1995), "Maximum Likelihood Estimation for Proportional Odds Regression Model with Current Status Data," *Lecture Notes-Monograph Series*, 27, 129âĂŞ145.

Huang, J. and Rossini, A. J. (1997), "Sieve Estimation for the Proportional-Odds Failure-Time Regression Model With Interval Censoring," *Journal of the American Statistical Association*, 92, 960-967.

Huang, J. (1996), "Efficient Estimation for the Proportional Hazards Model With Interval Censoring," *The Annals of Statistics*, 24, 540-568.

Huang, Y. and Louis, T. A. (1998), "Nonparametric Estimation of the Joint Distribution of Survival Time and Mark Variables," *Biometrika*, 85, 785-798.

Huang, J. and Wellner, J. A. (1997), "Interval Censored Survival Data: A Review of Recent Progress," In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Ed. D. Y. Lin and T. R. Fleming, pp. 123-169. New York: Springer-Verlag.

Hudgens, M. G., Li, C., and Fine, J. P. (2014), "Parametric Likelihood Inference for Interval Censored Competing Risks Data," *Biometrics*, 70, 1-9.

Hudgens, M. G., Longini, I. M., Vanichseni, S., Hu, D. J., Kitayaporn, D., Mock, P. A., *et al.*, and Mastro, T. D. (2002), "Subtype-specific Transmission Probabilities for Human Immunodeficiency Virus Type 1 among Injecting Drug Users in Bangkok, Thailand," *American journal of Epidemiology*, 155, 159-168.

Hudgens, M. G., Maathuis, M. H., and Gilbert, P. B. (2007), "Nonparametric Estimation of the Joint Distribution of a Survival Time Subject to Interval Censoring and a Continuous Mark Variable," *Biometrics*, 63, 372-380.

Hudgens, M. G., Satten, G. A., and Longini, I. M. (2001), "Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation," *Biometrics*, 57, 74-80.

Iacobelli, S. (2013), "Suggestions on the Use of Statistical Methodologies in Studies of the European Group for Blood and Marrow Transplantation," *Bone Marrow Transplantation*, 48, S1-S37.

Jeong, J. H. and Fine, J. (2006), "Direct Parametric Inference for the Cumulative Incidence Function," *Journal of the Royal Statistical Society,* Series C, 55, 187-200.

Jeong, J. H. and Fine, J. (2007), "Parametric Regression on Cumulative Incidence Function," *Biostatistics*, 8, 184-196.

Jewell, N. P., van der Laan, M., and Henneman, T. (2003), "Nonparametric Estimation from Current Status Data With Competing Risks," *Biometrika,* 90, 183-197.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd Edition. Hoboken: John Wiley.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004), "Robust Inference for Univariate Proportional Hazards Frailty Regression Models," *The Annals of Statistics*, 32, 1448-1491.

Kumar, S., Zhang, M. J., Li, P., Dispenzieri, A., Milone, G. A., Lonial, S., *et al.* (2011), "Trends in Allogeneic Stem Cell Transplantation for Multiple Myeloma: a CIBMTR Analysis," *Blood*, 118, 1979-1988.

Latouche, A., Porcher, R., and Chevret, S. (2005), "A Note on Including Time-Dependent Covariate in Regression Model for Competing Risks Data," *Biometrical Journal*, 47, 807-814.

Lau, B., Cole, S. R., and Gange, S. J. (2009), "Competing Risk Regression Models for Epidemiologic Data," *American Journal of Epidemiology*, kwp107.

Lawless, J. (2011), *Statistical Models and Methods for Lifetime Data*, New York, NY: Wiley.

Li, C. (2016), "The Fine-Gray Model under Interval Censored Competing Risks Data," *Journal of Multivariate Analysis*, 143, 327-344.

Li, C. and Fine, J. P. (2012), "Smoothed Nonparametric Estimation for Current Status Competing Risks Data," *Biometrika*, ass053.

Lin, H., Zhou, L., Li, C., and Li, Y. (2014), "Semiparametric Transformation Models for Semicompeting Survival Data," *Biometrics*, 70, 599-607.

Lu, W. and Peng, L. (2008), "Semiparametric Analysis of Mixture Regression Models With Competing Risks Data," *Lifetime Data Analysis*, 14, 231-252.

Lu, W. and Ying, Z. (2004), "On Semiparametric Transformation Cure Models," *Biometrika*, 91, 331-343.

Maathuis, M. H. and Wellner, J. A. (2008), "Inconsistency of the MLE for the Joint Distribution of IntervalâĂŘCensored Survival Times and Continuous Marks," *Scandinavian Journal of Statistics*, 35, 83-103.

Murphy, S. A. and Van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449-465.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximum Likelihood Estimation in the Proportional Odds Model," *Journal of the American Statistical Association*, 92, 968-976.

Peng, L. and Fine, J. P. (2007), "Regression Modeling of Semicompeting Risks Data," *Biometrics*, 63, 96-108.

Rabinowitz, D., Betensky, R. A., and Tsiatis, A. A. (2000), "Using Conditional Logistic Regression to Fit Proportional Odds Models to Interval Censored Data," *Biometrics*, 56, 511-518.

Robins, J. M. and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," In *AIDS Epidemiology*, pp. 297-331, Boston:Birkhäuser.

Rossini, A. J. and Tsiatis, A. A. (1996), "A Semiparametric Proportional Odds Regression Model for the Analysis of Current Status Data," *Journal of the American Statistical Association*, 91, 713-721.

Scheike, T. H. and Martinussen, T. (2004), "Maximum Likelihood Estimation for Cox's Regression Model under Case-cohort Sampling," *Scandinavian Journal of Statistics*, 31, 283-293.

Scheike, T. H. and Zhang, M. J. (2008), "Flexible Competing Risks rRegression Modeling and Goodness-of-fit," *Lifetime Data Analysis*, 14, 464-483.

Schick, A. and Yu, Q. (2000), "Consistency of the GMLE With Mixed Case Interval-Censored Data," *Scandinavia Journal of Statistics*, 27, 45-55.

Shen, X. (1998), "Proportional Odds Regression and Sieve Maximum Likelihood Estimation," *Biometrika*, 85, 165-177.

Sun, Y., Gilbert, P. B., and McKeague, I. W. (2009), "Proportional Hazards Models With Continuous Marks," *The Annals of Statistics*, 37, 394.

Sun, J. and Sun, L. (2005), "Semiparametric Linear Transformation Models for Current Status Data," *Canadian Journal of Statistics*, 33, 85-96.

Turnbull, B. W. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society,* Series B, 290-295.

van de Geer, S. A. (2000), *Applications of Empirical Process Theory*, Cambridge: Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Yin, G. and Zeng, D. (2006), "Efficient Algorithm for Computing Maximum Likelihood Estimates in Linear Transformation Models," *Journal of Computational and Graphical Statistics*, 15, 228-245.

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

Zeng, D. and Lin, D. Y. (2006), "Efficient Estimation of Semiparametric Transformation Models for Counting Processes," *Biometrika*, 93, 627-640.

Zeng, D. and Lin, D. Y. (2007), "Maximum Likelihood Estimation in Semiparametric Regression Models With Censored Data (With Discussion)," *Journal of the Royal Statistical Society,* Series B, 69, 507-564.

Zhang, Z., Sun, L., Zhao, X., and Sun, J. (2005), "Regression Analysis of Interval-Censored Failure Time Data With Linear Transformation Models," *Canadian Journal of Statistics*, 33, 61-70.

Zhang, Z. and Zhao, Y. (2013), "Empirical Likelihood for Linear Transformation Models With Interval-Censored Failure Time Data," *Journal of Multivariate Analysis*, 116, 398-409.