Title

Joint Modeling of Longitudinal and Survival Data: Censoring Robust Estimation, Influence Function Based Robust Variance and Shape Based Longitudinal Clustering

Permalink

https://escholarship.org/uc/item/26z7s8kh

Author

Chu, Cangao

Publication Date

2024

UNIVERSITY OF CALIFORNIA,
IRVINE

Joint Modeling of Longitudinal and Survival Data: Censoring Robust Estimation, Influence Function Based Robust Variance and Shape Based Longitudinal Clustering

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Cangao Chu

Dissertation Committee:
Chancellor's Professor Daniel Gillen, Chair
Assistant Professor Tianchen Qian
Professor Zhaoxiao Yu

2024

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Daniel Gillen, my dedicated and supportive advisor, for his invaluable guidance throughout the entire process of completing this thesis. Professor Gillen's unwavering commitment to excellence, keen insights, and willingness to share his wealth of knowledge have been instrumental in shaping the direction and quality of my research. His encouragement and support provided me with the motivation and confidence needed to navigate the challenges inherent in this academic journey. Without Professor Gillen's expertise and mentorship, I would not have been able to achieve this milestone in my academic career. I am truly fortunate to have had him as my advisor, and I extend my heartfelt thanks for his indispensable role in my success.

I would like to extend my sincere thanks to my dissertation committee members, Associate Professor Tianchen Qian and Professor Zhaoxiao Yu, for their invaluable contributions as members of my thesis committee. Their insightful feedback, constructive criticism, and expertise have played a crucial role in shaping the development and refinement of this research. I am grateful for their time, commitment, and dedication to ensuring the scholarly rigor and quality of my work. Their guidance has been instrumental in navigating the challenges of the research process.

I am deeply grateful to the NIH sponsored Titan T32 Training Program for their generous one-year support and mentorship on Alzheimer's Disease. Their assistance was instrumental in my research, offering both financial aid and valuable guidance. I appreciate the opportunity to learn from experienced professionals, and their commitment to Alzheimer's research has significantly influenced my academic success.

I would like to express my sincere gratitude to the Statistics Department for their unwavering support throughout the completion of this thesis. Their encouragement, resources, and collaborative environment provided a strong foundation for my research. The department's commitment to excellence and their willingness to assist me in navigating complex statistical analyses have been instrumental in shaping the outcome of this work. I am truly appreciative of the department's dedication to fostering a conducive academic environment, which has significantly contributed to the success of this thesis.

VITA

Cangao Chu

EDUCATION

Doctor of Philosophy **2024**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2018**
University of California, Riverside *Riverside, California*

Bachelor of Art in Statistics & Applied Mathematics **2015**
University of California, Berkeley *Berkeley, California*

RESEARCH EXPERIENCE

Graduate Research Assistant **2007–2012**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2019–2020**
University of California, Irvine *Irvine, California*

Teaching Assistant **2018–2018**
University of California, Riverside *Riverside, California*

ABSTRACT OF THE DISSERTATION

Joint Modeling of Longitudinal and Survival Data: Censoring Robust Estimation, Influence Function Based Robust Variance and Shape Based Longitudinal Clustering

By

Cangao Chu

Doctor of Philosophy in Statistics

University of California, Irvine, 2024

Chancellor's Professor Daniel Gillen, Chair

In quantifying heterogeneity in time-to-event data, potential biomarker information and demographic status reflecting key pathophysiology at the individual level are increasingly available. Typical analysis of time-to-event data relies on observed event data. To ensure adequate statistical power, longer follow-up periods or additional participants are required, adding enormously to study costs. Incorporating biomarker information can reduce the cost of data collection and maximize statistical power.

There are, however, several drawbacks to the direct incorporation of biomarkers for identification and analysis. Specifically, most models assume time-invariant effects, independent censoring, and complete information. Little work has been done to assess the robustness of the currently used models beyond these assumptions.

The Cox proportional hazards model (Cox, 1972) is the most commonly used model for time-to-event data. This model compares observed covariates to the weighted average of covariates in the risk set. The Cox model consistently estimates a weighted time-averaged effect under proportional hazards. However, in the field of Alzheimer's disease, the proportional hazard model often fails as the covariate effect diminishes due to disease progression. Directly applying the model ignores the potential changes in the underlying effect. In addi-

tion, the assumption of independent censoring may fail to acknowledge possible associations between biomarkers and the missing data mechanism. As a result, the observed event in later follow-up times may be underrepresented relative to the overall population due to censoring. Currently, there is no existing research to address the above issues simultaneously while being robust to possible violations of assumptions and adaptive to various types of biomarker information.

Given the increasing availability of repeated biomarker measurements, there is a growing interest in jointly modeling longitudinal and time-to-event data to maximize statistical information on the association between potential biomarkers and disease progression. Within longitudinal data, it is often observed that subgroups exist among populations, where the underlying development can be clustered into different patterns. These clustering patterns offer valuable insights into the natural history of diseases, including their progression, risk factors, and potential causes. In practice, longitudinal data frequently encounter missing values. While typical clustering methods handles intermittent missing values through imputation methods, in disease studies, monotone missingness often occurs in longitudinal data when measurements are lost after a specific time due to terminal events and censoring. Existing imputation methods may introduce bias when attempting to recover trajectories of similar lengths. However, there has been little examination of the currently used methods on longitudinal clustering with monotone missingness.

This dissertation aims to develop a flexible and robust statistical model to evaluate predictors of time-to-event data. The resulting estimator will be consistent and robust under violated assumptions and repeated measures. In Chapter 3, we proposed a reweighted censoring robust estimator using censoring weights and conditional covariate variance. In Chapter 4, we introduced an robust variance estimator based on the influence function for the estimator proposed in Chapter 3. In Chapter 5, we investigated the performance of the current longitudinal clustering method under the monotone missing censoring mechanism, and proposed a shape-based longitudinal partial mapping clustering method to complement the estimator proposed in Chapter 3.

Chapter 1

Introduction

1.1 Biomarker identification in Alzheimer's disease

Alzheimer's disease (AD) and related dementias (ADRD) are the most common types of dementia. Dementia is not a specific disease, but an overall term that covers medical conditions related to impaired ability to carry out daily activities. Currently, dementia affects more than 55 million people worldwide, with nearly 10 million new cases every year. Alzheimer's disease is the most common form of dementia, accounting for 60% to 70% of cases, according to the World Health Organization. The symptoms of Alzheimer's disease include occasional memory lapses, increased difficulty concentrating, thinking, making judgments, and performing routine activities. In addition, it can lead to changes in personality and behavior. There is currently no cure for Alzheimer's disease, which places an enormous burden on individuals, families and nations. Although AD is more commonly observed in older age groups, it is not a natural result of aging. The disease is thought to be related to the abnormal build-up of proteins in and around brain cells, but the exact cause of this process is unknown. Researchers have conducted extensive work to identify risk factors for the onset of Alzheimer's disease and to provide individual risk screening.

To accurately diagnose Alzheimer's disease, accumulating evidence has shown that progressive cerebral deposition of the 40- and 42-residue amyloid β-proteins in cerebrospinal fluid (CSF) serves as a core biomarker to AD detection (Selkoe, 1991). Another potential diagnostic biomarker is magnetic resonance imaging (MRI) (Blennow and Zetterberg, 2018), which provides spatial resolution to assess pathological development in the brain. However, both methods present challenges for direct interpretation due to the following reasons: (i) MRI tissue volume changes and increased CSF concentration is not specific to AD and require highly skilled labor to measure; (ii) acquiring CSF levels is invasive (De Leon et al., 2004). Therefore, it is crucial to develop statistical methods to model the association between AD and those risk factors in order to identify lower cost and lower burden biomarkers.

When quantifying heterogeneity in time-to-event data, it is increasingly common to have access to potential biomarker information and demographic data that reflect key pathophysiology at the individual level. Numerous studies have consistently shown that biomarkers can aid in the diagnostic process, early-stage screening, and risk analysis. Biomarkers provide diagnostic information that can be obtained in a more efficient and repeatable manner. Regular time-to-event data analysis relies mostly on observed event data. To increase statistical precision, elongated follow-up periods or additional participants are required, which tremendously escalates study costs. However, incorporating biomarker information can reduce the cost of data collection and maximize statistical information.

Due to the significance of biomarkers in disease research, there is increasing interest in jointly modeling longitudinal and time-to-event data to maximize statistical information regarding the association between potential biomarkers and disease progression. A typical strategy in joint modeling involves the two-stage model approach, wherein the longitudinal measurements are statistically modeled in the initial stage, followed by the analysis of survival data based on the first-stage model. However, there are several drawbacks to this approach when incorporating biomarkers for identification and analysis. Specifically, these settings are based on the assumptions of time-invariant effects, independent censoring, and complete information. However, little work has been done to assess the robustness of the current model outside of those assumptions. For the analysis of survival data, The Cox proportional hazards model (Cox, 1972) is the most commonly used model. The Cox proportional hazards model assumes proportional hazards, where the covariate effect remains constant over time. However, in practical applications, this strict assumption is often violated. In AD studies, for instance, the covariate effect diminishes over later follow-up times due to increased heterogeneity among patients, resulting in non-proportional hazards (NPH). Struthers and Kalbfleisch (1986) demonstrated that the Cox estimator yields a weighted time-average covariate effect dependent on an unknown censoring distribution under NPH. Consequently, the Cox estimator's reproducibility across studies under NPH is compromised, rendering the results less meaningful. Current methods aimed at mitigating dependence on censoring focus on reweighting the Cox estimator by the inverse of censoring probability, but these are restricted to time-independent covariates(Xu and O'Quigley, 2000; Boyd et al., 2012; Nguyen and Gillen, 2017). Currently, no existing research jointly models longitudinal and time-to-event data under NPH while simultaneously removing dependence on censoring distribution.

In the following section, we briefly introduce a study that exemplifies the application of joint modeling of longitudinal and survival data.

1.2 Alzheimer's disease neuroimaging initiative

The Alzheimer's Disease Neuroimaging Initiative (ADNI) brings together researchers and study data to establish the trajectory of AD. ADNI researchers gather, verify, and employ various types of data, such as Magnetic Resonance Imaging (MRI) and Positron emission tomography (PET) images, genetics, cognitive tests, cerebrospinal fluid (CSF) , and blood biomarkers, to predict the disease. This website provides access to study resources and data from the North American ADNI study, which includes information on Alzheimer's disease patients, individuals with mild cognitive impairment, and elderly controls.

This study comprises four phases. ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 are four phases of the study. The participant pool for each phase ranging from 700 to 2000 as the studies progress. Although each study has distinct goals, the overlapping objective among all four phases is to identify longitudinal risk factors and predict cognitive decline leading to the terminal event, typically the conversion to AD. It is noted that ADNI is a global research study that incorporates the tracking of AD progression over 63 sites in the US and Canada over several years. Such variation in location and time could lead to differences in unknown censoring mechanisms, even under the same set of standardized protocols. Without adjustment for possible censoring similarity, the resulting association with AD progression could be different from site to site, even if the underlying effects of covariates on the specific question remain the same. Given the global impact of ADNI, it is critical to develop censoring robust methods for identifying longitudinal risk factors and, in turn, predicting the risk of progression to AD.

1.3 Overview of this dissertation

This chapter presents a motivating example that illustrates the necessity and application of methodological developments presented in the rest of the dissertation. The ensuing chapter furnishes a concise overview of the statistical foundation, either as a foundational understanding within the discipline or as a preparatory groundwork for introducing the proposed methodologies. The topics include survival analysis, joint modeling of longitudinal and survival data, censoring robust estimation methods, and longitudinal clustering methods. Chapter 3 centers on the refinement of an unbiased censoring-robust estimator tailored to accommodate longitudinal covariates. Our investigation reveals that applying existing censoring-robust estimators to longitudinal covariates within a two-stage model yields estimates as weighted averages over time and conditional covariate variance, thereby making the results challenging to interpret. To render the results interpretable as a weighted time-average effect, we propose additional reweighting based on consistent estimates of the conditional covariate variances.Through numerical simulations, we assessed the performance of the proposed method compared to the naive Cox model and the existing reweighting algorithm. Subsequently, we apply the proposed method to the ADNI data, compared to alternative methodologies. In Chapter 4, we explore an alternative variance estimator for the estimator introduced in Chapter 3 for robust inference. The conventional robust variance estimator poses challenges in direct application and derivation. Consequently, we suggest employing a semi-parametric approach based on the influence function-based robust variance estimator. We derive a series of robust variance estimators applicable to the entire class of censoring-robust estimators. Simulation results confirm the validity of the proposed variance estimator and highlight its advantages compared to alternative variance estimators. In practical application, we apply the proposed variance estimator to ADNI data, drawing comparisons with the bootstrap method. In Chapter 5, inspired by our motivating example, we assess the effectiveness of various longitudinal clustering algorithms in addressing the unique challenges posed by monotone missingness. Furthermore, we introduce a nonparametric clustering algorithm that leverages partial mapping via the dynamic time warping (DTW) distance. The performance of the proposed algorithm is systematically compared with other clustering algorithms designed for imbalanced monotone missing scenarios under various settings. Additionally, we apply the proposed algorithm to ADNI data, demonstrating its utility in real-world applications. In the concluding section, we engage in a discussion, offering insights and reflections on the findings presented. Furthermore, we outline potential avenues for future research, identifying areas that warrant further exploration and investigation.

Chapter 2

Statistical background

2.1 Introduction

Due to the widespread occurrence of Alzheimer's disease, researchers have dedicated decades of effort to uncovering its causes and developing effective treatments. However, the current approach to treating Alzheimer's disease is predominantly centered on managing symptoms, as the root cause of the condition remains elusive. One potential causal hypothesis for Alzheimer's disease is the amyloid hypothesis (Selkoe, 1991), which centers around the abnormal accumulation of senile plaques. These plaques are characterized by the presence of amyloid fibrils, primarily composed of the amyloid β (Aβ) peptide (Glenner and Wong, 1984). In particular, Aβ 40 and Aβ 42, with the former containing 40 amino acid residues and the latter 42 amino acid residues, represent significant constituents of the accumulated Aβ. However, Kametani and Hasegawa (2018) point out that the onset of AD appears to be intricately connected to impairments in the metabolism of Amyloid Precursor Protein (APP) and the accumulation of APP C-terminal fragments, rather than the production of Aβ as efforts to develop drugs targeting Aβ for the treatment of AD have proven unsuccessful.

Moreover, the presence of the Aβ 40 and Aβ 42 proteins in cerebrospinal fluid (CSF) is often assessed through a procedure known as a lumbar puncture or "spinal tap." This involves inserting a needle into the space between two vertebrae in the lower spine to collect CSF for analysis. Unfortunately, this procedure is invasive, leading to a reluctance among patients to undergo lumbar punctures. Consequently, the limited availability of data poses challenges in obtaining comprehensive insights, as patients may refuse to undergo this procedure due to the associated discomfort. Total tau, a protein primarily found in the central nervous system, is a biomarker that reflects the extent of neuronal damage and degeneration (Hampel and Teipel, 2004). Despite the fact that the Aβ biomarker has not been conclusively established as the sole cause of AD, accumulating data from clinical research consistently supports the significance of biomarkers such as Aβ and total tau. These biomarkers are considered reflective of key elements in the pathophysiology of AD. Additionally, there is an ongoing quest for additional biomarkers and the enhancement of screening methods (Blennow and Zetterberg, 2018). This emphasis on biomarkers in AD exploration underscores the importance of continuous efforts to deepen our understanding and refine diagnostic approaches for this complex neurodegenerative condition.

Considering the challenges encountered by researchers in pinpointing the precise cause of AD, the development of robust statistical methodology becomes crucial. Such a methodology aims to identify and analyze potential biomarkers associated with AD without necessarily unraveling the intricate underlying mechanisms. This approach recognizes the complexity of AD and emphasizes the need for effective statistical tools to uncover and comprehend crucial biomarkers, even when the exact causal mechanisms remain elusive. The Cox model (Cox, 1972) stands as a widely adopted statistical method for assessing heterogeneity in time-to-event data while accounting for multiple covariates. Facilitating a regression type framework, it compares the risk of observed events to the average risk within the risk set. Its popularity is further enhanced by its semiparametric nature, avoiding the need to specify the baseline hazard explicitly. However, a notable limitation lies in its reliance on the proportional hazards assumption, which may not hold in real-world scenarios.

Under non-proportional hazards conditions (NPH), the Cox model's estimator is influenced by the censoring mechanism (Struthers and Kalbfleisch, 1986). To mitigate this dependency on censoring mechanisms, reweighted algorithms have been developed, addressing both categorical and continuous covariates (Boyd et al., 2012; Nguyen and Gillen, 2017). In the context of AD, where NPH and censoring due to dropouts may be pertinent, obtaining a censoring-robust estimator incorporating longitudinal covarites for biomarkers becomes imperative.

To build a censoring robust estimator incorporating longitudinal covariates, a joint modeling approach for longitudinal and survival data is proposed. This entails a joint modeling of longitudinal and survival data to yield an understanding of the biomarker's association with AD progression. By applying a censoring-robust estimator within the joint model, this dissertation aims to provide more accurate and reliable insights into the dynamics of biomarkers in the presence of potential non-proportional hazards and dropout effects in AD research.

In this chapter, we present a review of joint modeling techniques for longitudinal and survival data. The primary objective of these models is to assess longitudinal covariates while accommodating potential violations of the proportional hazards assumption and maintaining robustness against unknown censoring mechanisms. This review aims to contribute insights into advanced statistical approaches that can enhance the robustness and reliability of longitudinal and survival data analysis, especially in settings such as Alzheimer's disease research where non-proportional hazards and intricate censoring mechanisms may significantly influence the outcomes. We will dissect the fundamental elements of joint modeling, beginning with the rationale behind integrating longitudinal and survival data and the challenges associated with potential violations of the proportional hazards assumption. This introduction lays the foundation for a detailed comparison and discussion of major types of joint modeling approaches, elucidating their strengths, weaknesses, and applications.

The subsequent section delves into the realm of robust inference, a critical aspect in the context of joint modeling. A review of the current methods for robust variance estimation will be undertaken, focusing on their application to the Cox model and general linear regression-type estimators. Within this framework, we particularly explore the influence function-based approach, which is pivotal for our proposed estimator. This methodological choice aligns with the need for robustness in the presence of potential challenges such as non-proportional hazards and uncertain censoring mechanisms, especially relevant in the study of AD.

In the concluding part of this chapter, we shift our focus to longitudinal clustering methods. A review of these techniques will be provided, offering a foundational understanding to facilitate subsequent developments. This exploration is integral to our overarching goal of establishing a robust and comprehensive framework for the joint modeling of longitudinal and survival data in the context of AD research.

2.2 Review of survival analysis

2.2.1 Censoring

Survival analysis is a statistical method used to analyze the time until an event of interest occurs. However, the survival data include censoring, which occurs when the exact event time is unknown for some individuals. Censoring is an essential aspect of survival analysis and has been the subject of several research studies. For instance, Howe et al. (2016) focused on selection bias due to loss to follow up in cohort studies. They highlighted the potential impact of censoring on the validity of study results and emphasized the need to account for loss to follow up in the analysis of survival data. In real-world studies, censoring is a ubiquitous phenomenon. Censoring occurs when the complete information regarding the time until an event of interest is not available for all subjects in a study.In the context of diseases like Alzheimer's, where long-term observation is often necessary, and patient dropout or loss to follow-up can occur, dealing with censoring becomes particularly pertinent.

There are multiple types of censoring, including left censoring, interval censoring and right censoring. Left censoring occurs when the event of interest must have occurred for a subject before they are observed in the study. In other words, the subject enters the study already having experienced the event, but the exact time of the event is unknown because the observation starts after the event has occurred. In the context of left censoring, researchers are aware that the event of interest has taken place at some point, but due to the subject's entry into the study after the occurrence, the precise timing remains observed.

Another type of censoring is interval censoring, and it often occurs when a subject is followed for a certain period, experiences an interruption, such as being lost to follow-up, and then reengages with the study. This results in observed data that are represented as intervals, rather than precise event times. In the context of interval censoring, an uncensored observation of an event translates to an observed interval that essentially consists of a single point. This interval encompasses the time between the last known event-free point and the observed event time, providing a time range within which the event is known to have occurred.

Perhaps the most common form of censoring encountered is right censoring. For right censoring, a subject is observed up to a certain time point, and if the event of interest has not occurred by that time, the data for that subject are considered censored. In cases where the exact event time is known, it is observed only if the censoring time is greater than or equal to the exact event time. Right censoring is prevalent in long-term studies or studies with fixed observation periods where not all subjects experience the event of interest within the study duration. The data for these subjects are censored at the last observed time point, indicating that the event has not occurred up to that time. The major reasons for right censoring are diverse but often involve the termination of the study before the occurrence of the event of interest or a subject being unable to continue participation in the study.

The characterization of censoring can be further classified based on various attributes including random, double, independent, and non-informative censoring. Random censoring occurs when the likelihood of a subject being censored is unrelated to the subject's actual survival time or the occurrence of the event. It is considered a random process and is not influenced by any specific characteristics of the subjects. Double censoring refers to situations where both the beginning and end of the observation period are not precisely known. This can happen, for instance, when data collection starts after the event of interest has already occurred or when it continues after the study concludes. In the context described here, independent censoring differs from the common understanding of censoring and survival independence. Instead, it denotes that the censoring times for different subjects are not influenced by the occurrence of events in other subjects. Each subject's censoring time is independent of the others, allowing for separate analysis of their time-to-event data. Non-informative censoring suggests that the censoring mechanism does not carry information about the subjects' future event times. In other words, the censoring is not based on any knowledge or anticipation of when the event might occur. Indeed, assuming non-informative censoring is a common practice in survival analysis. Non-informative censoring implies that the decision to censor a subject's data is unrelated to the subject's future event times. By assuming non-informative censoring, researchers can treat censoring as a random process that does not carry information about when the event of interest might occur. This simplifies the analysis and helps avoid potential biases introduced by informative censoring, where the decision to censor is related to the subject's likelihood of experiencing the event. For further information, a detailed review of censoring in survival analysis can be found in Turkson et al. (2021).

2.2.2 Statistical functions of interest

We start with the scenario where there is no censoring, and the time to event of interest is denoted as T. In this context, there are several important functions related to survival data:

- Cumulative Distribution Function (CDF) $F(t)$:

  The cumulative distribution function, denoted as $F(t)$, represents the probability that the event occurs on or before time t. Mathematically, $F(t) = P(T \leq t) = \int_0^t f(t)dt$, providing the cumulative probability distribution of event times up to t.

- Probability Density Function (PDF) $f(t)$:

  The probability density function, denoted as $f(t)$, describes the instantaneous rate at which the event occurs at time t. Mathematically, it is the derivative of the cumulative distribution function, $f(t) = dF(t)/dt$. The PDF provides the probability of the event occurring in an infinitely small time interval around t.

- Survival Function $S(t)$: The survival function, denoted as $S(t)$, represents the probability that a subject survives beyond time t. Mathematically, it is defined as $S(t) = P(T > t) = 1 - F(t)$, which is the probability that the event has not occurred by time t. The survival function is fundamental in survival analysis and is used to estimate the time until an event occurs.

- Hazard Function $\lambda(t)$: The hazard function represents the instantaneous rate at which events occur at time t, given that the subject has survived up to that point. Mathematically, it is defined as $\lambda(t) = \lim_{\triangle t \to 0+}(1/\triangle t)\Pr[t \leq T < t + \triangle t | T \geq t] = f(t)/S(t)$. The survival function can be obtained with the transformation that $S(t) = \exp\{-\int \lambda(t)dt\}$.

- Cumulative Hazard Function $\Lambda(t)$: The cumulative hazard function is the integral of the hazard function up to time t. It represents the cumulative risk of experiencing event up to time t. Mathematically, it is defined as $\Lambda(t) = \int_0^t \lambda(t)dt$.

The above fundamental functions used to describe time-to-event data are interrelated. Once one of these functions is properly defined, others can be derived through transformations. In particular, the hazard function is often preferred over the survival function because it provides more flexibility in modeling, allows for capturing time-varying effects, facilitates comparing groups, and handles censored data more straightforwardly.

2.2.3 Parametric modeling

In early attempts to analyze survival data, researchers often employed parametric methods that assumed specific functional forms for the survival distribution. These parametric models were used as an initial approach to describe the underlying distribution of event times and estimate key parameters. While nonparametric and semiparametric methods have become more prevalent in recent times due to their flexibility and fewer assumptions, parametric models were instrumental in laying the groundwork for survival analysis. Some common parametric models are listed in Table 2.1.

Table 2.1: Parametric survival distributions and properties

Distribution	Hazard	Density	Cumulative Hazard	Survival Function	Mean	Chararteristic
Exponential	λ	$\lambda e^{-\lambda t}$	λt	$e^{-\lambda t}$	$1/\lambda$	constant hazard
Weibull	$\alpha\lambda t^{\alpha-1}$	$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$	λt^{α}	$e^{-\lambda t^{\alpha}}$	$\Gamma[1+1/\alpha]/\lambda^{1/\alpha}$	hazard as power of t
Gamma	$\frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma)-\Gamma_x(\gamma)}$	$\frac{(\frac{x-\mu}{\beta})^{\gamma-1}\exp(-\frac{x-\mu}{\beta})}{\beta\Gamma(\gamma)}$	$-\log(1-\frac{\Gamma_x(\gamma)}{\Gamma(\gamma)})$	$1-\frac{\Gamma_x(\gamma)}{\Gamma(\gamma)}$	γ	monotonic hazard
Log-normal	$\frac{(\frac{1}{x\sigma})\phi(\frac{\ln x}{\sigma})}{\Phi(\frac{-\ln x}{\sigma})}$	$\frac{e^{-((\ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}}$	$-\ln(1-\Phi(\frac{\ln(x)}{\sigma}))$	$1-\Phi(\frac{\ln(x)}{\sigma})$	$e^{0.5\sigma^2}$	unimodal, right skewed hazard and density
Log-logistic	$\frac{\lambda\kappa(\lambda x)^{\kappa-1}}{1+(\lambda x)^{\kappa}}$	$\frac{\lambda\kappa(\lambda x)^{\kappa-1}}{(1+(\lambda x)^{\kappa})^2}$	$\ln[1+(\lambda x)^{\kappa}]$	$\frac{1}{1+(\lambda x)^{\kappa}}$	$\int_0^{\infty}\frac{\lambda^{\kappa}x^{\kappa-1}}{(1+(\lambda x)^{\kappa})^2dx}$	similar to log-normal but heavier tail

Under the assumption of a parametric model, maximum likelihood theory is a powerful approach for estimating the survival distribution and functionals of the distribution. The likelihood function is constructed by considering the observed data, which includes information about both the observed event times and the censoring times. Suppose a sample of n subjects with underlying variable $T_i \sim F_T(\cdot)$ and $C_i \sim G_C(\cdot)$, where T_i represents the event times and C_i represents the censoring times, $i = 1, \ldots, n$. he observed time X_i can be defined as the minimum of the event time T_i and the censoring time C_i, mathematically expressed as $X_i = \min(T_i, C_i)$. Then, The event indicator δ_i is commonly defined as:

$$\delta_i = \begin{cases} 1 \quad \text{if} \quad T_i \leq C_i (\text{event occured}), \\ 0 \quad \text{if} \quad T_i \geq C_i (\text{censored}). \end{cases}$$

This event indicator is crucial in constructing the likelihood function for the analysis of survival data, as it helps distinguish between observed events and censored observations. The likelihood function is then formulated to account for the observed times and the event indicators, allowing for the estimation of parameters in the context of survival analysis. With censoring present, the full likelihood contribution for the ith subject assumming independent censoring is defined as:

$$L_i(x_i, \delta_i) = \{f_T(x_i)[1 - G_c(x_i)]\}^{\delta_i} \{g_c(x_i)[1 - F_T(x_i)]\}^{1-\delta_i},$$

where each subject's contribution to the likelihood is separated into two cases based on the observed event time. When the event time is observed, the likelihood contribution is based on the probability of the event at the time x_i, given that the censoring time is beyond the observed event time. When the event time is censored, the likelihood contribution is based on the probability of censoring at time x_i, given that the event time is beyond the observed censoring time. As mentioned before, for survival data analysis, it suffices to know one function of the distribution list in section 2.2.2, where the distribution parameter assumed can be estimated through the maximum likelihood procedure.

Since we focus on the survival data, the likelihood contribution that only contains censoring is omitted, and the likelihood function is reduced to

$$L(x, \delta) = \prod_{i=1}^{n} f_T(x_i)^{\delta_i} S_T(x_i)^{1-\delta_i},$$

with the score equation as

$$U(\beta) = \sum_{i=1}^{n} \left(\delta_i \frac{\lambda'(x_i)}{\lambda(x_i)} - \int_0^{x_i} \lambda'(u) du \right).$$

The maximum likelihood estimator is obtained by setting it equal to zero and solving the score equation. Define

$$\mathrm{I}(\beta) = -E[\frac{\partial}{\partial \beta} U(\beta)],$$

and it is shown that under standard regularity conditions (Van der Vaart, 2000),

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, I(\beta)^{-1}).$$

It is noteworthy that, due to the logarithmic property, the censoring distribution does not directly factor into the likelihood function in survival analysis. However, the influence of censoring on the Fisher information calculation is mediated through the event indicator. To address this, researchers commonly opt for using the observed Fisher information, wherein the expectation is replaced with the observed event indicator.

The above material briefly introduces the procedures and inference of parametric survival analysis, it would be efficient due to parametric assumptions; however, when the underlying distribution is misspecified, the result would be biased and not useful. Recognizing this limitation, there is a distinct interest in nonparametric estimation of the survival function $S(t)$ when possible.

2.2.4 Non-parametric estimation of the survival distribution

Nonparametric methods, such as the Kaplan-Meier estimator (Kaplan and Meier, 1958) provide a flexible approach that does not rely on specific distributional assumptions, making them robust in scenarios where the true underlying distribution is unknown or complex. This approach enhances the applicability and reliability of survival analysis, particularly in situations where parametric assumptions may not hold. In the following, We derive the Kaplan-Meier estimator intuitively.

In a complete case, consider that we have a life table where time is divided into K intervals $(0, \tau_1], (\tau_1, \tau_2], \ldots, (\tau_{K-1}, \tau_K]$ by observed event time τ_j. Then, the conditional probability for survival to τ_j upon survival to τ_{j-1} can be estimated as follows:

$$\hat{P}[T > \tau_i | T > \tau_{i-1}] = 1 - \frac{d_i}{n_i} = \frac{s_i}{n_i},$$

where d_i denotes the total number of events during the interval (τ_{i-i}, τ_i), n_i denotes the total number of subjects at risk, meaning those subjects that haven't experienced an event at the beginning of the interval, and s_i represents the number of subjects that did not experience an event during the interval. Given the definition of conditional probability, the probability of surviving past a particular interval can be estimated as:

$$\hat{P}[T > \tau_i] = \hat{P}[T > \tau_i | T > \tau_{i-1}] \times \ldots \times \hat{P}[T > \tau_2 | T > \tau_1] \times \hat{P}[T > \tau_1],$$

where $\hat{P}[T > \tau_1]$ is estimated similarly without considering the conditional probability. By multiplying conditional probabilities, we can introduce the Kaplan-Meier estimator. This estimator contemplates the probability of surviving a very small time interval, given that a subject is at risk at the beginning of the interval. The Kaplan-Meier estimator is defined as the limit of the life table estimator as the intervals shrink to zero. For small $\triangle t$, $\Lambda(t + \triangle t) \approx$

$P[t \leq T < t + \triangle t | T \geq t]$, then Nelson (1969) proposed the Nelson cumulative hazard estimator as

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} D_i / \bar{Y}_i,$$

where $\bar{Y}_i$ and D_i represent the number of subjects at risk and the event occurrence count at event times t_i respectively. Given the relationship between cumulative hazard function and survival probability, it can be shown that when $D_i/\bar{Y}_i \approx 0$,

$$\begin{aligned}\hat{S}(t) &= \prod_{i:t_i \leq t} \exp(-D_i/\bar{Y}_i) \\ &\approx \prod_{i:t_i \leq t} (1 - D_i/\bar{Y}_i),\end{aligned}$$

where the last line is the Kaplan and Meier (1958) product limit estimator. For a more formal representation of these estimators, they are often examined from the perspective of counting processes and martingales, providing insights into their properties. Let the counting processes $N_i(t)$ and $Y_i(t)$ be defined by

$$N_i(t) = I\{X_i \leq t, \delta_i = 1\}$$

$$Y_i(t) = I\{X_i \geq t\}.$$

Denote $\bar{N}(t) = \sum_i N_i(t)$ and $\bar{Y}(t) = \sum_i Y_i(t)$, then the cumulative hazard estimator can be written as

$$\hat{\Lambda}(t) = \int_0^t \frac{I\{\bar{Y}(u) > 0\}}{\bar{Y}(u)} d\bar{N}(u),$$

with the predicted process

$$\Lambda^*(t) = \int_0^t I\{\bar{Y}(u) > 0\} \lambda(u) du.$$

It can be shown that

$$\hat{\Lambda}(t) - \Lambda^*(t) = \sum_i \int_0^t \frac{I\{\bar{Y}(u) > 0\}}{\bar{Y}(u)} dM_i(u),$$

with

$$M_i(u) = N_i(u) - \int_0^u Y_i(s) d\Lambda(s),$$

where $M_i(u)$ represents the Martingale and plays central role in the formal development of asymptotic properties. Within the Martingale framework, Rebolledo's theorem demonstrates the normal distribution of the asymptotic distribution of the Kaplan-Meier estimator (Fleming and Harrington, 2013).

The Kaplan-Meier estimator can also be obtained via the maximum likelihood framework, where the likelihood takes the following form:

$$L = \prod_{i=1}^{N} \{[S(t_i^-) - S(t_i)]^{\delta_i} \prod_{j \in R(i)} S(t_{t_j})\},$$

where $R(i)$ represents all subjects who are at risk of experiencing an event at time t_i. By the relationship between the survival probability and the hazard function, the above likelihood can be further expressed as

$$L(\lambda) = \prod_{i=1}^{D} \lambda_i^{d_i} (1 - \lambda_i)^{n_i - d_i}.$$

This gives rise the score equation for each subject who experienced event as

$$U(\lambda) = \frac{d_i}{\lambda_i} - \frac{n_i - d_i}{1 - \lambda_i},$$

setting the score equation to zero and solving the equation gives

$$\hat{\lambda}_i = \frac{d_i}{n_i}.$$

Therefore, the MLE for survival probability is obtained as

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

which is exactly the Kaplan-Meier estimator.

For inference, the Greenwood formula (Greenwood et al., 1926) is a widely used method for estimating the variance of a survival function, particularly in the context of Kaplan-Meier survival analysis. This formula provides an estimate of the variance for the Kaplan-Meier estimator at each observed event time. To derive, the Kaplan-Meier estimator at t_i is defined as:

$$\hat{S}(t_i) = \prod_{j:t_j \leq t_i} \left(1 - \frac{D_j}{N_j}\right).$$

Now, the increment in the Kaplan-Meier estimator at t_i is given by the product:

$$\Delta\hat{S}(t_i) = \hat{S}(t_i) - \hat{S}(t_{i-1}) = \left(1 - \frac{D_i}{N_i}\right) \prod_{j=1}^{i-1} \left(1 - \frac{D_j}{N_j}\right).$$

Then, the variance at t_i can be expressed as the sum of the variances of these increments:

$$V_i = \text{Var}(\Delta\hat{S}(t_i)) = \sum_{j=1}^{i} \text{Var}\left(\left(1 - \frac{D_j}{N_j}\right)\right) \prod_{k=1}^{j-1} \left(1 - \frac{D_k}{N_k}\right)^2.$$

Assuming independence of increments, the variance of each increment is given by

$$\text{Var}\left(\left(1 - \frac{D_j}{N_j}\right)\right) = \frac{D_j}{N_j(N_j - D_j)}.$$

Substituting this into the expression for V_i , we get

$$V_i = \sum_{j=1}^{i} \frac{D_j}{N_j(N_j - D_j)} \prod_{k=1}^{j-1} \left(1 - \frac{D_k}{N_k}\right)^2.$$

Now, for all distinct event times up to t_i , we can get the standard error as

$$\hat{SE}_G\{\hat{S}_{KM}(t_i)\} = \hat{S}^2_{KM}(t_i)\sqrt{\sum_{j=1}^{i} \frac{d_j}{d_j(n_j - d_j)}}.$$

Note that the above formula can yield values outside the range $[0, 1]$ due to the support of normal distribution. To fix the problem, a better confidence can be obtained using the transformation $log\Lambda(t) = \log[-\log S(t)]$, so that the standard error can be estimated using delta method as

$$\hat{SE}_G\{\log[-\log\hat{S}_{KM}(t_i)]\} = \sqrt{\sum_{j=1}^{i} \frac{d_j}{d_j(n_j - d_j)}}/[-\log\hat{S}_{KM}(t_i)],$$

which results in a confidence interval in the range with both endpoints lying in the interval $[0, 1]$. It is noted that such construction may also speed up convergence (Borgan and Liestøl, 1990), and thus is a standard method for confidence interval construction.

2.2.5 Cox proportional hazards model

The Kaplan-Meier estimator in Section 2.2.4 is powerful in estimating the survival distribution. In observational studies, there is often a specific scientific question to address. In such cases, using a method that can quantify the heterogeneity in survival data concerning a prespecified variable of interest while adjusting for other potential confounding variables is preferable. This approach allows for a more targeted and nuanced analysis, providing insights into the impact of particular variables on survival outcomes while accounting for the influence of other relevant factors. The Cox proportional hazards model, often referred to as the Cox model or Cox regression, is a widely used statistical model for analyzing the relationship between the survival time of individuals and one or more predictor variables.

Sir David R. Cox, a statistician, first introduced it in 1972, and it has since become a crucial tool in survival analysis.

In the study by Cox (1972), the dataset consists of one or more measurements on each subject denoted as $Z_1, \ldots, Z_p$. For the j-th individual, the values of Z are represented as $Z_j = (Z_{1j}, \ldots, Z_{pj})$. Cox proposed assessing the relationship between the distribution of event time T and covariate Z by modeling the hazard function as:

$$\lambda(t|Z) = \lambda_0(t)\text{exp}(Z\beta),$$

where β represents the $p \times 1$ vector of unknown population parameters, and $\lambda_0(t)$ denotes an unknown function representing the baseline hazard when all covariates are equal to 0. It's essential to note that β is a vector of constants that does not depend on t, indicating that the relative risk between two subjects remains constant over time. This assumption is known as the proportional hazards assumption.

Following the established notations, we derive the partial likelihood by defining the data for subject i as tuples (T_i, C_i, δ_i), representing the true survival time, censoring time, and event indicator. The observed time is denoted as $X_i = \min\{T_i, C_i\}$.

Independence between the time to event and time to censoring is often assumed. This is expressed as $P(X > t) = P(T > t, C > t) = P(T > t)P(C > t)$. Alternatively, conditional independence between time to event and time to censoring can be assumed, where the independence is conditional on observed covariates Z. This leads to the expression $P(X > t|Z) = P(T > t, C > t|Z) = P(T > t|Z)P(C > t|Z)$.

Taking conditional independence as an example, the probability contribution for each individual can be defined as follows:

$$\begin{aligned}
P(X_i = t|Z_i, \delta_i) &= P(\min\{C_i, T_i\} = t|Z_i, \delta_i) \\
&= P(T_i = t, C_i > t|Z_i)^{\delta_i} P(T_i > t, C_i = t|Z_i)^{1-\delta_i} \\
&= P(T_i = t|Z_i)^{\delta_i} P(T_i > t|Z_i)^{1-\delta_i} P(C_i > t|Z_i)^{\delta_i} P(C_i = t|Z_i)^{1-\delta_i} \\
&= f_T(X_i|Z_i)^{\delta_i} S_T(X_i|Z_i)^{1-\delta_i} f_C(X_i|Z_i)^{\delta_i} S_C(X_i|Z_i)^{1-\delta_i},
\end{aligned}$$

where $f_*(\cdot)$ and $S_*(\cdot)$ represent the density function and survival function of T and C given the subscript. The third equivalence in the above equations is based on the conditional independence of T and C, and when independence is assumed, such equivalence is also valid.

As the association between time to event and covariates is of primary interest, we will omit the part where T is not involved. This results in the following likelihood from the contribution of independent triplets (X_i, δ_i, Z_i) for $i = 1, \ldots, n$,

$$L(\beta) = \prod_{i=1}^{n} f_T(X_i|Z_i)^{\delta_i} S_T(X_i|Z_i)^{1-\delta_i}.$$

Then the score function defined as $U(\beta) = \partial \text{log} L(\beta)$ will be as follows:

$$\begin{aligned}
U(\beta) &= \partial \left\{ \sum_{i=1}^{n} \delta_i \text{log}(f_T(X_i|Z_i)/S_T(X_i|Z_i)) + \text{log}(S_T(X_i|Z_i)) \right\} / \partial\beta \\
&= \partial \left\{ \sum_{i=1}^{n} \delta_i \text{log}\lambda(X_i) - \Lambda(X_i) \right\} / \partial\beta \\
&= \sum_{i=1}^{n} \left\{ \delta_i \frac{\lambda'(X_i)}{\lambda(X_i)} - \int_0^{X_i} \lambda'(u) du \right\}.
\end{aligned}$$

Evaluations of the equation above requires the full specification of the hazard function, which is often impractical. The Cox proportional hazards model is significant in such cases, as it assumes a baseline hazard for subgroups with all covariate values equal to 0, and focusing on quantifying the association between relative risk and covariates of interest. Since $\lambda_0(t)$ is an unknown function without parametric specification, but we fully specify the relative risk by adjusting for the covariate, the Cox proportional hazard model is categorized as a semi-parametric model.

By plugging the Cox proportional hazards into the score equation above, we get:

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j \in R_i} Z_j \text{exp}(Z_j \beta)}{\sum_{j \in R_i} \text{exp}(Z_j \beta)} \right\},$$

where $R_i(t)$, called the risk set, is the set of indices of individuals known to survive until time t, such that $j \in R_i(t)$ if $X_j \geq t$. Note that the above score equation can also be obtained by interpreting the likelihood as the probability that the i-th individual with covariate Z_i experiences an event, given that some individuals in the risk set $R_i(t)$ experience an event at time t. Taking the derivative with respect to β results in the same mathematical expression. The partial likelihood can be expressed as:

$$\begin{aligned} L(i) &= \text{P}\{\text{subject with } x_{(i)} \text{ fails at } t_{(i)} | \text{some subject failed at } t_{(i)}\} \\ &= \frac{\text{P}\{\text{subject with } x_{(i)} \text{ fails at } t_{(j)}\}}{\text{Pr}\{\text{some subject in } R_{(i)} \text{ failed at } t_{(i)}\}} \\ &= \frac{\lambda_{(i)}(t_{(i)}(\triangle t)}{\sum_{j \in R_{(i)}} \lambda_j(t_{(i)})(\triangle t)}. \end{aligned}$$

Since the $(\triangle t)$s cancel, the above equation is just

$$L_{(i)} = \frac{\text{risk for failed subject at } t_{(i)}}{\sum_{i \in \text{Risk Set}} \text{risk for subject } j \text{ at } t_{(i)}}.$$

Recall that $\lambda_i(t_{(j)}) = \exp(\beta^T x_i)\lambda_0(t_{(j)})$, so

$$L_{(i)} = \frac{\exp(\beta^T x)(i)\lambda_0(t_i))}{\sum_{j \in R_{(i)}} \exp(\beta^T x_j)\lambda_0(t_i)} = \frac{\exp(\beta^T x)(i)}{\sum_{j \in R_{(i)}} \exp(\beta^T x_j)}.$$

Moreover, we have the log-partial likelihood as

$$\log(L_P) = \sum_{\text{failure times } i} \log\left(\frac{\exp(\beta^T x)(i)}{\sum_{j \in R_{(i)}} \exp(\beta^T x_j)}\right).$$

If we take the derivative with respect to β, the resulting derivative would equal the score equation for the Cox proportional hazards. The popularity of the proportional hazard model lies in its simple format, where the baseline hazard has been completely removed from the problem.

It is observed that at each failure time, the partial likelihood compares the covariate values of individuals who experience an event to a weighted average of the covariate values for those still at risk. The Newton-Raphson technique is commonly employed to find the Maximum Partial Likelihood Estimates (MPLE). At each iteration, the first derivative (score vector) and the second derivative (information matrix) of the log-likelihood function with respect to the parameters are calculated. Then parameter estimates are updated using the formula:

$$\beta^{(k+1)} = \beta^{(k)} - [\frac{\partial}{\partial \beta} U(\beta)|_{\beta_k}]^{-1} \cdot U(\beta^{(k)}),$$

where $\frac{\partial}{\partial \beta} U(\beta)|_{\beta_k}$ is the Hessian matrix evalued at β_k, $U(\beta^{(k)})$ is the score vector, and k denotes the iteration. For inference, Rebolledo's Theorem showed that $\hat{\beta} \sim N(\beta_0, I^{-1}(\beta_0))$, where β_0 is the true value of β provided that the model specification is correct.

2.2.6 Censoring robust estimators

In survival analysis, dealing with censorship is a common challenge. The presence of censoring can impact the accuracy of parameter estimates and the validity of statistical inferences. Therefore, it becomes essential to use censoring-robust estimators to address these challenges.

First, using counting process notation, Cox's partial likelihood estimator can be written as the solution to

$$U(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left[Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j \exp(Z_j^T \beta)}{\sum_{j=1}^{n} Y_j(t) \exp(Z_j^T \beta)} \right] \mathrm{d}N_i(t) = 0, \tag{2.1}$$

where $Y_j(t) = I(C_i \geq t, T_i \geq t)$ and $N_i(t) = I(X_i \leq t, \delta_i = 1)$ denotes a counting process that counting the number of events in the interval $(0, t)$ for ith subject.

NPH effects are commonly observed in clinical research. For example, Kasten et al. (2015) conducted an analysis of Amyloid-β kinetics in 112 participants, examining the relationship between participant age and the amount of amyloid deposition. The study revealed a 2.5-fold longer half-life, indicating a significant correlation between increasing age and slowed amyloid-beta turnover rates. The findings suggested that due to the slowing of amyloid-beta turnover, the aging group faced a higher risk of progression to AD compared to the younger group, despite having a similar amyloid-beta deposition level. In our AD research context, when a biomarker that changes over time is solely evaluated at baseline, the association between the biomarker and the outcome may manifest a more notable influence around the measurement time due to within-subject changes in the biomarker over time. Although this leads to a non-proportional hazards biomarker effect, accurately predicting how the hazard ratio is expected to evolve over time poses a challenge. Consequently, there is a heightened interest in understanding what the Cox estimator estimates under NPH. For NPH, the hazard function is defined as

$$\lambda(t, Z) = \lambda_0(t) \exp(Z\beta(t)).$$

In an initial exploration, Xu and O'Quigley (2000) noted that the result of the Cox estimator under NPH conditions without censoring can be interpreted as an approximate average covariate effect over the observed survival times, $\int \beta(t)dF(t)$. It is important to highlight that in the absence of censorship, the average is determined by the time-varying function of β and the density of T. However, in the presence of censoring, the estimation becomes more intricate. In the presence of censorship, under the standard conditional independence between survival and censoring assumption, Struthers and Kalbfleisch (1986) demonstrated that the partial likelihood estimator under NPH estimates a quantity that is consistent with the solution of the following equation:

$$\int_0^\infty E\left\{f_T(t|Z)S_C(t|Z) \times \left[Z - \frac{E\{ZS_T(t|Z)S_C(t|Z)\exp(\beta Z)\}}{E\{S_T(t|Z)S_C(t|Z)\exp(\beta Z)\}}\right]\right\} \mathrm{d}t = 0, \tag{2.2}$$

where $f_T(t)$ represents the density function, $F_T(t)$ is the cumulative distribution function, and $S_T(t) = 1 - F_T(t)$ is the survival distribution of the true event time T. Additionally, the corresponding distribution functions for censoring time are denoted by changing the lower index to C. The above estimating equation suggests that the MPLE, denoted as $\hat{\beta}_{\text{cox}}$, is consistent for a quantity dependent on the distribution of covariates Z, the distribution of the true event time T, and the censoring distribution C.

The appearance of the censoring distribution in the estimating equation implies that when the true relative difference in hazards associated with a covariate of interest is heterogeneous, the resulting estimates from maximum partial likelihood integrates the coefficient over the study follow-up length with the density of survival distribution and the censoring distribution. As censoring mechanisms are unknown, the variability of $\hat{\beta}_{\text{cox}}$ across studies can be significant even with the same underlying covariate effect. This undesirable property renders the results irreproducible and introduces difficulties in biomarker validation.

To remove the dependency of $\hat{\beta}_{\text{cox}}$ on censoring, previous literature has proposed censoring-robust estimators under various censoring assumptions and covariate types using the reweighting method. Beginning with the scenario of independent censoring, where $S_C(t|Z) = S_C(t)$, Equation 2.2 reduces to

$$\int_0^\infty S_C(t)E\left\{f_T(t|Z) \times \left[Z - \frac{EZS_T(t|Z)\exp(\beta Z)}{ES_T(t|Z)\exp(\beta Z)}\right]\right\} \mathrm{d}t = 0, \tag{2.3}$$

where the dependence on the censoring distribution acts as a multiplier at each integrand.

To obtain an estimator robust to the censoring distribution, Xu and O'Quigley (2000) proposed reweighting the summands in the Cox estimator by a consistent estimate of the censoring probability. Consequently, the estimating equation becomes:

$$\sum_{i=1}^n \int_{t=0}^\infty W^a(t) \left[Z_i - \frac{\sum_{j=1}^n Y_j(t)Z_j\exp(Z_j^T\beta)}{\sum_{j=1}^n Y_j(t)\exp(Z_j^T\beta)}\right] \mathrm{d}N_i(t) = 0, \tag{2.4}$$

where $W^a(t) = n\hat{S}(t)/\sum_{i=1}^n Y_i(t)$, and $\hat{S}(t)$ is chosen to be the left continuous version of the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the marginal survival function, as required for a predictable process to retain the Martingale framework (Fleming and Harrington, 2011).

The estimator in Equation 2.4 consistently estimates the solution to the following equation:

$$\int_0^\infty E\left\{f_T(t|Z) \times \left[Z - \frac{EZS_T(t|Z)\exp(\beta Z)}{ES_T(t|Z)\exp(\beta Z)}\right]\right\} \mathrm{d}t = 0, \tag{2.5}$$

which no longer depends on the censoring distribution. For intuition, the expectation of risk set size divided by number of subjects equals the product of probabilities that both censoring time and survival time are greater than t. After removing the survival probability, the resulting weight is actually the inverse censoring probability.

The above scenario provides insight into the necessity of adjusting for censoring probability under the independent censoring assumption and NPH. When the relative difference in hazards associated with a covariate of interest is homogeneous, the comparison at each estimand of the partial likelihood estimator would be best leveraged with the constant corresponding to the parameter value. However, when covariate effects on the relative difference in hazards varies with respect to time, each comparison at event time would contribute statistical information regarding the parameter value as a function of event time. Thus, at best, we could obtain an estimator interpreted as the population average effect over time by integrating the coefficient over the study follow-up length with the density of survival distribution. By construction, a regular partial likelihood estimator assigns equal weights to events, which works fine under proportional hazards (PH); under NPH, treating each estimand with the same importance ignores the potential events that are unobservable due to censoring. Therefore, the average effect is further reweighted by the censoring probability. The principle for removing dependence on the censoring distribution in this case turns out to be adjusted for $S_C(t|Z_i)$, which in this case simplifies to $S_C(t)$.

Under the more standard assumption, where independence between the event time and censoring time is relaxed to independence conditional upon covariates, $S_C(t|Z)$ no longer simplifies. Consequently, the estimator from Equation (2.4) no longer fully removes the dependence on the unknown censoring distribution S_C. As a result, the estimated associations might differ across studies even when the covariates remain the same due to censoring adjustments on observed events.

In this dissertation, the considered data types for covariates are binary covariates for group indicators, continuous covariates for baseline measurements, and longitudinal covariates for repeated biomarker measurements. In the case of binary covariates, Boyd et al. (2012) proposed to remove the dependency of the estimator on the unobserved censoring distribution by reweighting the estimating equation as

$$\sum_{i=1}^{n} \int_{t=0}^{\infty} W_i^b(t) \left[Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j W_i^b(t) \exp(Z_j^T \beta)}{\sum_{j=1}^{n} Y_j(t) W_i^b(t) \exp(Z_j^T \beta)} \right] dN_i(t) = 0, \tag{2.6}$$

where $W_i^b(t) = \{\hat{S}_C(t|Z_j)\}^{-1}$ is the consistent estimate of survival probability for censoring time depending on the subject using the the left continuous version of Kaplan-Meier estimator at time t. It can be shown that the estimate from the Equation 2.6 is consistent for the root of Equation 2.5, which removes the dependence on the censoring distribution.

In the case of conditional independent censoring with only categorical covariates, removing dependence on the censoring distribution relies on consistent estimation of $S_C(t|Z)$ and reweighting the estimating equation accordingly. Reweighting the contribution by the inverse of the estimated censoring probability adds the potential events at the same event times and with the same covariate values, thus compensating for the missing data that would otherwise be unobservable due to censoring.

This approach extends Equation 2.4 to the case of conditional independent censoring since, when censoring does not depend on covariates, the equation simplifies to Equation 2.4 as $W_i^b(t) = W^b(t)$. In the case of independent censoring, the Kaplan-Meier estimator suffices to provide a consistent estimate. However, when the censoring distribution is conditionally independent of categorical covariates, applying the Kaplan-Meier estimator by group provides consistent estimates of censoring distribution by groups and removes the dependence on censoring distribution.

Incorporating continuous variables poses challenges as censoring groups are not readily available. To estimate the censoring probability by subject, parametric models on covariates and censoring distribution are prone to misspecification since, unlike group treatment where the relationship can be fully described by groups, parametric relationships with continuous variables can include possibilities of linear and nonlinear forms, making a correct parametric model unrealistic.

Nguyen and Gillen (2017) proposed discretizing the relationship between covariates and censoring distribution by identifying approximate censoring groups based on survival tree method. With approximate groups identified, one can extend the method by Boyd et al. (2012) to obtain the censoring-robust estimator. To cluster observations based on the closeness of censoring distribution, Nguyen and Gillen (2017) adopted a survival tree approach described in LeBlanc and Crowley (1993). When growing the tree, each iteration generates the best split considering all possible covariate space partitions based on chosen survival statistics measuring the goodness of the split. At the prune step, tree complexity is controlled by a parameter α so that a new node is only kept if the information provided outweighs the complexity it adds to the tree. Since choosing an appropriate tuning parameter is key to tree performance, cross-validation is used to approximate the real performance under different model complexities, and α is chosen for the best score.

Comparable to tuning parameters, goodness-of-split statistics are equally crucial for assessing the censoring differences between observations. Nguyen and Gillen (2017) considered several weighted log-rank statistics, Kolmogorov-Smirnov type statistics, Cramer-von Mises type statistics, and weighted Kaplan-Meier statistics. Lastly, they chose the KG^{ρ} statistic from

Fleming et al. (1987) as

$$\mathrm{KG}^{\rho} = \frac{\sup_{t \geq 0} \int_0^t \hat{S}_p^{\rho+1/2} \left(\frac{Y_1 Y_2}{Y_1+Y_2}\right)^{1/2} \left(\frac{\mathrm{d}N_1}{Y_1} - \frac{\mathrm{d}N_2}{Y_2}\right)}{\left[\int_0^{\infty} \hat{S}_p^{2\rho+1} \left(1 - \frac{\Delta N_1 + \Delta N_2 - 1}{Y_1+Y_2-1}\right) \times \mathrm{I}\{Y_1 Y_2 > 0\} \frac{\mathrm{d}(N_1+N_2)}{Y_1+Y_2}\right]^{1/2}}, \tag{2.7}$$

where $\hat{S}_p$ represents the pooled Kaplan-Meier estimator for the time outcome, and $\rho \geq 0$ serves as the tuning parameter for power at different alternatives. Since formal inference is not the primary focus for clustering, Nguyen and Gillen (2017) used $\rho = 0$ in their simulations. While most statistics are capable of detecting differences in relative risk under a proportional hazard structure, the KG^{ρ} statistic can detect differences between groups even when the relative hazard is non-proportional and crossing, as it record the maximum deviation.

Given the censoring robust estimator in previous scenarios, a robust variance estimator is also necessary for inference. A commonly used robust variance estimator for MLE is the sandwich estimator Freedman (2006).

Define $A(\theta) = -n^{-1} \sum \partial^2 l_i(\theta)/\partial\theta^2$, where $l_i(\theta)$ is the log-likelihood contribution from ith subject. A represents the variance under the correctly specified model by information theory of MLE. Additionally, define $B(\theta) = n^{-1} \sum U_i(\theta) U_i^T(\theta)$, where $U_i(\theta) = \partial l_i(\theta)/\partial\theta$ and $n^{-1/2} \sum U_i(\theta)$ is asymptotically zero-mean normal with covariance matrix $B(\theta)$ if $U_i(\theta)$ is *i.i.d.*. Then the robust variance is given by $V(\theta) = \lim_{n\to\infty} A^{-1}(\theta)B(\theta)A^{-1}(\theta)$ and the estimatd variance is given by $\hat{V}(\hat{\theta}) = \hat{A}^{-1}(\hat{\theta})\hat{B}(\hat{\theta})\hat{A}^{-1}(\hat{\theta})$, by replacing $A(\theta)$ and $B(\theta)$ by their estimate $\hat{A}(\hat{\theta})$ and $\hat{B}(\hat{\theta})$ respectively.

For the Cox proportional hazards model, if the hazard is accurately specified, according to theorem 2.1 of Struthers and Kalbfleisch (1986), $n^{1/2}(\hat{\beta}_{\text{cox}} - \beta_0)$ converges to a zero-mean

Gaussian distribution with variance $A(\beta)$ defined as

$$A(\beta) = \int_0^\infty \left\{ \frac{s^2(\beta, t)}{s^0(\beta, t)} - \frac{s^1(\beta, t)^{\otimes 2}}{s^0(\beta, t)^2} \right\} s^0(t)dt,$$

where $S^r(\beta, t) = n^{-1}\sum_{i=1}^n Y_i(t)Z_i^r \exp(\beta Z_j)$, $s^r(\beta, t) = E(S^r(\beta, t))$. The estimated variance $\hat{A}^{-1}(\beta)$ is given by replacing the $s^r(\beta, t)$ by $S^r(\beta, t)$ as

$$\hat{A}(\hat{\beta}) = \sum_{i=1}^n \int_0^\infty \left\{ \frac{S^2(\hat{\beta}, t)}{S^0(\hat{\beta}, t)} - \frac{S^1(\hat{\beta}, t)^{\otimes 2}}{S^0(\hat{\beta}, t)^2} \right\} dN_i(t).$$

For Cox proportional hazards model, Equation 2.1 can not be expressed as summation of *i.i.d* score contribution by subject for $B(\beta)$ due to the partial likelihood construction. To solve the issue, Lin and Wei (1989) showed in theorem 2.1 that $n^{-1/2}U(\beta^0)$ is asymptotically equivalent to $n^{-1/2}\sum w_i(\beta^0)$, where β_0 is the true covariate effect. $w_i(\beta)$ is defined as

$$w_i(\beta) = \int_0^\infty \left\{ Z_i(t) - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} dN_i(t) - \int_0^\infty \frac{Y_i(t)\exp(\beta Z_i(t))}{s^{(0)}(t, \beta)} \left\{ Z_i(t) - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} d\tilde{G}(t),$$

where $\tilde{G}(t) = E\{\sum N_i(t)/n\}$. Since w_i's are *i.i.d* contribution from subjects, we can construct $B(\beta) = n^{-1}\sum w_i(\beta)w_i(\beta)^T$ with estimation as

$$\begin{aligned} n\hat{B}(\hat{\beta}) = \sum_{i=1}^n \int_0^\infty & \left\{ Z_i(t) - \frac{S^{(1)}(t, \hat{\beta})}{S^{(0)}(t, \hat{\beta})} \right\} dN_i(t) \\ & - \sum_{i=1}^n \int_0^\infty \frac{Y_i(t)\exp(\hat{\beta} Z_i(t))}{S^{(0)}(t, \hat{\beta})} \left\{ Z_i(t) - \frac{S^{(1)}(t, \hat{\beta})}{S^{(0)}(t, \hat{\beta})} \right\} d\hat{G}(t), \end{aligned}$$

where $\hat{G}(t) = \sum N_i(t)/n$.

For censoring robust estimator with categorical and continuous covariates, the robust variance of censoring robust estimator can be obtained by adding consistent estimator of censoring probability as (Boyd et al., 2012; Nguyen and Gillen, 2017)

$$-n\hat{A}(\hat{\beta}) = \sum_{i=1}^{n} \int_{0}^{\infty} \hat{W}_i^b(t|Z_i) \left\{ \frac{S_W^2(\hat{\beta},t)}{S_W^0(\hat{\beta},t)} - \frac{S_W^1(\hat{\beta},t)^{\otimes 2}}{S_W^0(\hat{\beta},t)^2} \right\} dN_i(t),$$

and

$$\begin{aligned} n\hat{B}(\hat{\beta}) = &\sum_{i=1}^{n} \int_{0}^{\infty} \hat{W}_i^b(t|Z_i) \left\{ Z_i(t) - \frac{S_W^{(1)}(t,\hat{\beta})}{S_W^{(0)}(t,\hat{\beta})} \right\} dN_i(t) \\ &- \sum_{i=1}^{n} \int_{0}^{\infty} \frac{\hat{W}_i^b(t|Z_i)Y_i(t)\exp(\hat{\beta}Z_i(t))}{S_W^{(0)}(t,\hat{\beta})} \left\{ Z_i(t) - \frac{S_W^{(1)}(t,\hat{\beta})}{S_W^{(0)}(t,\hat{\beta})} \right\} d\hat{G}_W(t), \end{aligned}$$

where

$$S_W^r(t,\beta) = n^{-1} \sum_{j=1}^{n} W_j^b(t)Y_j(t)Z_j^r \exp(\beta Z_j), \qquad s_W^r(t,\beta) = E[S_W^r(t,\beta)],$$

and $\hat{G}_W(t) = \sum \hat{W}_i^b(t|Z_i)N_i(t)/n$, accounting for modifications in the estimating equation.

2.2.7 Joint modeling of longitudinal and survival data

As longitudinal biomarker information becomes increasingly available, the demand for joint models of longitudinal and survival data has grown in recent years. Wu et al. (2012) provided a thorough review on current joint modeling approaches, inference methods, and related issues concerning the analysis of longitudinal and survival data under the proportional hazard assumption.

There are two main approaches to joint modeling: (i) the full-likelihood method and (ii) the two-step method. In both approaches, longitudinal data are statistically modeled to maximize the available statistical information. For modeling longitudinal data, two primary models are commonly used: (i) the linear mixed-effect model (LME) and (ii) the generalized estimating equation (GEE) method. The LME assumes a parametric formulation that allows for subject deviation from the population average, while the GEE employs a marginal model focusing on the average covariate effect at the population level, which is robust to model misspecification.

In joint modeling of longitudinal and survival data, it's necessary to adjust for individual deviation from the population average to explain the variation at observed event time. This adjustment can be achieved by either predicting the longitudinal covariate at the event time on the subject level or by jointly modeling longitudinal and time-to-event data based on common latent random effects. In either case, the LME is a natural choice for joint modeling. This is because the GEE only facilitates population-wise prediction under marginal modeling, while joint modeling on latent random effects naturally assumes the use of LME.

The two stage method involves analyzing longitudinal and survival data separately in two stages (Wu et al., 2012). In the first stage, an LME model is fitted based on longitudinal data. In the second stage, a separate survival model analysis is conducted, where unobserved longitudinal covariate values at event times are replaced with predictions from the models in the first stage.

The main advantages of this approach include the flexibility of model construction as the two stages are separated. It also simplifies estimation, as algorithms for both the LME model and the Cox proportional hazards model are readily available. However, it is noted that the estimator from the two-stage model may suffer from biased prediction due to misspecification in the first-stage model, and failure to account for the uncertainty of the estimation in the first stage in the second stage of the Cox proportional hazards model.

To assess the impact of possible deviation between predicted covariate values and unobserved covariate values at event times, Tsiatis et al. (1995) suggest that bias can be greatly reduced if the predicted longitudinal covariate values at event times are close to the unobserved true values. However, this method does not fully address the potential deviation.

For inference, the Cox model in the second stage ignores the variation in estimates contributed by first-stage modeling and predictions. To incorporate the uncertainty of estimation in the first stage, Wu et al. (2012) proposed a parametric bootstrap method, which involves the following steps:

Step 1: Regard the assumed model as truth and generate covariate values for each subject based on the fitted model in the first stage.

Step 2: Simulate survival times from the fitted survival model based on the generated longitudinal covariate values.

Step 3: Fit models using the same two-stage method and obtain new parameter estimates using the consistent analyzing model for all generated datasets.

On the contrary, the full-likelihood method aims to jointly model longitudinal and survival data with the full likelihood to avoid covariate prediction, as in the two stage method. The full likelihood is decomposed into three parts using the conditional likelihood method. Following the notation from Wu et al. (2012), we define the observed data in tubes $(t_i, \delta_i, z_i, x_i)$, where $i = 1, \ldots, m$, and let $\theta = (\beta, \alpha, \sigma, A, \lambda_0(t))$ denote the collection of parameters in the models. Under conditional independent assumption, the full likelihood is given by

$$L(\beta) = \prod_{i=1}^{n} \int f(t_i, \delta_i | z_i^*, \lambda_0, \beta) f(z_i | a_i, \alpha, \sigma^2) f(a_i | A) da_i,$$

where

$$
\begin{aligned}
f(t_i, \delta_i | z_i^*, \lambda_0, \beta) &= \{\lambda_0(t_i)\text{exp}(z_i^*(t_i)\beta_1 + x_i^T\beta 2)\}^{\delta_i} \\
&\quad \times \text{exp}\left\{-\int_0^{t_i} \lambda_0(t)\text{exp}(z_i^*(t_i)\beta_1 + x_i^T\beta_2)dx\right\}, \\
f(z_i | a_i, \alpha, \sigma^2) &= (2\pi\sigma^2)^{-m/2}\text{exp}\left\{-\frac{(z_i - z_i^*)^T(z_i - z_i^*)}{2\sigma^2}\right\}, \\
f(a_i | A) &= (2\pi|A|)^{-1/2}\text{exp}\left\{-\frac{a_i^T A^{-1} a_i}{2}\right\}.
\end{aligned}
$$

It is assumed that $z_i = U_i\alpha + V_i a_i + \epsilon = z_i^* + \epsilon_i$, so that longitudinal covariate is modeled via LME model with design matrix U_i and V_i for fixed effect and random effect respectively. We further denote fixed effect coefficient α and random effect coefficient a_i assuming $a_i \sim N(0, A)$ and $\epsilon_i \sim N(0, \sigma^2 I)$. The full likelihood consists of three components: (i) the likelihood of time-to-event data conditional on the assumed true longitudinal covariate at event time, (ii) the conditional likelihood of the true longitudinal covariate at event time based on the LME model conditional on the latent random effect, and (iii) the likelihood of random effects for individuals. As the third likelihood contains the unobservable random effect, a_i's are integrated out.

It is observed that the full likelihood decomposition for survival data still involves the unknown baseline hazard $\lambda_0(t)$. In addition, the joint likelihood method introduces a prior distribution of the random effect that is integrated out. Such likelihood construction complicates parameter estimation. Kalbfleisch and Prentice (1980) proposed estimating the baseline hazard $\lambda_0(t)$ using a nonparametric approach by assigning mass only to discrete times of death. However, the infinitely dimensional nature of $\lambda_0(t)$ violates the assumption for developing standard asymptotic distributions of maximum likelihood estimators. To maximize the full likelihood over model parameters, one could directly maximize it or use the EM algorithm. Direct maximization of observed likelihood requires numerical integration due to the intractable integration over the random effects. On the other hand, the computation cost of the EM algorithm is comparable to that of the direct maximization method, even though the M step could use an efficient gradient descent algorithm. However, since the EM algorithm still requires numerical integration in the E step and the algorithm alternates between steps until convergence. On the contrary, the two stage method allows for the estimation of the longitudinal model through closed-form likelihood estimation and time-to-event data through the Cox semi-parametric partial likelihood method, avoiding the estimation of unspecified baseline hazards that could be of infinite dimensions, and greatly reduce the computational burden.

Despite the heavy computation, the full likelihood can incorporate longitudinal model uncertainty. However, Hsieh et al. (2006) noted that even with the full likelihood approach, inference using Fisher information encounters problems. Direct maximization is unavailable due to $\lambda_0(t)$. Assuming a discrete mass at event time transfers the nonparametric estimation of $\lambda_0(t)$ to the parametric model, but standard MLE asymptotic development arguments do not apply. Furthermore, since $\lambda_0(t)$ is replaced with estimated point mass $\hat{\lambda}_0(t_j)$, the expectation step involves β, which disallows the regular Fisher information method. Additionally, as the longitudinal data are censored, the Fisher information matrix also suffers from unrecoverable data loss, deviating from the true Hessian matrix. Therefore, bootstrap seems to be the best strategy for the full-likelihood approach.

2.3 Influence function-based robust variance estimator

2.3.1 Influence function

In statistical analysis, robust variance estimation plays a crucial role in providing reliable inference, especially in the presence of outliers, influential observations, or model misspecifications. One prominent method for robust variance estimation is the influence function-based approach, which offers robustness against various forms of data perturbations while providing valid standard errors for parameter estimates.

The influence function, introduced by Hampel in 1968 (Law et al., 1986), measures the impact of an individual observation on a statistical estimator. It quantifies how a small perturbation in the data affects the estimate of a parameter, thereby characterizing the sensitivity of the estimator to changes in the data. Leveraging the influence function, robust variance estimation methods assess the stability of parameter estimates by examining their behavior under data perturbations.

Robust variance estimators based on the influence function are particularly appealing due to their flexibility and robustness properties. Unlike traditional variance estimators that rely on specific distributional assumptions, influence function-based approaches provide robust estimates of variance without requiring stringent model assumptions. This makes them well-suited for analyzing data with complex or unknown distributions, where traditional methods may fail to capture the underlying variability accurately.

Moreover, influence function-based robust variance estimators offer robustness against outliers and leverage points, which can heavily influence the results of statistical analysis. By down weighting the influence of extreme observations, these methods mitigate the impact of outliers on parameter estimates and standard errors, leading to more reliable inference.

In recent years, the influence function-based approach has gained popularity across various fields of statistics, including regression analysis, survival analysis, and machine learning. Its versatility and robustness make it a valuable tool for researchers and practitioners seeking accurate and reliable statistical inference in the presence of data uncertainty and model misspecifications.

The influence function (IF) is an infinitesimal approach to accessing the robustness of an estimator. For a formal definition, suppose we have observations $X_1, \ldots, X_n$, which are independent and identically distributed (*i.i.d*). A parametric model contains a family of probability distributions F_θ on the sample space, where the unknown parameter θ belongs to some parameter space Θ. In robustness theory, $\{F_\theta; \theta \in \Theta\}$ is only an idealized approximation of reality. To relax the assumption, we define the sample $(X_1, \ldots, X_n)$ with its empirical distribution F_n, where F_n is given by $(1/n)\sum_{i=1}^{n} \triangle_{x_i}$, and $\triangle_x$ is the point mass 1 in X. Thus, the limiting distribution G, given by $\lim_{n\to\infty} F_n$, identify the real distribution contain shape deviation from specified model $\{F_\theta; \theta \in \Theta\}$. The estimators considered are functionals, which denote maps that map distributions to real numbers or vectors. It is also assumed that there exists a functional $T : \text{domain}(T) \to R$ so that $T_n(X_1, \ldots, X_n) \to_{n\to\infty} T(G)$, and $T(G)$ is referred as the asymptotic value of T_n at G. The influence function of a statistic that can be regarded as a functional is the first derivative of the functional evaluated at some point in the space of distribution functions. The differentiation of statistical functionals was originally proposed by Mises (1947), and a von Mises functional is a functional that is sufficiently smooth to allow a series expansion. If T is a von Mises functional, then there exists a real function a such that for all G in the domain of T, we have

$$\lim_{\epsilon \downarrow 0} \frac{T\{(1-\epsilon)F + \epsilon G\}}{t} = \frac{\partial}{\partial \epsilon}[T\{(1-\epsilon)F + \epsilon G\}]_{\epsilon=0} = \int a(x)dG(x).$$

Setting $G = F$ in the previous equation to get that

$$\int a(x)dF(x) = \frac{\partial}{\partial \epsilon}[T\{(1-\epsilon)F + \epsilon F\}]_{\epsilon=0} = 0.$$

This kernel function a is the influence function with the definition given by

Definition 2.1. *The influence function (IF) of T at F is given by*

$$IF(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T\{(1-\epsilon)F + \epsilon G\}}{t} \tag{2.8}$$

for those x in support and where this limit exists.

For now, the IF is a function to evaluate the robustness of estimator. For influence function-based robust variance estimator, more derivations are needed. If some distribution G is near F, then the first-order von Mises expansion of T at F evaluated in G is given by:

$$T(G) = T(F + G - F) = T(F) - \int \mathrm{IF}(x, T, F) d(G - F)(x) + \text{remainder}, \tag{2.9}$$

provided that the distribution G is in small neighborhood of distribution F. Then, the behavior of $T(G)$ may be described by the first two terms, which is the basis for further applications. For application, when $X_i's$ are $i.i.d$ from F, the empirical distribution F_n will converge to G for sufficiently large n. Therefore, we can substitute G by F_n in Equation 2.9 for sufficiently large n. We also assume that $T_n(F_n)$ can be sufficiently approximated by $T(F_n)$. Then, we can rewrite Equation 2.9 as

$$T_n(F_n) \approx T(F_n) = T(F + F_n - F) = T(F) + \int \mathrm{IF}(x; T, F) dF_n(x) + \text{reminder}. \tag{2.10}$$

Note that we applied the property that $\int \mathrm{IF}(x; T, F) dF(x) = 0$ to reduce $d(F_n - F)(x)$ to $dF_n(x)$ in the last equality in the above equation.

Utilizing the above property, the influence function can be conceptualized as the change functional affected by minor contamination, standardized by the quantity of contamination. Returning to Equation 2.9, then evaluating the integral over the empirical distribution F_n, we obtain

$$\sqrt{n}(T_n - T(F)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(x_i; T, F) + \text{remainder}.$$

Hence, the distribution on the left side of the equation can be approximated by the right-hand side, which tends to become asymptotically normal according to the central limit theorem. Furthermore, $\sqrt{n}(T_n - T(F))$ converges in probability to $N\{0, V(T, F)\}$, where the asymptotic variance is given by:

$$V(T, F) = \int \mathrm{IF}(x; T, F)^2 dF(x),$$

which can be estimated by

$$\hat{V}(T, F) = \frac{1}{n}\sum_{i=1}^{n} IF(x_i; T, F)^2.$$

The distribution G has remained arbitrary up to this point. Therefore, a practical expression for the influence function is obtained by replacing the arbitrary G with $\triangle_x$.

Example 2.1. *Consider the simple example where the underlying distribution is the standard normal distribution with the density function $f(x) = (2\pi)^{-1/2} exp(-\frac{1}{2}x^2)$. To estimate the mean, we consider the arithmetic mean where $T_n = (1/n)\sum_{i=1}^{n} x_i$. Following equation 2.8, we can derive the IF of arithmetic mean as*

$$\begin{aligned} IF(x; T, F) &= \lim_{\epsilon \downarrow 0} \frac{\int ud[(1-\epsilon)F + \epsilon\triangle_x](u) - \int udF(u)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon)\int udF(u) + t\int ud\triangle_x(u) - \int udF(u)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{\epsilon x}{\epsilon} \\ &= x. \end{aligned}$$

Thus $V(T, F) = \int IF(x; T, F)^2 dF(x) = \int x^2 dF(x) = 1$ since $F(x)$ is standard normal distribution.

In the previous example, deriving the influence function was straightforward because the functional could be explicitly expressed. However, in more general cases, the functional may not be directly expressible, especially with M-estimators. In 1964, Peter J. Huber introduced the concept of M-estimators, defined as solutions to the minimization problem of an objective function $\sum_{i=1}^{n} \rho(x_i; \theta)$, where the solutions $\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} \rho(x_i, \theta)$ are termed

M-estimators, with M standing for maximum likelihood-type (Huber, 2011). Although maximum likelihood is one type of M-estimators, least squares estimators also fall under this category. For M-estimators, the estimator $\hat{\theta}$ can be defined as the maximization over the objective function.

$$\hat{\theta} = \text{argmax}_{\theta} E[G(X, \theta)].$$

If assuming G is differentiable with respect to θ, the the functional $T_n(F_n)$ can be written as the solution to the following equations

$$E[g(X, \theta)] = 0,$$

where $g(X, \theta) = \nabla_{\theta} G(X, \theta)$. In this setting, functional expression is not directly available, but the IF can be obtained similarly as mentioned in multiple sources Kahn (2022); Law et al. (1986). For contaminated distribution, the functional θ is the solution to the following equation:

$$\begin{aligned} \int g(u, \theta) d[(1-\epsilon)F + \epsilon \triangle_x](u) = 0 \\ (1-\epsilon) \int g(u, \theta) dF(u) + \epsilon \int g(u, \theta) d\triangle_x(u) = 0 \\ (1-\epsilon) \int g(u, \theta) dF(u) + \epsilon g(x, \theta) = 0 \end{aligned}$$

To obtain IF, we can apply total differentiation to figure out the effect of changing ϵ. It should also be aware that θ also varies as the solution to the new equation when ϵ changes.

$$\begin{aligned} \frac{d}{d\theta}(1-\epsilon) \int g(u, \theta) dF(u) + \frac{d}{d\epsilon} \epsilon g(x, \theta) = \frac{d}{d\epsilon} 0 \\ -\int g(u, \theta) dF(u) + (1-\epsilon) \int \nabla_{\theta} g(u, \theta) dF(u) \frac{d\theta}{d\epsilon} + g(x, \theta) + \epsilon \nabla_{\theta} g(x, \theta) = 0 \\ [-(1-\epsilon) \int \nabla_{\theta} g(u, \theta) dF(u)]^{-1} [-\int g(u, \theta) dF(u) + g(x, \theta) + \epsilon \nabla_{\theta} g(x, \theta)] = \frac{d\theta}{d\epsilon} \end{aligned}$$

Thus, the IF of M-estimator can be defined as

$$\begin{aligned}\text{IF}(x;\theta,F) &= \lim_{\epsilon\downarrow 0}\frac{d\theta}{d\epsilon}\\ &= \lim_{\epsilon\downarrow 0}[-(1-\epsilon)\int \nabla_\theta g(u,\theta)dF(u)]^{-1}[-\int g(u,\theta)dF(u)+g(x,\theta)+\epsilon\nabla_\theta g(x,\theta)]\\ &= -[\int \nabla_\theta g(u,\theta)dF(u)]^{-1}g(x,\theta),\end{aligned} \tag{2.11}$$

since $\epsilon \downarrow 0$ and $g(u,\theta)dF(u) = 0$ by the definition of functional.

2.3.2 Influence function for MLE and comparison to sandwich estimator

To derive the influence function (IF) for the Maximum Likelihood Estimator (MLE) and compare it to the robust sandwich estimator, another common alternative for robust variance estimation, let's start by considering the MLE as one type of M-estimator, widely used in practice. To derive the IF of the MLE, we can explicitly define $G(x,\theta)$ as the log-likelihood, typically denoted as $l(\theta;x)$, where $l(\theta;x) = \log[f(\theta;x)]$ and $f(\theta;x)$ is the likelihood function. Substituting the above expression into the equation:

$$\text{IF}(F,\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}l(\theta;x_i)\right]$$

we obtain:

$$\begin{aligned}\text{IF}(x;\theta,F) &= -[\int \nabla_\theta g(u,\theta)dF(u)]^{-1}g(x,\theta)\\ &= [-\int \frac{d^2}{d\theta^2}\{l(\theta;u)\}dF(u)]^{-1}\nabla_\theta l(\theta,x).\end{aligned} \tag{2.12}$$

The result can be decomposed into two components, where the first term in the bracket is the fisher information, and the second term is the score function, the first derivative of the log-likelihood function, evaluated at the covariate value x. Then to give an estimate, we use the weak law of convergence by

$$\begin{aligned}\sum_{i=1}^{n} \hat{\mathrm{IF}}(x_i;\theta,F)^T \hat{\mathrm{IF}}(x_i;\theta,F) &= \sum_{i=1}^{n} \hat{E}_F[\frac{d^2}{d\theta^2} l(\theta;u)]^{-1} \nabla_\theta l(\theta,x_i)^T \nabla_\theta l(\theta,x_i) \hat{E}_F[\frac{d^2}{d\theta^2} l(\theta;u)]^{-1} \\ &= \hat{E}_F[\frac{d^2}{d\theta^2} l(\theta;u)]^{-1} [\sum_{i=1}^{n} \nabla_\theta l(\theta,x_i)^T \nabla_\theta l(\theta,x_i)] \hat{E}_F[\frac{d^2}{d\theta^2} l(\theta;u)]^{-1},\end{aligned}$$

where the expectation is with respect to the assumed distribution F, and the fisher information matrix is approximated by observed fisher information as common practice. If define $A = \hat{E}_F[\frac{d^2}{d\theta^2} l(\theta;u)]$ and $B = \sum_{i=1}^{n} \nabla_\theta l(\theta,x_i)^T \nabla_\theta l(\theta,x_i)$, then the robust variance estimator from IF approach can be written as $A^{-1}BA^{-1}$, which is exactly the Huber-White robust sandwich variance estimator.

2.3.3 Influence function for censored data

In the case of survival analysis, where times to terminal events, also called survival time, serve as the response variable, the associations between the survival time and potential predictors are often of interest. In disease studies, it is expected that accurate survival times are not observed. Among frequent types of censoring, we often deal with right censoring, where we only know that the subject survives up to the last observation and thus, the survival time is only partially known. In the presence of censoring, complete case analysis by disregarding censored cases would waste statistical information. Thus, it is significant to incorporate censored cases and draw statistical information for maximum efficiency.

For survival probability estimation, the Kaplan-Meier estimator estimates the conditional probability of survival at time t incorporating censored cases (Kaplan and Meier, 1958).

When all true survival times are observed, the survival probability can be estimated easily. Let $T_1, T_2, \ldots, T_n$ be $i.i.d$ random variables with distribution function F_T. Then, by the weak law of convergence, the empirical distribution function defined as $F_n(t) = 1/n \sum_{i=1}^{n} I\{T_i \le t\}$ consistently estimates the underlying distribution function F_T. With the introduction of censoring, the survival time is obscured by censoring time and is thus not directly observable. Similarly, Define $C_1, C_2, \ldots, C_n$ as $i.i.d.$ random variable with distribution function F_C.The observations are n pairs $(X_1, \delta_1).(X_2.\delta_2), \ldots, (X_n, \delta_n)$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I\{T_i \le C_i\}$, the indication function for the event $T_i \le C_i$, namely that the true survival time is observed for subject i. For new observed data, the empirical distribution $F_n(t) = 1/n \sum_{i=1}^{n} I\{X_i \le t\}$ is consistent for underlying distribution of X_i's, where the conditional survival probability equals $[1-F_T(t)][1-F_C(t)]$. To estimate the probability that the survival time is longer than the specified time, the Kaplan-Meier estimator is given by:

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i is a time when at least one event happened, d_i describes the number of events that happened at timet_i, and n_i records the number of individuals known to have not yet had an event or been censored up to time t_i. When the number of individual in the study approach infinity, we can have the following approximation by Taylor approximation of the exponential function:

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \approx \prod_{i:t_i \le t} \exp(-d_i/n_i)$$

The last part of the equation is referred to as the Nelson cumulative hazard estimator (Nelson, 1969). To explain, we first define the hazard function as

$$\begin{aligned}\lambda(t) &= \lim_{\triangle t \downarrow 0} \frac{1}{\triangle t} P\{t \leq T < t + \triangle t | T \geq t\} \\ &= -\left[\frac{d}{dt}\{S(t)\}\right] / S(t) \\ &= f(t)/S(t).\end{aligned}$$

The hazard function can be viewed as the instantaneous rate of the event at time t, given that the individual was known to survive until that time. Then the function $\Lambda(t) = \int_0^t \lambda(u)du$ is called the cumulative hazard function for T. and Thus, it can be shown by differential equation fitting that for continuous T, $S(t) = \exp\{-\Lambda(t)\}$. When the hazard function is constant, the distribution of T simplifies to an exponential distribution with the rate parameter being that constant. Nelson's cumulative hazard function assumed a constant rate of hazard between events. It estimated the instantaneous rate of the event by the fraction of the number of events divided by the number of individuals known to survive until the time. Kaplan-Meier estimator is preferred over Nelson cumulative hazard function as it is less restrictive on distribution assumptions, and it is interesting to note that since two equations are approximately equal when the number of individuals in the study approaches infinity, it is called product limit estimator.

With censoring in presence, the derivation of IF can also be complicated. If T_i's are observed, then conditional distribution probability can be estimated by empirical distribution function $F_n(t)$, then the infinitesimal representation can be expressed as $T(F,t) = \int_0^t dF(u)$. Applying definition of IF gives $\mathrm{IF}(x; T, F, t) = \mathrm{I}(x \leq t)$. Under right censoring, IF calculate involves a chain rule for two subdistribution functions(Reid, 2007). The two subdistribution functions F^u and F^c are defined as

$$F^u(t) = P\{X_i \leq t, \delta_i = 1\},$$

and

$$F^c(t) = P\{X_i \leq t, \delta_i = 0\},$$

where superscripts are used to avoid confusion with lower scripts for survival time and censoring time. The relationship between sub-distribution function F^u and unobserved distribution function F_T and F_C can be summarized as

$$1 - F = (1 - F_T)(1 - F_C),$$

and

$$F^u(t) = \int_0^t [1 - F_C(u)] dF_T(u).$$

Given such a relationship, the Kaplan-Meier estimator can be written in another format for IF derivation (Breslow and Crowley, 1974). The empirical cumulative hazard process of the Kaplan-Meier estimator can be written as

$$\Lambda_n(t) = \int_0^t [1 - F_n(u)]^{-1} dF_n^u(u),$$

where F_n and F_n^u are empirical distribution function that converge to F and F^u. By the previous relationships, $\Lambda_n(t)$ converge to

$$\begin{aligned}\Lambda(t) &= \int_0^t [1 - F(u)]^{-1} dF^u(u) \\ &= \int_0^t [1 - F(u)]^{-1} [1 - F_C(u)] dF_T(u) \\ &= \int_0^t [1 - F_T(u)]^{-1} dF_T(u) \\ &= -\text{log}[1 - F_T(t)].\end{aligned}$$

The above equation established the consistency of the Kaplan-Meier estimator to true survival time distribution and that the sub-distribution representation that the Kaplan-Meier estimator is consistent to can be written as

$$1 - F_T(t) = \exp[\int_0^t \frac{d[1 - F^u(u)]}{1 - F^u(u) + 1 - F^c(u)}].$$

It is pointed out that the above equation holds only when the underlying distribution functions F_T and F_C are continuous. If unfortunately, F_T and F_C have jumps, then the above expression does not hold as the integrand and integrator may have common jump points. To include possible jumps in underlying distribution, the representation is extended to (Peterson, 1977)

$$S_T(t) = \exp\int_0^t \frac{dS^u(u)}{S^u(u) + S^c(u)} \times \exp\sum_{u \leq t} \log\frac{S^u(u^+) + S^c(u^+)}{S^u(u^-) + S^c(u^-)}, \tag{2.13}$$

where $S.(t) = 1 - F.(t)$ to simplify equation. The connection to the previous representation can be seen that if the underlying distribution function is continuous, then the second term in Equation 2.13 equals 0 and the expression back to the original representation. Under this representation, the functional for cumulative hazard would be

$$T(S^u, S^c, t) = -\int_0^t \frac{dS^u}{S^u(u) + S^c(u)} + \sum_{u \leq t} -\log\frac{S^u(u^+) + S^c(u^+)}{S^u(u^-) + S^c(u^-)}. \tag{2.14}$$

With the presence of two distribution functions, the influence functions are defined through bivariate von Mises expansion for an arbitrary bivariate functional $T(F_1, F_2)$

$$\begin{aligned} T(G_1, G_2) = T(F_1, F_2) &+ \int \mathrm{IF}_1(u; T, F_1, F_2) d(G_1 - F_1)(u) \\ &+ \int \mathrm{IF}_2(u; T, F_1, F_2) d(G_2 - F_2)(u) + \text{higher order terms}, \end{aligned} \tag{2.15}$$

and the two IFs are defined by

$$\frac{\partial}{\partial \epsilon} T(F_1 + \epsilon(G_1 - F_1), F_2 + \delta(G_2 - F_2))|_{\epsilon=0;\delta=0} = \int \mathrm{IF}_1(u; T, F_1, F_2) d(G_1 - F_1)(u)$$
$$\frac{\partial}{\partial \delta} T(F_1 + \epsilon(G_1 - F_1), F_2 + \delta(G_2 - F_2))|_{\epsilon=0;\delta=0} = \int \mathrm{IF}_2(u; T, F_1, F_2) d(G_2 - F_2)(u)$$

Then, following similar logistics of von Mises approximation in the case of single distribution, the approximation works similarly in the bivariate case

$$\begin{aligned}\sqrt{n}[T_n(G_1, G_2) - T(F_1, F_2)] \approx & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}_1(x_i; T, F_1, F_2) \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}_2(x_i; T, F_1, F_2) + \text{higher order terms.}\end{aligned} \tag{2.16}$$

For application toward the Kaplan-Meier estimator, F_1 and F_2 are substituted with sub-distributions F^u and F^c, and then G_1 and G_2 are replaced with corresponding empirical distribution functions S_n^u and S_n^c. It is shown in further detail that the two IFs can be written as (Reid, 2007)

$$\mathrm{IF}_1(x; T, S^u, S^c)(t) = \int_0^{x \wedge t} \frac{dS^u(u)}{[S^u(u) + S^c(u)]^2} + \frac{\mathrm{I}\{x \le t\}}{S^u(x) + S^c(x)}$$
$$\mathrm{IF}_2(x; T, S^u, S^c)(t) = \int_0^{x \wedge t} \frac{dS^u(u)}{[S^u(u) + S^c(u)]^2}.$$

Since $\mathrm{IF}_1(x; T, S^u, S^c)$ is with respect to sub-distribution S^u and $\mathrm{IF}_2(x; T, S^u, S^c)$ is with respect to sub-distribution S^c, the two IFs can be combined to give

$$\mathrm{IF}(x; T, F)(t) = \int_0^{x \wedge t} \frac{dS^u(u)}{[S(u)]^2} - \frac{\delta_x \mathrm{I}\{x \le t\}}{S(x)}, \tag{2.17}$$

where the first term of Equation 2.17 corresponds to the summation of first term of $\mathrm{IF}_1(x; T, S^u, S^c)(t)$ and the $\mathrm{IF}_2(x; T, S^u, S^c)(t)$, and the second term of equation 2.17 corresponds to the second term of $\mathrm{IF}_1(x; T, S^u, S^c)(t)$ adjusted for sub-distribution function. The justification can be given by Equation 2.15

$$\begin{aligned}
T(F_n^u, F_n^c) - T(F^u, F^c) &\approx \int\int_0^{x\wedge t} \frac{dS^u(u)}{[S^u(u)+S^c(u)]^2} + \frac{\mathrm{I}\{x \leq t\}}{S^u(x)+S^c(x)} dF_n^u \\
&+ \int\int_0^{x\wedge t} \frac{dS^u(u)}{[S^u(u)+S^c(u)]^2} dF_n^c \\
&= \int\int_0^{x\wedge t} \frac{dS^u(u)}{[S(u)]^2} - \frac{\delta_x \mathrm{I}\{x \leq t\}}{S(x)} d(F_n - F),
\end{aligned}$$

Using the fact that $F_n^u(t) = \frac{1}{n}\sum_{i=1}^n \mathrm{I}\{x_i \leq t, \delta_i = 1\} = \frac{1}{n}\sum_{i=1}^n \mathrm{I}\{x_i \leq t\}\mathrm{I}\{\delta_i = 1\}$. Then the IF for the Kaplan-Meier estimator can be found using $\hat{S}_T(t) = \exp - \hat{\Lambda}(t)$ as

$$\mathrm{IF}(x; T, F)(t) = \hat{S}_T(t) \left\{ \int_0^{x\wedge t} \frac{dF^u(u)}{[S(u)]^2} - \frac{\delta_x \mathrm{I}\{x \leq t\}}{S(x)} \right\}. \tag{2.18}$$

As we can see with the involvement of censoring, the IF becomes much more complicated, indicating possible sources of variation due to model misspecification and outliers.

2.3.4 Influence function for proportional hazards regression

In terms of proportional hazards regression, the Cox model is perhaps the most used among all life-table models. A robust variance estimator has been derived using an approximation to the Huber-White sandwich estimator (Lin and Wei, 1989). For the convenience of explanation, define:

$$\begin{aligned}
S^{(r)}(t) &= n^{-1}\sum_{i=1}^n Y_i(t)\lambda_i(t)Z_i(t)^{\otimes r} \\
s^{(r)}(t) &= E\{S^{(r)}(t)\} \\
S^{(r)}(\beta, t) &= n^{-1}\sum_{i=1}^n Y_i(t)Z_i(t)^{\otimes r}\exp\{\beta Z_i(t)\} \\
s^{(r)}(\beta, t) &= E\{S^{(r)}(\beta, t)\}
\end{aligned}$$

for $r = 0, 1, 2$, where $\lambda_i(t)$ is the true hazard function, $Y_i(t) = \mathrm{I}\{X_i \geq t\}$ and expectations are taken with respect to the triple tube of (X_i, δ_i, Z_i), $i = 1, \ldots, n$. Following the notation, the logarithm of the partial likelihood function can be written as

$$l(\beta) = \sum_{i=1}^{n} \delta_i [\beta Z_i(X_i) - \log\{S^{(0)}(\beta, X_i)\}],$$

and the score function as

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i(X_i) - \frac{S^{(1)}(\beta, X_i)}{S^{(0)}(\beta, X_i)} \right\}.$$

It has been shown that the solution of the Cox estimator under non-proportional hazards has a solution consistent to the root of the following equation(Struthers and Kalbfleisch, 1986):

$$\int_0^\infty s^1(t)dt - \int_0^\infty \frac{s^1(\beta, t)}{s^0(\beta, t)} s^0(t)dt = 0.$$

When the underlying hazard function is correctly specified, the asymptotic variance can be described by

$$A(\beta) = \int_0^\infty \left\{ \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \frac{s^{(1)}(\beta, t)^{\otimes 2}}{s^{(0)}(\beta, t)^2} \right\} s^{(0)}(t)dt$$

For the usual Huber-White sandwich estimator, $B(\beta)$ consists of the independent observed score function. For the partial likelihood estimator, the score contributions at event times are correlated for $S^{(r)}(\beta, t)$. Thus, they are replaced with asymptotically equivalent term $\sum w_i(\beta)$, where $w_i(\beta)$s are independent,

$$w_i(\beta) = \int_0^\infty \left\{ Z_i(t) - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} dN_i(t) - \int_0^\infty \frac{Y_i(t)\exp(\beta Z_i(t))}{s^{(0)}(t, \beta)} \left\{ Z_i(t) - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} d\tilde{G}(t),$$

and $\tilde{G}(t) = E\{\sum N_i(t)/n\}$.

As an alternative for robust variance estimator, IF for proportional hazards regression is derived (Reid and Crépeau, 1985). Let $H(X, \delta, Z)$ be the joint cumulative distribution function. Let $H(X, Z)$ be the joint marginal distribution function of (X, Z). Denote the empirical distribution functions as $H_n(x, z, \delta)$ and $H_n(x, z)$. Then the Cox estimator can be expressed

$$\int \delta \left\{ z - \frac{\int \tilde{z} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH_n(\tilde{x}, \tilde{z})\}}{\int e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH_n(\tilde{x}, \tilde{z})\}} \right\} dH_n(x, z, \delta) = 0$$

. Then, the infinitesimal functional $T(H)$ is the solution to

$$\int \delta \left\{ z - \frac{\int \tilde{z} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}}{\int e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}} \right\} dH(x, z, \delta) = 0$$

The derivation of IF follows the same principle and can be done through the sub-distribution technique. Given the complexity of the functional, further complications through the sub-distribution functions shall be avoided unless necessary. The derivation through implicit derivation is outlined(Knafl, 1978) for simplicity. To derive the IF, it suffices to calculate the limit of $\epsilon^{-1}[T(H+\epsilon\mathrm{I}\{x, \delta, z\}) - T(H)]$ as $\epsilon \to 0$. Using the implicit function theoremFilippova (1962), we take derivation with respect to terms related to ϵ. Define

$$\Psi(\beta, w, H) = \delta \left[z - \frac{E_H(Ze^{\beta z}\mathrm{I}\{\tilde{x} \geq x\})}{E_H(e^{\beta z}\mathrm{I}\{\tilde{x} \geq x\})} \right],$$

where β is the functional $T(H)$ and δ_w represents the point mass distribution at (x, δ, z). Then define $H_\epsilon = (1-\epsilon)H + \epsilon\delta_w$. Then, β_ϵ represents the functional under distribution function H_ϵ. Thus, the infinitesimal representation can be expressed as

$$\int \Psi(\beta_\epsilon, w, H_\epsilon) dH_\epsilon = 0$$

. Then it becomes clear that differentiation with respect to ϵ gives

$$\int \Psi(\beta, w, H) d[\delta_{\bar{w}} - H](w) + \int \frac{\partial}{\partial \beta} \Psi(\beta, w, H) dH(w) \frac{\partial \beta}{\partial \epsilon} + \int \frac{\partial}{\partial H_\epsilon} \Psi(\beta, w, H_\epsilon) dH(w) = 0,$$

where the three parts correspond to implicit differentiation with respect to ϵ for H_ϵ in the out integral, β_ϵ in $\Psi(\beta_\epsilon, w, H_\epsilon)$ and H_ϵ in $\Psi(\beta_\epsilon, w, H_\epsilon)$. It can be seen that $\int \frac{\partial}{\partial \beta} \Psi(\beta, w, H) dH(w)$ is $A(\beta)$, the variance matrix assuming model correctness and that $\lim_{\epsilon \downarrow 0} \partial \beta / \partial \epsilon$ is the IF. Then the calculation of $\frac{\partial}{\partial H_\epsilon} \Psi(\beta, w, H_\epsilon)$ gives

$$\begin{aligned} \frac{\partial}{\partial H_\epsilon} \Psi(\beta, w, H_\epsilon) &= \frac{\partial}{\partial H_\epsilon} \delta \left\{ z - \frac{\int z e^{\beta z} \mathrm{I}(\tilde{x} \geq x) d[H + \epsilon(\delta_w - H)]}{\int e^{\beta z} \mathrm{I}(\tilde{x} \geq x) d[H + \epsilon(\delta_w - H)]} \right\} \\ &= -\delta \left\{ \frac{E_{\bar{w}} z e^{\beta z} \mathrm{I}(\tilde{x} \geq x)}{E_H e^{\beta z} \mathrm{I}(\tilde{x} \geq x)} - \frac{E_H z e^{\beta z} \mathrm{I}(\tilde{x} \geq x) E_{\bar{w}} e^{\beta z} \mathrm{I}(\tilde{x} \geq x)}{[E_H e^{\beta z} \mathrm{I}(\tilde{x} \geq x)]^2} \right\} \\ &= -\tilde{\delta} \left\{ \frac{z e^{\beta z} \mathrm{I}(\tilde{x} \leq x)}{s^{(0)}(\tilde{x}, \beta)} - \frac{s^{(1)}(\tilde{x}, \beta) e^{\beta z} \mathrm{I}(\tilde{x} \leq x)}{[s^{(0)}(\tilde{x}, \beta)]^2} \right\}. \end{aligned}$$

IF are obtained by arranging terms as

$$A(\beta) \times \mathrm{IF}(w; T, H) = \delta \left\{ z - \frac{s^{(1)}(x, \beta)}{s^{(0)}(x, \beta)} \right\} - \left[e^{\beta z} \int \frac{\tilde{\delta} \mathrm{I}(\tilde{x} \leq x)}{s^{(0)}(\tilde{x}, \beta)} \left\{ z - \frac{s^{(1)}(\tilde{x}, \beta)}{s^{(0)}(\tilde{x}, \beta)} \right\} \right] dH(\tilde{x}, \tilde{\delta}, \tilde{z}), \tag{2.19}$$

where the first term is a regular Huber-White robust variance estimator if the terms are independent and similar to the IF of M-estimators. The second term represents the influence from the risk sets. Comparing IF with the previous approximation to Huber-White robust variance estimator, they are the same even if the approaches are quite different.

2.3.5 Influence function for two-stage model

Many estimators are obtained through a two-stage model, in which the first stage provides input for the second stage. For instance, in our proposed censoring robust estimator, the first stage model helps approximate the exact hazard by predicting longitudinal covariate value at event time. For mathematical representation, the two-stage estimation problem can be characterized by (Hardin, 2002)

$$\text{Model 1} : E\{y_1|x_1,\theta_1\}$$

$$\text{Model 2} : E\{y_2|x_2,\theta_2,E\{y_1|x_1,\theta_1\}\},$$

where the parameter vectors to be estimated are θ_1 and θ_2. Instead of working on the full information maximum likelihood $f(y_1, y_2|x_1, x_2, \theta_1, \theta_2)$, θ_1 can be directly estimated independent from θ_2. Then, conditional on the estimations from the first stage, θ_2 can be estimated and thus finish the estimation process in two steps. For robust variance estimation, the influence function has been derived for the special cases where both stages consist of estimation equations (Hardin, 2002). The result is found to be similar to the Huber-White robust variance estimator. Among more general M-estimators, the Huber-White robust variance estimator may not exist as estimating equation representation is unavailable. The influence function still exists, and a general framework for the analysis of the robustness property was given considering the class of two-stage models defined as (Zhelonkin et al., 2012)

$$\begin{array}{l} E_F\left[\Psi_1\left(z^{(1)}; S(F)\right)\right]=0 \\ E_F\left[\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right)\right]=0 \end{array},$$

Where $\Psi_1(\cdot;\cdot)$ and $\Psi_2(\cdot;\cdot,\cdot)$ denote the score functions of the first and second stage estimators, respectively, $h(\cdot;\cdot)$ is a given continuously piecewise differentiable function in the second variable. Here S is the functional for the parameters of the first stage, such that $S\left(F_N\right)=\hat{\beta}_1$ and at the model $S(F)=\beta_1$, while T is the functional for the second stage, such that $T\left(F_N\right)=\hat{\beta}_2$ and at the model $T(F)=\beta_2$. $T(F)$ depends directly on F and indirectly on F through $S(F)$. Define $F_\epsilon=(1-\epsilon)F+\epsilon\Delta_z$ and Δ_z is the probability measure which puts mass one at the point z. Then the infinitesimal representation of functional of the second stage can be written as

$$\int \Psi_2\left(z^{(2)}; h\left(z^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) dF_\epsilon = 0.$$

Taking the derivative with respect to ϵ gives

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon}(1-\epsilon)\int \Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) dF(\tilde{z})\bigg|_{\epsilon=0} \\
&+\frac{\partial}{\partial\epsilon}\epsilon\int \Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) d\Delta_z\bigg|_{\epsilon=0} = 0.
\end{aligned}$$

The second terms give the point mass distribution, and the first term can be obtained through implicit differentiation with respect to the functionals $S(F)$ and $T(F)$. Arrange terms gives the following general equation for the influence function of the second stage estimator:

$$\begin{aligned}
\mathrm{IF}(z; T, F) = M^{-1}(&\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right) \\
&+\int \frac{\partial}{\partial\theta}\Psi_2\left(\tilde{z}^{(2)}; \theta, T(F)\right)\frac{\partial}{\partial\eta}h\left(\tilde{z}^{(1)}; \eta\right) dF(\tilde{z})\cdot \mathrm{IF}(z; S, F)),
\end{aligned}$$

where $M = -\int \frac{\partial}{\partial\xi}\Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S(F)\right), \xi\right) dF(\tilde{z})$.

It can be noticed that the above formula does not consider the complex estimator, but the method still applies towards our proposed two-stage estimator. The above formula consists of three parts. The multiplier M outside the parenthesis is the second derivative common to influence function for score equations. The first term in the parenthesis denotes the point mass at the observed point. The second term represents the differentiation with respect to ϵ through the functional of the first stage functional, where the chain rule starts from function $h(\cdot;\cdot)$ to functional $S(F)$, and finally involves the influence function of first stage estimator. Given the decomposition, the influence function proposed without considering influence from the first stage would be $M^{-1}(\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right)$. This simple influence function may come from a simple MLE in the second-stage model. Still, the influence function can be more complex after including the influence from factors such as partial likelihood and other potential variations. Nevertheless, given the influence function of the second-stage model without considering the variation from the first stage, the modification to include the variation from the first-stage model is similar. It is expected that there will be additional terms corresponding to the partial differential with respect to the functionals of the first stage. When the dependence over the first-stage model relies on function $h(\cdot;\cdot)$, the last term in the parenthesis can be added for direct application. For dependency on multiple functions, the chain rule can be further applied so that if the class of the two-stage models is defined as:

$$\begin{aligned} &E_F\left[\Psi_1\left(z^{(1)}; S(F)\right)\right]=0 \\ &E_F\left[\Psi_2\left(z^{(2)}; h_1\left(z^{(1)}; S(F)\right), \ldots, h_k\left(z^{(1)}; S(F)\right), T(F)\right)\right]=0 \end{aligned},$$

where there are k continuously piecewise differentiable functions defined similarly. Then the influence function of the second stage estimator can be represented as:

$$\begin{aligned} \mathrm{IF}(z; T, F)= & M^{-1}(\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right) \\ & +\sum_{j=1}^k \int \frac{\partial}{\partial \theta} \Psi_2\left(\tilde{z}^{(2)}; \theta, T(F)\right) \frac{\partial}{\partial \eta} h_j\left(\tilde{z}^{(1)}; \eta\right) d F(\tilde{z}) \cdot \mathrm{IF}(z; S, F)), \end{aligned}$$

2.4 Longitudinal clustering

Longitudinal clustering is a data analysis technique used to uncover latent structures and patterns within longitudinal data, where multiple measurements are collected from the same individuals over time. This approach goes beyond traditional clustering methods by considering the temporal nature of the data and identifying clusters based on similarity in longitudinal trajectories rather than static features. By segmenting individuals into distinct subgroups with similar longitudinal profiles, longitudinal clustering facilitates the exploration of complex temporal dynamics, heterogeneity, and underlying patterns within the data.

One of the primary challenges in longitudinal clustering is the high dimensionality of the data, which arises from the large number of variables measured at multiple time points. Analyzing such high-dimensional longitudinal data requires robust statistical methods capable of handling the complexity and extracting meaningful information from the temporal structure of the data.

Temporal dynamics play a crucial role in longitudinal clustering, as observations collected at different time points are often correlated within individuals. Clustering algorithms must account for these temporal dependencies and capture the dynamic nature of longitudinal trajectories. Methods such as dynamic time warping, functional data analysis, or latent trajectory models are commonly used to model and cluster longitudinal data while preserving the temporal relationships between observations.

Furthermore, longitudinal data often exhibit heterogeneity, where different individuals may follow distinct trajectory patterns or exhibit varying rates of change over time. Longitudinal clustering aims to identify and characterize this heterogeneity by partitioning the data into homogeneous subgroups with similar trajectory patterns. However, determining the optimal number of clusters and selecting appropriate clustering algorithms and distance measures are critical tasks in longitudinal clustering analysis.

Several clustering algorithms have been adapted or developed specifically for longitudinal data analysis, including hierarchical clustering, k-means clustering, mixture models, and trajectory-based methods such as growth mixture modeling and latent class trajectory analysis. These methods offer different approaches to partitioning the data and identifying clusters based on various criteria, such as similarity in trajectory shapes, cluster compactness, and interpretability of cluster profiles.

In practice, longitudinal clustering has applications across various domains, including healthcare, psychology, sociology, economics, and ecology. In healthcare, longitudinal clustering can help identify subgroups of patients with similar disease progression patterns or treatment responses, leading to more personalized and effective healthcare interventions. In social sciences, longitudinal clustering can uncover distinct developmental trajectories, behavioral patterns, or life course pathways among individuals, shedding light on important social and demographic phenomena.

Overall, longitudinal clustering is a powerful tool for exploring and uncovering hidden structures within longitudinal data, providing researchers with valuable insights into the dynamic nature of temporal processes and individual differences over time. By effectively clustering longitudinal data, researchers can better understand complex longitudinal phenomena, inform decision-making, and tailor interventions or policies to specific subpopulations. In the subsequent sections, we will provide a more detailed examination of various longitudinal clustering methods. Here, we offer a concise overview of these methods, setting the stage for deeper exploration in later sections.

K-means clustering, a widely-used technique, partitions data into a predetermined number of clusters based on similarity. In longitudinal data analysis, K-means extends to group trajectories by their shapes or profiles. The algorithm iteratively assigns data points to the nearest cluster centroid and updates centroids until convergence, minimizing within-cluster sum of squares. While intuitive and computationally efficient, K-means assumes clusters of similar size and shape, which may not hold in longitudinal data. Moreover, determining the appropriate number of clusters remains a challenge. Nonetheless, K-means serves as a foundational method in clustering analysis (Hartigan and Wong, 1979).

Two-step longitudinal clustering using random effect variables with K-means is a long-standing approach for clustering longitudinal data that incorporates subject-specific random effects first mentioned in MacQueen et al. (1967). This method combines the flexibility of mixed-effects modeling with the simplicity and efficiency of K-means clustering to identify meaningful subgroups within longitudinal datasets. In the first step, individual longitudinal trajectories are modeled using mixed-effects models. These models account for the overall trend of each subject's trajectory while capturing subject-specific random effects that represent deviations from the population trend. The random effects capture the unique characteristics of each subject's data that cannot be explained by the population-level trend. Once the mixed-effects models are fitted to the data, the subject-specific random effects are extracted. These random effects represent the unique variability in each subject's trajectory and serve as the basis for clustering. With the subject-specific random effects extracted, traditional K-means clustering is applied to group subjects into clusters based on the similarities in their random effects profiles. K-means clustering aims to minimize the within-cluster sum of squares by iteratively assigning subjects to clusters and updating the cluster centroids until convergence. By including subject-specific random effects in the clustering process, two-step K-means can capture individual variability in longitudinal trajectories, leading to more accurate and interpretable cluster assignments.

Group-based longitudinal clustering is a powerful analytical approach used to identify distinct subgroups or clusters within longitudinal data, where individuals share similar trajectories over time. This method is invaluable for uncovering heterogeneity in longitudinal datasets and understanding how different groups of individuals evolve over time. One seminal work in group-based longitudinal clustering is the study by Nagin (1999), where the author introduced the Group-Based Trajectory Modeling (GBTM) framework. GBTM is widely used for identifying latent groups with distinct developmental trajectories in longitudinal data.

The Growth Mixture Model (GMM) serves as an extension to the GBTM method which allows researchers to capture the heterogeneity in developmental trajectories that may exist among individuals within a population. One seminal work in the development and application of Growth Mixture Models is the study by Muthén and Shedden (1999), where they introduced GMM as a flexible approach for modeling longitudinal data with latent classes.

Spline-based longitudinal clustering methods provide a powerful framework for analyzing longitudinal data by capturing complex and nonlinear patterns of change over time. Unlike traditional parametric models that assume specific functional forms for growth trajectories, spline-based methods offer more flexibility and adaptability to diverse longitudinal data structures. In spline-based clustering, piecewise polynomial functions, known as splines, are fitted to the observed longitudinal trajectories. These splines allow for local flexibility in modeling the data, enabling the identification of distinct subgroups with unique trajectory shapes. By partitioning the time axis into intervals and fitting splines within each interval, spline-based clustering methods can effectively capture the heterogeneity in growth patterns across individuals or groups. One of the key advantages of spline-based methods is their ability to accommodate irregularly spaced and unequally sampled longitudinal data, which is common in longitudinal studies. Moreover, spline-based clustering approaches are robust to missing data and can handle data with varying observation times.

Chapter 3

Censoring robust estimation in joint longitudinal and survival modeling

3.1 Introduction

In many studies, particularly in the context of Alzheimer's disease (AD) research, understanding the timing of terminal events has become a primary focus. Survival analysis techniques provide a robust framework for investigating the intricate relationship between covariates and the risk of these terminal events under PH. This work is motivated by the need to explore such relationships, especially in the realm of AD biomarkers.

Longitudinally sampled AD biomarkers play a crucial role in understanding the disease's progression and identifying individuals at risk of developing AD or transitioning through its various stages over time. However, acquiring and analyzing these biomarkers, such as phosphorylated tau (P-tau) and amyloid-beta (Aβ), can be both costly and burdensome for study participants. These proteins, associated with the formation of plaques and tangles in the brain, are hallmark features of AD pathology (Selkoe, 1991).

Given the expense and complexity associated with obtaining biomarker data, it is imperative to employ statistical methods that robustly identify potential biomarkers associated with AD. By doing so, researchers can streamline disease pathology research and identify new avenues for potential intervention. The application of statistical methods, particularly joint longitudinal and survival analysis techniques, allows researchers to identify longitudinally sampled biomarkers that are most strongly associated with AD risk and progression.

The Cox proportional hazards model is frequently employed to investigate the association between potential biomarkers and the progression of AD. Its semi-parametric framework allows for the adjustment of multiple confounding covariates while also remaining robust to mis-specification of unknown baseline hazard functions and censoring mechanisms, all under the assumption of proportional hazards. However, despite its merits, the assumption of proportional hazards is frequently violated in practice, leading to scenarios characterized by (non-proportional hazards) NPH. It has been demonstrated that under NPH the Cox model consistently estimates a quantity that can be interpreted as the average covariate effect over the range of observed event times, weighted by both the survival and censoring distributions (Struthers and Kalbfleisch, 1986). This can lead to statistical conclusions that are not replicable across studies solely due to changes in patient censoring patterns. In an effort to restore robustness to the censoring distribution under NPH, Chapter 2 introduced several attempts that have been made to remove the dependency on censoring distribution. These attempts are based on settings ranging from independent censoring to conditional independent censoring with categorical and continuous covariates (Xu and O'Quigley, 2000; Boyd et al., 2012; Nguyen and Gillen, 2017). All the aforementioned methods utilize a similar idea of reweighting the estimands of Cox proportional hazards by the inverse of censoring probability to remove the dependency on the censoring mechanism. Our aim in Chapter 3 is to extend such methods to incorporate longitudinal covariates, thereby developing a joint modeling of longitudinal and survival data that can be robust to censoring mechanisms under NPH. In this chapter we develop a two-stage model using the censoring robust estimators in the second stage that incorporates longitudinal covariates. The complete algorithm is provided in Section 3.2. Section 3.3 presents simulated results to illustrate the performance of the proposed estimators compared to the direct application of previous methods. In Section 3.4, we apply the estimators to ADNI data where we consider the estimation of longitudinally measured biomarker effects on the risk of progression of dementia. Section 3.4.1 concludes with a discussion of the scientific relevance of the methodology and avenues for future research.

3.2 Method

To develop a censoring-robust estimator incorporating longitudinal covariates under NPH, we opt for a two-stage approach over a full-likelihood approach due to its flexibility and ease of implementation. When comparing the two methods, the two-stage method offers great flexibility with independent choices of models in each stage. For instance, one can construct the two-stage method with a parametric LME model in the first stage and a semi-parametric Cox model in the second stage. In exchange for model flexibility, the full-likelihood method incorporates the uncertainty of the longitudinal model with the conditional likelihood, which is ignored in the regular Cox variance estimator in the second stage. However, such a construction inevitably leads to a tremendous computational burden, even after approximating the hazard through numerical techniques.

Using the predicted longitudinal covariate value at each observed event time, direct application of previous censoring-robust estimators in the second stage introduces additional dependence on conditional covariate variance, complicating the interpretation of the estimator. Xu and O'Quigley (2000) suggests that even in the absence of censoring and with the true longitudinal covariate at event time, the Cox estimator represents a weighted average of $\beta(t)$ over both the true survival function $F(t)$ and the conditional covariate variance $v(t)$.

Under categorical and baseline continuous variables, where the conditional covariate variance is approximately constant over time t, the interpretation of the estimator can be approximated by $\hat{\beta}_{\text{Cox}} \approx \int_0^\infty \beta(t) \mathrm{d}F(t)$. In this scenario, the estimator approximate the weighted average covariate effect depending on the survival function $S(t)$, with the interpretation of the weighted time-average coefficient. However, for time-varying variables with varying variance with respect to time, the change conditional covariate variance is non-negligible, the Cox estimates is a weighted average covariate effect based on the true survival function and the conditional covariate variance function over time, making the estimator difficult to interpret. As the estimates are no longer approximated by a weighted time-average effect, the resulting findings may lack clinical significance.

Our goal is to develop a clinically meaningful estimator that is censoring-robust with the interpretation as the weighted time-average covariate effect. To focus on the interpretation of the censoring-robust estimator in the second stage, we will treat the substituted longitudinal covariate as the truth and consider similar techniques in Xu and O'Quigley (2000) to investigate the approximation of the quantity to which the censoring robust estimator in the second stage converges. Xu and O'Quigley (2000) provided the estimator interpretation under independent censorship, and we will extend the interpretation to the context of conditional independent censorship with time-varying covariates.

3.2.1 Censoring robust estimation with a time-varying covariate

Recall the hazard function under NPH setting:

$$\lambda(t, Z(t)) = \lambda_0(t)\text{exp}(\beta(t)Z(t)). \tag{3.1}$$

Let us define:

$$\pi_i(\beta(t), t) = \frac{W_i^b(t) Y_i(t) \exp(\beta(t) Z_i(t))}{\sum_{j=1}^n W_j^b(t) Y_j(t) \exp(\beta(t) Z_j(t))}, \tag{3.2}$$

where $W_i^b(t)$ is a consistent estimator of censoring survival probability of ith subject at time t. The follwoing proposition states that $\{\pi_i(\beta(t), t)\}_i$ provides a consistent estimate of the conditional distribution of $Z(t)$ given $T = t$ under Model 3.1.

Proposition 1. *Under Model 3.1 and a conditional independent censorship, assuming $\beta(t)$ known, given a consistent estimator of censoring survival probability at time t as $W_\cdot^b(t)$, the conditional distribution function of $Z(t)$ given $T = t$ is consistently estimated by*

$$\hat{P}(Z(t) \leq z | T = t) = \sum_{\{j : Z_j(t) \leq z\}} \pi_j(\beta(t), t).$$

Proof. The theorem presented above is an extension of Theorem II.1 from Xu and O'Quigley (2000) to accommodate censoring dependent on covariates. In the original derivation, independent censoring is assumed, so that the censoring probability remains the same at time t regardless of the covariate and thus cancels out in the derivation. Under the conditional independent assumption, however, the censoring probability does not cancel out due to its dependence on covariates. Nevertheless, with the provision of a consistent estimate $W_i^b(t)$ of $1/S_C(t|Z)$, the censoring probability is canceled out by $W_i^b(t)$. The subsequent derivation then follows exactly as in Xu and O'Quigley (2000). □

Proposition 1 is an extension of Theorem 1 in Xu and O'Quigley (2000) for conditional independent censorship. As $\beta(t)$ is often unknown, the major implication of Proposition 1 is the of expectation and variance conditional upon time t. Define $S_W^{(r)}(\beta(t), t) = \frac{Z^r W_i^b(t) Y_i(t) \exp(\beta(t) Z_i(t))}{\sum_{j=1}^n W_j^b(t) Y_j(t) \exp(\beta(t) Z_j(t))}$, $S_W^{(r)}(\beta, t) = \frac{Z^r W_i^b(t) Y_i(t) \exp(\beta Z_i(t))}{\sum_{j=1}^n W_j^b(t) Y_j(t) \exp(\beta Z_j(t))}$, and $s_W^{(r)}(\cdot, t) = E[S_W^{(r)}(\cdot, t)]$ for $r = 0, 1, 2$. Then we have $s_W^{(1)}(\beta(t), t)/s_W^{(0)}(\beta(t), t) = E\{Z(t)|T = t\}$ and $s_W^{(1)}(\beta, t)/s_W^{(0)}(\beta, t)$

is what we obtain when we force a constant β in place of $\beta(t)$. Both do not involve the censoring distribution, which leads to the following proposition:

Proposition 2. *Define estimator $\hat{\beta}$ as the solution to the estimating equation*

$$\sum_{i=1}^{n}\int_{t=0}^{\infty} W_i^b(t)\left[Z_i(t)-\frac{\sum_{j=1}^{n} W_j^b(t)Y_j(t)Z_j(t)\,exp(Z_j(t)^T\beta)}{\sum_{j=1}^{n} W_i^b(t)Y_j(t)\,exp(Z_j(t)^T\beta)}\right]dN_i(t)=0. \tag{3.3}$$

Then under model 3.1 and conditional independent censorship, the estimator $\hat{\beta}$ converges in probability to a constant β_0, where β_0 is the unique solution to the following equation

$$\int_0^{\infty}\left\{\frac{s_W^{(1)}(\beta(t),t)}{s_W^{(0)}(\beta(t),t)}-\frac{s_W^{(1)}(\beta,t)}{s_W^{(0)}(\beta,t)}\right\}dF(t)=0. \tag{3.4}$$

Proof. Given a consistent estimator $W_i^b(t)$ of $1/S_C(t|Z_i(t))$, which is the left continuous version of Kaplan-Meier estimate of censoring survival distribution, following similar techniques as in Struthers and Kalbfleisch (1986) gives that our estimator in Equation 3.3 is consistent for the solution to

$$\begin{aligned}&\int_0^{\infty}\frac{1}{S_C(t|Z(t))}S_C(t|Z(t))E\Big\{f_T(t|Z(t))\times\\&\left[Z(t)-\frac{EZ(t)\frac{1}{S_C(t|Z(t))}S_C(t|Z)S_T(t|Z(t))\exp(\beta Z(t))}{E\frac{1}{S_C(t|Z(t))}S_C(t|Z(t))S_T(t|Z(t))\exp(\beta Z(t))}\right]\Big\}\mathrm{d}t\\&\qquad=\int_0^{\infty}E\left\{f_T(t|Z(t))\times\left[Z(t)-\frac{EZ(t)S_T(t|Z(t))\exp(\beta Z(t))}{ES_T(t|Z(t))\exp(\beta Z(t))}\right]\right\}\mathrm{d}t=0,\end{aligned}$$

which does not depend on censoring distribution. To show that $\hat{\beta}$ is consistently for β_0 ,Since the only change is the additional reweighting of consistent estimate of inverse censoring probability, it suffices to show the conditions hold for derivation in Ringhui (1996), which is similar to those in Andersen and Gill (1982). This requires that there exists an open neighborbood $\mathcal{B}$ of β_0 and $s^{(r)}(\beta,t), r=0,1,2$ defined on $\mathcal{B}\times[0,\tau]$ that satisfy the following:

1. $\int_0^t \lambda_0(u)du<\infty$;

2. Z_i is bounded $\forall t \in [0, \tau]$;

3. $\sum(\beta, t) = \int_0^\tau v(\beta, u) s_W^{(0)}(\beta, u)\lambda_0(u)\mathrm{d}u$ is positive definite $\forall \beta \in \mathcal{B}$;

4. $\sup_{\beta \in \mathcal{B}, t \in [0,\tau]} ||S_W^{(r)}(\beta, t) - s_W^{(r)}(\beta, t)|| \xrightarrow{p} 0$ as $n \to \infty$;

5. $s_W^{(0)}(\beta, t)$ is bounded away from 0 for $t \in [0, \tau]$;

6. For $r = 0, 1, 2$, $s_W^{(r)}(\beta, t)$ is a continuous function of β uniformly in $t \in [0, \tau]$, $s_W^{(1)}(\beta, t) = \partial/\partial\beta\{s_W^{(0)}(\beta, t)\}$ and $s_W^{(2)}(\beta, t) = \partial/\partial\beta\{s_W^{(1)}(\beta, t)\}$.

The assumption that there is positive probability that subject i is at risk over the support interval implies that the conditions 1 and 5 holds. The conditions 2 and 3 are assumed. Then condition 2 together with the dominated convergence theorem shows that condition 6 holds. The assumption of *i.i.d* observations together with condition 2 implies that the $S_W^{(r)}(\beta, t)$ are comprised of independent terms so that by the strong law of large numbers $S_W^{(r)}(\beta, t)$ converges to $s_W^{(r)}(\beta, t)$. Given that the conditions 1-6 hold, application of theorem II.3 in Ringhui (1996) gives that $\hat{\beta} \xrightarrow{p} \beta_0$ by replacing $S^{(r)}(\cdot)$ with $S_W^{(r)}(\cdot)$ and $s^{(r)}(\cdot)$ with $s_W^{(r)}(\cdot)$. □

Proposition 2 gives a result similar to that of Theorem 2 in Xu and O'Quigley (2000) but under conditional independent censoring and time-varying covariates. With the aid of Proposition 1, Proposition 2 implies that the direct application of the previous censoring robust estimator removes the dependence on censorship as both fractions do not involve the censoring distribution. Next we consider the interpretation of the estimand targeted by solving Equation 3.4.

Proposition 3. *Under the same conditions as in Proposition 3.1 and 3.2, define* $g(x) = s_W^{(1)}(x, t)/s_W^{(0)}(x, t)$. *Then equation 3.4 can be written as* $\int_0^\infty \{g(\beta(t)) - g(\beta)\}\, dF(t) = 0.$

Applying a first-order Taylor series approximation to the integrand gives

$$\int_0^\infty v(t)\{\beta(t) - \beta_0\}dF(t) \approx 0,$$

where $v(t) = v(\beta(t), t) = \frac{\partial g(x)}{\partial x}_{|\beta(t)} = Var(Z(t)|T = t)$.

Proof. Given the condition 6, $s_W^{(1)}(\beta, t) = \partial/\partial\beta\{s_W^{(0)}(\beta, t)\}$ and $s_W^{(2)}(\beta, t) = \partial/\partial\beta\{s_W^{(1)}(\beta, t)\}$, applying the first order taylor expansion of $g(\beta)$ around $\beta(t)$ gives

$$g(\beta) \approx g(\beta(t)) + \left\{ \frac{s_W^{(2)}(\beta, t)}{s_W^{(0)}(\beta, t)} - \left[\frac{s_W^{(1)}(\beta, t)}{s_W^{(0)}(\beta, t)} \right]^2 \right\} \{\beta(t) - \beta\},$$

where the second term is, by definition, in Proposition 1 the conditional covariate variance at time t, denoted by $v(t) = Var(Z(t)|T = t)$. □

Rearranging terms, we obtain $\beta_0 \approx \int_0^\infty v(t)\beta(t)dF(t)/\int_0^\infty v(t)dF(t)$, representing a weighted average of $\beta(t)$ over time by both the survival distribution and the conditional covariate variance. Importantly, the interpretability of the resulting estimand, β_0, is complicated by its reliance on the conditional covariate variances.

The extent of dependence is negligible in cases where only time-invariant covariates are considered, as the conditional covariance remains approximately constant (Xu and O'Quigley, 2000). Consequently, $\beta_0 \approx \int \beta(t)dF(t)$ can be interpreted as a weighted time average covariate effect. It is, however, common for the variance of conditional covariates for biomarkers to exhibit temporal variability. Hence, it is possible for two studies that share the same underlying covariate effect function over time to yield disparate outcomes as a result of variations in conditional covariate variances in longitudinal covariates.

To better understand the above ideas, we designed a sequence of numerical studies. First, we aim to demonstrate that we can retrieve the treatment effect in the case of constant conditional covariance. Subsequently, we will investigate the scenario of changing conditional covariance by incorporating random effects. To demonstrate the necessity of reducing bias in estimation through predicting the longitudinal covariate at event time, we designed a simulation study where the longitudinal covariate is measured at fixed intervals of 6 months, 3 months, and 0.01 years, with a maximum follow-up length of 4 years.

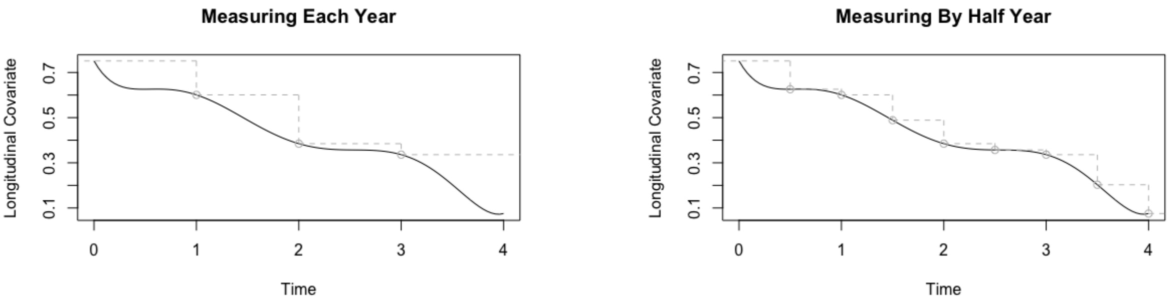


Figure 3.1: Effect of measuring frequency by treating longitudinal covariate constant between measurements

From Figure 3.1, we observe that as the measurement frequency decreases, there is less bias in the longitudinal covariate of using last observation carried forward method for imputing the missing longitudinal covariate values at event times, resulting in reduced bias overall. In our simulation study, we assume that the data comes from the following hazard model:

$$\lambda_i(t) = \lambda_0 \exp(\beta_1 X_{1i} + \beta_2 X_{2i}(t))$$

Here, X_{1i} represents the baseline categorical covariates for subject i following Bernoulli distribution with probability of $(0.4, 0.6)$, and $X_{2i}(t)$ denotes the underlying longitudinal covariate value at time t for subject i, consisting of two groups underlying longitudinal trajectories as depicted in Figure 3.2. The randomness of $X_{2i}(t)$ is introduced with $i.i.d$ measurement error following $N(0, 0.02)$. For simplicity, we set $\lambda_0(t)$ to a fixed constant λ_0. The population parameters are specified as follows: $\lambda_0 = 0.3$, $\beta_1 = 1$, and $\beta_2 = -1$.

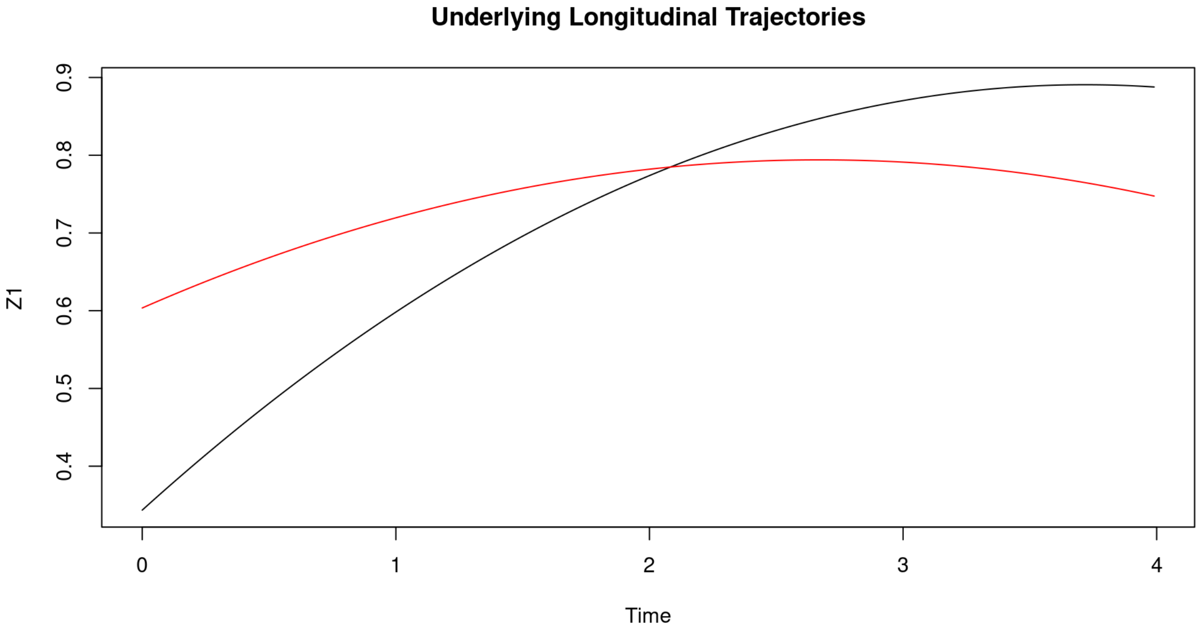


Figure 3.2: Two groups with underlying longitudinal trajectories are presented, where one group exhibits a higher intercept with slower increases, while the group with a lower intercept shows faster increases.

Long. Meas.	Mean		Model SE		Emp. SE		Cov. Prob.	
Spacing	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Baseline Only	-0.74	0.75	0.10	0.10	0.11	0.09	0.00	0.00
6 Months	0.99	-1.07	0.11	0.05	0.12	0.06	0.91	0.70
3 Months	1.00	-1.02	0.10	0.04	0.12	0.04	0.94	0.92
0.01 Year	1.01	-1.00	0.11	0.05	0.11	0.05	0.94	0.97

Table 3.1: Simulated coefficient estimates from the Cox proportional hazards model when a longitudinal covariate is included as a time-varying covariate (last observation carried forward). Density of longitudinal measures are varied from having only a single baseline measurement up to every 0.01 years. In each case, the true value of β_1 is 1.0 and the true value of β_2 is -1.0.

From Table 3.1, we observe that when only the baseline measurement of the longitudinal covariate is used, the bias in the coefficient is significant enough to alter its direction. However, as the longitudinal measurement become frequent, the imputed longitudinal covariate value approaches the true value at event times, the bias diminishes, and the coverage probability returns to 0.95. It's noteworthy that the model-based standard error remains similar to the truth, as the robust variance estimator is incorporated in the `coxph()` function in R. Therefore, it is crucial to predict the longitudinal covariate to reduce bias.

In the subsequent simulation, we maintain the same model setup but introduce a longitudinal covariate effect under the NPH scenario. Specifically, we set $\beta(t)$ to 0 if $t < T_0$ and -1 if $t \geq T_0$. For censoring, we employ the power function distribution with a cumulative distribution function defined as $F(x) = \left(\frac{x}{\theta}\right)^r$, where θ determines the maximum follow-up length and r governs the location of predominant censoring. A higher r implies more censoring in later stages. To ensure a minimum 1-year follow-up duration for simulation stability, we define the cumulative density function as $F^*(x) = \left(\frac{x-1}{\theta}\right)^r$, where $1 \leq x \leq 4$ and $\theta = 3$. We considered the two-stage modeling to exhibit that we can mitigate the effect of longitudinal measurement frequency and the censoring distribution on regular Cox estimator using last observation carrying forward method in NPH setting. Specifically, in the first stage, longitudinal data is modeled with the LME model, giving subject-wise predictions of longitudinal covaraite at event time. Then in the second stage, coefficient is estimated via the censoring robust estimator in Equation 3.3.

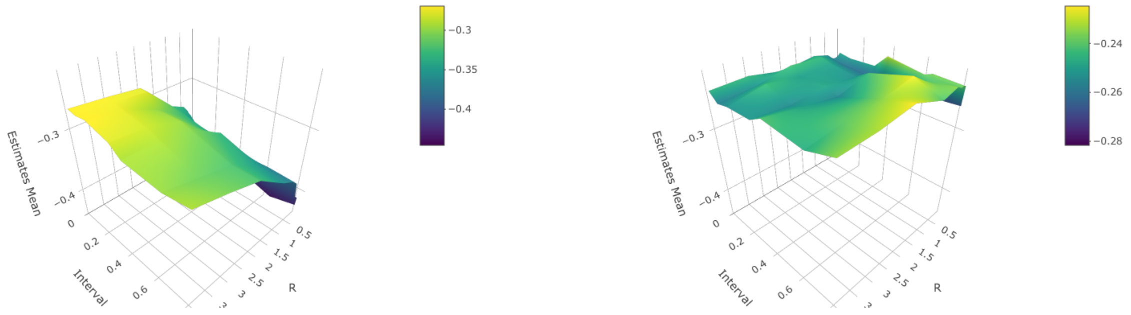


Figure 3.3: Surface plot comparing longitudinal coefficient estimates from naive Cox model and reweighted method after longitudinal prediction. $T_0 = 1$

One can observe from Figure 3.3-3.6 that in the left panels, which depict results from the naive Cox proportional hazards model, the surface tilts with respect to both the measurement interval size and the choice of r in the power function distribution. Conversely, in the right

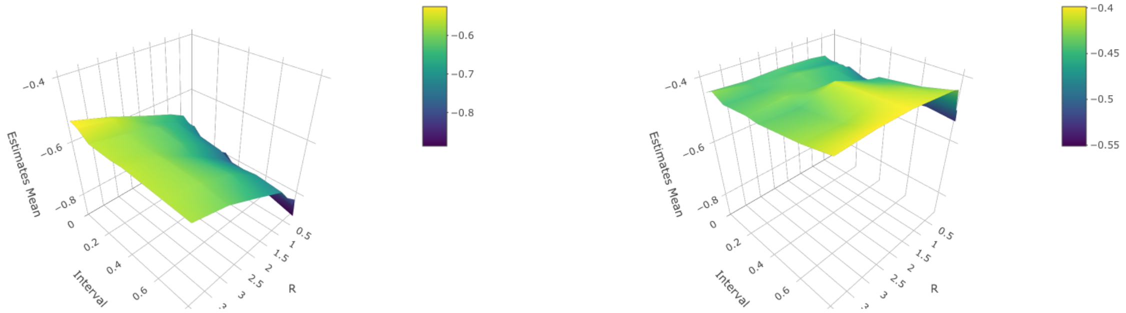


Figure 3.4: Surface plot comparing longitudinal coefficient estimates from naive Cox model and reweighted method after longitudinal prediction. $T_0 = 2$

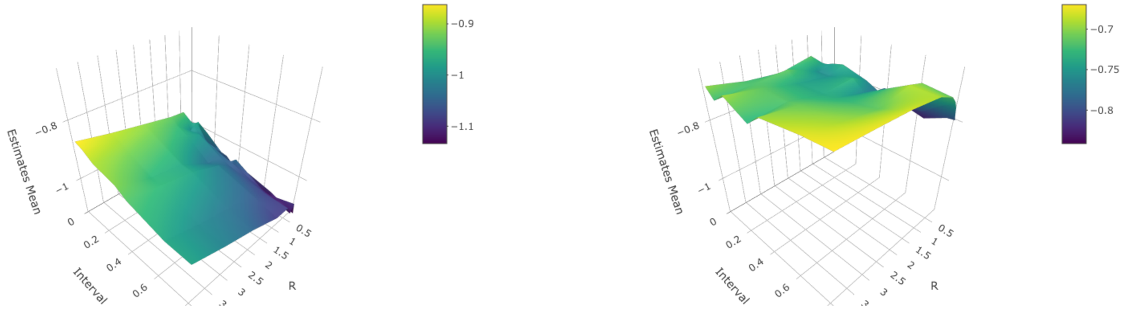


Figure 3.5: Surface plot comparing longitudinal coefficient estimates from naive Cox model and reweighted method after longitudinal prediction. $T_0 = 3$

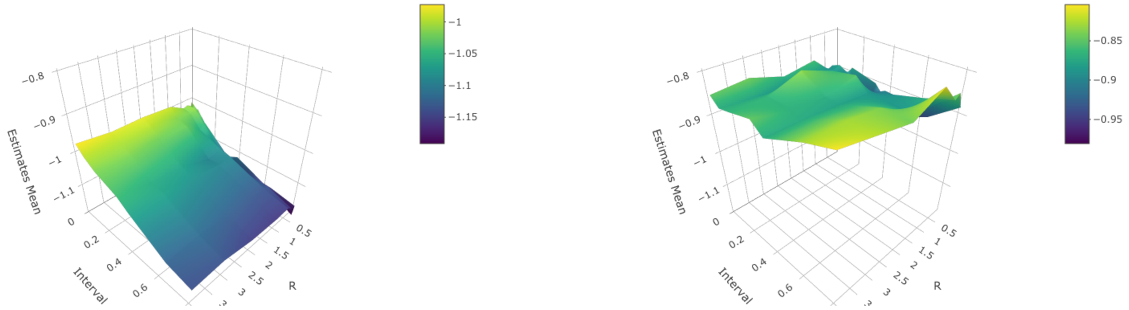


Figure 3.6: Surface plot comparing longitudinal coefficient estimates from naive Cox model and reweighted method after longitudinal prediction. $T_0 = 3.5$

panels, our estimator is approximately a flat surface, demonstrating robustness to both censoring distribution and measurement interval, except in cases of extreme censoring where most subjects are censored before $T = 2$, resulting in less information being available for the longitudinal covariate effect in later follow-up times.

With the addition of a random effect in the underlying longitudinal trend, the subsequent simulation demonstrated that previous estimated results are also influenced by the conditional covariate variance.

Consider the following effect of the longitudinal covariate value

$$\beta_2(t) = \begin{cases} 0, & \text{for } 0 \leq t < 1 \\ log(0.3), & \text{for } 1 \leq t \leq 4, \end{cases}$$

where the diverging effect manifests as a negative effect after $t = 1$. The data is generated through the hazard function as

$$\lambda_i(t) = 0.3\text{exp}(X_{1i} + \beta_2 X_{2i}(t)).$$

The censoring distribution is set to be $C \sim \text{Power}(4, r),\ F(x) = \left(\frac{x}{\theta}\right)^r$. We estimate the coefficient via the same two-stage model with the except that the underlying longitudinal trajectories follow the LME model with random slope that follows $N(0, v_{slope})$. The covariates are generated such that

$$\begin{aligned} X_1 &\sim \text{Bin}(1, 0.4), \\ X_2(t)|b &\sim N(f(t) + b_1 + b_2 * t^2, v_{error} * I_D), \\ \begin{bmatrix} b_1 \\ b2 \end{bmatrix} = b &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} v_{inter} & 0 \\ 0 & v_{slope} \end{bmatrix}\right), \end{aligned} \tag{3.5}$$

where X_1 is generated through $\text{Bin}(1, 0.4)$ and $X_2(t)$ is generated using underlying longitudinal trajectories with the addition of independent random intercept with $v_{inter} = 0.01$, random slope with $v_{slope} = 0.08$ and normal measurement error with $v_{error} = 0.02$. The underlying longitudinal trajectories without random effect follows the function

$$f(t) = \begin{cases} -0.03956718 * t^2 + 0.2942684 * t + 0.343545, & \text{for Group1}, \\ -0.026718 * t^2 + 0.142684 * t + 0.603545, & \text{for Group2}, \end{cases}$$

as depicted in Figure 3.2, where the two subgroups have the probabilities of $(0.4, 0.6)$. For this simulation, instead of varying the measuring frequency and the censoring distribution, we vary the random slope variance ranging from 0 to 0.9.

The result from Figure 3.7 illustrates that the mean estimated effect of the longitudinal covariate fluctuates according to the underlying random slope variance. It is observed that as the random slope variance increases, the estimated mean longitudinal covariate effect converges towards log(0.3), suggesting that more weight is attributed to the covariate effect value in the later follow-up time.

To mitigate the dependency on conditional covariance variance and restore interpretability as approximately average time-weighted covariate effect, variance reweighting and scaling methods were explored as potential solutions. In case of previous simulation, with the random intercept of variance v_{inter}, random slope of variance v_{slope} and measurement error of variance v_{error}, the conditional variance can be expressed as $v(t) = v_{error} + v_{inter} + v_{slope} * t^2$, which increase quadratically with respect to time, rendering the average time-weighted covariate effect with significant more focus in later follow-up times. By utilizing an inverse of a consistent estimator of conditional covariate variance, we can reweight the estimand, effectively removing the influence of $v(t)$ in the interpretation.

Proposition 4. *Define estimator $\hat{\beta}_R$ as the solution to the estimating equation*

$$\sum_{i=1}^{n} \int_{t=0}^{\infty} \hat{V}_i(t)^{-1} W_i^b(t) \left[Z_i(t) - \frac{\sum_{j=1}^{n} W_j^b(t) Y_j(t) Z_j(t) exp(Z_j(t)^T \beta)}{\sum_{j=1}^{n} W_j^b(t) Y_j(t) exp(Z_j(t)^T \beta)} \right] dN_i(t) = 0, \quad (3.6)$$

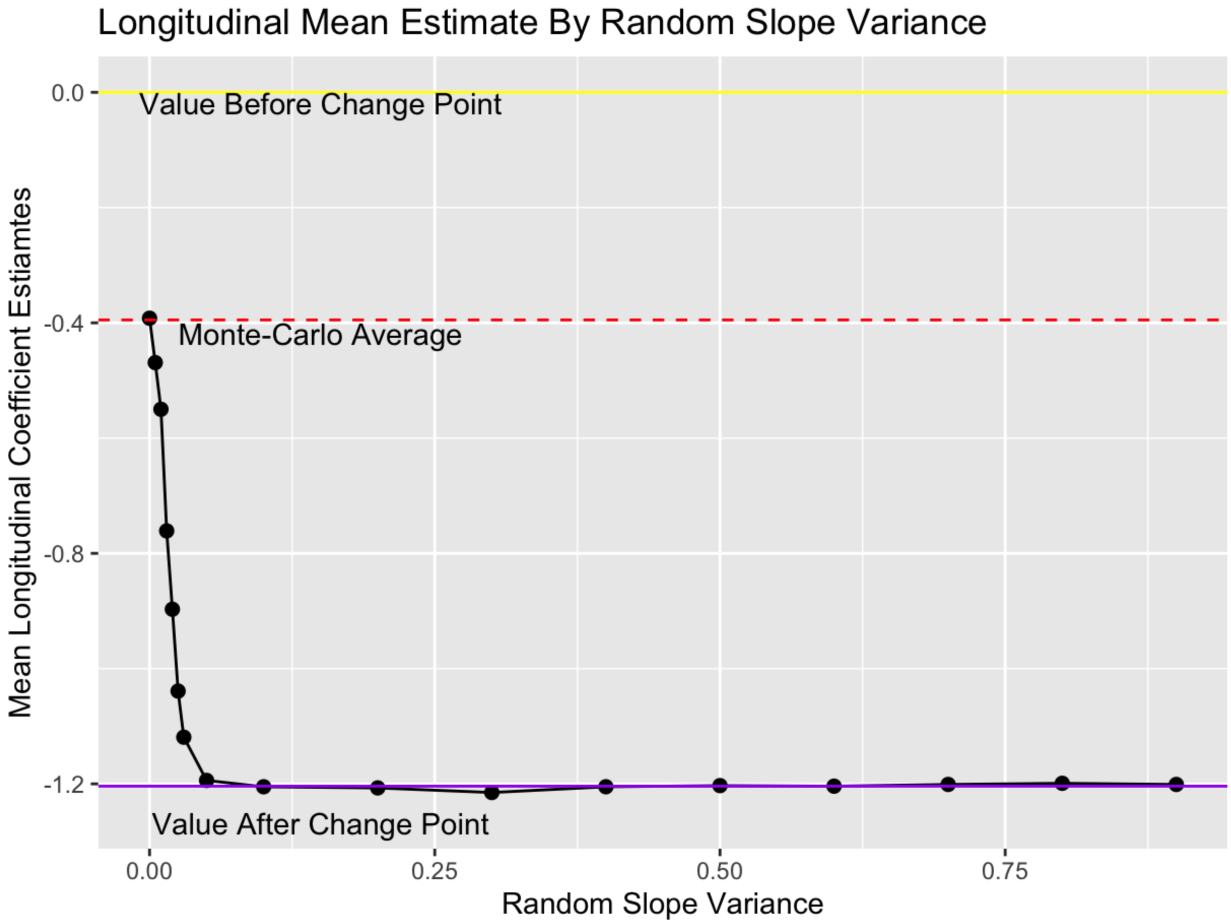


Figure 3.7: The mean treatment effect using regular censoring robust estimator with longitudinal prediction by varying random slope variance

where $\hat{V}(t)$ is a consistent estimator of conditional covariate variance $v(t)$. Under model 3.1 and conditional independent censorship, the estimator $\hat{\beta}_R$ converges in probability to a constant β_0, where β_0 is the unique solution to the following equation

$$\int_0^\infty v(t)^{-1} \left\{ \frac{s_W^{(1)}(\beta(t),t)}{s_W^{(0)}(\beta(t),t)} - \frac{s_W^{(1)}(\beta,t)}{s_W^{(0)}(\beta,t)} \right\} dF(t) = 0. \tag{3.7}$$

Applying the first-order Taylor series approximation gives $\int_0^\infty (\beta(t) - \beta) dF(t) \approx 0$.

Proof. Under similar conditions, since $\hat{V}(t)$ is a consistent estimator of conditional covaraite variance $v(t)$, applying the continuous mapping gives that $\hat{V}(t)^{-1}$ is a consistent estimator of conditional covariate variance $v(t)^{-1}$. Following the proof in Ringhui (1996), the $\hat{\beta}_R$ is consistent to β_0. When applying the first-order Taylor series approximation of $g(x) = s_W^{(1)}(x,t)/s_W^{(0)}(x,t)$, the approximation is given as

$$\int_0^\infty v(t)^{-1}v(t)(\beta(t)-\beta)dF(t) = \int_0^\infty (\beta(t)-\beta)dF(t) \approx 0$$

□

The reweighting estimator $\hat{\beta}_R$ is censoring robust with interpretation as average covaraite effect. It is, however, evident that the weight W_i^b is assigned exclusively to time-varying covariates, while the weight is set to 1 for all time-invariant covariates. The discrepancy in the multiplier results in an asymmetric Hessian matrix, thereby disrupting the fundamental likelihood framework that facilitates variance estimation via the likelihood information theorem. It should be noted that the positive definiteness of the Hessian matrix is maintained, as the Hessian matrix without multipliers is positive definite. Consequently, the gradient descent method is anticipated to converge towards a local minimum. Still, we prefer the scaling method due to the preservation of the likelihood framework for variance estimation.

Before moving on to the scaling method, we want to first demonstrate that the proposed variance reweighting estimator, β_R actually removes the dependency on longitudinal conditional variance through a simulation study. The simulation setting is the same as in the previous simulation study with covariates generated as depicted in Equation 3.5. For this simulation, the random slope variance and censoring distribution will be varied.

The results from the naive Cox model, the censoring robust reweight method, and the censoring robust reweight method adjusted for conditional variance are shown in Figure 3.8 as contour plots. The color in the graph indicates the mean estimates from three estimators. In the left panel, the results from the Cox model show sloped steps, indicating a dependency

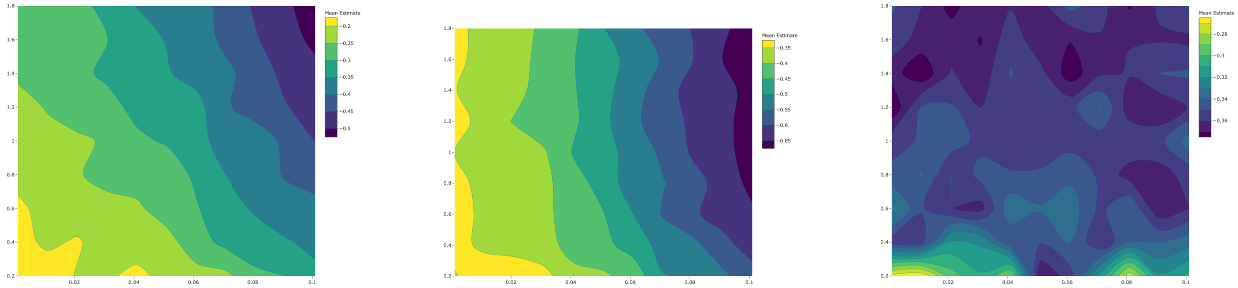


Figure 3.8: Contour plots comparing naive Cox model, censoring robust reweight method, and censoring robust reweight method adjusted for conditional variance

on both the censoring distribution and the random slope variance. The regular reweight method results, which can be seen in the middle panel, are shown as vertical strips. These show that the method is robust to variation in censoring distribution, but it still depends on conditional covariate variance. In the right panel, the results from our proposed method exhibit no obvious trend upon changes in the censoring distribution or the random slope variance.

The scaling method reinstates the interpretation of the average covariate effect by multiplying the longitudinal covariate by the inverse of a consistent estimate of conditional covariate variance. By scaling the longitudinal covariate with conditional covariate variance, the partial likelihood is equivalent to the previous censoring robust estimator using scaled longitudinal covarite, which maintains the likelihood framework. While both the scaling method and the reweighting method adjust the estimating equations based on the weight assigned to the covariate at risk, they solely differ in terms of this weighting adjustment. Both methods incorporate a scaling factor of $V_i(t)^{-1}$ to mitigate the impact of conditional covariate variance at the front of the integrand. However, in the scaling procedure, the weights in $S^{(r)}(\beta, t)$ are further adjusted by the conditional covariate variance at time t by rescaling the predicted covariate value with estimates of conditional covariate variance. The estimating equation for scaling method is

$$\sum_{i=1}^{n}\int_{t=0}^{\infty} W_i^b(t)\Bigg[Z_i(t)\hat{V}_i(t)^{-1} - \frac{\sum_{j\in R_{(i)}} W_j^b(t)Y_j(t)Z_j(t)\hat{V}_i(t)^{-1}\exp(Z_j(t)^T\hat{V}_i(t)^{-1}\beta)}{\sum_{j\in R_{(i)}} W_j^b(t)Y_j(t)\exp(Z_j(t)^T\hat{V}_i(t)^{-1}\beta)}\Bigg]\mathrm{d}N_i(t) = 0. \tag{3.8}$$

In comparison to the reweighting method, the only distinction in the estimating equation lies in the weight assigned to the covariate value in the risk set. The weight for $Z_j(t)$ in the risk set changed from $\frac{W_j^b(t)Y_j(t)\exp(Z_j(t)\beta)}{\sum_{k=1}^n W_k^b(t)Y_k(t)\exp(Z_k(t)\beta)}$ to $\frac{W_j^b(t)Y_j(t)\exp(Z_j(t)\hat{V}_j(t)^{-1}\beta)}{\sum_{k=1}^n W_k^b(t)Y_k(t)\exp(Z_k(t)\hat{V}_j(t)^{-1}\beta)}$. Thus, at each failure time, the replacement of Z_j with $\hat{V}_j(t)^{-1}Z_j$ results in standardizing the covariate effect contribution, so that the estimated coefficient contribution at time t with original longitudinal covariate is scaled to $\beta\hat{V}_j(t)^{-1}$. Furthermore, since the weighting scheme employed in Equation 3.8 assigns weights to each integrand by $W_i^b(t)\hat{V}_j(t)^{-1}$. Consequently, the results obtained from the scaling method expected to be a weighted average of $\beta V_i(t)$. To obtain an estimator that is consistent with the average time-weighted effect as in Equation 3.6, one can multiply the estimates from Equation 3.8 by $\sum_i W_i^b(t)\hat{V}_j(t)^{-1}/\sum_i W_i^b(t)$. Currently, we possess an estimator that exhibits robustness against censoring and allows for the interpretation of average regression effects for the single longitudinal group scenario.

When confronted with multiple longitudinal groups, it becomes necessary to stratify the data according to these groups and then combine the estimating equations to determine the relationship between estimates from the scaling method and the reweighting method. Recalling that all previous estimators for survival data compare the observed covariate with the expected covariate at the event time, when multiple longitudinal groups are present, the subjects in the risk set are likely to consist of subjects from different longitudinal groups with different longitudinal trajectories. Comparing the expected covariate from a mixture of groups with the observed covariate value from a specific group would introduce bias in the contribution to the covariate effect at each observed event. Even if the underlying longitudinal trajectory is similar, at each event time, the first-order approximation of the quantity that the estimator is consistent to rely on consistent estimation of the conditional covariate variance of Z_j, and a mixture of covariates from different groups would compromise the interpretation. Through simulation, the outcome exhibited significant bias and instability as a consequence of mismatched conditional covariate variance. In our simulation, when using all subjects in the risk set, we experimented with reweighting conditional covariate variance by longitudinal group and by the entire population. Both methods exhibited bias in the mean estimated covariate coefficients. Specifically, the biases were found to be -0.15 (equivalent to a 38% bias) for variance calculation by group and 0.25 (equivalent to a 64% bias) for variance calculation by entire population, respectively. It is important to note that the true value of the main estimated covariate coefficient is -0.39, far from the estimated mean covariate effect. To address this issue, we suggest implementing a data stratification approach based on the longitudinal group so that the risk set is modified to be the subjects not experiencing an event or censoring and is within the same longitudinal group as the subject experiencing the event at the event time. Since the covariate effect at each time would be the same across the stratum, we propose to combine the estimating equation of strata in the usual fashion. This method implicitly recognizes the potential variation in the coefficient across different strata. So far, we have obtained a two-stage censoring estimator for longitudinal covariates with the interpretation of a weighted time-average covariate effect, as shown in the following proposition.

Proposition 5. *Define the estimator for longitudinal covariate $\hat{\beta}_S$ as the solution to the estimating equation*

$$\sum_{i=1}^{n} \int_{t=0}^{\infty} W_i^b(t) \Bigg[Z_i(t)\hat{V}_i(t)^{-1} - \frac{\sum_{j=1}^{n} I(G_i = G_j) W_j^b(t) Y_j(t) Z_j(t) \hat{V}_i(t)^{-1} exp\{Z_j(t)^T \hat{V}_i(t)^{-1}\beta\}}{\sum_{j=1}^{n} I(G_i = G_j) W_j^b(t) Y_j(t) exp\{Z_j(t)^T \hat{V}_i(t)^{-1}\beta)\}} \Bigg] dN_i(t) = 0, \quad (3.9)$$

where $\hat{V}_i(t)$ is a consistent estimator of conditional covariate effect $v_i(t)$ and G_i represents the longitudinal group that the ith subject belongs to. Under Model 3.1 and conditional independent censorship, the estimator $\hat{\beta}_S$ converges in probability to a constant β_0, where β_0 is the unique solution to the following equation

$$\int_0^\infty v_i(t)^{-1}\left\{\frac{s_W^{(1)}(\beta(t),t,Gi)}{s_W^{(0)}(\beta(t),t,Gi)}-\frac{s_W^{(1)}(v_i^{-1}(t)\beta_0,t,Gi)}{s_W^{(0)}(v_i^{-1}(t)\beta_0,t,Gi)}\right\}dF(t)=0, \tag{3.10}$$

and β_0 can be approximated by $\int_0^\infty[\beta(t)-v_i^{-1}(t)\beta_0]dF(t)\approx 0$.

Proof. Define

$$S_W^{(r)}(\beta,t,G_i)=n^{-1}\sum_{j=1}^{n}\mathrm{I}(G_i=G_j)Z_j^r(t)W_j^b(t)Y_i(t)\mathrm{exp}(Z_j(t)\beta)$$

and

$$s_W^{(r)}(\beta,t,G_i)=E[S_W^{(r)}(\beta,t,G_i)],$$

where the $S_W^{(r)}(\beta,t,G_i)$ and $s_W^{(r)}(\beta,t,G_i)$ are defined for the data strata that the ith subject belongs to for stratification. Since the modification is based on proper longitudinal clustering, the conditions 1-6 still hold. Note that the above estimating equation could also be written as

$$\begin{aligned}\sum_{i=1}^{n}\int_{t=0}^{\infty}W_i^b(t)\hat{V}_i(t)^{-1}\Bigg[&Z_i(t)\\ &-\frac{\sum_{j=1}^n\mathrm{I}(G_i=G_j)W_j^b(t)Y_j(t)Z_j(t)\mathrm{exp}\{Z_j(t)^T\hat{V}_i(t)^{-1}\beta\}}{\sum_{j=1}^n\mathrm{I}(G_i=G_j)W_j^b(t)Y_j(t)\mathrm{exp}\{Z_j(t)^T\hat{V}_i(t)^{-1}\beta)\}}\Bigg]\mathrm{d}N_i(t)\\ &=\sum_{i=1}^{n}\int_{t=0}^{\infty}W_i^b(t)\hat{V}_i(t)^{-1}\left[Z_i(t)-\frac{S_W^{(r)}(\hat{V}_i(t)^{-1}\beta,t,G_i)}{S_W^{(r)}(\hat{V}_i(t)^{-1}\beta,t,G_i)}\right]\mathrm{d}N_i(t)=0.\end{aligned}$$

Applying similar techniques as in Proposition 2, it can be shown that $\hat{\beta}_S$ converges to β_0. The result from Proposition 1 still applies to $s_W^{(r)}(\beta(t),t)/s_W^{(0)}(\beta(t),t)$, which consistently estimates $E(Z^r|T=t)$. It follow that the first-order Taylor series expansion of $g(v_i(t)^{-1}\beta) = \frac{s_W^{(1)}(v_i(t)^{-1}\beta,t,G_i)}{s_W^{(0)}(v_i(t)^{-1}\beta,t,G_i)}$ around β_0 gives

$$g(v_i(t)^{-1}\beta) \approx \frac{s_{WV}^{(1)}(\beta(t),t,Gi)}{s_{WV}^{(0)}(\beta(t),t,Gi)} + v_i(t)[\beta(t) - v_i(t)^{-1}\beta_0].$$

Applying the approximation to integrand of Equation 3.10 gives

$$\int_0^\infty v_i(t)^{-1}\left\{\frac{s_W^{(1)}(\beta(t),t,Gi)}{s_W^{(0)}(\beta(t),t,Gi)} - \frac{s_W^{(1)}(v_i^{-1}(t)\beta_0,t,Gi)}{s_W^{(0)}(v_i^{-1}(t)\beta_0,t,Gi)}\right\}dF(t)$$
$$= \int_0^\infty v_i(t)^{-1}\left\{\frac{s_W^{(1)}(\beta(t),t,Gi)}{s_W^{(0)}(\beta(t),t,Gi)} - \frac{s_W^{(1)}(\beta(t),t,Gi)}{s_W^{(0)}(\beta(t),t,Gi)} - v_i(t)[\beta(t) - v_i(t)^{-1}\beta_0]\right\}dF(t)$$
$$= \int_0^\infty [\beta(t) - v_i(t)^{-1}\beta_0]dF(t) = 0$$

□

After rearranging terms, the approximation in Proposition 5 indicates $\int_0^\infty \beta(t)dF(t) = \beta_0\int_0^\infty v_i(t)^{-1}dF(t)$. Thus, to retrieve an estimator approximated by the weighted time-average effect, we need to multiply the $\hat{\beta}_S$ by estimates of $\int_0^\infty v_i(t)^{-1}dF(t)$. To adjust for censoring, the resulting estimate is $\sum W_i^b(t)\hat{V}_i^{-1}(t)/\sum W_i^b(t)$.

Proposition 6. *For the modified partial likelihood model in Proposition 5, assume that the Anderson and Gill regularity conditions hold. Under Model 3.1 and conditional independent censorship,* $n^{1/2}(\hat{\beta}_S - \beta_0)$ *has an asymptotic normal distribution with mean zero.*

Proof. The consistency of $\hat{\beta}_S$ to β_0 is established in Proposition 5; it remains to show the asymptotic normality of $\hat{\beta}_S$. Define

$$S_W^{(r)}(\beta,t,G_i,Z^*) = n^{-1}\sum_{j=1}^n \mathrm{I}(G_i = G_j)[Z_j^*(t)]^r W_j^b(t)Y_i(t)\exp(Z_j^*(t)\beta)$$

and

$$s_W^{(r)}(\beta, t, G_i, Z^*) = E[S_W^{(r)}(\beta, t, G_i, Z^*)],$$

where

$$Z_i^*(t) = Z_i(t)/\hat{V}(t).$$

It can be seen that $\hat{\beta}_S$ is exactly the same as the estimator in Boyd et al. (2012) using scaled longitudinal covariate values. The conditions 1-6 are also the assumptions for Theorem 5.3 of Kalbfleisch and Prentice (2011), which implies Rebolledo's martingale central limit theorem. Using a similar argument to that given in Boyd et al. (2012), it can be shown that the estimating equation in Equation 3.9 can be written as a sum over stochastic integrals of a predictable process with respect to a martingale, giving the asymptotic normality of $\hat{\beta}_S$. □

3.2.2 Estimating the conditional variance of a longitudinally sampled covariate

Returning to the first stage of the two-stage model, since longitudinal covariate values at each event time are typically missing, we propose to predict covariate values using the LME model and estimate conditional covariate variance through empirical variance. For our study, the prediction of $Z_i(t)$ at event time is not of direct scientific interest, but it is necessary to reduce potential bias due to incorrect hazards. Our goal for this prediction problem is to develop a prediction procedure that retrieves a longitudinal covariate value reasonably close to the unknown truth at the event time. As mentioned, two approaches toward longitudinal data modeling are the LME model and the GEE model. For our specific purpose of predicting the longitudinal value at the individual level with the imbalanced data, the GEE model may not be appropriate. Even though this semi-parametric model has the advantage of a robust variance estimator using the sandwich estimator, the GEE model is based on the marginal model, where only population-level predictions are available. The LME model is preferred since it also captures individual characteristics through random effects, and those random effects can be estimated through the Naive Bayes method.

To estimate the conditional longitudinal variance, we considered two empirical methods: (i) empirical variance at a small window using the observed covariate value; and (ii) empirical variance using the predicted covariate value. For the first method, we propose to estimate the conditional covariate variance at time t using the available observed covariate value within a small window around t for those in the risk set. The first method is expected to be robust to model misspecification since it does not rely on predicted longitudinal covariate values. However, it suffers from the limited amount of data available. In addition, borrowing information from covariates at nearby times relies on the strong assumption that the covariate variance remains similar. In data simulations, this method shows sensitivity to window size selection and abrupt changes in conditional variance. The second method calculates empirical variance using the predicted covariate values of all subjects in the risk set. Even though it is subject to model misspecification, the second method is preferred in data simulation for its stability from the tenfold of samples compared to the numbers of the first method. In data simulations, the conditional longitudinal variance from the first method deviates from the true longitudinal variance, and the results are biased. We also conducted a simulation with a true longitudinal value at event time for each subject in comparison to predict the longitudinal value using the Naive-Bayes method under the LME model. We found that when the conditional variance is estimated by empirical methods, as mentioned above, the mean estimated variance centers closely around the Monte Carlo average. In contrast, the mean estimated variance using the unobserved true longitudinal value exhibited consistent bias based on the underlying longitudinal design.

3.2.3 An algorithm for computing censoring robust estimators with longitudinally sampled covariates

Given the above, we propose an algorithm for censoring robust estimation that incorporates longitudinal covariates in Algorithm 1.

Algorithm 1 Censoring Robust Algorithm For Estimating Longitudinal Covariate Average Effect

```
Specify the prior scientific model:
    $\lambda(t|Z) = \lambda_0(t)\exp(\boldsymbol{Z}^T\beta)$.
Identify Longitudinal Clusters:
    $i : i = 1, \ldots, n \rightarrow$ LongGroup j : $j = 1, \ldots, m$ .    ▷ $m < n$
Identify censoring-specific groups using survival trees with longitudinal covariates replaced by a dummy variable indicating group assignment:
    $M_C(Z) \rightarrow \{1, \ldots, k\}$.    ▷ $k < n$
Estimate $S_C(\cdot|M_C(Z))$ using the left-continuous Kaplan-Meier estimator for each group $\{1, \ldots, k\}$:
    $\hat{S}_C(\cdot|M_C(Z))$.
Estimate $V(t_{(i)}|M_C(Z))$ using individual-level prediction from linear mixed effect model built from only subject in the longitudinal group where the individual experienced an event at $t_{(j)}$ belongs to:
    $\hat{V}(t_{(i)}|M_C(Z))$.
Plug in the inverse probability of censoring, inverse conditional variance, and redefined risk set consisting of people who have not experienced the event nor truncation from the same longitudinal group as the one who experienced the event to form a weighted estimating equation:
    $$U_W(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \hat{W}(t|Z_i)\hat{V}_i(t)^{-1} \Bigg[ Z_i(t) - \frac{\sum_{j=1}^{n} \mathrm{I}(G_i = G_j)\hat{W}(t|Z_j)Y_j(t)Z_j(t)\exp\{Z_j(t)^T\hat{V}_i(t)^{-1}\beta\}}{\sum_{j=1}^{n} \mathrm{I}(G_i = G_j)\hat{W}(t|Z_j)Y_j(t)\exp\{Z_j(t)^T\hat{V}_i(t)^{-1}\beta)\}} \Bigg] \mathrm{d}N_i(t),$$
    where $\hat{W}(t|Z_i) = 1/\hat{S}_C(t|M_C(Z_i))$ and $G_i(t)$ defines longitudinal group that subject $i$ belongs to.
Solve $U_W(\beta) = 0$ and rescale the longitudinal estimator by $\sum \hat{V}(t|Z_i)^{-1}\hat{W}(t|Z_i)/\sum \hat{W}(t|Z_i)$ to obtain the censoring-robust estimator, $\hat{\beta}_S$.
```

In the context of inference, directly applying the robust variance estimator, as discussed in the previous section, may underestimate the potential variance resulting from variation in longitudinal prediction, even if the longitudinal model is correctly specified. Moreover, when dealing with multiple longitudinal groups, misclassification of longitudinal covariates can further exacerbate the variation in the resulting estimates. Since there is no assurance of proper clustering, a prudent approach would be to employ the widely-used nonparametric robust variance estimator, namely the bootstrap method (Tibshirani and Efron, 1993). While some literature, such as Hsieh et al. (2006), recommend a parametric bootstrap two-stage joint modeling approach under the proportional hazard model, which involves simulating a large number of copies of data B by fitting the longitudinal component using the LME model and subsequently proceeding with Cox model estimation. The parametric bootstrap method may offer efficiency but could suffer from potential misspecification of the longitudinal model. Therefore, we propose utilizing a case resampling bootstrap for inference. In this approach, each copy comprises random draws from the original data with replacement, maintaining the same sample size. By employing case resampling bootstrap, we can effectively incorporate variance resulting from LME model misspecification and longitudinal clustering errors into our analysis.

3.3 Numerical studies

3.3.1 Simulation setup

In this section, we compare the performance of three estimators: (i) the naive Cox estimator β_{Cox} proposed by Cox (1972); (ii) the censoring robust estimator β_{CR} as introduced by Boyd et al. (2012) and Nguyen et al. (2017); and (iii) the proposed revised censoring robust estimator β_{S}. Since all three estimators approximate the hazard functions by substituting unknown longitudinal covariates at event times with predicted values, they assume the same linear mixed-effects (LME) model in the first stage. Therefore, we will not delve further into the first modeling stage, as the focus of this study is on the second modeling stage, specifically the development of an estimator with censoring robust properties under misspecified hazards and an average covariate effect for time-varying covariates, even under time-varying conditional variance.

To distinguish the covariate interpretations between our proposed estimator and the other two estimators under various scenarios, three specific scenarios are designed to exemplify the boundaries that previous estimators could handle. We are particularly interested in quantifying the association between the time to the event and the longitudinal covariate $Z_1 \in \mathbb{R}$, with β_1 being the corresponding parameter of interest. To assess the performance of the estimators, a confounding covariate $Z_2 \sim \text{Bernoulli}(0.4)$ is included to simulate a more general regression situation.

The simulation studies presented here consider a combination of one categorical and one longitudinal covariate, although the algorithm has the capability to handle combinations of categorical, continuous, and longitudinal covariates. Incorporating multiple covariates increases computational complexity significantly. Since the proposed estimator involves a survival tree algorithm to identify approximate censoring groups by optimal splitting, adding another continuous covariate can lead to a computational complexity increase of $\mathcal{O}(n)$. Considering these computational constraints, we focus solely on longitudinal covariates in this study.

For comparison purposes, we evaluate the estimators based on bias and coverage probability of 95% confidence intervals under each scenario. We also vary the censoring distributions to assess dependence on the censoring mechanism.

Due to the specific interest, we focus on longitudinal covariates in this study. We assess the performance of the estimators based on bias and coverage probability of 95% confidence intervals under each scenario. Additionally, we vary the censoring distributions to assess dependence on the censoring mechanism.

Following the brief overview, we introduce the logistic for the simulations. We assume there are two longitudinal subgroups with probabilities of $(0.4, 0.6)$, and these subgroups follow underlying longitudinal trajectories where the group with a higher starting biomarker value generally experiences slow increments and vice versa, as depicted in Figure 3.2. Both covariates are generated as in Equation 3.5.

It's worth noting that Figure 3.2 only displays the underlying trajectory up to time four since it's common practice to limit the length of the study, also known as the maximum follow-up length. Such a predetermined maximum study length is a form of administrative censoring or truncation. In the following simulations, we set $\tau = 4$. The three scenarios we consider correspond to the proportional hazard model, the non-proportional hazard model with a constant marginal variance longitudinal covariate, and the non-proportional hazard model where the longitudinal covariate varies with respect to time. The first scenario is the proportional hazard model, where we generate survival time according to

$$\lambda(t|Z) = \exp\{-0.8 \times Z_1(t) - 0.4 \times Z_2\}.$$

For this scenario, our intention is to compare the performances of the naive Cox estimator β_{Cox} and that of the proposed revised censoring robust estimator β_S. Since the naive Cox estimator β_{Cox} is censoring robust with correct interpretation under the proportional hazard model, the above comparison aims to establish the effectiveness of the proposed revised censoring robust estimator β_{CRV} when the model is not misspecified. The comparison of performance between the censoring robust estimator β_{CR} and the naive Cox estimator β_{Cox} was discussed in Nguyen and Gillen (2017) and thus is not of primary interest in this scenario.

The second scenario is the non-proportional hazard model with constant marginal variance longitudinal covariate, where we generate the survival time according to

$$\lambda(t|Z) = \exp\{\log(0.3)I\{t > 1\} \times Z_1(t) - 0.1 \times Z_2\}.$$

In this scenario, we compared the performance of the proposed revised censoring robust estimator β_S to that of the censoring robust estimator β_{CR} and the naive Cox estimator β_{Cox}. Ideally, we expect to observe similar results as shown in Nguyen and Gillen (2017), where the naive Cox estimator β_{Cox} estimates a quantity different from those estimated by the censoring robust estimator β_{CR} and the proposed revised censoring robust estimator β_S. Additionally, we anticipate that the performance of the censoring robust estimator β_{CR} and that of the proposed revised censoring robust estimator β_S will be similar due to the longitudinal covariate with constant marginal variance. However, the interpretation of the covariate is modified by the conditional covariate variance instead of the underlying covariate variance. Thus, the unknown censoring mechanism can affect the distribution of observed covariate values conditional upon time, leading to variation in conditional variance even if the underlying longitudinal variance is constant over time. Nevertheless, this second scenario is designed to contrast the differences between estimated quantities comparing β_S to β_{CR} and β_S to β_{Cox}.

The third scenario involves a non-proportional hazard model with longitudinal covariate variance varying over time. We generate the survival time according to the same hazard as the second scenario, with the difference being the addition of a random slope to the longitudinal covariate. In this scenario, we compare the performance of the proposed revised censoring robust estimator β_S to that of the censoring robust estimator β_{CR}. With the addition of the random slope, the variance of the longitudinal covariate increases quadratically over time, placing more emphasis on the later follow-up time covariate effect. As we formally introduce variation in conditional covariate variance, we expect that the difference between quantities estimated by the two estimators will increase significantly compared to the previous scenario.

Case	Censoring Distribution
C:1	Power(4,0.6+0.8*I$\{G_i = 1\}$
C:2	Power(4,0.6+0.8*I$\{Z_2 = 1\}$
C:3	Power(4,0.6+0.8*I$\{G_i$*$Z_2 = 1\}$

Table 3.2: The three scenarios of censoring distributions are considered based on longitudinal grouping, categorical grouping, and their interactions.

For each hazard scenario, three censoring scenarios are utilized to generate censoring times, aiming to test the effect of censoring distribution on estimators. Table 3.2 provides a description of the censoring scenarios: censoring by categorical covariate (Case 1), censoring by longitudinal covariate (Case 2), and censoring by the interaction between covariates (Case 3). In all cases, administrative censoring at time $\tau = 4$ is required to maintain a constant observable time support and ensure meaningful comparison of the average covariate effect between scenarios. For all cases, the censoring times by group are generated according to the power-function distribution with parameters (b, r), where $0 < C < b$ and $r \geq 0$. When $r = 1$, C follows a uniform distribution over $(0, b)$. When $r \to 0$, more weight is concentrated toward 0; when $r \to \infty$, more probability is concentrated toward b. We repeated simulations on each combination of the data-generating scenario and censoring scenario 1000 times for replicable results.

Next, we will provide further details about the simulation setting. For non-proportional hazards scenarios, due to a lack of an analytic expression for $\int_0^\infty \beta_1(t)dF(t)$, we approximate it with the Monte Carlo average of β_{Cox} (with administrative censoring at $\tau = 4$) for $n = 2000$. To clarify, we use the average of the maximum partial likelihood estimator for large samples without intermittent censoring as the target quantity in each scenario.

For the longitudinal covariate, we incorporate variance through the mixed effect model. In the first two scenarios, we assume a random intercept following $\mathbf{N}(0, 0.08)$ so that the conditional variance of the longitudinal covariate remains constant without censoring. Moving on to the last scenario, we add an random slope following $\mathbf{N}(0, 0.08)$ and assume independence between the random intercept and random slope for simplicity. Even if the random intercept and the random slope follow the same distribution, the contribution of conditional variance from the random slope outweighs that from the random intercept. The conditional covariate variance is proportional with a multiplier of $0.08t^2$ due to the random slope, so the conditional variance increases quadratically with respect to time, whereas the conditional covariate variance only increases by a constant of 0.08 due to the random intercept. Lastly, we assume identically and independently distributed measurement error following $\mathbf{N}(0, 0.02)$ throughout the study.

Even with the model explicitly written out, mentioning the data generation mechanism is worthwhile. When generating time-to-event data according to specified hazards that depend on a longitudinal covariate, we cannot simply utilize existing functions to make random draws from the exponential distribution as the hazard changes constantly. Nevertheless, it is also nearly impossible to draw directly from the target distribution due to the lack of an analytic expression. Thus, the closest solution involves using a grid approximation of the target hazard.

For grid approximation, we partitioned the entire follow-up length into intervals of a grid size. We assumed that the hazard within each interval remains constant, so a piecewise exponential distribution can approximate the survival distribution. There are two methods to draw from piecewise exponential distributions, where one may derive the cumulative density function of the target distribution and make draws using the probability integration theorem. The other approach utilizes the memoryless property of the exponential distribution, where $P(T_i > t + s | T_i > t) = P(T_i > s) \sim \text{Exponential}(\lambda_i(t))$. By discretizing continuous time into small intervals of 0.01 years, we can easily calculate the survival probability by $P(T_i > t) = \prod_{t_i}^{t-0.01} P(T_i > t_i + 0.01|T_i > t_i)$. To generate the data, we draw randomly from the exponential distribution according to the grid size and the corresponding hazard at the start time point of each interval. Then, the survival time is obtained through the first random draw that does not exceed the grid length. Furthermore, note that the longitudinal covariate values are generated before survival time generation so that the simulated covariate is external and allows for the prediction of survival probability.

In this method, the grid size controls the performance of the approximation, and a smaller grid size gives a better approximation but requires more computational power. In our case, the grid size was chosen to be 0.01 depending on the available computational power and used for the entire simulations for consistency.

For calculating coverage probability using 95% confidence intervals, in addition to the estimated mean, the variance of $\hat{\beta}_{\text{S}}$ is estimated through the bootstrap method for reasons discussed in the Discussion section. When applying the bootstrap method for variance estimation, it is inevitable to encounter ties in the data. Common approaches for handling tied survival times include the average partial likelihood, Breslow approximation, and Efron's formula.

The average partial likelihood naturally arises under the assumption of true continuous survival time and treats all possible combinations equally. This method provides the most precise likelihood representation but involves computationally intensive bookkeeping. The Breslow approximation aims to reduce computational burden by approximating the denominator of each likelihood by the complete case, but it introduces bias as the denominator is always larger than the truth. To mitigate such bias, Efron's formula deducts the average contribution to the denominator by a single subject from the denominator.

In practice, when the number of ties is manageable, the average partial likelihood method is preferred for its accuracy. However, when ties are common, Efron's formula is preferred. In our case, we choose to use the average partial likelihood approach, even though a moderate number of ties exists. The bookkeeping process can be avoided as the tied survival times are associated with subjects of the same covariate value at any time, i.e., the bootstrap data.

The variances of the other two estimators are calculated through the robust variance formula in original paper mentioned in Chapter 2 (Lin and Wei, 1989; Nguyen and Gillen, 2017). While using the bootstrap method to estimate the variance of the other two estimators is also possible, we would like to assess the performance based on the original estimators without modification.

3.3.2 Simulation results

Table 3.3: Results of the simulation study with random intercepts and random slopes under non-proportional hazards, n=800. Each scenario is repeated 1000 times. The Truth refers to the Monte Carlo average of $\hat{\beta}_{\text{Cox}}$ under simulation with administrative censoring at time year 4 and subject in each trial increased to $n = 2000$.

	Naive Cox Estimator				Robust Est. - Survival Trees				Robust Est. - Variance Adj.			
Scenario	$\hat{\beta}_{\text{Cox}}$				$\hat{\beta}_{\text{CR}}$				$\hat{\beta}_{\text{S}}$			
(Truth:-0.395)	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP
C:1	-0.216	0.351	0.347	0.864	-0.592	0.388	0.384	0.849	-0.404	0.537	0.553	0.938
C:2	-0.218	0.333	0.344	0.856	-0.362	0.381	0.379	0.902	-0.385	0.556	0.586	0.931
C:3	-0.195	0.340	0.337	-0.819	-0.407	0.385	0.372	0.890	-0.390	0.516	0.520	0.919

ESE = empirical standard error of the coefficient of interest; MSE = model-based standard error of the coefficient of interest; CP = coverage probability of 95% confidence intervals based on the MSE.

The remaining parts of this subsection focus on comparing the performances of the three estimators. We assess their performance based on desired properties such as unbiasedness, consistency, and type-I error across three types of hazard functions, as per the simulation design.

When the underlying model is misspecified and there are fluctuations in conditional covariate variance, we observe higher bias in $\hat{\beta}_{\text{Cox}}$ and $\hat{\beta}_{\text{CR}}$ compared to $\hat{\beta}_{\text{S}}$, as expected. As demon-

Table 3.4: Results of the simulation study with random intercepts only under non-proportional hazards, n=800. Each scenario is repeated 1000 times. The Truth refers to the Monte Carlo average of $\hat{\beta}_{\text{Cox}}$ under simulation with administrative censoring at time year 4 and subject in each trial increased to $n = 2000$.

	Naive Cox Estimator				Robust Est. - Survival Trees				Robust Est. - Variance Adj.			
Scenario	$\hat{\beta}_{\text{Cox}}$				$\hat{\beta}_{\text{CR}}$				$\hat{\beta}_{\text{S}}$			
(Truth:-0.381)	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP
C:1	-0.116	0.348	0.350	0.864	-0.186	0.380	0.376	0.846	-0.387	0.801	0.789	0.945
C:2	-0.130	0.350	0.348	0.856	-0.501	0.380	0.378	0.887	-0.377	0.709	0.689	0.934
C:3	-0.130	0.339	0.342	0.805	-0.287	0.368	0.367	0.885	-0.365	0.625	0.642	0.958

ESE = empirical standard error of the coefficient of interest; MSE = model-based standard error of the coefficient of interest; CP = coverage probability of 95% confidence intervals based on the MSE.

Table 3.5: Results of the simulation study with random intercepts and random slopes under proportional hazards, n=800. Each scenario is repeated 1000 times. The Truth refers to the Monte Carlo average of $\hat{\beta}_{\text{Cox}}$ under simulation with administrative censoring at time year 4 and subject in each trial increased to $n = 2000$.

	Naive Cox Estimator				Robust Est. - Survival Trees				Robust Est. - Variance Adj.			
Scenario	$\hat{\beta}_{\text{Cox}}$				$\hat{\beta}_{\text{CR}}$				$\hat{\beta}_{\text{S}}$			
(Truth:-0.800)	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP	Mean	ESE	MSE	CP
C:1	-0.825	0.368	0.365	0.942	-0.828	0.456	0.442	0.921	-0.811	0.534	0.596	0.929
C:2	-0.808	0.375	0.370	0.941	-0.802	0.477	0.444	0.935	-0.798	0.544	0.525	0.932
C:3	-0.792	0.362	0.360	0.942	-0.819	0.380	0.378	0.890	-0.791	0.520	0.552	0.951

ESE = empirical standard error of the coefficient of interest; MSE = model-based standard error of the coefficient of interest; CP = coverage probability of 95% confidence intervals based on the MSE.

strated in previous sections, with the presence of random intercept and random slope in the longitudinal covariate, the conditional covariate variance tends to increase as time progresses. Without adjustment, $\beta(t)$ is weighted according to the conditional covariate variance, leading to a higher contribution from $\beta(t)$ in the later follow-up time to the estimated quantity.

As shown in Table 3.3, for all three censoring scenarios, $\hat{\beta}_{\text{S}}$ has mean estimates around the truth, while the estimated means of $\hat{\beta}_{\text{Cox}}$ are biased from the truth by 0.20, almost 50% of the truth. Such a significant deviation from the truth can also be attributed to the failure to adjust for censoring weight. The equal weight assigned by $\hat{\beta}_{\text{Cox}}$ to each event fails to acknowledge the uneven number of unobserved events due to censoring. Even though the censoring mechanism is generally unknown, the censoring weight would only increase with respect to time, with differences existing only in the rate of increment. As a result of the high rate of censoring in the later stages, $\beta(t)$ in the later follow-up time is underestimated, further exaggerating the bias.

This influence is supported by the results of $\hat{\beta}_{\mathrm{CR}}$, where the deviation from the truth decreases. However, the mean estimates are still biased from the truth, indicating that adjusting for the unobserved events due to censoring alone is not sufficient. For the coverage probability (CP), the three estimators perform similarly across the censoring scenarios. While $\hat{\beta}_{\mathrm{S}}$ achieves almost 95% coverage for all three scenarios, $\hat{\beta}_{\mathrm{CR}}$ achieves coverage between 85% and 90%, and $\hat{\beta}_{\mathrm{Cox}}$ only achieves coverage between 82% and 86%.

If we remove the random slope from longitudinal trajectories, we initially expected that $\hat{\beta}_{\mathrm{CR}}$ would perform similarly to $\hat{\beta}_{\mathrm{S}}$, as the longitudinal covariate variance remains marginally constant. However, the result differs from our expectations. While $\hat{\beta}_{\mathrm{S}}$ still achieved nearly unbiasedness and 95% coverage, $\hat{\beta}_{\mathrm{CR}}$ showed better performance but still deviated from the truth and had lower coverage.

The presence of bias and low coverage of $\hat{\beta}_{\mathrm{CR}}$ contradicts our expectation. Further exploration indicates the difference between conditional covariate variance and covariate variance. Even though the marginal covariate variance is constant, the conditional covariate variance still varies due to survival, where $E(Z(t)|T = t)$ is conditional on T, and the expectation builds on the triple (δ, z, x). Even after adjusting for the censoring mechanism, when patients with lower initial longitudinal covariate values experience events earlier, subjects that remain in the study in the later follow-up times have lower variation in biomarker values. Thus, although the difference in variance decreases, fluctuations still exist, and it remains necessary to adjust for conditional covariate variance.

In this scenario, the random effect of data is magnified, as the subtle differences in conditional covariate variance can be quite distinct between samples, compared to the variation brought by the random slope. The mean estimate of $\hat{\beta}_{\mathrm{CR}}$ varies significantly across the scenarios, and the variance of $\hat{\beta}_{\mathrm{S}}$ is also higher compared to that of the other two cases. This incidence indicates the possible dependence between conditional covariate variance and longitudinal variance if not adjusted simultaneously.

When the model is correctly specified, the results in Table 3.5 indicate that all three estimators are unbiased and attain consistency. When the underlying covariate effect is constant, all methods should estimate the correct quantity, as all adjustments are made to reweight the constant coefficient, resulting in the same constant. It is still observable that the variances of $\hat{\beta}_{\mathrm{S}}$ are larger than those of $\hat{\beta}_{\mathrm{CR}}$ in all three cases, and similarly, the variances of $\hat{\beta}_{\mathrm{CR}}$ are larger than those of $\hat{\beta}_{\mathrm{Cox}}$. Such a trade-off between bias and variance is beneficial considering the reduction in bias from the unknown censoring mechanism and the retrieval of the weighted time-average covariate effect interpretation.

Lastly, it is also noted that $\hat{\beta}_{\mathrm{CR}}$ does not achieve a 95% coverage probability in all scenarios. In addition to the limitation of sample size, the failure to attain nominal coverage could also be due to possible variations from censoring group clustering and longitudinal group clustering. From the tables, the coverage probability is still above 90% and is expected to increase with a larger sample size. Additionally, the proposed procedure is flexible to changes in the clustering method. Given a more precise longitudinal clustering method, it can be seamlessly integrated into the existing framework with minimal adjustments.

3.4 Application

In this section, we apply the proposed estimator to ADNI data to explore the relationship between longitudinal cortical thickness and the progression of Alzheimer's disease (AD). The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal study aimed at developing clinical, imaging, genetic, and biochemical markers for the early detection and monitoring of Alzheimer's disease. The dataset used here combines data from the initial five-year study (ADNI-1), the two-year follow-up study (ADNI-GO), and the subsequent renewal study (ADNI-2). More information about the study designs and data collection methods can be found at http://ida.loni.ucla.edu.

The ADNI-1 dataset consists of a significant number of participants and a wide range of clinical and biological data. Initially, it included 1213 subjects scheduled for assessments at the initial visit and subsequent exams every 6 months, with a maximum follow-up duration of 120 months. However, some subjects may have missed appointments or discontinued their participation in the study.

The dataset encompasses demographic information such as age, gender, education level, race, and marital status, recorded at the baseline visit. Additionally, genetic data, specifically the APOE-ϵ4 carrier status, is included.

Clinical variables cover neuropsychological assessments, functional and behavioral evaluations, neuroimaging data from magnetic resonance imaging (MRI), and cerebrospinal fluid (CSF) biomarkers. Neuropsychological assessments consist of various tests including the Alzheimer Disease Assessment Scale-Cognitive (ADAS-Cog), Rey Auditory Verbal Learning Test (RAVLT), Montreal Cognitive Assessment (MoCA), Mini Mental State Examination (MMSE), and Clinical Dementia Rating Sum of Boxes (CDR-SB). Notably, ADAS-Cog 11 and ADAS-Cog 13 are available, with ADAS-Cog 13 assessing additional domains relevant to treatment targets. RAVLT scores are transformed to represent immediate recall, learning curve, and forgetting score. Functional and behavioral assessments include the Functional Assessment Questionnaire (FAQ) and Everyday Cognition questionnaires filled out by both patients (ECogPT) and their study partners (ECogSP), covering various domains.

MRI data provide volume measurements segmented by brain regions such as Ventricles, Hippocampus, Entorhinal cortex, Fusiform gyrus, middle temporal gyrus, intracerebral volume (ICV), and the whole brain.

Unfortunately, the CSF biomarker data only contain baseline measurements for amyloid-beta, t-tau, and p-tau.

Table 3.6: Table of number of visits comparing full ADNI data and adjusted ADNI data

Full Data															
Number of Visits	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Counts	41	72	200	239	246	213	79	26	23	24	28	16	4	1	1
Adjusted Data															
Number of Visits	3	4	5	6	7	8	9	10	11	12	13				
Counts	144	115	114	161	59	5	6	5	7	5	1				

We narrow our focus to a subset of the original data to address the specific scientific question at hand. Table 3.6 indicates the number of visits for each subject. Among these subjects, 41 have only baseline measurements, while 72 have only one follow-up visit. Since our scientific inquiry revolves around the association between longitudinal measurements and the progression to AD diagnosis, subjects with only baseline measurements do not contribute statistical information to determine this association. Similarly, subjects with only one follow-up visit are not suitable for prediction with a random intercept and a random slope in the LME model, as they pose a risk of singular systems. As a result, subjects with two or fewer visits were excluded from the data analysis. Additionally, visits with incomplete cortical thickness measurements were omitted. If the first or last visit contained missing MRI data, the subject was removed to avoid extrapolation. Furthermore, if the total number of data entries per subject fell below 3 due to these procedures, the subject was also excluded from the analysis. Subjects diagnosed with AD at the study's onset were also excluded, leaving 478 subjects for further analysis. A comparison between the two tables highlights that subjects with a high number of visits predominantly consist of healthy individuals at the initial visit. Conversely, subjects diagnosed with MCI, including early mild cognitive impairment

(EMCI), late mild cognitive impairment (LMCI), and significant memory concern (SMC), typically had fewer visits, with a median of four visits.

It is worth noting that the outcome covariate for diagnosis results (DX) at each visit was recorded in the format of transitions from the previous stage. Six subjects experienced transitions from MCI to dementia, immediately followed by transitions from dementia back to MCI. To simplify the analysis, these subjects were classified as having dementia, as the diagnosis transition to dementia contains pertinent statistical information regarding AD conversion. Additionally, observations after dementia diagnosis were removed from the dataset, as they do not align with our definition of terminal events.

For comparison, the application to the ADNI dataset is intentionally designed to align with a similar study that adopts a two-stage model for joint modeling of longitudinal and survival data (Li et al., 2017). The original study seeks to assess the predictive capacity of markers for AD conversion within the framework of joint modeling. This joint model comprises two stages:

1. The longitudinal stage models the repeated measurements over time using a Linear Mixed Effects (LME) model with random intercept and random slope. This stage adjusts for baseline age and the presence of the APOE-$\epsilon 4$ allele.

2. The survival stage utilizes the predicted longitudinal covariate at the event time as a time-dependent covariate and get coefficient estimate via naive Cox estimator. This stage adjusts for gender, years of education, baseline age, and APOE-$\epsilon 4$.

Although not explicitly stated, the model approximates the hazard using the predicted longitudinal covariate at event time, and assumes proportional hazard for all biomarkers. Additionally, the model implicitly assumes a linear relationship between repeated measurements and time, with individual variation in baseline measurements and the rate of change.

In our analysis, we employ precise time intervals between dates to reduce tied survival times. However, in instances where ties persist despite these efforts, we introduce adaptive noise jittering. This method effectively breaks the ties while maintaining the chronological order of survival times.

To facilitate comparison, we apply our proposed variance reweighted two-stage model to the data with a matched regression setup. Then, we replace the model-based standard deviation used in the original paper with bootstrap standard deviation. This adjustment allows us to uncover any differences in the mean and standard error of estimates between the two models.

The dataset comprises additional ongoing studies, ADNI-GO and ADNI-2, and Table 3.7 presents the fundamental characteristics of the study participants. Participants remained enrolled in the study until experiencing the terminal event or being censored, with an average duration of 3.03 years (SD 1.62). Consistent with previous research on risk factors for AD conversion, individuals who progressed to AD during the study were older (mean=73.45, SD=7.43), exhibited a higher prevalence of positive APOE-ϵ4 carrier status (n=111, 63%), had fewer years of education (mean=15.95, SD=2.77), and included a lower proportion of females (n=70, 40%).

Table 3.7: Baseline characteristics of ADNI participants with mild cognitive impairment (MCI). Continuous covariates are summarized via mean (SD). Discrete covariates are summarized via frequency (%).

	Progressed to AD	Did not progress to AD	Combined
	(n=177, 28%)	(n=445, 72%)	(n=622)
Female Sex	70 (40%)	208 (47%)	278 (45%)
Age (yrs)	73.45 (7.43)	71.74 (7.26)	72.23 (7.34)
APOE4 Present	111 (63%)	177 (40%)	288 (46%)
Education (yrs)	15.95 (2.77)	16.30 (2.70)	16.20 (2.72)
Time before progression (yrs)	2.40 (1.62)	3.28 (1.54)	3.03 (1.62)

Table 3.8 illustrates the association of biomarkers with AD progression, contrasting our proposed method with the particular model procedure outlined previously. Comparing to the naive two-stage model in Li et al. (2017), the proposed method generally yields smaller abso-

lute Z-values. Notably, the naive Cox model produces larger Z-values for all predictors, while the proposed model offers more moderate values that align with consensus. For instance, predictors like MMSE, being a non-AD-specific questionnaire, show lower associations with AD progression risk. Similarly, ICV exhibits no statistically significant association with AD progression due to its non-specific nature. Regarding estimates, there exists a discrepancy between the two methods, with an average percentage difference exceeding 10% across all predictors. Although both approaches may lead to similar conclusions, the discrepancy is likely to persist across repeated studies as they estimate distinct quantities. Results from the naive method may lack replicability due to censoring mechanisms and potentially different estimated quantities. In contrast, results from the proposed method can be more reliably tested across studies and locations. Figure 3.9 provides a visual comparison of the estimated log-relative risk and the corresponding 95% confidence interval for each estimator and variable combination, showcasing similar observations.

Table 3.8: Association with AD progression: model results (controlled for age, APOE in the longitudinal model and controlled for age, gender, years of education, and APOE in the survival model)

			Naive Cox Estimator			Proposed Estimator		
Variable	Participants	Events	Estimate	95% CI	$\|Z\|$ value	Estimate	95% CI	$\|Z\|$ value
RAVLT.learning	622	177	-0.63	(-0.75,-0.52)	10.81	-0.64	(-0.81,-0.48)	7.68
MidTemp*	532	151	-2.80	(-3.40,-2.25)	9.58	-3.01	(-3.75,-2.27)	7.34
Entorhinal*	532	151	-10.31	(-12.76,-7.86)	8.24	-9.99	(-12.73,-7.26)	7.16
Hippocampus*	558	156	-7.72	(-9.22,-6.21)	10.06	-8.19	(-10.20,-6.13)	7.14
CDRSB	622	177	0.68	(0.62,0.76)	18.79	0.78	(0.55,0.99)	6.78
RAVLT.immediate	622	177	-0.14	(-0.16,-0.12)	13.45	-0.14	(-0.18,-0.09)	6.01
Fusiform*	532	151	-3.25	(-3.97,-2.53)	8.87	-3.22	(-4.28,-2.16)	5.63
ADAS13	622	177	0.042	(0.030,0.054)	6.66	0.021	(0.013,0.029)	5.35
Ventricles*	574	173	0.16	(0.11,0.21)	5.74	0.19	(0.11,0.26)	4.76
ADAS11	622	177	0.14	(0.13,0.16)	15.67	0.15	(0.09,0.22)	4.57
FAQ	622	177	0.20	(0.18,0.22)	17.70	0.30	(0.17,0.43)	4.45
WholeBrain*	588	175	-0.059	(-0.078,-0.040)	6.14	-0.065	(-0.095,-0.035)	4.25
FDG	287	81	-9.58	(-11.66,07.50)	9.03	-8.76	(-13.44,-4.09)	3.68
MMSE	622	177	-0.26	(-0.29,-0.23)	15.28	-0.27	(-0.48,-0.06)	2.50
RAVLT.forgetting	622	177	0.17	(0.095,0.240)	4.53	0.18	(-0.0043,0.3678)	1.91
ICV*	602	176	0.013	(0.0020,0.0245)	2.31	0.005	(-0.014,0.025)	0.56

* variable value is multiplied by 10000 for the ease of displaying

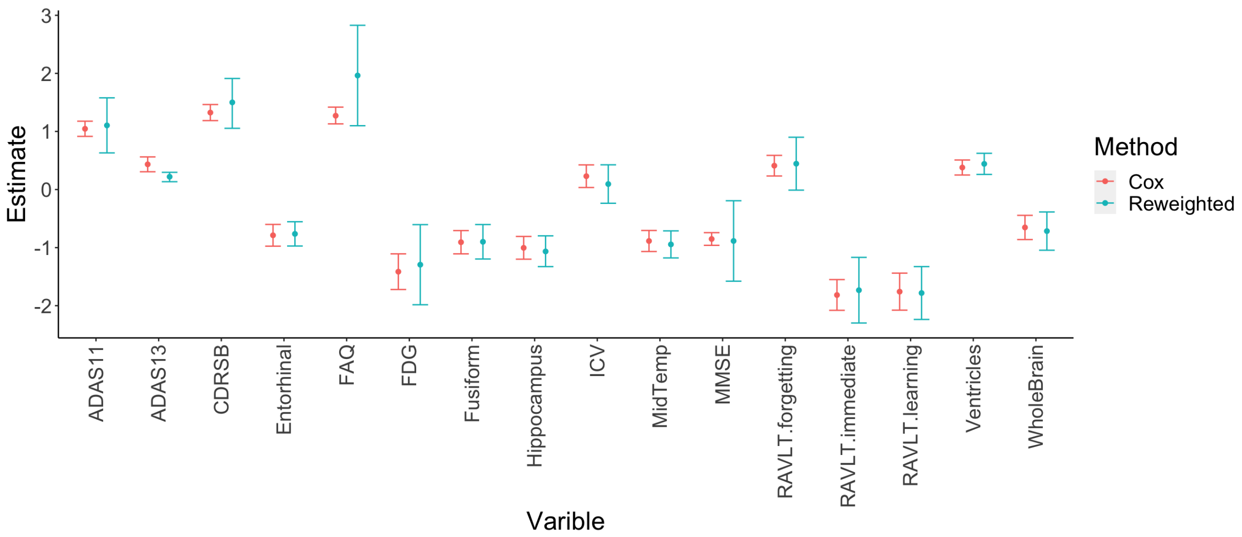


Figure 3.9: Confidence interval side-by-side comparison for all longitudinal covariates. Variables are scaled by baseline empirical standard error for comparable scales.

3.4.1 Discussion

Under NPH conditions, the traditional maximum partial likelihood estimator (MPLE) has been found to consistently estimate a quantity that relies on survival distribution, censoring distribution, and conditional covariate variance. This leads to two deficiencies with this estimator: i) Results can vary across studies even when the underlying effect remains the same, making them difficult to replicate and rendering them meaningless; ii) Even without dependency on the censoring distribution, the estimator consistently estimates a quantity that can be approximated by the average longitudinal covariate effect over the support of observed event time weighted by the conditional covariate effect. This renders the results challenging to interpret and not clinically meaningful.

In this chapter, we have introduced a two-stage reweighted censoring robust estimator that incorporates categorical, continuous, and longitudinal covariates. Our proposed estimator addresses potential bias resulting from not observing longitudinal covariate values at event times for other subjects in the risk set by predicting these values using a linear mixed effects

(LME) model. In the second stage, we apply regular censoring robust estimator techniques to remove dependence on censoring by reweighting using the inverse of censoring weights.

To enhance the interpretability of the resulting estimator, we add additional reweighting by the inverse of the estimated conditional covariate variance at the integrand and scaled the covariate by the inverse of the estimated conditional covariate variance. This adjustment allows our estimator to be interpreted as the weighted time-average covariate effect for the longitudinal covariate.

In the numerical simulations, we validate our method under the assumption of independent censoring and demonstrated its necessity for restoring an interpretable quantity. Applying the method to the ADNI data yields different results compared to conventional approaches. However, deriving a robust variance estimator poses nontrivial challenges, and the bootstrap method imposes a heavy computational burden. Furthermore, accurate longitudinal clustering is essential when multiple underlying longitudinal groups exist to estimate longitudinal conditional covariate variance accurately.

These challenges motivate the exploration in the following chapters. Chapter 4 aims to derive a robust variance estimator for the entire class of censoring robust estimators using a different approach than the sandwich estimator. Chapter 5 review current longitudinal clustering methods, with a focus on addressing the unique challenge of monotone missingness.

Chapter 4

Influence function-based robust variance estimator for censoring-robust reweight algorithms

4.1 Introduction

Survival analysis is a prevalent statistical method employed across diverse fields, such as medicine and epidemiology. With the advancement of information collection and management systems, researchers now have easier access to repeated measurements collected over time from individual subjects. Consequently, there is a growing emphasis on integrating longitudinal and survival data through joint modeling approaches. Henderson et al. (2000) offer insights into scenarios where simultaneous modeling of longitudinal and survival data is required, providing a comprehensive overview of existing methodologies developed for such models.

Ibrahim et al. (2010) highlight the significance of employing joint modeling techniques in longitudinal studies focusing on biomarkers. Their research sheds light on the presence of bias in the naive Cox model and reveals that both full likelihood and two-stage models can exhibit such bias. They emphasize that the integration of survival and longitudinal analysis inherits the challenges associated with survival data, particularly the potential misspecification issues arising from the assumption of proportional hazards in the Cox model.

As discussed in preceding chapters, the standard Cox estimator faces challenges under NPH, primarily due to its reliance on unknown censoring mechanisms. To address this dependency, reweighted partial likelihood estimators have been proposed, aiming to reweight subjects in the risk set by the inverse of their censoring weights(Xu and O'Quigley, 2000; Boyd et al., 2012; Nguyen and Gillen, 2017). This approach seeks to emulate the scenario of complete data, thereby mitigating bias introduced by censoring. However, a key challenge lies in accurately estimating the censoring weights. For categorical covariates, estimating the censoring distribution by categorical group suffices. However, for continuous covariates, an approach involving the identification of approximate censoring groups using survival tree methods has been proposed.

In Chapter 3, we extended the previously introduced censoring robust framework to incorporate longitudinal covariates. To align with existing literature, we adopted a two-step joint modeling approach for longitudinal covariates. In the first stage, we predicted the longitudinal covariate values at event times, and in the second stage, we substituted these predicted values into the model. However, we discovered that the interpretability of the resulting estimator was compromised by variations in conditional covariance variance.

To restore clinical meaningfulness to the estimator, we proposed scaling the longitudinal covariate at each event time by the inverse of the estimated conditional covariate variance. This method yielded a flexible censoring robust estimator with an interpretation as the average covariate effect incorporating longitudinal covariate variables. However, deriving robust inference for such an estimator proved challenging, necessitating the use of bootstrap methods.

The bootstrap method involves creating a large number of random copies of the original data using case resampling and generating a distribution of the estimator to estimate variance. While effective, this method is computationally intensive, particularly when dealing with a large number of subjects. To address this issue, we sought an alternative approach to derive a robust variance estimator to alleviate computational burden.

The original approach to robust variance estimation for censoring robust estimators relies on the Huber-White robust variance estimator, which is based on two methods of calculating variance(Freedman, 2006). The variance estimated using the product of independent score equations is sandwiched by the variance estimated by the model-based Fisher information matrix. This method has been demonstrated as a powerful robust variance estimator in the context of linear regression-related methods.

In survival analysis, however, the Cox model introduces correlations between score equations arising from the weighted expectation of covariate values at event times. To address this issue, Lin and Wei (1989) approximated the score equations using independent summation asymptotically, forming the basis of the censoring robust variance estimator. Previous censoring robust variance estimators directly apply this method. However, in our case, we have additional variance information that would induce further correlation, making the derivation of the robust variance estimator a nontrivial task. For other types of robust variance estimators, Hardin (2002) established a relationship between the Murphy-Topel variance estimator Webel (2011) and the Huber-White robust variance estimator. However, the Murphy-Topel variance estimator does not account for partial correlation. Another approach to robust variance estimation is the influence function-based variance estimation introduced in Law et al. (1986). The influence function method offers the advantage of being easily applicable and not being constrained by partial correlation.

The influence function has been previously computed for various models, including the fixed effect of linear mixed effect model (Demidenko and Stukel, 2005), censored data (Reid, 2007), and regular Cox regression model (Reid and Crépeau, 1985). The modification to the new estimation equation is relatively straightforward compared to previous robust methods and has been implemented in scenarios of independent censoring (Xu and Harrington, 2001).

Regarding the two-stage model, Zhelonkin et al. (2012) presents the formula for the uncorrelated case in our two-stage model, where the two stages of the model are connected solely through a single term. However, there is still a gap in the establishment of a robust variance estimator for a two-stage model with correlated score equations, particularly concerning the set of censoring robust variance estimators.

For the remaining part of this study, Section 4.2 focuses on the development of robust variance estimators using the influence function based method. Subsequently, a numerical investigation is conducted to evaluate the performance in various scenarios, while also comparing it to an existing method. In the Section 4.4, a data analysis is conducted using the influence function method and compared to an analysis performed using the bootstrap method.

4.2 Method

4.2.1 Statistical setup

As discussed in Chapter 2 (Section 2.3.2), we focus on an infinitesimal representation of the partial likelihood estimator, where the Cox estimator can be expressed as a functional of the following:

$$\Psi(\beta, w, H) = \delta \left[z - \frac{E_H(Ze^{\beta z}\mathrm{I}\{\tilde{x} \geq x\})}{E_H(e^{\beta z}\mathrm{I}\{\tilde{x} \geq x\})} \right],$$

where β is the functional $T(H)$ and δ_w represents the point mass distribution at (x, δ, z). Then define $H_\epsilon = (1-\epsilon)H + \epsilon\delta_w$. Then, β_ϵ represents the functional under distribution function H_ϵ. Thus, the infinitesimal representation can be expressed as

$$\int \Psi(\beta_\epsilon, w, H_\epsilon) dH_\epsilon = 0.$$

Then the calculation of $\frac{\partial}{\partial H_\epsilon}\Psi(\beta, w, H_\epsilon)$ gives

$$\begin{aligned}
\frac{\partial}{\partial H_\epsilon}\Psi(\beta, w, H_\epsilon) &= \frac{\partial}{\partial H_\epsilon}\delta \left\{ z - \frac{\int ze^{\beta z}\mathrm{I}(\tilde{x} \geq x) d[H + \epsilon(\delta_w - H)]}{\int e^{\beta z}\mathrm{I}(\tilde{x} \geq x) d[H + \epsilon(\delta_w - H)]} \right\} \\
&= -\delta \left\{ \frac{E_{\tilde{w}} ze^{\beta z}\mathrm{I}(\tilde{x} \geq x)}{E_H e^{\beta z}\mathrm{I}(\tilde{x} \geq x)} - \frac{E_H ze^{\beta z}\mathrm{I}(\tilde{x} \geq x) E_{\tilde{w}} e^{\beta z}\mathrm{I}(\tilde{x} \geq x)}{[E_H e^{\beta z}\mathrm{I}(\tilde{x} \geq x)]^2} \right\} \\
&= -\tilde{\delta} \left\{ \frac{ze^{\beta z}\mathrm{I}(\tilde{x} \leq x)}{s^{(0)}(\tilde{x}, \beta)} - \frac{s^{(1)}(\tilde{x}, \beta) e^{\beta z}\mathrm{I}(\tilde{x} \leq x)}{[s^{(0)}(\tilde{x}, \beta)]^2} \right\}.
\end{aligned}$$

IF are obtained by arranging terms:

$$A(\beta) \times \mathrm{IF}(w; T, H) = \delta \left\{ z - \frac{s^{(1)}(x, \beta)}{s^{(0)}(x, \beta)} \right\} - \left[e^{\beta z} \int \frac{\tilde{\delta}\mathrm{I}(\tilde{x} \leq x)}{s^{(0)}(\tilde{x}, \beta)} \left\{ z - \frac{s^{(1)}(\tilde{x}, \beta)}{s^{(0)}(\tilde{x}, \beta)} \right\} \right] dH(\tilde{x}, \tilde{\delta}, \tilde{z}). \tag{4.1}$$

With the fundamental framework of the influence function for the proportional hazard function established, we will now proceed with the derivation of the entire class of censoring robust variance estimators in the remaining part of this section.

4.2.2 Influence function for censoring robust estimators under independent censoring

Under independent Censoring, given the minor change to the estimating equation, the infinitesimal representation can be written as

$$\int \delta S(x)^{-1}\left\{z - \frac{\int \tilde{z}e^{\tilde{z}\beta}\mathrm{I}\{\tilde{x} \geq x\}dH(\tilde{x},\tilde{z})\}}{\int e^{\tilde{z}\beta}\mathrm{I}\{\tilde{x} \geq x\}dH(\tilde{x},\tilde{z})\}}\right\} dH(x,z,\delta) = 0.$$

Given that the estimating equation was only multiplied by $S(x)^{-1}$, we expect that only $\int \frac{\partial}{\partial\epsilon}\Psi(\beta, w, H_\epsilon)\, dH(w)$ will change and have an additional term related to the influence of $S(x)$. Applying the chain rule along with the previous result of the influence function for the Kaplan-Meier estimator, we have:

$$\begin{aligned}
\mathrm{IF}(w; S(x)^{-1}, H) &= -S(\tilde{x})^{-2}\mathrm{IF}(w; S(x), H) \\
&= S(\tilde{x})^{-1}\left\{-\int_0^{x\wedge\tilde{x}} \frac{dF^u(u)}{[S(u)]^2} + \frac{\delta\mathrm{I}\{x \leq \tilde{x}\}}{S(\tilde{x})}\right\} \\
&= S(\tilde{x})^{-1}\left\{-\int_0^{x\wedge\tilde{x}} \frac{dF^u(u)}{[s^{(0)}(u,0)]^2} + \frac{\delta\mathrm{I}\{x \leq \tilde{x}\}}{s^{(0)}(x,0)}\right\},
\end{aligned}$$

where $S(x) = s^{(0)}(x, 0)$. Combining the previous result of IF for regular Cox estimator, we obtain IF for censoring robust estimator under independent censoring as

$$\begin{aligned}
A(\beta) \times \mathrm{IF}(w; T, H) &= \delta S(x)^{-1}\left\{z - \frac{s^{(1)}(x,\beta)}{s^{(0)}(x,\beta)}\right\} \\
&- \int \tilde{\delta} S(\tilde{x})^{-1}\left[\frac{e^{\beta z}\mathrm{I}(\tilde{x} \leq x)}{s^{(0)}(\tilde{x},\beta)}\left\{z - \frac{s^{(1)}(\tilde{x},\beta)}{s^{(0)}(\tilde{x},\beta)}\right\}\right. \\
&+ \left.\left\{\tilde{z} - \frac{s^{(1)}(\tilde{x},\beta)}{s^{(0)}(\tilde{x},\beta)}\right\} \times \left\{\int_0^{x\wedge\tilde{x}} \frac{dF^u(u)}{[s^{(0)}(u,0)]^2} - \frac{\delta\mathrm{I}\{x \leq \tilde{x}\}}{s^{(0)}(x,0)}\right\}\right] dH(\tilde{x},\tilde{\delta},\tilde{z}).
\end{aligned} \tag{4.2}$$

Comparing Equation 4.2 with Equation 4.1, the first two terms are similar with slight modifications according to the alternation in estimand, while the additional term accounts for uncertainty in censoring weight estimation.

Substituting $\hat{\beta}$ and the empirical distribution function into the expression for the influence function, we obtain the empirical influence function. The empirical influence function evaluated at (x_i, δ_i, z_i) can be written as

$$I_i = \hat{A}(\hat{\beta})^{-1} \times \left[\delta_i \hat{S}(x_i)^{-1} \left\{ z_i - \frac{S^{(1)}(x_i, \hat{\beta})}{S^{(0)}(x_i, \hat{\beta})} \right\} - D_i(\hat{\beta}) \right],$$

where

$$\begin{aligned} D_i(\hat{\beta}) = \sum_{j=1}^{n} \delta_j \hat{S}(x_j)^{-1} \times & \left[\frac{e^{\beta' z_i} \mathrm{I}(x_j \leq x_i)}{S^{(0)}(x_j, \hat{\beta})} \left\{ z_i - \frac{S^{(1)}(x_j, \hat{\beta})}{S^{(0)}(x_j, \hat{\beta})} \right\} \right. \\ & \left. + \left\{ z_j - \frac{S^{(1)}(x_j, \hat{\beta})}{S^{(0)}(x_j, \hat{\beta})} \right\} \times G_{ij}(\hat{\beta}) \right], \end{aligned}$$

and

$$G_{ij}(\hat{\beta}) = \sum_{x_l \leq (x_i \wedge x_j)} \frac{\delta_l}{n S^{(0)}(x_l, 0)^2} - \frac{\delta_i \mathrm{I}\{x_i \leq x_j\}}{S^{(0)}(x_j, 0)}.$$

Then the estimated covariance matrix of $n^{1/2}(\hat{\beta} - \beta^*)$ is $n^{-1} \sum I_i I_i'$. The empirical influence function consists of three layers of summations.

4.2.3 Influence function for censoring robust estimators under conditional independent censoring without longitudinal covariates

Under conditional independent censoring, the estimator from the previous subsection encounters difficulty in the direct estimation of censoring distribution and fails to completely remove dependency on censoring. To address this, when only categorical covariates exist, the conditional independent censoring distribution can be categorized by the power set of categorical covariates. Within each categorical combination, the censoring distribution is homogeneous and can be estimated using the Kaplan-Meier estimator. To fully eliminate the dependency on censoring in the weight used to estimate the expected covariate value, the covariate weight in the fraction is further reweighted by the inverse of censoring weights. In the case of continuous covariates, approximate censoring groups are identified through the survival tree approach, and the estimation proceeds in the same manner. With further alternation in the estimating equation, the infinitesimal representation can be written as

$$\int \delta S_z(x)^{-1}\left\{z - \frac{\int S_{\tilde{z}}(x)^{-1}\tilde{z}e^{\tilde{z}\beta}\mathrm{I}\{\tilde{x}\geq x\}dH(\tilde{x},\tilde{z})\}}{\int S_{\tilde{z}}(x)^{-1}e^{\tilde{z}\beta}\mathrm{I}\{\tilde{x}\geq x\}dH(\tilde{x},\tilde{z})\}}\right\}dH(x,z,\delta)=0,$$

where $S_z(x)$ denotes the censoring distribution conditional on the covariate value z.

Even though the estimating equation becomes more complicated, estimating equation can still be expressed as

$$\int \Psi^*(\beta_\epsilon, w, H_\epsilon)dH_\epsilon = 0.$$

. Then implicit differentiation with respect to ϵ still gives

$$\int \Psi^*(\beta,w,H)d[\delta_{\bar{w}}-H](w)+\int \frac{\partial}{\partial\beta}\Psi^*(\beta,w,H)dH(w)\frac{\partial\beta}{\partial\epsilon}+\int \frac{\partial}{\partial H_\epsilon}\Psi^*(\beta,w,H_\epsilon)dH(w)=0,$$

where the only changes to first two terms are the addition of censoring weights to remove the dependence on censoring. The major change due to additional censoring weight lies in third term. For independent censoring, an additional term corresponds to influence from censoring weights estimation in the multiplication. When censoring weights are incorporated into fractions, it is expected that the influence on the risk set would be further adjusted for the influence from Incorporated censoring weights. Combining previous result of IF for $S_z(x)$, we have

$$\begin{aligned}\frac{\partial}{\partial H_\epsilon}\Psi^*(\beta, w, H_\epsilon) =& \frac{\partial}{\partial \epsilon} S_z(x)^{-1} \times \left\{ z - \frac{\int S_{\tilde{z}}(x)^{-1} \tilde{z} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}}{\int S_{\tilde{z}}(x)^{-1} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}} \right\} \\ &+ S_z(x)^{-1} \times \frac{\partial}{\partial \epsilon} \left\{ z - \frac{\int S_{\tilde{z}}(x)^{-1} \tilde{z} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}}{\int S_{\tilde{z}}(x)^{-1} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}} \right\}.\end{aligned}$$

where the first part is similar to the previous independent censoring case. For the second part, we have

$$\begin{aligned}\frac{\partial}{\partial H_\epsilon}\Psi^{**}(\beta, w, H_\epsilon) =& \frac{\partial}{\partial \epsilon} \left\{ z - \frac{\int S_{\tilde{z}}(x)^{-1} \tilde{z} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} d[H + \epsilon(\delta_w - H)]}{\int S_{\tilde{z}}(x)^{-1} e^{\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} d[H + \epsilon(\delta_w - H)]} \right\} \\ =& - \frac{E_{\bar{w}} S_{\tilde{z}}(\tilde{x})^{-1} z e^{\beta z} \mathrm{I}(\tilde{x} \geq x) + E_H \mathrm{IF}(\tilde{w}, S, H) z e^{\beta z} \mathrm{I}(\tilde{x} \geq x)}{E_H S_{\tilde{z}}(\tilde{x})^{-1} e^{\beta z} \mathrm{I}(\tilde{x} \geq x)} \\ &+ \frac{E_H S_{\tilde{z}}(\tilde{x})^{-1} z e^{\beta z} \mathrm{I}(\tilde{x} \geq x) [E_{\bar{w}} e^{\beta z} \mathrm{I}(\tilde{x} \geq x) + E_H \mathrm{IF}(\tilde{w}, S, H) e^{\beta z} \mathrm{I}(\tilde{x} \geq x)]}{[E_H S_{\tilde{z}}(\tilde{x})^{-1} e^{\beta z} \mathrm{I}(\tilde{x} \geq x)]^2} \\ =& - \frac{S_z(x)^{-1} z e^{\beta z} \mathrm{I}(\tilde{x} \leq x) + E_H \mathrm{IF}(\tilde{w}, S, H) z e^{\beta z} \mathrm{I}(\tilde{x} \geq x)}{s_W^{(0)}(\tilde{x}, \beta)} \\ &+ \frac{s^{(1)}(\tilde{x}, \beta)[S_z(x)^{-1} e^{\beta z} \mathrm{I}(\tilde{x} \leq x) + E_H \mathrm{IF}(\tilde{w}, S, H) e^{\beta z} \mathrm{I}(\tilde{x} \geq x)]}{[s_W^{(0)}(\tilde{x}, \beta)]^2},\end{aligned}$$

where $\mathrm{IF}(\tilde{w}, S, H)$ stands for the influence function of censoring weight function, and $s_W^{(j)}(x, \beta)$ is defined as same as in the previous section with the addition of censoring weights. Combining all parts of the equation, we can derive the influence function for the censoring robust estimator under categorical covariates as

$$\begin{aligned}
A(\hat{\beta}) \times \mathrm{IF}(w; T, H) = {} & \delta S_z(x)^{-1} \left\{ z - \frac{s_W^{(1)}(x, \hat{\beta})}{s_W^{(0)}(x, \hat{\beta})} \right\} \\
& - \int \tilde{\delta} S_{\tilde{z}}(x)^{-1} \left[\frac{S_z(x)^{-1} e^{\hat{\beta} z} \mathrm{I}(\tilde{x} \leq x)}{s_W^{(0)}(\tilde{x}, \hat{\beta})} \left\{ z - \frac{s_W^{(1)}(\tilde{x}, \hat{\beta})}{s_W^{(0)}(\tilde{x}, \hat{\beta})} \right\} \right. \\
& + \frac{1}{s_W^{(0)}(\tilde{x}, \hat{\beta})} \left\{ E_H \mathrm{IF}(\tilde{w}, S, H) \tilde{z} e^{\hat{\beta} \tilde{z}} \mathrm{I}(\tilde{x} \geq x) \right. \\
& \left. - \frac{s_W^{(1)}(\tilde{x}, \hat{\beta}) E_H \mathrm{IF}(\tilde{w}, S, H) e^{\hat{\beta} \tilde{z}} \mathrm{I}(\tilde{x} \geq x)]}{s_W^{(0)}(x, \hat{\beta})} \right\} \\
& \left. + \left\{ \tilde{z} - \frac{s_W^{(1)}(\tilde{x}, \hat{\beta})}{s_W^{(0)}(\tilde{x}, \hat{\beta})} \right\} \times \left\{ \int_0^{x \wedge \tilde{x}} \frac{dF^u(u)}{[s^{(0)}(u, 0)]^2} - \frac{\delta \mathrm{I}\{x \leq \tilde{x}\}}{s^{(0)}(x, 0)} \right\} \right] dH(\tilde{x}, \tilde{\delta}, \tilde{z}),
\end{aligned} \tag{4.3}$$

where $A(\beta)$ is updated with the addition of censoring weight. Terms in equation 4.3 are similar to equation 4.2 with slight changes in notation. The first term remains analogous to the usual influence function for M-estimators, while the fourth term represents the influence from reweighting the estimand through censoring weights. The middle terms represent the influence of the ith observation on the risk sets when combined. Such influence can be further separated into the influence of reweighting from the ith observation on the integrand, as in the third term, and the influence on the risk set as in the second term.

For conditional independent censoring distribution on the combination of categorical and continuous covariates, assuming the correctness of approximate grouping, the influence function remains unchanged as the estimating equation is not altered. The approximate grouping is assumed to be correct as the grouping procedure cannot be expressed in an infinitesimal format and therefore cannot be incorporated into the influence function through implicit differentiation. However, the effect of misclassification on the censoring group is expected to be reflected in the influence function.

When deriving the influence function for the censoring robust estimator, the influence function for the Kaplan-Meier estimator is utilized which account for misspecified censoring distribution. The influence function for censoring weights is included for both reweighting the entire integrand and reweighting for the expected covariate value at event time. Since these are the locations where censoring weights occur in the estimating equation, they are vulnerable to misspecification of censoring weights. By adding the influence function of the Kaplan-Meier estimator at these locations, the robust property is expected to be preserved.

Substituting $\hat{\beta}$ and the empirical distribution function into the expression for the influence function, we obtain the empirical influence function. The empirical influence funciton evaluated at (x_i, δ_i, z_i) can be written

$$I_i = \hat{A}(\hat{\beta})^{-1} \times \left[\delta_i \hat{S}(x_i)^{-1} \left\{ z_i - \frac{S^{(1)}(x_i, \hat{\beta})}{S^{(0)}(x_i, \hat{\beta})} \right\} - D_i(\hat{\beta}) \right],$$

where

$$\begin{aligned} D_i(\hat{\beta}) = \sum_{j=1}^{n} \delta_j \hat{S}(x_j)^{-1} \times \Bigg[& \frac{e^{\beta' z_i} \mathrm{I}(x_j \leq x_i)}{S_W^{(0)}(x_j, \hat{\beta})} \left\{ z_i(x_j) - \frac{S_W^{(1)}(x_j, \hat{\beta})}{S_W^{(0)}(x_j, \hat{\beta})} \right\} \\ & + \left\{ z_j(x_j) - \frac{S_W^{(1)}(x_j, \hat{\beta})}{S_W^{(0)}(x_j, \hat{\beta})} \right\} \times G_{ij}(\hat{\beta}) \\ & - \frac{G_{ij}(\hat{\beta}) S_{z_k}^{-1}(x_j)}{S_W^{(0)}(x_j, \hat{\beta})} \Bigg\{ \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} z_k(x_j) e^{\hat{\beta} z_k(x_j)} \mathrm{I}(x_k \geq x_j) \\ & - \frac{S_W^{(1)}(x_j, \hat{\beta}) \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} e^{\hat{\beta} z_k(x_j)} \mathrm{I}(x_k \geq x_j)}{S_W^{(0)}(x_j, \hat{\beta})} \Bigg\} \Bigg], \end{aligned}$$

$$G_{ij}(\hat{\beta}) = \sum_{x_l \leq (x_i \wedge x_j)} \frac{G_l^j (1 - \delta_l)}{n_l^c S^{(0)}(x_l, 0)^2} - \frac{G_i^j (1 - \delta_i) \mathrm{I}\{x_i \leq x_j\}}{S^{(0)}(x_j, 0)},$$

where G_m^i is an indicator function that gives value one only if subject m and subject i belong to the same censoring group, n_l^c indicates the number of subjects of the censoring group that the subject l belongs to, and n_{z_k} indicates the number of subjects of the longitudinal variance group that the subject k belongs to.

4.2.4 Influence function for censoring robust estimators under conditional independent censoring with a longitudinal covariate

In the case of conditional independent censoring with longitudinal covariates, it is useful to initially consider a heuristic scenario where the exact longitudinal covariate values at event time are available. With the inclusion of longitudinal covariates, the complexity of the influence function doubles due to the uncertainty stemming from unknown longitudinal covariate values at event time. To address this uncertainty, we employ a two-stage model where we predict the longitudinal covariate values at event time using a LME model.

The estimator derived from this two-stage model inherits the influence from the previous censoring robust estimator while also facing potential influence from the correctness of the model used in the first stage. As observed in previous cases, the influence function accumulates from various sources of misspecification. Therefore, beginning with the assumption of known longitudinal values at event time lays the groundwork for the final result and streamlines the process.

Without considering the subgroups of longitudinal trajectories, the estimating equation for longitudinal covariate can be written in an infinitesimal form as

$$\int \delta S_z(x)^{-1} \left\{ V(x)^{-1} z - \frac{\int S_{\tilde{z}}(\tilde{x})^{-1} \tilde{z} V(x)^{-1} e^{V(x)^{-1}\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}}{\int S_{\tilde{z}}(\tilde{x})^{-1} e^{V(x)^{-1}\tilde{z}\beta} \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})\}} \right\} dH(x, z, \delta) = 0,$$

where the $V(x)$ represents the conditional variance of longitudinal covariate with respect to the distribution H. To derive the influence function given covariate value at event time, it is foreseeable that the influence function for the conditional variance of longitudinal covariate is required. The influence function of $V(x)$ can be obtained through a slight modification of the variance function, which is readily available (Zhang et al., 2019). Since $\mathrm{Var}[X|T] = E[X^2|T] - E[X|T]^2$, then by the chain rule for influence function we have

$$\begin{aligned} IF(w, V, H) &= \frac{\partial}{\partial \epsilon} \left\{ \int \tilde{x}^2 \mathrm{I}\{\tilde{x} \geq x\} dH_\epsilon - \left(\int \tilde{x} \mathrm{I}\{\tilde{x} \geq x\} dH_\epsilon \right)^2 \right\} \\ &= x^2 \mathrm{I}\{x \geq \tilde{x}\} - E_H[X^2|T] - 2E_H[X|T](x\mathrm{I}\{x \geq \tilde{x}\} - E_H[X|T]) \\ &= (x\mathrm{I}\{x \geq \tilde{x}\} - E_H[X|T])^2 - Var[X|T], \end{aligned}$$

where $E_H^j[X|T]$ are just regular moment with indicator in risk set.

The empirical influence function for the empirical variance estimator can be obtained by

$$I_i\{V(x_j)\} = \left\{ z_i(x_j)\mathrm{I}(x_i \geq x_j) - \hat{E}_H[X|T] \right\}^2 - V\hat{a}r[X|T],$$

where

$$\hat{E}_H[X|T] = \sum_l z_l \mathrm{I}(x_l \geq x_i),$$

and

$$\hat{E}_H[X^2|T] = \sum_l z_l^2 \mathrm{I}(x_l \geq x_i).$$

To derive IF, two more terms will be associated with the influence from the variance estimator. The derived IF for the censoring robust estimator with longitudinal covariate assuming covariate at event time known is

$$\begin{aligned}
A(\beta) \times \mathrm{IF}(w; T, H) = {} & \delta S(x)^{-1} \left\{ V(x)^{-1} z - \frac{s^{(1)}_{WV}(x,\beta)}{s^{(0)}_{WV}(x,\beta)} \right\} \\
& - \int \tilde{\delta} S(\tilde{x})^{-1} \left[\frac{e^{\beta z} \mathrm{I}(\tilde{x} \leq x)}{s^{(0)}_{WV}(\tilde{x},\beta)} \left\{ V(x)^{-1} z - \frac{s^{(1)}_{WV}(\tilde{x},\beta)}{s^{(0)}_{WV}(\tilde{x},\beta)} \right\} \right. \\
& - \frac{S(\tilde{x})}{s^{(0)}_{WV}(x,\beta)} \left\{ E_H \mathrm{IF}_1(\tilde{w}) z e^{\beta z} \mathrm{I}(\tilde{x} \geq x) \right. \\
& \left. - \frac{s^{(1)}_{WV}(\tilde{x},\beta) E_H \mathrm{IF}_1(\tilde{w}) e^{\beta z} \mathrm{I}(\tilde{x} \geq x)]}{s^{(0)}_{WV}(x,\beta)} \right\} \\
& + \left\{ V(\tilde{x})^{-1} \tilde{z} - \frac{s^{(1)}_{WV}(\tilde{x},\beta)}{s^{(0)}_{WV}(\tilde{x},\beta)} \right\} \times \left\{ \int_0^{x \wedge \tilde{x}} \frac{dF^u(u)}{[s^{(0)}_V(u,0)]^2} - \frac{\delta \mathrm{I}\{x \leq \tilde{x}\}}{s^{(0)}_V(x,0)} \right\} \\
& + IF_2(\tilde{w}) V(\tilde{x})^{-2} \begin{pmatrix} 0 \\ \tilde{z}_L \end{pmatrix} \\
& - \frac{1}{s^{(0)}_{WV}(\tilde{x},\beta)} \{ E_H IF_2(\tilde{w}) V(\tilde{x})^{-2} \begin{pmatrix} 0 \\ \tilde{z}_L \end{pmatrix} S(\tilde{x})^{-1} e^{\beta z} \mathrm{I}(\tilde{x} \geq x) \\
& + E_H V(\tilde{x})^{-1} \tilde{z} S(\tilde{x})^{-1} e^{\beta \tilde{z}} \beta' IF_2(\tilde{w}) V(\tilde{x})^{-2} \begin{pmatrix} 0 \\ \tilde{z}_L \end{pmatrix} \mathrm{I}(\tilde{x} \geq x) \\
& \left. - \frac{s^{(1)}_{WV}(\tilde{x},\beta) E_H S(\tilde{x})^{-1} e^{\beta \tilde{z}} \beta' IF_2(\tilde{w}) V(\tilde{x})^{-2} \begin{pmatrix} 0 \\ \tilde{z}_L \end{pmatrix} \mathrm{I}(\tilde{x} \geq x)}{s^{(0)}_{WV}(\tilde{x},\beta)} \} \right] dH(\tilde{x}, \tilde{\delta}, \tilde{z}),
\end{aligned} \tag{4.4}$$

where IF_1 denotes the influence function for $S(x)$ and IF_2 denotes the influence function for $V(x)$ respectively for simplicity. Even with many terms, the additional items come from the influence of reweighting the observed covariate value and the influence of reweighting the expected covariate value with the conditional covariance variance.

Substituting $\hat{\beta}$ and the empirical distribution function into the expression for the influence function, we obtain the empirical influence function. The empirical influence function evaluated at (x_i, δ_i, z_i) can be written as

$$I_i = \hat{A}(\hat{\beta})^{-1} \times \left[\delta_i \hat{S}(x_i)^{-1} \left\{ z_i \hat{V}^{-1}(x_i) - \frac{S^{(1)}_{WV}(x_i, \hat{\beta})}{S^{(0)}_{WV}(x_i, \hat{\beta})} \right\} - D_i(\hat{\beta}) \right],$$

where

$$\begin{aligned}
D_i(\hat{\beta}) = \sum_{j=1}^{n} \delta_j \hat{S}(x_j)^{-1} \times \Biggl[& \frac{e^{\beta' z_i^*(x_i)} \mathrm{I}(x_j \le x_i)}{S^{(0)}_{WV}(x_j, \hat{\beta})} \left\{ z_i^*(x_j) - \frac{S^{(1)}_{WV}(x_j, \hat{\beta})}{S^{(0)}_{WV}(x_j, \hat{\beta})} \right\} \\
& + \left\{ z_j^*(x_j) - \frac{S^{(1)}_{WV}(x_j, \hat{\beta})}{S^{(0)}_{WV}(x_j, \hat{\beta})} \right\} \times G_{ij}(\hat{\beta}) \\
& - \frac{G_{ij}(\hat{\beta}) S^{-1}_{z_k}(x_j)}{S^{(0)}_{WV}(x_j, \hat{\beta})} \Biggl\{ \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} z_k^*(x_j) e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \ge x_j) \\
& \quad - \frac{S^{(1)}_{W}(x_j, \hat{\beta}) \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \ge x_j)}{S^{(0)}_{W}(x_j, \hat{\beta})} \Biggr\} \\
& + \begin{pmatrix} 0 \\ \\ z_{jL}^*(x_j) \end{pmatrix} \hat{V}^{-1}(x_j) * I_i\{V(x_j)\} \\
& - \frac{I_i\{V(x_j)\} \hat{V}^{-1}(x_j)}{S^{(0)}_{WV}(x_j, \hat{\beta})} \Biggl\{ \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S^{-1}_{z_k}(x_j) \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \ge x_j) \\
& + \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S^{-1}_{z_k}(x_j) z_k^*(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} \mathrm{I}(x_k \ge x_j) \\
& - \frac{S^{(1)}_{WV}(x_j, \hat{\beta}) \times \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S^{-1}_{z_k}(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} \mathrm{I}(x_k \ge x_j)}{S^{(0)}_{WV}(x_j, \hat{\beta})} \Biggr\} \Biggr],
\end{aligned}$$

and

$$G_{ij}(\hat{\beta}) = \sum_{x_l \leq (x_i \wedge x_j)} \frac{G_l^j(1-\delta_l)}{n_l^c S^{(0)}(x_l, 0)^2} - \frac{G_i^j(1-\delta_i)\mathrm{I}\{x_i \leq x_j\}}{S^{(0)}(x_j, 0)}.$$

where $z_k^*(x_j)$ indicates the scaled covariate value of subject k at the event time of subject j, and $z_{kL}^*(x_j)$ specifies the scaled longitudinal covariate value.

4.2.5 Influence function for two-stage model

Many estimators are obtained through a two-stage model, in which the first stage provides input for the second stage. For instance, in our proposed censoring robust estimator, the first stage model helps approximate the exact hazard by predicting longitudinal covariate value at event time. For mathematical representation, the two-stage estimation problem can be characterized by Hardin (2002):

$$\text{Model 1} : E\{y_1|x_1, \theta_1\}$$
$$\text{Model 2} : E\{y_2|x_2, \theta_2, E\{y_1|x_1, \theta_1\}\},$$

Instead of dealing with the full information maximum likelihood $f(y_1, y_2|x_1, x_2, \theta_1, \theta_2)$, where the parameter vectors to be estimated are θ_1 and θ_2, θ_1 can be directly estimated independently of θ_2. Then, conditional on the estimations from the first stage, θ_2 can be estimated, thus completing the estimation process in two steps.

For robust variance estimation, the influence function has been derived for special cases where both stages consist of estimation equations (Hardin, 2002). The result is akin to the Huber-White robust variance estimator. However, for more general M-estimators, the Huber-White robust variance estimator may not exist as the representation as estimating equations may not unavailable. Nonetheless, the influence function still exists, and a general framework for the analysis of the robustness property was provided, considering the class of two-stage models as defined by Zhelonkin et al. (2012):

$$\begin{aligned} E_F\left[\Psi_1\left(z^{(1)}; S(F)\right)\right] &= 0 \\ E_F\left[\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right)\right] &= 0 \end{aligned},$$

Where $\Psi_1(\cdot;\cdot)$ and $\Psi_2(\cdot;\cdot,\cdot)$ denote the score functions of the first and second stage estimators, respectively, and $h(\cdot;\cdot)$ is a given continuously piecewise differentiable function in the second variable. Here, S is the functional for the parameters of the first stage, such that $S\left(F_N\right) = \hat{\beta}_1$ and at the model $S(F) = \beta_1$, while T is the functional for the second stage, such that $T\left(F_N\right) = \hat{\beta}_2$ and at the model $T(F) = \beta_2$. Here, $T(F)$ depends directly on F and indirectly on F through $S(F)$.

Define $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_z$ and Δ_z as the probability measure which puts mass one at the point z. Then, the infinitesimal representation of the functional of the second stage can be written as

$$\int \Psi_2\left(z^{(2)}; h\left(z^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) dF_\epsilon = 0.$$

Taking the derivative with respect to ϵ gives

$$\begin{aligned} &\frac{\partial}{\partial\epsilon}(1-\epsilon)\int \Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) dF(\tilde{z})\bigg|_{\epsilon=0} \\ &+ \frac{\partial}{\partial\epsilon}\epsilon\int \Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S\left(F_\epsilon\right)\right), T\left(F_\epsilon\right)\right) d\Delta_z\bigg|_{\epsilon=0} = 0. \end{aligned}$$

The second terms give the point mass distribution, and the first term can be obtained through implicit differentiation with respect to the functionals $S(F)$ and $T(F)$. Arrange terms gives the following general equation for the influence function of the second stage estimator:

$$\begin{aligned} \mathrm{IF}(z; T, F) = M^{-1}(&\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right) \\ &+ \int \frac{\partial}{\partial\theta}\Psi_2\left(\tilde{z}^{(2)}; \theta, T(F)\right)\frac{\partial}{\partial\eta}h\left(\tilde{z}^{(1)}; \eta\right) dF(\tilde{z}) \cdot \mathrm{IF}(z; S, F)), \end{aligned}$$

where $M = -\int \frac{\partial}{\partial \xi} \Psi_2\left(\tilde{z}^{(2)}; h\left(\tilde{z}^{(1)}; S(F)\right), \xi\right) dF(\tilde{z})$.

It can be observed that the above formula does not account for the complexity of the estimator, but the method remains applicable to our proposed two-stage estimator. The formula consists of three parts. The multiplier M outside the parentheses is the second derivative common to the influence function for score equations. The first term within the parentheses denotes the point mass at the observed point. The second term represents the differentiation with respect to ϵ through the functional of the first stage, where the chain rule starts from the function $h(\cdot;\cdot)$ to the functional $S(F)$, and ultimately involves the influence function of the first stage estimator.

Given this decomposition, the influence function proposed without considering influence from the first stage would be $M^{-1}(\Psi_2\left(z^{(2)}; h\left(z^{(1)}; S(F)\right), T(F)\right)$. Such simple terms differ from our previous result of the censoring robust estimator given longitudinal covariate values. This simplified influence function may arise from a simple MLE in the second-stage model. However, the influence function can become more complex after including the influence from factors such as partial likelihood and other potential variations.

Nevertheless, given the influence function of the second-stage model without considering the variation from the first stage, the modification to include the variation from the first-stage model is similar. It is expected that there will be additional terms corresponding to the partial differential with respect to the functionals of the first stage. When the dependence over the first-stage model relies on single function $h(\cdot;\cdot)$, the last term within the parentheses can be added for direct application. For dependencies on multiple functions, the chain rule can be further applied, so that if the class of the two-stage models is defined as:

$$\begin{aligned} &E_F\left[\Psi_1\left(z^{(1)}; S(F)\right)\right] = 0 \\ &E_F\left[\Psi_2\left(z^{(2)}; h_1\left(z^{(1)}; S(F)\right), \ldots, h_k\left(z^{(1)}; S(F)\right), T(F)\right)\right] = 0 \end{aligned},$$

where there are k continuously piecewise differentiable functions defined similarly. Then the influence function of the second stage estimator can be represented as:

$$\begin{aligned}\mathrm{IF}(z;T,F) = M^{-1}(&\Psi_2\left(z^{(2)};h\left(z^{(1)};S(F)\right),T(F)\right)\\&+\sum_{j=1}^{k}\int\frac{\partial}{\partial\theta}\Psi_2\left(\tilde{z}^{(2)};\theta,T(F)\right)\frac{\partial}{\partial\eta}h_j\left(\tilde{z}^{(1)};\eta\right)dF(\tilde{z})\cdot\mathrm{IF}(z;S,F)).\end{aligned}$$

Simple linear regression

To derive the IF for our two-stage model, it's beneficial to begin with the IF for the simple regression, which involves a single-variable linear regression with a slope and an intercept. Understanding this derivation will provide insights into extending it to the LME model used in the first stage of our two-stage model. In this case, $\beta = T(F) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$. Given the infinitesimal representation, the chain rule gives the influence function of the regression estimator of SLR:

$$\begin{aligned}\psi_{\hat{\theta}}(x,y) &= \frac{(x-\mathrm{E}[X])(y-\mathrm{E}[Y])-\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} - \frac{((x-\mathrm{E}[X])^2-\mathrm{Var}(X))\,\mathrm{Cov}(X,Y)}{(\mathrm{Var}(X))^2}\\&= \frac{(x-\mathrm{E}[X])(y-\mathrm{E}[Y])-\beta(x-\mathrm{E}[X])^2}{\mathrm{Var}(X)}\\&= \frac{(x-\mathrm{E}[X])}{\mathrm{Var}(X)}[(y-\mathrm{E}[Y])-\beta(x-\mathrm{E}[X])]\end{aligned}$$

Multiple linear regression

In multiple regression, it is assumed that

$$y_i = X\beta + \epsilon_i,$$

where instead of single covariate x, $X = (x_1,\ldots,x_k)$, and that $\beta = (\beta_1,\ldots,\beta_k)$. When adjusting for multiple covariates, we cannot simplify to the previous formulation as the covariance between y and $(x_1, \ldots, x_k)$ is not defined. To derive the influence function under multiple regression, the derivative of the functional β would suffice. To estimate β for multiple regression, it is common to adopt the least square method or the maximum likelihood method. For the estimation of β, both methods result in the same target function, and the following will proceed with the maximum likelihood approach.

For MLE of multiple linear regression, we seek to minimize the following objective function:

$$\sum_{i=1}^{n}(y_i - X_i\beta)^2.$$

Thus, it is an M-estimator with $G(X, y, \beta) = (y - X\beta)^2$ as the infinitesimal representation of objective function is $E[y - X\beta^2]$. We can further obtain that

$$\begin{aligned} g(X, y, \beta) &= \nabla_\beta G(X, y, \beta) = -2X'(y - X\beta) \\ \nabla_\beta g(X, y, \beta) &= 2X'X \end{aligned}$$

Using the previous formula, we obtained that

$$\text{IF}(X, y; \beta, F) = -E[\nabla_\beta g(X, y, \beta)]^{-1} g(X, y, \beta) = E[X'X]^{-1}X'(y - X\beta).$$

Weighted multiple linear regression

When the homoscedasticity assumption doesn't hold, there is a variation of variance. It is common to assume the same formulation of y_i, but ϵ doesn't follow the identical normal distribution. With the presence of heteroscedasticity, the original estimator would still be unbiased but not the best linear predictor. Instead, according to the Gauss-Markov theorem, the best unbiased linear estimator would be the original predictor weighted by the inverse of the variance. Define w_i be the inverse of $\text{Var}(y_i) = \text{Var}(\epsilon_i)$, then the modified objective function is

$$\sum_{i=1}^{n} w_i(y_i - X_i\beta)^2.$$

Using similar techniques, we find that

$$\begin{aligned}
&G(X, y, w, \beta) = w(y - X\beta)^2 \\
&g(X, y, w, \beta) = \nabla_\beta G(X, y, w, \beta) = -2X'w(y - X\beta) \\
&\nabla_\beta g(X, y, \beta) = 2X'wX
\end{aligned}$$

The influence function of weighted multiple linear regression is obtained as

$$\text{IF}(X, y; \beta, F) = -E[\nabla_\beta g(X, y, \beta)]^{-1} g(X, y, \beta) = E[X'wX]^{-1} X'w(y - X\beta),$$

where $E[X'wX] = 1/n \sum_{i=1}^{n} X_i' w_i X_i$. This result is similar to the one obtained in Jann (2019), and the only difference would be the degree-of-freedom correction term $\sqrt{\frac{n-1}{n-k-1}}$, where k represents the number of covariates. Since the influence function is an infinitesimal approach, the focus would be on a situation with a large number of observations. As the number of observations increases, the correction term will approach 1 quickly. Thus, the effect of ignoring the correction terms would only be significant with an insufficient degree of freedom.

Influence function for the linear mixed effects model

The influence function for the full LME model is quite complex to derive. Assuming the general format that

$$
\begin{aligned}
y &= X\beta + Zu + \epsilon, \\
u &\sim N(0, G), \\
\epsilon &\sim N(0, \Sigma), \\
V &= \Sigma + Z'GZ.
\end{aligned}
$$

Then the distribution of y has a density function given by:

$$p(y|\beta, V) = (\frac{1}{2\pi})^{-\frac{n}{2}}|V|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\right\},$$

so the likelihood and the log-likelihood functions can be expressed respectively as

$$L(\beta, V) \propto |V|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\right\},$$

and

$$l(\beta, V) \propto -\frac{1}{2}|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta).$$

For parameter estimation, there is no closed-form solution. Assuming V is known, then taking the first derivative of log-likelihood with respect to β and setting it to zero gives the solution as

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y.$$

To estimate the random effect, we can use the joint distribution of y and u as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{ZG} \\ \mathbf{GZ}^{\mathrm{T}} & \mathbf{G} \end{bmatrix} \right).$$

Given information about β and V, then an estimate of random effect can be obtained using the property of conditional expectation of multivariate normal distribution.

$$\begin{aligned} \mathrm{E}[\mathbf{u} \mid \mathbf{y}] &= \mathrm{E}[\mathbf{u}] + \mathrm{Cov}\left[\mathbf{u}, \mathbf{y}^{\mathrm{T}}\right] \mathrm{Var}^{-1}[\mathbf{y}](\mathbf{y} - \mathrm{E}[\mathbf{y}]) \\ &= \mathbf{GZ}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{GZ}^{\mathrm{T}}\left(\mathbf{ZGZ}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned}$$

then replace fixed effect β by estimator$\hat{\beta}$ to have the prediction

$$\hat{\mathbf{u}} = \mathbf{GZ}^{\mathrm{T}}\left(\mathbf{ZGZ}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

It can be seen that the estimators of β and u both depend on V. Currently, solutions to variance components still rely on the iterative or numerical method. So far, all the derivation indicates the complexity of the functional. Due to iterative estimation, the influence function becomes even harder to derive as it includes all possible sources of variation. To reduce the complexity, we propose to assume V is known by substituting it with the estimates. We argue that the change in influence function would be minimal as the fluctuation of the β estimate resulting from the misspecified variance-covariance matrix is small compared to that from the misspecified mean structure. A simulation has been done to assess the impact of the misspecified variance component on fixed effect estimates, and it is found that the fixed effect coverage remains stable with a moderate number of observations, whereas the coverage probability deviates far from standard as soon as the mean form is misspecified.

Given $\hat{V}$ as true V, we returned to the multivariate case of weighted regression, where the objective function is:

$$\sum_{i=1}^{n}(y_i - X_i\beta)V_i^{-1}(y_i - X_i\beta),$$

giving the derivation to the influence function of β as

$$G(X, y, V, \beta) = (y - X\beta)V^{-1}(y - X\beta),$$
$$g(X, y, V, \beta) = \nabla_\beta G(X, y, V, \beta) = -2X'V^{-1}(y - X\beta),$$
$$\nabla_\beta g(X, y, \beta) = 2X'V^{-1}X,$$

and thus

$$\text{IF}(X, y; \beta, F) = -E[\nabla_\beta g(X, y, \beta)]^{-1}g(X, y, \beta) = E[X'V^{-1}X]^{-1}X'V^{-1}(y - X\beta).$$

Notice that even though X_i's are of various dimensions due to the number of observations and that even if the number of observations is the same for two subjects, the data may not be comparable as the measurement time may differ. Still, the $X'V^{-1}X$ may be recognized as leverage standardized by variance at each measurement so that it is still of dimension $p \times p$, where p is the number of fixed covariates. Then it acts as the weight for standardized covariate value $X'V^{-1}$ to assess the influence of a specific subject.

4.2.6 Influence function for censoring robust estimators under conditional independent censoring with a predicted longitudinal covariate

With the addition of the predicted longitudinal covariate, it is expected that the influence function would have additional terms corresponding to fluctuations in predicted values. For derivation, the implicit differentiation indicates the necessity of deriving the influence function for the predicted longitudinal covariate value as the second stage model only depends on the first stage model through the predicted longitudinal covariates. For an arbitrary predicted longitudinal covariate value for subject j, we define the covariate $X^{(i)}$ to indicate the fixed effect covariate for prediction at event time i corresponding to ith subject and let $Z^{(i)}$ to indicate the random effect covariate for prediction at event time i corresponding to ith subject. Then the predicted value at the event time would be

$$
\begin{aligned}
\hat{Z}_i(X_j) &= X_i^{(1)}(X_j)\beta + Z_i^{(1)}(X_j)\hat{u}_i \\
&= X_i^{(1)}(X_j)\boldsymbol{\beta} + Z_i^{(1)}(X_j)\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}}\left(\mathbf{Z^{(1)}}\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}(\mathbf{y} - \mathbf{X^{(1)}}\boldsymbol{\beta}) \\
&= \left\{X_i^{(1)}(X_j) - Z_i^{(1)}(X_j)\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}}\left(\mathbf{Z^{(1)}}\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{X^{(1)}}\right\}\boldsymbol{\beta} \\
&\quad + Z_i^{(1)}(X_j)\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}}\left(\mathbf{Z^{(1)}}\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{y},
\end{aligned}
$$

where only the first term is associated with β in the first stage. Thus, by vector differentiation, the implicit differentiation down through $\hat{Z}_{ij}^{(2)}$ would be

$$
\begin{aligned}
\frac{\hat{Z}_{ij}^{(2)}}{\partial\epsilon} &= \frac{\partial\hat{Z}_{ij}^{(2)}}{\partial\beta} \times \frac{\partial\beta}{\partial\epsilon} \\
&= \left\{X_i^{(2)}(X_j) - Z_i^{(2)}(X_j)\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}}\left(\mathbf{Z^{(1)}}\mathbf{G}\mathbf{Z^{(1)}}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{X^{(1)}}\right\} \times \mathrm{IF}(X, y; \beta, F).
\end{aligned}
$$

To reduce redundancy in representation, we focus on the additional terms corresponding to influence from the first stage model. From the previous derivation, it is expected that the additional term can be obtained through the partial derivative of integrand through the predicted longitudinal covariate values at event time.

$$
\frac{\partial}{\partial H_\epsilon}\Psi^{**}(\beta, w, H_\epsilon) = \frac{\partial}{\partial H_\epsilon}\left\{V(x)^{-1}\tilde{z} - \frac{\int S_{\tilde{z}}(\tilde{x})^{-1}\tilde{z}V(x)^{-1}e^{V(x)^{-1}\tilde{z}\beta}\mathrm{I}\{\tilde{x} \geq x\}dH(\tilde{x}, \tilde{z})\}}{\int S_{\tilde{z}}(\tilde{x})^{-1}e^{V(x)^{-1}\tilde{z}\beta}\mathrm{I}\{\tilde{x} \geq x\}dH(\tilde{x}, \tilde{z})\}}\right\}.
$$

In the above equation, z represents the observed covariate value at event time, while $\tilde{z}$ represents the predicted longitudinal covariate values at event time. It is commonly assumed that the exact event time is known to avoid further complications. However, obtaining the longitudinal covariate value at event time is rare in practice, so it is typically predicted from the fitted LME model. If only $\tilde{z}$ values involve imputations through the LME model, then the partial derivative only applies to the $\tilde{z}$ values for the expected covariate value at event time.

It's important to note that the censoring weight estimator $S(x_j)$ doesn't contain influence through predicted longitudinal covariate values, as it only uses the observed time and status. However, the conditional variance estimator $\hat{V}(x_j)$ involves influence from the ith subject through the predicted values in the risk set. Since $V(x_j)$ follows $\tilde{z}_j$ at every occurrence, pre-calculating $IF(x_i; V(x_j))$ in advance can reduce computation complexity.

$$\begin{aligned}
IF(x_i; \hat{V}(x_j)) &= \frac{\partial \hat{V}(x_j)}{\partial \epsilon} \\
&= \frac{\partial}{\partial \epsilon}\{E(Z^2|T) - [E(Z|T)]^2\} \\
&= \frac{\partial}{\partial \epsilon}\{\frac{1}{n}\sum z^2(x_j) - [\frac{1}{n}\sum z(x_j)]^2\} \\
&= \frac{2}{n}\sum z(x_j)\frac{\partial z(x_j)}{\partial \epsilon} - \frac{2}{n}\sum z(x_j) \times \frac{1}{n}\sum \frac{\partial z(x_j)}{\partial \epsilon} \\
&= \frac{2}{n}\left\{\sum \left[z(x_j) - \frac{1}{n}\sum z(x_j)\right]\frac{\partial z(x_j)}{\partial \epsilon}\right\}.
\end{aligned}$$

Then the additional terms related to influence from first stage model prediction are as follows

$$\begin{aligned}
\frac{\partial}{\partial H_\epsilon}\Psi^{**}(\beta, w, H_\epsilon) =& IF_3 \\
& - \frac{1}{s^{(0)}_{WV}(\tilde{x}, \beta)}\bigg\{\int S_{\tilde{z}}(x)^{-1} IF_3 \exp(\beta' z^*(x)) \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z}) \\
& + \int S_{\tilde{z}}(x)^{-1} z^*(x) \exp(\beta' z^*(x)) \beta' IF_3 \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z}) \\
& - \frac{s^{(1)}_{WV}(\tilde{x}, \beta) \int S_{\tilde{z}}(x)^{-1} \exp(\beta' z^*(x)) \beta' IF_3 \mathrm{I}\{\tilde{x} \geq x\} dH(\tilde{x}, \tilde{z})}{s^{(0)}_{WV}(\tilde{x}, \beta)}\bigg\},
\end{aligned}$$

where

$$IF_3 = \begin{pmatrix} 0 \\ V^{-1}(x)\frac{\partial z_L(x)}{\partial \epsilon} - V^{-2}(x) z_L(x) \frac{\partial V(x)}{\partial \epsilon} \end{pmatrix}.$$

When predicted longitudinal covariates at event time are not treated as provided, the influence function contribution from the previous section can also be updated. As predicted longitudinal covariates at event time can deviate from the true distribution of longitudinal covariates at event time, the influence function contributed from the inverse of variance would have additional terms including the fluctuation of the first-stage model. The influence function from the inverse of variance needs to be updated since predicted covariate values depend on the first-stage model, whereas the Kaplan-Meier estimate only takes inputs of the observed times and thus possesses no dependence on the first-stage model.

Substituting $\hat{\beta}$ and the empirical distribution function into the expression for the influence function, we obtain the empirical influence function. The empirical influence funciton evaluated at (x_i, δ_i, z_i) can be written

$$I_i = \hat{A}(\hat{\beta})^{-1} \times \left[\delta_i \hat{S}(x_i)^{-1} \left\{ z_i \hat{V}^{-1}(x_i) - \frac{S^{(1)}_{WV}(x_i, \hat{\beta})}{S^{(0)}_{WV}(x_i, \hat{\beta})} \right\} - D_i(\hat{\beta}) \right],$$

where

$$
\begin{aligned}
D_i(\hat{\beta}) = \sum_{j=1}^{n} \delta_j \hat{S}(x_j)^{-1} \times \Bigg[& \frac{e^{\beta' z_i^*(x_i)} \mathrm{I}(x_j \leq x_i)}{S_{WV}^{(0)}(x_j, \hat{\beta})} \left\{ z_i^*(x_j) - \frac{S_{WV}^{(1)}(x_j, \hat{\beta})}{S_{WV}^{(0)}(x_j, \hat{\beta})} \right\} \\
& + \left\{ z_j^*(x_j) - \frac{S_{WV}^{(1)}(x_j, \hat{\beta})}{S_{WV}^{(0)}(x_j, \hat{\beta})} \right\} \times G_{ij}(\hat{\beta}) \\
& - \frac{G_{ij}(\hat{\beta}) S_{z_k}^{-1}(x_j)}{S_{WV}^{(0)}(x_j, \hat{\beta})} \Bigg\{ \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} z_k^*(x_j) e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \geq x_j) \\
& \quad - \frac{S_W^{(1)}(x_j, \hat{\beta}) \sum_{k=1}^{n_{z_k}} \frac{G_i^k}{n_{z_k}} e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \geq x_j)}{S_W^{(0)}(x_j, \hat{\beta})} \Bigg\} \\
& + \begin{pmatrix} 0 \\ \\ z_{jL}^*(x_j) \end{pmatrix} \hat{V}^{-1}(x_j) * I_i\{V(x_j)\} \\
& - \frac{I_i\{V(x_j)\} \hat{V}^{-1}(x_j)}{S_{WV}^{(0)}(x_j, \hat{\beta})} \Bigg\{ \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} e^{\hat{\beta}' z_k^*(x_j)} \mathrm{I}(x_k \geq x_j) \\
& + \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) z_k^*(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} \mathrm{I}(x_k \geq x_j) \\
& - \frac{S_{WV}^{(1)}(x_j, \hat{\beta}) \times \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \begin{pmatrix} 0 \\ \\ z_{kL}^*(x_j) \end{pmatrix} \mathrm{I}(x_k \geq x_j)}{S_{WV}^{(0)}(x_j, \hat{\beta})} \Bigg\} \\
& - \hat{IF}_{ij}^{(3)}(x_j) \\
& + \frac{1}{S_{WV}^{(0)}(x_j, \hat{\beta})} \Bigg\{ \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{IF}_{ik}^{(3)}(x_j) \mathrm{I}(x_k \geq x_j) \\
& + \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) z_k^*(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \hat{IF}_{ik}^{(3)}(x_j) \mathrm{I}(x_k \geq x_j) \\
& - \frac{S_{WV}^{(1)}(x_j, \hat{\beta}) \times \sum_{k=1}^{n_{z_k}} \frac{1}{n_{z_k}} S_{z_k}^{-1}(x_j) e^{\hat{\beta}' z_k^*(x_j)} \hat{\beta}' \hat{IF}_{ik}^{(3)}(x_j) \mathrm{I}(x_k \geq x_j)}{S_{WV}^{(0)}(x_j, \hat{\beta})} \Bigg\} \Bigg],
\end{aligned}
$$

with

$$G_{ij}(\hat{\beta}) = \sum_{x_l \leq (x_i \wedge x_j)} \frac{G_l^j(1-\delta_l)}{n_l^c S^{(0)}(x_l, 0)^2} - \frac{G_i^j(1-\delta_i)\mathrm{I}\{x_i \leq x_j\}}{S^{(0)}(x_j, 0)},$$

$$\hat{IF}_{ik}^{(3)}(x_j) = \begin{pmatrix} 0 \\ \hat{V}^{-1}(x_j)\left[\frac{\partial z_{kL}(x_j)}{\partial \epsilon} - z_{kL}^*(x_j)\frac{\partial \hat{V}_i(x_j)}{\partial \epsilon}\right] \end{pmatrix},$$

$$\frac{\partial \hat{V}_i(x_j)}{\partial \epsilon} = \sum_{k=1}^{n_{z_k}} \frac{2}{n_{z_k}}\left[z_k(x_j) - \sum_{k=1}^{n_{z_k}} \frac{z_k(x_j)}{n_{z_k}}\right]\frac{\partial z_{kL}(x_j)}{\partial \epsilon},$$

$$\frac{\partial z_{kL}(x_j)}{\partial \epsilon} = \left\{X_k^{(2)}(x_j) - Z_k^{(2)}(x_j)G_k Z_k^{(1)^T}\left(Z_k^{(1)}G_k Z_k^{(1)^T} + \Sigma_k\right)^{-1} X_k^{(1)}\right\} \times \hat{IF}(x_i, \hat{\beta}),$$

$$\hat{IF}(x_i, \hat{\beta}) = \left[\sum_{k=1}^{n_{z_k}} X_k^{(1)^T}\left(Z_k^{(1)}G_k Z_k^{(1)^T} + \Sigma_k\right)^{-1} X_k^{(1)}\right]^{-1} \times$$
$$X_i^{(1)^T}\left(Z_i^{(1)}G_i Z_i^{(1)^T} + \Sigma_i\right)^{-1}(y_i - X_i^{(1)}\hat{\beta}).$$

4.2.7 Influence calculation

The computational efficiency of the R programming language may be compromised when executing complex or nested loop functions, leading to longer running times. Therefore, computation is conducted using RcppArmadillo, which integrates C++ within the R environment. Through simulation, it has been observed that the average execution duration has significantly decreased from approximately two hours to just 10.31 seconds, compared to the R *forloop*.

4.3 Numerical studies

4.3.1 Simulation setup

In this section, we compare the performances of the three variance estimators: (i)the naive variance estimator using Fisher's information V_{cox} by Cox (1972); (ii) the robust variance estimator V_{Robust} as in Boyd et al. (2012); Nguyen and Gillen (2017) modified from Lin and Wei (1989);(iii) the proposed robust variance estimator V_{IF}. The application of all three variance estimators will be utilized in the context of the censoring robust estimator within the framework of joint modeling of survival and longitudinal data, and the simulation setup will be replicated from the previous paper in order to maintain consistency. In accordance with the established framework, three distinct scenarios have been devised to facilitate a comparative analysis of the performance exhibited by the three variance estimators. For the covariate setting, we are interested in quantifying the association between the time to the event and longitudinal covariate $Z_1 \in \mathbf{R}$ for each scenario. Subsequently, β_1 shall serve as the parameter corresponding to our evaluations and the primary focus of our analysis and only a confounding covariate $Z_2 \sim Bernoulli(0.4)$ was adjusted to simulate a general regression situation. The covariates are generated such that

$$
\begin{aligned}
Z_2 &\sim \text{Bin}(1, 0.4),\\
Z_1(t)|b &\sim N(f(t) + b_1 + b_2 * t^2, v_{error} * I_D),\\
\begin{bmatrix} b_1 \\ b2 \end{bmatrix} = b &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} v_{inter} & 0 \\ 0 & v_{slope} \end{bmatrix}\right),
\end{aligned}
$$

where Z_2 is generated through $\text{Bin}(1, 0.4)$ and $Z_1(t)$ is generated using underlying longitudinal trajectories with the addition of independent random intercept with $v_{inter} = 0.01$, random slope with $v_{slope} = 0.08$ and normal measurement error with $v_{error} = 0.02$. The underlying longitudinal trajectories without random effect follows the function

$$f(t) = \begin{cases} -0.03956718 * t^2 + 0.2942684 * t + 0.343545, & \text{for Group1}, \\ -0.026718 * t^2 + 0.142684 * t + 0.603545, & \text{for Group2}, \end{cases}$$

as depicted in Figure 3.2. For comparison, we evaluate the variance estimators based on the mean of empirical standard error and coverage probability of 95 % confidence intervals under each scenario. In each given scenario, it is assumed that there exist two longitudinal subgroups, with respective probabilities of $(0.4, 0.6)$. These subgroups adhere to an underlying longitudinal pattern, wherein the group with a higher initial biomarker value typically exhibits a slower rate of increase, while the group with a lower initial biomarker value generally experiences a faster rate of increase. The three scenarios we considered correspond to the proportional hazard model, the non-proportional hazard model with constant variance longitudinal covariate, and the non-proportional hazard model, where the longitudinal covariate varies with respect to time. The first scenario is the proportional hazard model, where we generate survival time according to

$$\lambda(t|Z) = \exp\{-0.8 \times Z_1(t) - 0.4 \times Z_2\}.$$

The second scenario is the non-proportional hazard model with constant variance longitudinal covariate, where we generate the survival time according to

$$\lambda(t|Z) = \exp\{\log(0.3) I\{t > 1\} \times Z_1(t) - 0.1 \times Z_2\}.$$

The third scenario is the non-proportional hazard model with longitudinal covariate variance varying over time. The survival time is generated based on the same hazard as the second scenario, except that a random slope is incorporated into the longitudinal covariate. To examine the impact of censoring distribution on estimators, three censoring scenarios are employed to generate the censoring time for each scenario. The censoring scenarios differ as follows : censoring based on a categorical covariate (case 1), censoring based on a longitudinal covariate (case 2), and censoring based on the interaction between covariates (case 3). The censoring time is generated through power function distribution with a maximum follow-up time of 4 years. For replicable results, the data-generating scenario and the censoring scenario were each run 1000 times.

For non-proportional hazard, due to a lack of an analytic expression for $\int_0^\infty \beta_1(t)dF(t)$, we approximate with the Monte Carlo average of β_{Cox} with administrative censoring at $\tau = 4$) for $n = 2000$. For the longitudinal covariate, the LME model incorporates the desired variance structure. In the first two scenarios, we assumed a random intercept following $\mathbf{N}(0, 0.08)$. In the last scenario, we add an independent random slope following $\mathbf{N}(0, 0.04)$ and assume the independence between the random intercept and random slope for simplicity. Finally, we assume that our study's measurement error is identically and independently distributed, following $\mathbf{N}(0, 0.02)$ throughout the study. For NPH cases, survival time was generated by discretizing continuous time to small intervals of 0.01 years using the memoryless property of exponential distribution.

4.3.2 Simulation results

Table 4.1: Results of the simulation study with random intercept and random slope under non-proportional hazard, $n = 800$

		Naive Estimator		Regular Robust Variance		Influence Based Variance	
Scenario	Mean	$\hat{V}_{\text{cox}}$		$\hat{V}_{\text{Robust}}$		$\hat{V}_{\text{Inf}}$	
	(Truth:-0.395)	ESE	CP	ESE	CP	ESE	CP
C:1	-0.396	0.469	0.90	0.454	0.89	0.534	0.95
C:2	-0.392	0.471	0.92	0.456	0.90	0.530	0.94
C:3	-0.398	0.474	0.92	0.459	0.91	0.530	0.97

The subsequent sections of this subsection are dedicated to the comparative analysis of the performances exhibited by the three estimators. Given the objective of developing a robust

Table 4.2: Results of the simulation study with random intercept under non-proportional hazard, n=800

Scenario	Mean	Naive Estimator $\hat{V}_{\text{cox}}$		Regular Robust Variance $\hat{V}_{\text{Robust}}$		Influence Based Variance $\hat{V}_{\text{Inf}}$	
	(Truth:-0.381)	ESE	CP	ESE	CP	ESE	CP
C:1	-0.379	0.432	0.90	0.445	0.88	0.598	0.96
C:2	-0.380	0.434	0.89	0.433	0.87	0.561	0.92
C:3	-0.422	0.474	0.93	0.459	0.91	0.530	0.97

Table 4.3: Results of the simulation study under proportional hazard, n=800

Scenario	Mean	Naive Estimator $\hat{V}_{\text{cox}}$		Regular Robust Variance $\hat{V}_{\text{Robust}}$		Influence Based Variance $\hat{V}_{\text{Inf}}$	
	(Truth:-0.800)	ESE	CP	ESE	CP	ESE	CP
C:1	-0.812	0.474	0.86	0.459	0.89	0.530	0.95
C:2	-0.838	0.420	0.88	0.446	0.90	0.614	0.94
C:3	-0.823	0.445	0.87	0.463	0.88	0.543	0.95

variance estimator, it is natural to evaluate the three estimators by examining their coverage probability.

We initiate our comparison by assessing the influence-based robust variance estimator ($\hat{V}_{IF}$) against the bootstrap technique, following the approach of prior research. Our examination reveals that both variance estimators displayed similar magnitudes for the mean standard error across all scenario combinations. However, when it came to coverage probability, $\hat{V}_{IF}$ outperforms the bootstrap method by consistently achieving higher coverage probabilities, tightly centered around the target value of 0.95 across all scenarios.

In contrast, the preceding study, which evaluates the coverage probabilities of the bootstrap variance estimator in the context of NPH with random slopes, reports slightly lower coverage probabilities ranging from 0.92 to 0.94 across all censoring cases. However, our analysis of $\hat{V}_{IF}$ demonstrate notably improved and more consistent coverage probabilities, ranging from 0.94 to 0.97 across the three censoring scenarios. This enhanced consistency in achieving the target coverage probability of 0.95 underscores the robustness of $\hat{V}_{IF}$ across diverse scenarios.

Before delving into the comparison involving $\hat{V}_{IF}$ and the other two variance estimators, it's essential to elucidate the disparity between the naive estimator $\hat{V}_{cox}$ and the standard robust variance estimator $\hat{V}_{Robust}$. Previous investigations (Nguyen and Gillen, 2017; Boyd et al., 2012) have indicated that $\hat{V}_{IF}$ typically yields higher estimates compared to $\hat{V}_{cox}$. However, our specific analysis reveals this trend primarily in the proportional hazard model scenario. In cases where the hazard coefficient varies, we observed that $\hat{V}_{Robust}$ tends to yield smaller estimates than $\hat{V}_{cox}$.

This discrepancy in the magnitudes of $\hat{V}_{Robust}$ and $\hat{V}_{cox}$ hinges on the correlation between the parameter and the covariate. When this relationship remains stable, the inclusion of additional observations may not proportionally augment statistical information. Consequently, the naive variance estimator may overestimate the information contribution, leading to an underestimation of the actual underlying variance.

In the context of NPH, where the instantaneous parameter is solely influenced by the covariate value at the event time, the statistical information provided is typically greater compared to the proportional hazards case. Hence, it's anticipated that $\hat{V}_{Robust}$ would be lower than $\hat{V}_{cox}$ in scenarios employing the NPH model with either only a random intercept or with a random slope. Conversely, in the case of the PH model, a higher estimated variance is expected.

It's noteworthy that $\hat{V}_{Robust}$ is akin to the variance estimator obtained by assuming that the subject's influence is confined to partial correlation. It's presumed that the population parameter encapsulates the predicted longitudinal covariate value, the estimated censoring weight, and the estimated conditional covariate variance, all considered independent of individual influences.

In our evaluation of the proposed variance estimator $\hat{V}_{IF}$, we observed notable enhancements in coverage probability and a noticeable increase in the average estimated standard error across nine different scenarios. The coverage probabilities achieved using $\hat{V}_{IF}$ were consistently close to 0.95 across all three scenarios. Notably, the scenario featuring proportional hazards exhibited the least variability in coverage, whereas the scenario with non-proportional hazards and only a random intercept displayed the highest variability. This variability can be attributed to the challenge of distinguishing between the variance stemming from the random intercept and that resulting from measurement error. Despite some slight variation observed within individual variance estimators, it is evident that $\hat{V}_{Inf}$ demonstrated substantial improvement compared to the other two estimators.

The coverage of $\hat{V}_{Cox}$ was centered around 0.91, while that of $\hat{V}_{Robust}$ was centered around 0.89. It's important to emphasize that $\hat{V}_{Robust}$ is essentially equivalent to the proposed variance estimator, with the exception of potential variations in censoring weight estimation, covariate prediction, and variance estimation.

The average estimated standard error of $\hat{V}_{IF}$ was observed to be 18% higher than that of $\hat{V}_{Robust}$. This increase primarily stems from the considerations mentioned above. The significance of accounting for the potential impact of these variations is apparent and is reflected in the coverage probability.

4.4 Application

In this section, we apply the proposed variance estimator to ADNI data to assess the association between longitudinal cortical thickness and progression to AD compared to the previous study using the bootstrap variance estimator. A detailed introduction to the dataset can be found in Section 3.4. In Table 4.4, the outcomes obtained from the proposed robust variance estimator generally align with those from the bootstrap variance estimator.

Table 4.4: Association with AD progression: model results (controlled for age, APOE in the longitudinal model and controlled for age, gender, years of education, and APOE in the survival model)

				Influence Variance Estimator			Bootstrap Variance Estimator		
Variable	Participants	Events	Estimate	SE	95% CI	$\|Z\|$ value	SE	95% CI	$\|Z\|$ value
MidTemp*	532	151	-3.01	0.39	(-3.77,-2.25)	9.58	0.41	(-3.75,-2.27)	7.72
RAVLT.learning	622	177	-0.64	0.076	(-0.79,-0.49)	8.39	0.083	(-0.81,-0.48)	7.68
Hippocampus*	558	156	-8.19	1.20	(-10.54,-5.84)	6.83	1.15	(-10.20,-6.13)	7.14
Entorhinal*	532	151	-9.99	1.53	(-12.99,-6.99)	6.53	1.40	(-12.73,-7.26)	7.16
RAVLT.immediate	622	177	-0.14	0.024	(-0.19,-0.09)	5.83	0.023	(-0.18,-0.09)	6.01
CDRSB	622	177	0.78	0.14	(0.50,1.06)	5.48	0.12	(0.55,0.99)	6.78
Fusiform*	532	151	-3.22	0.60	(-4.40,-2.04)	5.36	0.57	(-4.28,-2.16)	5.63
ADAS13	622	177	0.021	0.0040	(0.013,0.029)	5.32	0.0041	(0.013,0.029)	5.35
FAQ	622	177	0.30	0.060	(0.18,0.42)	4.97	0.067	(0.17,0.43)	4.45
Ventricles*	574	173	0.19	0.04	(0.10,0.27)	4.75	0.04	(0.11,0.26)	4.76
ADAS11	622	177	0.15	0.036	(0.08,0.22)	4.17	0.033	(0.09,0.22)	4.57
FDG	287	81	-8.76	2.46	(-13.58,3.92)	4.03	2.38	(-13.44,-4.09)	3.68
WholeBrain*	588	175	-0.065	0.017	(-0.098,-0.032)	3.83	0.015	(-0.095,-0.035)	4.25
MMSE	622	177	-0.27	0.055	(-0.38,-0.16)	2.49	0.055	(-0.38,-0.16)	2.50
RAVLT.forgetting	622	177	0.18	0.14	(-0.0944,0.4544)	1.29	0.09	(-0.0043,0.3678)	1.91
ICV*	602	176	0.005	0.011	(-0.021,0.022)	0.45	0.009	(-0.014,0.025)	0.56

* variable value is multiplied by 10000 for the ease of displaying

4.5 Discussion

The current Chapter provides a comprehensive overview and comparison of existing methods for estimating robust variance in survival analysis, particularly in scenarios involving both adherence to and deviation from the proportional assumption. We observed a correspondence between the prevailing approaches when considering scenarios with categorical and continuous covariates exclusively. Additionally, we proposed an extension of the existing variance estimator to accommodate a censoring robust estimator while incorporating longitudinal covariates. Our simulation study results showcased that our proposed estimator demonstrated improved consistency in accurately estimating coverage probability. Moreover, it exhibited enhanced stability compared to the bootstrap method. The robust variance method displayed resilience against misspecification of the hazard function and potential errors in longitudinal prediction. Nonetheless, there remains a potential concern regarding the potential misclassification of longitudinal and censoring clusters. However, the infinitesimal approach demonstrates a semiparametric property that may confer robustness to our estimator in the face of misclassification, as evidenced by its superior performance in coverage probability relative to the bootstrap variance estimator.

Chapter 5

Review of longitudinal clustering method and comparison for application in monotone missing data

5.1 Introduction

Longitudinal clustering, a subset of clustering analysis, addresses the challenge of extracting meaningful insights from datasets characterized by temporal dependencies and evolving trajectories. Unlike traditional clustering methods that operate on static snapshots of data, longitudinal clustering considers the time dimension, enabling the identification of temporal patterns, trends, and subgroups within longitudinal datasets. Longitudinal datasets often exhibit intricate temporal dynamics and evolutionary patterns, where observations at different time points are interrelated and display dependencies over time. For instance, in healthcare data, patient trajectories may evolve as diseases progress or treatment regimens change. Longitudinal clustering facilitates the identification of meaningful patterns or groups of trajectories based on their temporal behavior or evolution (Golub et al., 1979).

By clustering trajectories over time, researchers can discern latent structures or subpopulations with similar longitudinal profiles. Longitudinal clustering finds applications across diverse fields, including healthcare, finance, ecology, and social sciences. In healthcare, longitudinal clustering aids in patient stratification, disease progression modeling, and treatment response prediction (Teuling et al., 2021). Unlike traditional clustering methods, which treat data as static entities, longitudinal clustering integrates time as a key variable. This integration allows for the detection of patterns that evolve over time or exhibit specific temporal characteristics. In summary, longitudinal clustering offers a powerful framework for uncovering temporal patterns and structures in longitudinal datasets, thereby enhancing our understanding of complex phenomena that evolve over time.

The necessity for longitudinal clustering is also pervasive within the realm of AD research. For instance, a study was conducted to compare the progression of memory, general cognition tasks, and functional scales over time between patients with behavioral-variant frontotemporal dementia (bvFTD) and AD (Schubert et al., 2016). Despite similar baseline performance, bvFTD patients exhibited a more rapid functional decline and greater cognitive deterioration compared to AD patients. This underscores the challenge of accurately distinguishing between these conditions based solely on neuropsychological profiles. Utilizing longitudinal clustering methods could offer a more precise means of categorizing patients, facilitating tailored care strategies that address the distinct needs of bvFTD and AD patients. In our proposed method, where we aim to derive a censoring-robust estimator with an interpretation akin to the average treatment effect even under non-proportional hazards (NPH), the necessity of proper longitudinal clustering becomes apparent. This is crucial for accurately estimating the conditional covariate variance. Studies have shown that in scenarios with multiple longitudinal groups, an erroneous longitudinal clustering approach can yield incorrectly estimated conditional covariate variance, ultimately leading to biased results based on the characteristics of longitudinal clustering.

The utilization of shape-based longitudinal clustering is imperative due to its ability to capture the nuanced variations in longitudinal trajectories over time. Traditional clustering methods often rely solely on endpoint measurements or summary statistics, which may overlook important temporal patterns in the data. Shape-based clustering, on the other hand, considers the entire longitudinal profile, enabling the detection of subtle changes and complex patterns in individual trajectories. This approach is particularly valuable in scenarios where the timing and rate of change in longitudinal measurements hold diagnostic or prognostic significance, such as in neurodegenerative diseases like Alzheimer's or in monitoring disease progression over time. Shape-based longitudinal clustering thus offers a more comprehensive and nuanced understanding of longitudinal data, making it essential for uncovering meaningful insights and informing personalized interventions or treatments.

In joint modeling of longitudinal and survival analysis, we encounter monotone missingness in longitudinal data resulting from censoring events. The challenge of monotone missingness in longitudinal clustering arises when data are systematically missing in a consistent pattern over time for each individual. This poses a significant hurdle in longitudinal data analysis because traditional methods for handling missing data may not be suitable due to their assumptions of randomness or ignorable missingness. Monotone missingness can introduce bias into parameter estimates and lead to incorrect cluster assignments if not appropriately addressed. When missing data are systematically related to the underlying longitudinal trajectories, failing to account for this pattern can result in biased clustering solutions and inaccurate inference.

In the upcoming section of this chapter, we will provide a comprehensive overview of current clustering algorithms designed for longitudinal data. Subsequently, we will delve into the challenges and potential solutions regarding the handling of monotone missing data within these methods. Following the discussion of existing clustering approaches, we will introduce a novel shape-based longitudinal clustering method tailored specifically to address imbalanced monotone missingness in longitudinal data. In the numerical study section, we will conduct simulations to compare the performance of each method across various scenarios. Furthermore, we will apply the algorithm to the ADNI dataset to demonstrate its practical utility. Finally, we will conclude with a brief discussion on potential future directions for longitudinal clustering methods dealing with monotone missingness.

5.2 Method

5.2.1 Review of longitudinal clustering methods

K-means

The k-means algorithm, a fundamental clustering technique, partitions data into k clusters based on their feature similarity. First proposed by MacQueen et al. (1967) and later refined by Lloyd (1982), it remains one of the most widely used clustering methods due to its simplicity and effectiveness.

At the core of the k-means algorithm is the minimization of the within-cluster sum of squares (WCSS), also known as inertia. Given a dataset X consisting of n observations and p features, and an initial set of k cluster centroids, the algorithm iteratively assigns each observation to the nearest centroid and updates the centroids to minimize the WCSS. This process continues until convergence, typically defined by either a maximum number of iterations or when the centroids no longer change significantly.

Despite its simplicity, k-means has shown remarkable performance in various applications, including image segmentation, customer segmentation, and anomaly detection. However, it has some limitations, such as sensitivity to initial centroid positions, dependence on the number of clusters specified (k), and the assumption of spherical clusters with equal variance.

Numerous extensions and variations of the k-means algorithm have been proposed to address these limitations and adapt it to specific data characteristics and applications. These include k-means++, which improves the selection of initial centroids, and k-medoids, which uses actual data points as centroids to enhance robustness to outliers.

In summary, the k-means algorithm serves as a cornerstone in the field of clustering, providing a simple yet powerful tool for partitioning data into coherent groups based on their similarity, with numerous applications across various domains.

KML

Longitduinal K-means (KML) is a variant of the traditional K-means clustering algorithm designed specifically for longitudinal data analysis. Introduced by Genolini and Falissard (2010), KML extends the standard K-means algorithm to handle longitudinal data, which consists of repeated measurements taken over time for each individual or subject.

The KML algorithm operates by clustering individuals based on the trajectories of their longitudinal measurements rather than on individual data points. It leverages the temporal structure inherent in longitudinal data to identify clusters of individuals with similar patterns of change over time.

Mathematically, the KML objective function can be formulated as follows:

$$\min_{C,Z} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{k} z_{itj} \|x_{it} - c_j\|^2$$

where $C = \{c_1, c_2, ..., c_k\}$ are the cluster centroids representing the average trajectory for each cluster, $Z = \{z_{itj}\}$ is an indicator matrix indicating the assignment of each individual i at time t to cluster j, x_{it} represents the longitudinal measurement for individual i at time t, k is the number of clusters, n is the number of individuals, and T is the number of time points.

The advantages of KML include its ability to capture complex longitudinal patterns and its flexibility in handling irregularly sampled longitudinal data. By clustering individuals based on their longitudinal trajectories, KML can identify distinct subgroups with similar patterns of change over time, facilitating the exploration of underlying trends and heterogeneity in longitudinal data. When it comes to defining cluster centroids, the conventional approach typically involves computing the average of subject covariate values at each time point j. However, there's an alternative method that incorporates kernel functions. By leveraging kernel functions, KML can capture nonlinear relationships and intricate patterns within longitudinal data, resulting in more precise and expressive cluster representations. This flexibility equips KML to handle diverse data distributions and effectively uncover the inherent structure of longitudinal trajectories.

However, KML also has limitations. The choice of distance metric and kernel function can influence the clustering results, and selecting appropriate parameters for these functions may require careful tuning. Additionally, KML may be sensitive to outliers and missing data, which can affect the quality of the clustering solution. KML typically requires complete data at balanced time points to format the data into a matrix, where each column represents the data for all subjects at a specific time. However, missing data can often occur at specific time points. In such cases, traditional methods are employed to handle the missing data. These methods include multiple imputation, last observation carried forward (LOCF), baseline observation carried forward (BOCF), mean substitution, and regression imputation. Multiple imputation involves imputing multiple sets of plausible values for missing data (Little and Rubin, 2019). LOCF involves carrying forward the last observed value to replace missing values at subsequent time points (Mallinckrodt et al., 2001). BOCF involves carrying forward the baseline (initial) observation to replace missing values at subsequent time points (Lachin, 2000). Mean substitution involves replacing missing values with the mean value of the variable across all participants at the same time point (Schafer, 1999). Regression imputation involves building a regression model based on observed values of other variables and using this model to predict missing values (Enders, 2022). The missing values are then filled in with either the mean of the predicted values or with random values generated from the conditional distribution of the model. These methods are particularly useful for addressing the challenge of missing data in longitudinal studies, especially when the number of missing values is relatively small compared to the total number of observations in the dataset. However, as the number of missing data increases, biases in the results may also increase. Moreover, none of these methods effectively handle the issue of monotone missing data, where data is missing due to censoring or events occurring after a certain time point, primarily because of the block of missing data. Additionally, a limitation of KML is its requirement for balanced data, which is often not met in practice. As a result, various k-means-based methods have been developed in an attempt to address this issue.

Two-step K-means

A direct and intuitive extension to regular KML is the two-step k-means mentioned in Twisk and Hoekstra (2012) and Shek et al. (2011). The two-step longitudinal K-means clustering algorithm, incorporating random effects from linear mixed-effects models (LME), is a powerful technique used in the analysis of longitudinal data. This method involves two key steps: (i) fitting an LME model to the longitudinal data to estimate subject-specific random effects capturing individual variability, and (ii) performing K-means clustering on the estimated random effects to identify clusters of subjects with similar longitudinal trajectories.

In the first step, the LME model is specified as:

In the context of linear mixed-effects models (LME), the model can be represented in matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is the $n \times 1$ vector of observed outcomes, $\mathbf{X}$ is the $n \times p$ design matrix for fixed effects, where p is the number of fixed effects parameters, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects coefficients, $\mathbf{Z}$ is the $n \times q$ design matrix for random effects, where q is the number of random effects parameters, $\mathbf{b}$ is the $q \times 1$ vector of random effects, $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of residual errors.

The random effects are assumed to follow a multivariate normal distribution:

$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}),$$

where $\mathbf{D}$ is the variance-covariance matrix for the random effects. The model parameters $\boldsymbol{\beta}$ and $\mathbf{D}$ are estimated by maximizing the likelihood function based on the observed data $\mathbf{y}$.

In the second step, the subject-specific random effects ($\hat{\mathbf{b}}_i$) estimated from the LME model are used as input for K-means clustering. The K-means algorithm aims to partition the subjects into K clusters based on the similarity of their random effects. The centroids of the clusters represent typical trajectories of the underlying longitudinal process.

This two-step approach combines the flexibility of LME models in capturing complex longitudinal trajectories with the simplicity and interpretability of K-means clustering. More importantly, the use of LME enables the accommodation of imbalanced data where observations are scattered across different time points. Additionally, it facilitates handling data with monotone missingness, ensuring that the random effect outcomes of equal size can be utilized for clustering purposes regardless of variations in data length. As a supplementary note, if prior knowledge of grouping at a higher stage is available, it is also feasible to define further subgroupings by incorporating both the fixed effects and random effects.

Group-based trajectory modeling

Group-based trajectory modeling is a statistical technique used to identify distinct subgroups, or trajectories, within a population based on their longitudinal data. It aims to characterize the heterogeneity in longitudinal trajectories of individuals over time, often in the context of behavioral or developmental studies. The method involves fitting a finite mixture model to the longitudinal data, where each mixture component represents a distinct trajectory group. Mathematically, the model can be represented as follows. Define $\mathbf{Y}$ is matrix of observed longitudinal data with dimensions $N \times J$, where N is the number of individuals and J is the number of time points. $\boldsymbol{C}$ is vector of trajectory group assignments with length N, and $\boldsymbol{\mu}$ is vector of group-specific means with length $K \times J$, where K is the number of trajectory groups. Then, further define Σ represent the variance-covariance matrix in a presepcified format, and that $\boldsymbol{\pi}$ is the vector of group probabilities with length K. Then the probability density function (PDF) for the observed data $\mathbf{Y}$ can be written in matrix form as:

$$\mathbf{f}(\mathbf{Y}|\mathbf{C}) = \frac{1}{(2\pi)^{J/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}\boldsymbol{C})^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}\boldsymbol{C})\right).$$

where $\boldsymbol{\mu C}$ represents the matrix multiplication of $\boldsymbol{\mu}$ and an indicator matrix $\mathbf{C}$, which has dimensions $N \times J$ and is constructed such that each row corresponds to the mean trajectory for the assigned group.

The joint likelihood function for the entire sample is then given by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot \mathbf{f}(\mathbf{Y}|\mathbf{C}),$$

where $\boldsymbol{\theta}$ represents the set of parameters including $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\pi}$.

The estimation of parameters $\boldsymbol{\theta}$ involves maximizing the joint likelihood function $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\pi}$, typically using iterative optimization techniques such as expectation-maximization (EM) algorithm.

This mathematical framework forms the basis of group-based trajectory modeling, allowing researchers to uncover latent longitudinal patterns within their data and make inferences about the underlying population dynamics. Due to this characteristic, this method is commonly referred to as latent class growth analysis (LCGA). For a more in-depth understanding, Nagin (1999) and Jones and Nagin (2007) offer comprehensive introductions to these methods. Additionally, it's worth mentioning that a polynomial formulation with random error can also be employed, as demonstrated by Nagin and Land (1993).

The fundamental concept of the group-based trajectory approach is that individuals within each group share a common trajectory. Still, it is capable of accommodating imbalanced data with monotone missingness through the group characteristics. The modeling method employed to describe these groups can take various forms, including parametric, non-parametric, and semi-parametric approaches, offering flexibility in capturing the underlying trajectory patterns.

Growth mixture modeling

The growth mixture model (GMM) is a statistical framework used to identify latent subgroups within a population that exhibit distinct trajectories of change over time(Marcoulides and Schumacker, 2001; Ram and Grimm, 2009). Mathematically, the GMM can be expressed as:

$$y_{ij} = \boldsymbol{\gamma}_{ki} z_{ij} + \boldsymbol{\beta}_k \cdot \boldsymbol{x}_{ij} + \boldsymbol{\epsilon}_{ij}$$

where y_{ij} represents the observed outcome for individual i at time j, $\boldsymbol{\gamma}_{ki}$ denotes the random effect realization for the ith subject in the kth latent subgroup, $\boldsymbol{z}_{ij}$ represents the vector of covariates correspond to random effect for ith subject at time j, $\boldsymbol{\beta}_k$ represents the vector of regression coefficients for the kth subgroup, $\boldsymbol{x}_{ij}$ is the vector of covariates for individual i at time j, and $\boldsymbol{\epsilon}_{ij}$ is the error term.

The growth mixture model allows for the estimation of parameters specific to each latent subgroup, capturing heterogeneity in trajectory shapes and developmental patterns within the population. This method is particularly useful when the population consists of distinct subgroups with different developmental trajectories over time. The GMM serves as an extension of previous GBTM. However, unlike GBTM, GMM not only accommodates group-specific trajectories but also permits individuals to deviate from these trajectories, thus offering a more comprehensive explanation for within-cluster heterogeneity.

The GMM method inherits the advantages of the GBTM approach in handling imbalanced data and monotone missingness. Furthermore, it offers improved clustering by accounting for within-cluster variation. However, specifying the prior format of the cluster distribution is still required.

Spline-based clustering

In recent years, there has been growing interest in utilizing spline-based methods for longitudinal clustering, which offer flexibility in capturing nonlinear trajectories and identifying distinct patterns of change over time.

Spline-based longitudinal clustering methods leverage the concept of splines, which are piecewise polynomial functions that provide a flexible framework for modeling smooth curves. By fitting spline models to individual trajectories, these methods enable the detection of underlying patterns and heterogeneity within longitudinal datasets.

One of the key advantages of spline-based longitudinal clustering is its ability to capture complex relationships and non-linear trends in longitudinal data. Unlike traditional linear models, spline-based methods can accommodate irregularly sampled data, handle missing values, and effectively model trajectories with nonlinear patterns or abrupt changes.

The general formula for spline-based longitudinal clustering can be expressed as follows:

$$Y_{ij} = f_j(t_{ij}) + \epsilon_{ij},$$

whereY_{ij} represents the observed response for the i-th subject at time t_{ij}, $f_j(t_{ij})$ is the smooth function capturing the trajectory for the j-th cluster, ϵ_{ij} is the random error term.

Spline-based methods typically involve fitting spline models to individual trajectories, where the smooth function $f_j(\cdot)$ is estimated using basis functions such as B-splines, natural splines, or penalized splines. These basis functions allow for the flexible representation of the underlying trajectories while ensuring smoothness and continuity.

B-splines are piecewise polynomial functions defined on a set of knots. They offer flexibility in modeling smooth curves by dividing the range of the predictor variable into segments and fitting low-degree polynomials within each segment. The general formula for B-splines can be expressed as follows (De Boor and De Boor, 1978)

$$S(x) = \sum_{i=1}^{K} \beta_i B_i(x),$$

where $S(x)$ represents the B-spline function, β_i are the coefficients to be estimated, $B_i(x)$ are the basis functions, often defined recursively using Cox-de Boor recursion formula, and K is the number of basis functions.

Natural splines are a variation of B-splines that impose additional boundary constraints, ensuring smoothness at the endpoints of the curve. These constraints eliminate the need for specifying boundary knots, leading to a more stable and interpretable model. The formula for natural splines is similar to that of B-splines, with the additional boundary constraints (Hastie, 2017) as

$$S(x) = \sum_{i=1}^{K} \beta_i B_i(x) + \lambda \cdot (f''(x_1) + f''(x_K)),$$

where λ is the penalty parameter controlling the degree of smoothness, and $f''(x_1)$ and $f''(x_K)$ represent the second derivatives of the spline function at the boundary points x_1 and x_K.

Penalized splines, also known as P-splines, combine the flexibility of B-splines with the idea of penalization to control the smoothness of the fitted curve. The penalty term is added to the likelihood function, penalizing deviations from smoothness. The formula for penalized splines can be expressed as follows (Eilers and Marx, 1996):

$$S(x) = \sum_{i=1}^{K} \beta_i B_i(x) + \lambda \cdot \sum_{j=3}^{K} (\Delta^2 \beta_j)^2,$$

where Δ^2 represents the second-order difference operator and λ is the smoothing parameter controlling the amount of penalization.

Various approaches have been developed for longitudinal clustering based on spline methods. Abraham et al. (2003) utilized B-splines as a basis and employed the k-means algorithm on the basis parameters. Similarly, Coffey et al. (2014) applied the P-spline technique as a basis expansion method within the linear mixed-effect model framework, conducting clustering based on the resulting model . Additionally, the P-spline method has shown increasing utility by empowering researchers to smooth individual trajectories, cluster groups, and concurrently ascertain the number of groups (Zhu and Qu, 2018). In summary, spline-based methods excel in efficiently handling irregularly sampled data and aiding in subgroup identification. However, owing to their nonparametric nature, clustering methods based on splines can pose computational challenges and may be sensitive to the selection of knots or basis functions. Additionally, violations of the smoothness assumption or small sample sizes can result in inadequate fitting.

Shape-based clustering

All previous methods assumed uniform time progression across all subjects. In practice, this assumption may not hold true, especially in longitudinal studies where subjects may progress at different rates or speeds despite sharing similar underlying trajectories. This phenomenon is particularly evident in diseases like AD, where individuals may exhibit similar patterns of cognitive decline but progress at different rates. For instance, subjects may follow similar cognitive decline trajectories, but the pace at which this decline occurs varies from one individual to another.

To tackle this issue, two approaches are worth considering. The first approach involves addressing time shifting before performing clustering, while the second approach involves addressing time shifting simultaneously with the clustering process. The k-mean alignment algorithm introduced by Sangalli et al. (2010) simultaneously conducts clustering and alignment by integrating a warping function. This function facilitates alignment during the cluster assignment step. On the other hand, Liu and Yang (2009) proposed an alternative method to address the issue simultaneously. They apply Taylor expansion to the shifted B-spline basis, obtaining a standard time scale. In two-step methods, the emphasis typically lies on utilizing shape-respecting distances, with the most commonly employed ones being the Frechet distance and the Dynamic Time Warping distance (DTW).

The Frechet distance measures the similarity between two curves by considering the ”walk” of two points along each curve while maintaining their temporal ordering. It computes the minimum leash length required for a dog (representing one curve) and its owner (representing the other curve) to traverse their respective paths simultaneously. Mathematically, the Frechet distance between two curves P and Q is defined as:

$$F(P, Q) = \inf_{\gamma \in \Gamma} \max_{t \in [0,1]} \{\text{dist}(P(\gamma(t)), Q(t))\},$$

where Γ is the set of all possible parameterizations of the curves, $P(\gamma(t))$ represents the position of the point on curve P parameterized by γ, and dist denotes the Euclidean distance between two points.

DTW is another shape-respecting distance measure that accommodates variations in the alignment and pacing of temporal sequences. Unlike traditional distance measures, DTW allows for local stretching and compressing of the time axis, enabling more flexible matching of temporal patterns. The DTW distance between two sequences A and B is computed by finding the alignment that minimizes the total distance between corresponding points:

$$DTW(A, B) = \min_{\text{path}} \sum_{(i,j) \in \text{path}} d(A[i], B[j])$$

where $d(\cdot, \cdot)$ represents the local distance between points, and the minimum is taken over all possible alignments.

The Frechet distance and DTW distance are significant in shape-respecting longitudinal data analysis due to their ability to capture the intrinsic shape similarity between temporal trajectories. By incorporating temporal dependencies and considering variations in the alignment of sequences, these distances enable more accurate comparisons and clustering of longitudinal data. A generalized shape distance-based method is proposed by Genolini et al. (2016) using the generalized Fréchet distance. The approach based on shape-respecting distances holds promise in addressing potential time-stretching issues in longitudinal clustering. However, it encounters a challenge wherein the comparison of curves assumes that the starting and ending points of the curves are consistent, a condition often violated due to censoring.

5.2.2 Partial-mapping DTW clustering method

DTW and Frechet distance have been employed for clustering based on shape. Although both methods can accommodate missing data and irregularly sampled data, they share a fundamental assumption: the trajectories should represent the same longitudinal pattern but at different progression rates. In AD studies, the complete longitudinal trend is frequently unobservable due to terminal events or censoring. In such scenarios, we propose a partial-mapping DTW clustering method, which explores the mapping of curves onto a target curve for clustering purposes. In the process of selecting the distance metric for clustering following partial mapping, we take into account insights from Agarwal et al. (2015). The DTW distance is noted for its sensitivity to sampling, as it solely considers vertices. Conversely, the Frechet distance often demands substantially more computational resources and is prone to sensitivity towards outliers, given its computation involving the minimum of the maximum. Furthermore, considering the method of clustering given the distance, Genolini et al. (2016) utilizes k-means, which necessitates the calculation of the cluster centroid. In cases of time stretching, however, determining the cluster mean becomes more intricate than merely averaging the cluster. As outlined by Genolini et al. (2016), the mean of two curves is defined as the midpoint of the leash. Extending this concept to multiple curves introduces complexity, rendering exact computation impractical. Consequently, a compromise is reached, and only a subset of the data is computed, resulting in an approximation of the true cluster centroid. Given the computational burden associated with this approach, we opt to employ hierarchical clustering, pre-computing the distance matrix for all pairs of curves in advance to avoid cluster mean calculation.

To determine the segment of a curve onto the target curve, we assume that the starting point of all trajectories is the same, a characteristic common in AD studies. However, the endpoint may vary due to censoring and terminal events. It's also important to note that precisely determining the time point to be mapped onto is challenging since the underlying trajectory is unknown. Therefore, a technique similar to that depicted in Witowski et al. (2011) is employed. In the referenced paper, a shorter curve is glided through a longer curve with offset to find the best match, ensuring that the mapped curve maintains the same length as the original curve. However, in our context, we propose stretching the original curve with a multiplier to match the length of the longer curve, thus exploring possible endpoint variations. To prevent overfitting to the data, we opt to employ a different similarity measure to evaluate the fitting, namely the area between the curves. The proposed algorithm is listed below with reference to Figure 5.1.

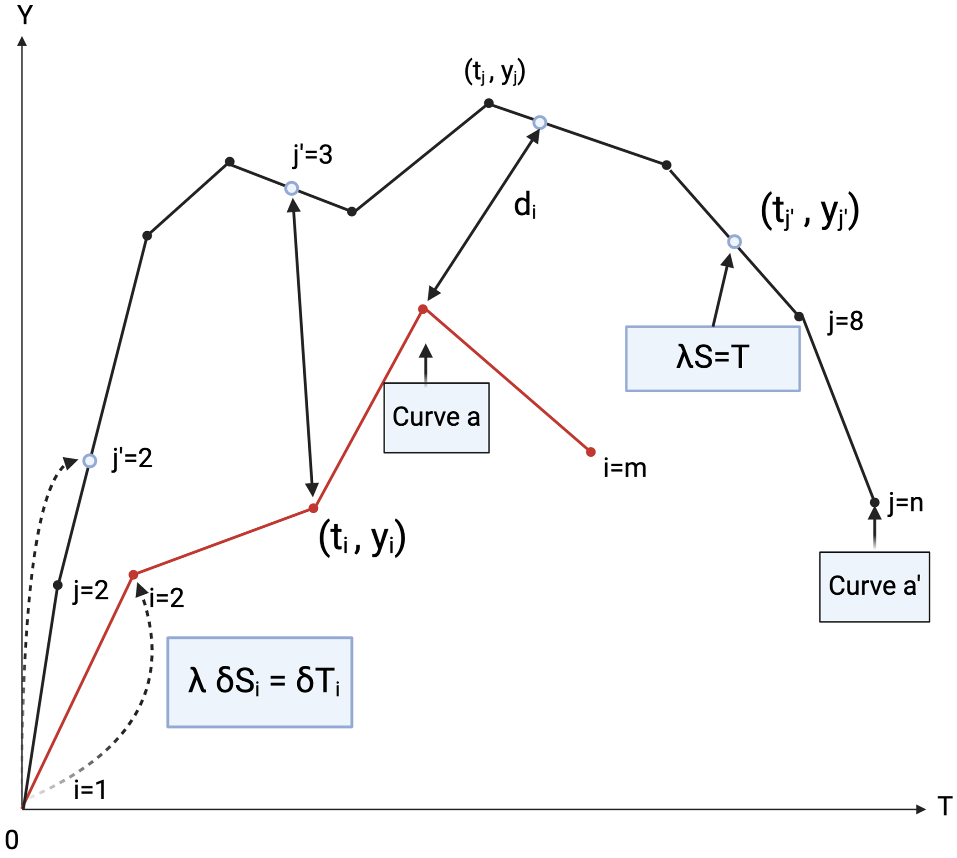


Figure 5.1: Partial curve mapping of curve a onto curve a' with multiplier λ.

5.3 Numerical studies

5.3.1 Simulation setup

Models compared

To evaluate the efficacy of our proposed partial-mapping DTW longitudinal clustering algorithm, we conducted a comparative analysis against several established methods previously mentioned in the literature. These methods are specifically designed to handle monotonous missing data with irregular sampling and include two-step k-means, growth mixture model,

Algorithm 2 Parital-mapping DTW longitudinal clustering algorithm

```
For each pair of longitudinal curves, a and a', ensure that curve a is shorter than mapped curve a'; if not, swap two curves.
Calculate the total length of the shorter polygon $a$, denoted as $S$. Additionally, compute the length of each segment, represented as $\delta S_i$:
    $\delta S_i = \sqrt{(t_{i+1} - t_i)^2 + (y_{i+1} - y_i)^2} \quad \text{for} \quad i = 2, \ldots, m$
Calculate the ratio of each segment to the total length:
    $\tilde{S}_i = \delta S_i / S \quad \text{for} \quad i = 1, \ldots, m-1$
Calculate the total length of the longer polygon $a'$, denoted as $T$.
Define a stretch parameter $\lambda_p$ to use only part of the longer longitudinal curve for comparison. $\lambda_p$ searches a uniformly spaced grid for the difference between two curves. For $p = 1 \ldots P$, $\lambda_p = \left[S + \frac{p(T-S)}{P}\right] / S$.
Given the stretch parameter $\lambda_p$, set the new point on the curve $a'$ mapped from the curve $a$ so that $\delta T_i = \lambda_p \delta S_i$ for each new segment.
Calculate the pairwise distance between curve $a$ and mapped curve $a''$ by
    $d_i = \sqrt{(t_i - t_{j'})^2 + (y_i - y_{j'})^2}$
    , then calculate the total mismatch measurement
    $\epsilon_p = \sum_{i=1}^{m-1} \frac{(d_i + d_{i+1}) * \tilde{S}_i}{2}$
Find $p$ that minimize $\epsilon_p$ and obtain the best matching partial curve.
Obtain the distance matrix for all pairs of curves.
Perform hierarchical clustering algorithm utilizing the distance matrix and trim the result based on the desired number of clusters.
```

and spline-based longitudinal clustering. Two-step k-means is often considered a simplistic approach to longitudinal clustering, yet we included it as a baseline for comparison with other methods. For our implementation, we utilized the *lmer* package in *R* to model the initial stage, followed by passing the random effect estimates to the *k-means* package in *R* for clustering. When employing finite mixture modeling, we favor the growth mixture model over the group trajectory-based model. This preference stems from the growth mixture model serving as an extension to the latter, with the key distinction being that the growth mixture model permits subject-wise deviations from the group trajectory. For implementation, we utilize the *lcmm* package, which relies on the *hlme* package. This method falls under the growth mixture model within the framework of linear mixed model theory. For the spline-based approach, we opted to utilize the *clustra* package. This package is specifically designed for clustering longitudinal trajectories (time series) on a shared time axis. It accommodates observations that are unequally spaced, of unequal length, and only partially overlapping. The clustering process involves an EM algorithm, which iteratively switches between fitting a thin plate spline (TPS) to combined responses within each cluster (M-step) and reassigning cluster membership based on the nearest fitted B-spline (E-step).

Simulated longitudinal trajectories

For simulations, we considered various types of commonly encountered trajectories: quadratic longitudinal trajectory, normal density, normal cumulative distribution function (CDF), and log-normal distribution. The quadratic longitudinal trend follows the Linear Mixed Effects (LME) model, where the fixed effect is in the form $f(x) = ax^2 + bx + c$, and for the random effect, we assume random slope and random intercept.

The longitudinal trajectory with a normal distribution density follows the form $f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)$, where Φ represents the CDF of the normal distribution.

On the other hand, we have the longitudinal trajectory with log-normal distribution, which follows the density function as $f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$. The log-normal distribution is a log transformation of the normal distribution. While the normal distribution is symmetric, the log-normal distribution exhibits skewness and a heavier tail. It is especially suitable for cases where the increase in the distribution yields unproportioned returns.

While the probability density function (PDF) and CDF of the normal distribution tend to diverge later, comparing the normal PDF to the log-normal PDF can serve as an excellent example of shape change.

To introduce randomness into our data, we apply distortion to all longitudinal trajectories. In the LME model, randomness arises naturally through the use of random intercept and random slope. For all other types of trajectories, we introduce multiple distortions using the formula $f^*(x) = a_2 f(a_1 x + b_1) + b_2$, where $a_1, a_2 \sim \mathrm{U}(1-\sigma, 1+\sigma)^2$ and $b_1, b_2 \sim \mathrm{U}(-\sigma, \sigma)$. Here, (a_1, a_2) serves as the scaling parameter, and (b_1, b_2) serves as the location shift. The magnitude of their distortion is commonly controlled by the tuning parameter σ. For comparison, we considered five case as depicted in Figure 5.2.

- **Case 1**: two groups A and B, with

  $f_A(x) = -0.0396x^2 + 0.294x + 0.344$ and $f_B(x) = -0.0267x^2 + 0.143x + 0.604$.

- **Case 2**: two groups C and D, with

  $f_C(x) = \Phi(x, 1.5, 1) \times 2.5$ and $f_D = \phi(x, 2, 1) \times 2.5$.

- **Case 3**: three groups C, D, and E, with

  $f_E(x) = \Phi(x, 1.5, 1) \times 2.5$.

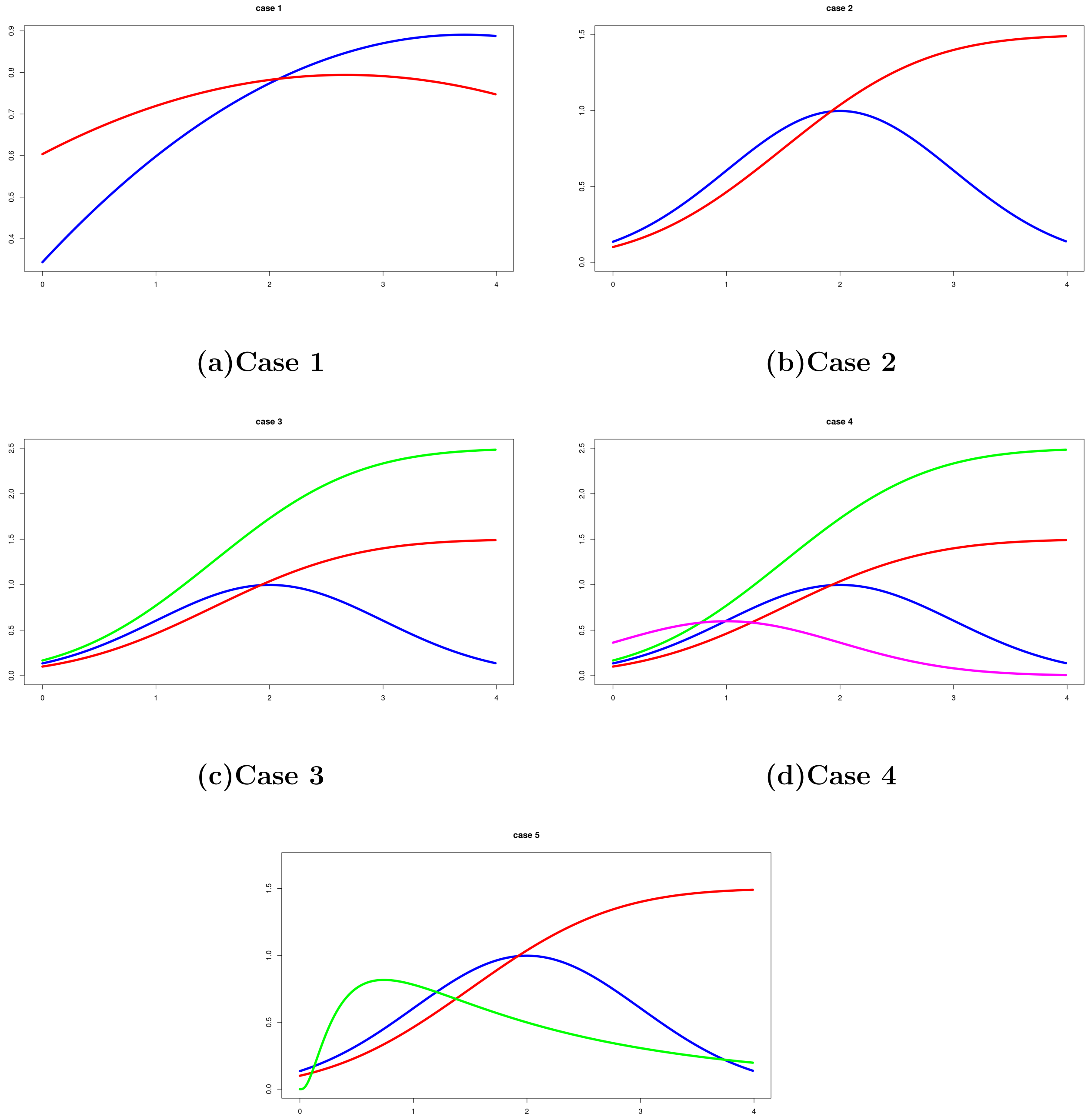
- **Case 4**: four groups C, D, E, and F, with

  $f_F(x) = \phi(x, 2, 1) \times 1.5$.

- **Case 5**: three groups C, D, and G, with

  $f_G(x) = \text{log-normal}(0.7, 1) \times 2.5$.

For groups A and B, we assume that the random intercept follows a normal distribution with mean 0 and standard deviation 0.1, while the random slope follows a normal distribution with mean 0 and standard deviation 0.08. For other types of longitudinal trajectories, we set the tuning parameter σ to 0.1. To induce monotone missingness, we employ the power function distribution with the cumulative distribution function (CDF) given by $F(x) = (x/\theta)^r$, where the support is determined by the parameter θ and the parameter r governs the mode of censoring. Following the simulation, we set $\theta = 4$ for a common support range and generated r from a uniform distribution $\mathrm{U}(0.8, 1.4)$ for each group.



(a)Case 1

(b)Case 2

(c)Case 3

(d)Case 4

(e)Case 5

Figure 5.2: **(a)** Case 1:f_A is in red, f_B in blue; **(b)** Case 2:f_C is in red, f_D in blue; **(c)** Case 3:f_C is in red, f_D in blue, f_E in green; **(d)** Case 4:f_C is in red, f_D in blue, f_E in green, f_F in magenta; **(e)** Case 5:f_C is in red, f_D in blue, f_G in green.

Performance assessment

To evaluate the effectiveness of longitudinal clustering methods during simulation, we employ two commonly used indices from the literature (Genolini et al., 2016; Verboon and Pat-El, 2022; Den Teuling et al., 2023): i) the percentage error; and ii) the adjusted Rand index. The correct classification rate is defined as the percentage of agreement between the found partitions and the true partitions, considering the true clustering and the clustering result. In scenarios involving multiple groups, the correct classification rate is computed by considering the highest number of true partitions within each clustered partition. Fhe the method, it is noted that the adjusted randon index (Hubert and Arabie, 1985) is a variation of random index (Rand, 1971). Given two partitions, the Rand Index calculates the percentage of agreements between pairs of samples with respect to their clustering assignments. It is defined as:

$$RI = \frac{a+b}{\binom{n}{2}},$$

where a represents the number of pairs of elements that are in the same cluster in both the true and predicted partitions, b represents the number of pairs of elements that are in different clusters in both the true and predicted partitions and n is the total number of samples.

The Adjusted Rand Index (ARI) adjusts the Rand Index for chance. It corrects for the expected similarity between two random clusterings, producing a score between -1 and 1, where 1 indicates perfect similarity between the clusterings, 0 indicates the expected similarity under random chance, and negative values indicate dissimilarity. The formula for

Adjusted Rand Index is given by:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where n_{ij} is the number of pairs of elements that are in the same cluster in the true partition and in the same cluster in the predicted partition for clusters i and j, a_i is the number of elements in cluster i in the true partition, b_j is the number of elements in cluster j in the predicted partition, and n is the total number of samples.

In terms of their relationship, the Adjusted Rand Index (ARI) can be seen as a standardized version of the Rand Index (RI). It's calculated as follows:

$$ARI = \frac{RI - E[RI]}{Max(RI) - E[RI]},$$

where RI is the Rand Index, $E[RI]$ is the expected Rand Index under random chance, and $Max(RI)$ is the maximum possible Rand Index. This standardization ensures that the ARI eliminates cases where the solution yields a Rand Index lower than expected, preventing the Rand Index from becoming negative.

5.3.2 Simulation results

The correct classification rates for the four methods across the five cases are presented in Table 5.1. In Case 1, both the shape-based method and the spline-based method exhibit lower performance compared to the growth mixture model and the two-step k-means method. This can be attributed to the underlying trajectory fitting well within the domain of parametric models, where clustering based on parametric models proves to be more effective. Starting from Case 2, the shape-based and spline-based methods perform better than the growth mixture model as well as the two-step method. It is worth noting that in Case 2, where the normal PDF and CDF are compared, the parametric model-based method performed much worse than the nonparametric-based method. However, in Case 3, with the addition of another PDF of normal distribution, the growth mixture model obtained much higher accuracy, perhaps due to the fact that the added curve is well separated from the first two curves, making the task easier for the growth mixture model. Such conjecture may be proven in Case 4, where another normal density is added, resulting in two PDFs and two CDFs. Within each pair of PDF and CDF, they are hard to distinguish, but it is easy to distinguish the two sets of pairs. As we can see, the decrease in performance of the shape-based method is much less than the increase in error rate in the growth mixture model compared to the previous case. In Case 5, which should be the most challenging case to distinguish, as the difference not only lies in the later stage but also in the first stage, the nonparametric model performs remarkably well, even better than in Case 1. However, the parametric-based model performs much worse compared to its performance in Case 1. Note that in all cases, the methods within each realm share similar performance compared to other realms, but the shape-based method is slightly better than the spline-based method, and the growth mixture model shows more robust results compared to the two-step k-means method. In Table 5.2, the Adjusted Rand Index (ARI) for each method under each case is shown. The results generally align with Table 5.1, but the ARI shows larger differences, where small differences in percentage error become much bigger differences in ARI.

	Method			
Trajectory	Shape Based	Growth Mixture Model	Two-Step K-means	Spline Method
Type	Mean Error Rate	Mean Error Rate	Mean Error Rate	Mean Error Rate
Case 1	0.096	0.066	0.047	0.111
Case 2	0.163	0.330	0.254	0.215
Case 3	0.215	0.224	0.317	0.241
Case 4	0.239	0.276	0.336	0.290
Case 5	0.079	0.192	0.265	0.085

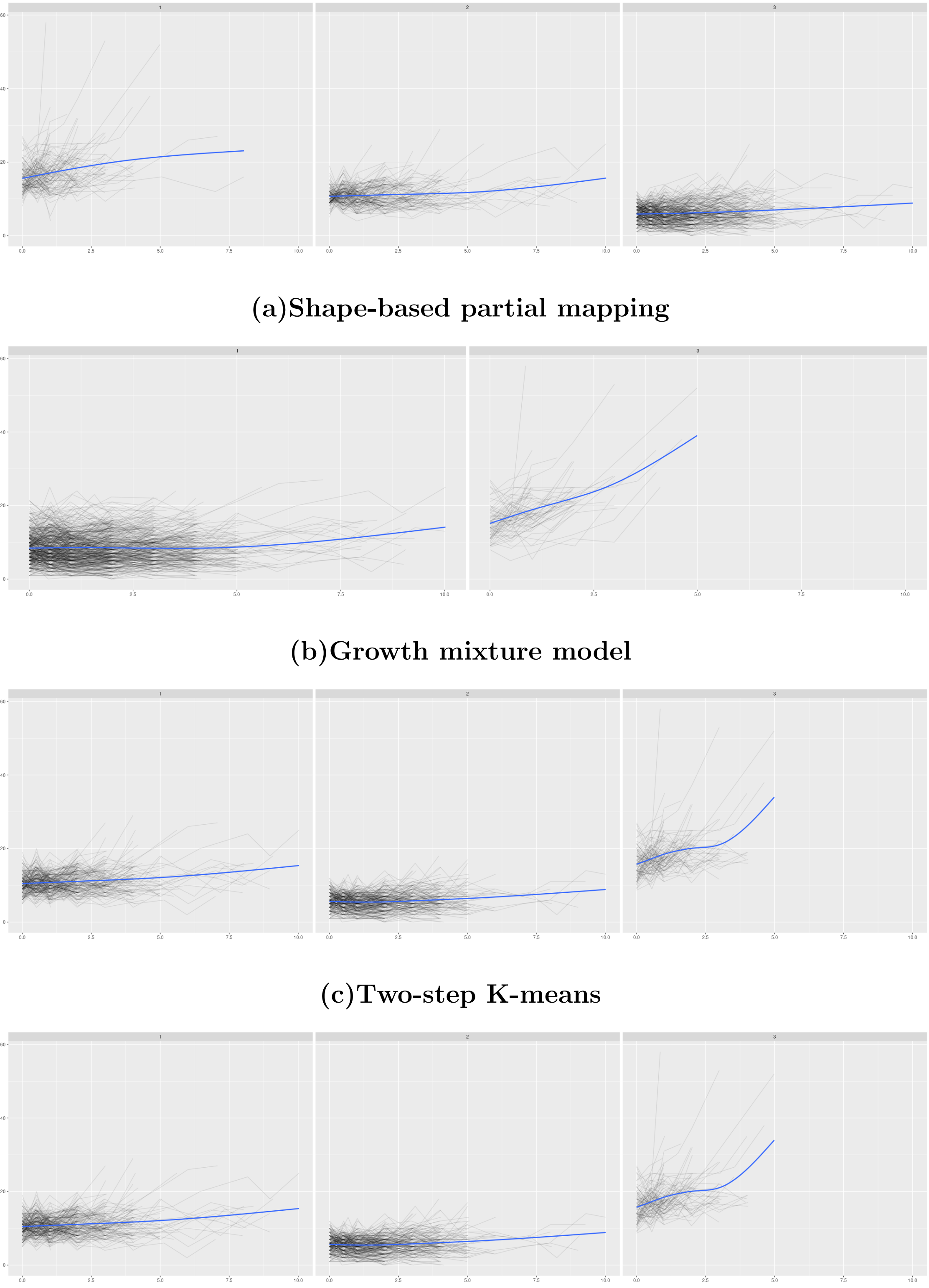
Table 5.1: Correct classification rate of shape-based partial mapping, growth mixture model, two-step k-means, and spline-based clustering methods.

	Method			
Trajectory	Shape Based	Growth Mixture Model	Two-Step K-means	Spline Method
Type	Mean ARI	Mean ARI	Mean ARI	Mean ARI
Case 1	0.659	0.810	0.819	0.609
Case 2	0.454	0.070	0.243	0.376
Case 3	0.468	0.354	0.269	0.443
Case 4	0.476	0.397	0.336	0.449
Case 5	0.792	0.385	0.363	0.761

Table 5.2: ARI of shape-based partial mapping, growth mixture model, two-step k-means, and spline-based clustering methods.

5.4 Application

For practical application, we employ the methods on the ADNI dataset as outlined in previous sections. To determine the number of groups, we utilize gap statistics. Figure 5.3 illustrates the resulting clustering based on the ADAS13 covariate. It is evident that all methods detect one group with an obvious upward trend, comprising a small portion of the entire dataset. However, for the remaining subjects, the growth mixture model only identifies two groups. On the other hand, all three other methods are able to detect two subgroups, which, upon visual inspection, differ by their intercept and exhibit slight flat trends.



(a)Shape-based partial mapping

(b)Growth mixture model

(c)Two-step K-means

(d)Spline-based clustering

Figure 5.3: ADNI clustering based on ADAS13

5.5 Discussion

In this chapter, we have compiled and compared various types of longitudinal clustering algorithms, discussing their capabilities in handling monotonous missing data with irregular sampling. Additionally, we introduce a novel shape-based partial mapping method. This approach enables us to group individuals whose trajectories exhibit similar shapes but with shifts in their time positions due to interruptions in data collection.

When compared to other methods applicable to longitudinal data with irregular sampling and monotone missingness, the shape-based partial mapping method demonstrates superior performance, particularly when the underlying trajectory contains distortion noise. Although the selection of partial mapping lacks conclusive assurance due to informational constraints, our method remains flexible as the indices utilized at the partial mapping stage can be adapted to other criteria based on domain knowledge. Compared to existing shape-based methods, our approach offers the advantage of effectively handling irregular data with monotone missingness. Furthermore, in contrast to shape-based methods that utilize Frechet distance, our method imposes significantly less computational burden. Local Variables: ***

Chapter 6

Discussion

Motivated by the potential irreproducible outcomes stemming from multi-center Alzheimer's disease studies due to potential model misspecification, we have developed a censoring robust estimator for jointly modeling longitudinal and survival data. In Chapter 3, we introduce a censoring robust estimator for this joint modeling task, with its interpretation centered around the average covariate effect under the non-proportional hazards assumption. For our inference process, we utilize a bootstrap method, as detailed in Chapter 3. Chapters 4 and 5 are dedicated to implementing the methodology proposed in Chapter 3. In Chapter 4, we examine the constraints of the robust variance estimator when applied to the proposed censoring robust estimator. We then develop a comprehensive range of robust variance estimators for censoring robust estimators using an influence function-based approach. The estimator discussed in Chapter 4 can significantly alleviate the computational burden caused by the bootstrap method, thereby enhancing the utility of our proposed approach in handling large datasets. In Chapter 5, we identify the necessity for a shaped-based longitudinal clustering method in the context of irregular sampling and monotone missingness, drawing inspiration from our proposed approach. We examine the existing approach of longitudinal clustering and analyze its capacity to handle distinct types of longitudinal data. In addition, we have developed a shape-based partial mapping longitudinal clustering technique capable of managing monotone missingness caused by censoring, which is frequently encountered in disease studies.

There are numerous promising avenues for future research stemming from the proposed work outlined here. Our proposed censoring robust estimator exhibits flexibility owing to its two-stage modeling approach. While much of the emphasis in existing censoring robust estimators lies on the second stage, there is potential for enhancing the precision of longitudinal covariate value prediction at the event time by modifying the first stage model. The utilization of spline methods shows considerable promise in accurately fitting longitudinal curves, prompting interest in substituting the current first-stage model with a spline-based approach. Moreover, the method of generating the censoring group currently involves using fixed and random effects from the linear mixed model as grouping factors in the survival tree method. However, this approach may lead to information loss. Therefore, exploring alternative methods for incorporating longitudinal data into censoring clustering is warranted. Additionally, our simulations in Chapter 5 have revealed that spline-based longitudinal clustering methods perform admirably in shape detection tasks. This prompts further exploration into the feasibility of employing penalized splines with specific penalties for partial mapping and shape recognition. Each of these possibilities warrants thorough investigation to advance our understanding and application of censoring robust estimation methodologies.

Bibliography

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics* **30,** 581–595.

Agarwal, P. K., Fox, K., Pan, J., and Ying, R. (2015). Approximating dynamic time warping and edit distance for a pair of point sequences. *arXiv preprint arXiv:1512.01876* .

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics* pages 1100–1120.

Blennow, K. and Zetterberg, H. (2018). Biomarkers for alzheimer's disease: current status and prospects for the future. *Journal of internal medicine* **284,** 643–663.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics* pages 35–41.

Boyd, A. P., Kittelson, J. M., and Gillen, D. L. (2012). Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Statistics in medicine* **31,** 3504–3515.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics* pages 437–453.

Coffey, N., Hinde, J., and Holian, E. (2014). Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis* **71,** 14–29.

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

De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

De Leon, M., DeSanti, S., Zinkowski, R., Mehta, P., Pratico, D., Segal, S., Clark, C., Kerkman, D., DeBernardis, J., Li, J., et al. (2004). Mri and csf studies in the early diagnosis of alzheimer's disease. *Journal of internal medicine* **256,** 205–223.

Demidenko, E. and Stukel, T. A. (2005). Influence analysis for linear mixed-effects models. *Statistics in Medicine* **24,**. influence function for lme, but only use yi or xi but not distribution misspecification.

Den Teuling, N., Pauws, S., and van den Heuvel, E. (2023). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics-Simulation and Computation* **52,** 621–648.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science* **11,** 89–121.

Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

Filippova, A. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory of Probability & Its Applications* **7,** 24–57.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*. John Wiley & Sons.

Fleming, T. R. and Harrington, D. P. (2013). *Counting processes and survival analysis*, volume 625. John Wiley & Sons.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association* **82,** 312–320.

Freedman, D. A. (2006). On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician* **60,** 299–302.

Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., and Subtil, F. (2016). kmlshape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *Plos one* **11,** e0150738.

Genolini, C. and Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics* **25,** 317–328.

Glenner, G. G. and Wong, C. W. (1984). Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and biophysical research communications* **120,** 885–890.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

Greenwood, M. et al. (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.* .

Hampel, H. and Teipel, S. J. (2004). Total and phosphorylated tau proteins: evaluation as core biomarker candidates in frontotemporal dementia. *Dementia and geriatric cognitive disorders* **17,** 350–354.

Hardin, J. W. (2002). The robust variance estimator for two-stage models. *The Stata Journal: Promoting communications on statistics and Stata* **2,**. replace morphy by sandwitch, but assume model right, but mention morphy require full likelihood.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* **28,** 100–108.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

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

Howe, C. J., Cole, S. R., Lau, B., Napravnik, S., and Eron Jr, J. J. (2016). Selection bias due to loss to follow up in cohort studies. *Epidemiology (Cambridge, Mass.)* **27,** 91.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics* **62,** 1037–1043.

Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2,** 193–218.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28,** 2796.

Jann, B. (2019). Influence functions for linear regression (with an application to regression adjustment). *University of Bern Social Sciences Working Paper* **41,**.

Jones, B. L. and Nagin, D. S. (2007). Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological methods & research* **35,** 542–571.

Kahn, J. (2022). Influence function for fun and profit.

Kalbfleisch, J. D. and Prentice, R. (1980). Survival analysis.

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

Kametani, F. and Hasegawa, M. (2018). Reconsideration of amyloid hypothesis and tau hypothesis in alzheimer's disease. *Frontiers in neuroscience* **12,** 25.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53,** 457–481.

Kasten, T., Zhang, L., Goate, A., Morris, J. C., Holtzman, D., and Bateman, R. J. (2015). Age and amyloid effects on human cns amyloid-beta kinetics.

Knafl, G. (1978). *Regression with censored data.* PhD dissertation, Northwestern University.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled clinical trials* **21,** 167–189.

Law, J., Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). Robust statistics-the approach based on influence functions. *The Statistician* **35,**.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association* **88,** 457–467.

Li, K., Chan, W., Doody, R. S., Quinn, J., Luo, S., Initiative, A. D. N., et al. (2017). Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. *Journal of Alzheimer's Disease* **58,** 361–371.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association* **84,** 1074–1078.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, X. and Yang, M. C. (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis* **53,** 1361–1376.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory* **28,** 129–137.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mallinckrodt, C. H., Clark, W. S., and David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of biopharmaceutical statistics* **11,** 9–21.

Marcoulides, G. A. and Schumacker, R. E. (2001). *New developments and techniques in structural equation modeling*. Psychology Press.

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics* **18,** 309–348.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics* **55,** 463–469.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods* **4,** 139.

Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology* **31,** 327–362.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1,** 27–52.

Nguyen, V. Q. and Gillen, D. L. (2017). Censoring-robust estimation in observational survival studies: Assessing the relative effectiveness of vascular access type on patency among end-stage renal disease patients. *Statistics in biosciences* **9,** 406–430.

Peterson, A. V. (1977). Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association* **72,**.

Ram, N. and Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International journal of behavioral development* **33,** 565–576.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66,** 846–850.

Reid, N. (2007). Influence functions for censored data. *The Annals of Statistics* **9,**.

Reid, N. and Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72,**.

Ringhui, X. (1996). *Inference for the proportional hazards model.* PhD thesis, University of California, San Diego.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics & Data Analysis* **54,** 1219–1233.

Schafer, J. (1999). Multiple imputations: A primer. statistical methods in medical research.

Schubert, S., Leyton, C. E., Hodges, J. R., and Piguet, O. (2016). Longitudinal memory profiles in behavioral-variant frontotemporal dementia and alzheimer's disease. *Journal of Alzheimer's Disease* **51,** 775–782.

Selkoe, D. J. (1991). The molecular pathology of alzheimer's disease. *Neuron* **6,** 487–498.

Shek, D. T., Ma, C., et al. (2011). Longitudinal data analyses using linear mixed models in spss: concepts, procedures and illustrations. *The scientific world journal* **11,** 42–76.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73,** 363–369.

Teuling, N. D., Pauws, S., and van den Heuvel, E. (2021). Clustering of longitudinal data: A tutorial on a variety of approaches.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability* **57,**.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association* **90,** 27–37.

Turkson, A. J., Ayiah-Mensah, F., and Nimoh, V. (2021). Handling censoring and censored data in survival analysis: a standalone systematic literature review. *International journal of mathematics and mathematical sciences* **2021,** 1–16.

Twisk, J. and Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of clinical epidemiology* **65,** 1078–1087.

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

Verboon, P. and Pat-El, R. (2022). Clustering longitudinal data using r: A monte carlo study. *Methodology* **18,** 144–163.

Webel, K. (2011). Greene, wh, econometric analysis. *Statistical Papers* **52,** 983.

Witowski, K., Feucht, M., and Stander, N. (2011). An effective curve matching metric for parameter identification using partial mapping. In *8th European LS-DYNA, Users Conference Strasbourg, pgs*, pages 1–12.

Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics* **2012,**.

Xu, R. and Harrington, D. P. (2001). A semiparametric estimate of treatment effects with censored data. *Biometrics* **57,**. influence function variance.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1,** 423–439.

Zhang, S., Martin, R. D., and Christidis, A.-A. (2019). Influence functions for risk and performance estimators. *Available at SSRN 3415903* .

Zhelonkin, M., Genton, M. G., and Ronchetti, E. (2012). On the robustness of two-stage estimators. *Statistics and Probability Letters* **82,**.

Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups.