UNIVERSITY OF CALIFORNIA
RIVERSIDE

Joint Models and A Study of Missing Data Mechanisms: New Statistical Methods and Novel Applications

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Debaleena Sain

September 2021

Dissertation Committee:

Dr. Esra Kürüm, Chairperson
Dr. Weixin Yao
Dr. Brandon Brown

The Dissertation of Debaleena Sain is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to begin by thanking my dissertation advisor Dr. Esra Kürüm. This research would not have been possible without her knowledge, invaluable feedback, and constant support. She has spent countless hours on this manuscript, and her contribution to this work is immeasurable. I cannot imagine having a better advisor for my PhD research. Thank you Dr. Kürüm for bringing the best out of me.

I extend my deepest gratitude to my remaining committee members: Drs. Weixin Yao, Brandon Brown, Shemra Rizzo, and Thomas Girke. You all have provided valuable time and kind support. Additionally, I am extremely grateful to Dr. Chioun Lee for the research opportunities, which led to the motivation behind selected parts of this dissertation. I would also like to thank Dr. Rachel Wu for her collaborative projects, which helped me to grow as an applied statistician.

Drs. Analisa Flores and Linda Penas, I thank you from the bottom of my heart for being the best teaching supervisors. Your mentorship deeply influenced my teaching style and has improved my overall communication skills.

I am grateful to every teacher in my life from primary school to my graduate education. You all have provided me a quality education, taught me invaluable life-skills, pushed me towards excellence, inspired me to pursue higher education, and believed in me.

I want to thank my fellow graduate students, Isaac Quintanilla Salinas, Samantha VanSchalkwyk, Mi (Bibby) Zhou, and Nathan Robertson. Life would have been difficult in an unknown country without your friendship. I am especially grateful to Isaac for your valuable insight and help in completing this project.

It cannot be put into words how much my parents have sacrificed for me. Thank you for supporting me at every decision of my life. I am also grateful to my life-long friends Purba and Susweta for always being there for me. This dissertation would have been a far-fetched dream without your unconditional love and support throughout my life.

In loving memory of my grandfather,

Rabindranath Ray.

ABSTRACT OF THE DISSERTATION

Joint Models and A Study of Missing Data Mechanisms: New Statistical Methods and Novel Applications

by

Debaleena Sain

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, September 2021
Dr. Esra Kürüm, Chairperson

Motivated by Womens' Interagency HIV Study (WIHS), we propose an intuitive time-varying joint model (TV-JM) for longitudinal and time-to-event outcomes. In this model, conditional on a set of random effects, a joint likelihood is constructed which accounts for the dependence of the two outcomes and correlation among the repeated measurements. We allow all the coefficients in both longitudinal and survival submodels to vary as smooth functions of time, and hence, this method will allow researchers to explore the dynamic response-predictor as well as response-response relationships in a longitudinal data efficiently and accurately. For estimation of the model parameters, we employ an Expectation-Maximization algorithm. In the E-step of the algorithm, the underlying random effects are estimated and in the M-step, we employ local linear regression techniques to fit the time-varying coefficients. The finite sample performance of the proposed method is illustrated via extensive simulation studies. The proposed method is demonstrated by jointly analyzing CD4 cell percentage and time to death outcomes from WIHS.

In the second part of this dissertation, we study the performance of generalized varying coefficient models (GVCM) under missing data mechanisms via extensive simulation studies. This work was motivated by the Midlife in the United States (MIDUS) data, where significant number of missing observations exist and our main goal is to perform a novel application of GVCM to provide impactful insights to research on aging. We present the results of our simulation studies and apply GVCM to analyze data from MIDUS.

Key words and phrases: Expectation-Maximization; Gauss-Hermite quadrature; Local linear fitting; Varying-coefficient models; Joint models

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In follow-up studies implemented in several applied fields, namely, medicine, epidemiology, and sociology, subjects provide two types of outcomes: longitudinal repeated measurements and time to an event of interest such as time to death or time to a disease. In these types of studies, the common goal is to investigate the effect of risk factors on the survival and longitudinal outcomes in addition to identifying the association between these processes. A typical example is from HIV studies, where the repeated measurements on biomarkers for disease progression such as CD4 cell counts or the estimated viral load are collected to predict time to AIDS or death. The traditional approach to analyze these outcomes would be to fit separate regression models: the longitudinal outcome (CD4 cell count or viral load) using a mixed effects regression model and the time-to-event outcome (time to AIDS or death) using a survival model such as the time-dependent Cox model, where the longitudinal outcome is included as a time-dependent predictor. However, it is shown that this approach ignores possible dependence among these outcomes and ignores

the underlying assumption in survival models with time-dependent predictors. More specifically, these models assume that the time-varying predictors are exogenous, that is, their values at any time point after the failure time are not affected by the occurrence of the event. However, this assumption does not hold for longitudinal outcomes, which are endogenous time-dependent covariates, that is, the value of the covariate is directly related to the failure status and thus, this type of predictors requires the survival of the subject for their existence. As a result, the traditional approach of fitting separate models to each outcome leads to inefficient estimates (Sweeting and Thompson, 2011; Tsiatis and Davidian, 2004). In order to overcome these challenges, a new class of models, namely, joint modeling of longitudinal and survival outcomes were introduced.

The major challenge in joint modeling of longitudinal and survival outcomes is the lack of a natural joint distribution for these types of responses. To overcome this challenge, shared-parameter models have been proposed. This class of models assumes that the longitudinal and the time-to-event outcomes are conditionally independent given a set of underlying latent variables shared by both submodels. In most cases, these latent variables are included in the form of random effects, which account for the association between the outcomes and the within subject correlation among the repeated measurements. Since the outcomes become conditionally independent given the random effects, the joint likelihood of the two outcomes consists of conditional submodels for each outcome. A frequent choice of submodel for the longitudinal component is a linear mixed effects model (Laird and Ware, 1982) and that of the survival component is the Cox hazard model (Cox, 1972). Early works on joint models were postulated by De Gruttola and Tu (1994), Wulfsohn and

Tsiatis (1997), Henderson et al. (2000), and Tsiatis and Davidian (2001). De Gruttola and Tu (1994) used the shared-parameter approach described above, where they linked the two outcomes by including the same set of random effects in both submodels, whereas Wulfsohn and Tsiatis (1997) used a set of random effects to model the longitudinal process only and used this longitudinal trajectory as a predictor in the proportional hazard model. Instead of using random effects, Henderson et al. (2000) proposed a latent class joint model, where a latent bivariate random process connects the two submodels. All of these methods assumed that the underlying unknown shared parameters (or variables) are normally distributed and used Expectation-Maximization algorithm to estimate the model parameters. Tsiatis and Davidian (2001), on the other hand, proposed a joint modeling technique, where no distributional assumptions were imposed on the random effects and estimated the model parameters via a conditional score approach. Extensions of these joint models include consideration of multiple longitudinal processes (Rizopoulos and Ghosh, 2011) and multiple failure times (Elashoff et al., 2008).

In a longitudinal study, the association between the responses (survival and longitudinal) and the relationships between each response and their corresponding predictors may change over time. The aforementioned shared-parameter models that employ traditional longitudinal and survival models cannot capture this dynamic structure of longitudinal data sets. Therefore, Song and Wang (2008) introduced a joint model for multiple continuous longitudinal processes, where the associations of these processes with the hazard of the subjects were allowed to be flexible functions of time. They proposed two semiparametric estimators, namely, local corrected score and local conditional score, to accurately

estimate the time-varying associations between the outcomes in their hazard model. Ye et al. (2015) postulated a time-varying association joint model where the values of the longitudinal outcome belong to a canonical exponential family, which allowed these values to be both discrete and continuous. They employed functional principal component analyses to model the canonical parameter in their longitudinal submodel and hence, this model cannot investigate the effects of exploratory variables on the longitudinal outcome. However, Andrinopoulou et al. (2018) proposed models that can be employed to explore the response-predictor relationships on both submodels and they estimated the time-dependent association between the true longitudinal biomarker trend and the survival of the subjects via Bayesian P-splines. This model was extended by Piulachs et al. (2021) for zero-inflated longitudinal count data.

Although these authors have flexibly modeled the mean longitudinal trajectories across time, none of these methods allow us to observe the time-varying effects of the risk factors on the longitudinal biomarker while allowing the association between the outcomes to be time-varying. In this project, we fulfill this gap in literature by proposing a flexible and intuitive joint modeling framework, namely, time-varying joint models (TV-JM), that allows all parameters in both submodels to be flexible time-dependent functions. To achieve this goal, we introduce time-varying coefficient models to the joint modeling framework. This novel method will allow researchers to uncover complex dynamic patterns of association between the outcomes and response-predictor relationships, that is, unlike the aforementioned dynamic joint modeling methods, our approach is not limited to exploring time-varying association between the outcomes. In addition, varying coefficient models

can enhance the flexibility of ordinary regression models and reduce the modeling bias by fully capturing the dynamic trends that exist in longitudinal studies (Cleveland et al., 1991; Hastie and Tibshirani, 1993; Fan and Zhang, 2008). Furthermore, our method can accommodate time-varying exogenous predictors in the hazard model.

In our proposed approach, we employ the above mentioned shared-parameter framework, that is, we assume that random effects shared by both outcomes account for the association between the two outcomes as well as the within subject correlation among the repeated measurements of the longitudinal process. Treating these random effects as missing data, we propose an Expectation-Maximization (EM; Dempster et al., 1977) algorithm to estimate the model parameters. In the E-step of the algorithm, we target the random effects and compute the conditional expectation of the complete log-likelihood given the observed data. Since the expected log-likelihood is intractable, we approximate it via a second-degree Taylor's expansion around the estimated mean of the random effects. In the M-step, we apply an iterative Newton-Raphson algorithm to maximize the approximate conditional expected log-likelihood with respect to the parameters. At this step, for the estimation of the time-varying regression coefficients, we employ a local linear regression technique (Fan and Gijbels, 1996). In terms of inference, we study the performance of model-based standard errors in TV-JM and provide practical guidance.

Our motivating data for this project comes from the Women's Interagency HIV Study (WIHS) which started during 1994 and became the largest HIV cohort of women in the United States till date. This study indicated several alarming trends, for example, the cases of AIDS among women increased three times (89%) that of men (29%) between

1990 and 1994. The 1994 rate of AIDS cases among African-American women was twice that of Hispanic and seventeen times that of Caucasian women. By 1995, HIV infection had become the third leading cause of death among U.S. women between ages 25 and 44 and the leading cause of death among African-American women in this age group. This increment in HIV infection among women, especially from the under-represented groups, needed further research to monitor the progression of HIV, time to AIDS, and survival after AIDS, particularly among women. For this purpose, WIHS was established and one of the many objectives of this study was to investigate nutritional, sociodemographic, and behavioral risk factors that may be associated to the rate of the disease progression and time to death. It is known from previous HIV studies (Abrams et al., 1994) that the percentage of CD4+ T cells is a biomarker of HIV progression. The number of these immune cells declines in HIV infected patients, leaving them vulnerable to opportunistic infections, which eventually may lead them to acquired immunodeficiency syndrome (AIDS) and even death. Therefore, it is important to monitor the changes in CD4 with time, conditioned on potential risk factors. Furthermore, it will be of interest to investigate how these factors affect the survival of the patients in long run. Since CD4 percentage and time to death are associated, exploring the changes in this association over years would be of utmost importance as well. In order to accomplish these goals, marginal models were employed in literature, where a linear mixed model was used to fit the CD4 measurements, and the time-to-death outcome was fitted using a Cox model with observed CD4 percentage as time-dependent predictor (Kalish et al., 1999). As CD4 percentage is an endogenous variable, that is, the values of this variable are generated from within the subject and the

survival of the subject is essential in order to obtain these measurements, in other words, since the relationship between the two outcomes is bidirectional, this approach may result in inefficient and biased estimates (Tsiatis and Davidian, 2004; Sweeting and Thompson, 2011). In order to avoid these drawbacks of marginal modeling, we propose joint modeling of CD4 percentage and time to death as the longitudinal and the time-to-event outcomes, respectively.

In addition to estimating the association and response-predictor relationships, we aim to explore the dynamic pattern of these features. The existing joint modeling methodologies for the analysis of WIHS data may provide us with the information on which factors are possibly associated with the progression of HIV or death, but fails to fully capture their dynamic effects on the two outcomes. Our main contribution to statistical literature is to bring two modeling techniques together, that is, we combine varying coefficient models with the joint modeling framework. Our joint modeling approach will provide an estimation framework to obtain both response-predictor and response-response associations efficiently, and the time-varying coefficients will capture the complex dynamic nature of these relationships. Precisely, the introduction of this new class of joint models, that is, time-varying joint models, will significantly improve the flexibility of the existing joint model methodologies. Though the motivation of our model came from the WIHS data, the applicability of these methods go well beyond the study of HIV/AIDS.

In the second part of this dissertation, we perform a novel application of the generalized varying coefficient models (GVCM) to investigate the effects of socioeconomic mobility and major later-life events on the purpose in life during midlife and older adulthood.

This project was motivated by the Midlife in the United States (MIDUS) study. This data consists of longitudinal responses from three waves on purpose in life (PIL) scores of English-speaking adults in the U.S. along with their demographic variables, namely, gender, race, age, socioeconomic status during childhood and adulthood, and major life events such as retirement and widowhood. The data has a significant amount of missing observations (46%) and before application of the GVCM, it is necessary to investigate the performance of this method under missing data. Therefore, we perform extensive simulation studies to show that the estimation procedure under GVCM produces accurate and efficient estimates. Finally, we analyze MIDUS data to explore the time-varying changes in PIL scores based on social mobility, demographic status, and major life events.

This dissertation is organized as follows. Chapter 2 provides a detailed review of the statistical concepts that are relevant to the development of our modeling scheme. In Chapter 3, we introduce our new time-varying joint model (TV-JM) for longitudinal and time-to-event processes. We describe our estimation procedure based on the EM algorithm and local linear regression techniques, and discuss the practical issues related to the estimation. We conduct simulation studies to demonstrate the finite sample behavior of our estimators and we further illustrate the proposed methodology by analyzing the WIHS data. In Chapter 4, we briefly discuss our motivation behind the analysis of purpose in life and describe the MIDUS data in details. We present the estimation procedure of generalized varying coefficient models (GVCM), and provide a brief literature review on missing data mechanisms to show how these challenges are overcome by the GVCM. We conduct simulation studies to demonstrate the performance of GVCM under the missing at random

mechanism. Finally, we present the analysis results of the MIDUS data. In Chapter 5, we present our conclusions and describe some future research topics.

Chapter 2

Literature Review

In this chapter, we briefly review statistical concepts that are relevant to the development of our novel methodology, namely, methods for longitudinal and time-to-event (survival) data, varying-coefficient models (VCMs), and joint modeling techniques for longitudinal and survival outcomes. In Sections 2.1 and 2.2, we review literature on longitudinal and survival data analyses, respectively. A brief summary on VCMs along with estimation procedures and inference are presented in Section 2.3. Finally, Section 2.4 focuses on existing methods developed for joint modeling of longitudinal and survival outcomes.

2.1 Longitudinal Data Analysis

Longitudinal data arises mostly in health and medical sciences, where measurements are taken from the same set of subjects repeatedly over time. There are several challenges in modeling longitudinal data. First, even though the subjects are assumed to be independent of each other, within subject dependence exists due to the repeated

measurements on the same subject. Therefore, traditional regression models, where all the observations are assumed to be independent, cannot be applied to analyze longitudinal data. Additionally, measurement times and number of observations may differ among subjects either due to the design of the study or subjects dropping out or missing their scheduled visits during the study.

A useful parametric statistical tool that accounts for the aforementioned challenges is the linear mixed effects model

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \\ \boldsymbol{\xi}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_\xi), \\ \boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{2.1}$$

where, for the ith subject, $i = 1, \ldots, n$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$ is the response vector, $\mathbf{X}_i$ is the corresponding design matrix for the fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\mathbf{Z}_i$ is the design matrix for the corresponding random effects $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iq})^{\mathrm{T}}$, $\boldsymbol{\varepsilon}_i$ is the biological error, and $\mathbf{I}_{n_i}$ is an identity matrix of order n_i. Note that, in these models, $\boldsymbol{\xi}_i$ and $\boldsymbol{\varepsilon}_i$ are assumed to be independent. One advantage of these models is, due to inclusion of the random effects, they are flexible enough to accommodate different intercepts and slopes for different subjects, allowing us to observe subject-specific response profiles over time. Another flexibility of this model is that it can be employed to analyze longitudinal data with irregular measurement times and number of measurements, that is, subjects can be measured at different time points and different number of times. In the case that the specified random effects model is not sufficient to handle the dependence structure among the repeated measurements, an extension of this model can be considered by assuming $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an

$n_i \times n_i$ covariance matrix, which depends on the ith subject only through its dimension. In this way, we can specify different correlation structures for different subjects. Several well-known correlation structures, such as, uniform, exponential, first order autoregressive, Gaussian, and even an unstructured correlation matrix can be used in this situation.

Estimation of the parameters is often performed via the maximum likelihood estimation principles. As the longitudinal observations of the ith subject are marginally correlated due to the vector of shared random effects $\boldsymbol{\xi}_i$, we rely on the assumption that the repeated measurements for the ith subject are conditionally independent given their random effects,

$$f(\mathbf{Y}_i|\boldsymbol{\xi}_i, \boldsymbol{\theta}) = \prod_{j=1}^{n_i} f(Y_{ij}|\boldsymbol{\xi}_i, \boldsymbol{\theta}).$$

The marginal density for the ith subject $f(\mathbf{Y}_i, \boldsymbol{\theta}) = \int f(\mathbf{Y}_i|\boldsymbol{\xi}_i, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\xi}_i, \boldsymbol{\theta}_\xi) d\boldsymbol{\xi}_i$ has a closed form solution and marginally the response for the ith individual $\mathbf{Y}_i \sim N_{n_i}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{V}_i)$, where $\mathbf{V}_i = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\Sigma}_\xi \mathbf{Z}_i + \sigma^2 \mathbf{I}_{n_i}$. The log-likelihood for the full parameter vector $\boldsymbol{\theta}$ is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{Y}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int f(\mathbf{Y}_i|\boldsymbol{\xi}_i, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\xi}_i, \boldsymbol{\theta}_\xi) d\boldsymbol{\xi}_i, \tag{2.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2, \boldsymbol{\theta}_\xi^{\mathrm{T}})^{\mathrm{T}}$, with $\boldsymbol{\theta}_\xi$ as a vector containing the parameters from $\boldsymbol{\Sigma}_\xi$. If $\mathbf{V}_i$ is known, maximizing (2.2) gives us the maximum likelihood estimates (MLEs) of the fixed effects, which are same as the generalized least squares estimates obtained via

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{V}_i^{-1} \mathbf{X}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{V}_i^{-1} \mathbf{Y}_i. \tag{2.3}$$

In case $\mathbf{V}_i$ is unknown, it can be substituted by its estimate $\widehat{\mathbf{V}}_i$ in (2.3), but the maximum likelihood estimator obtained via maximizing $\ell(\boldsymbol{\theta}_\xi, \sigma^2)$ produces a biased estimator when the sample size is small. A restricted maximum likelihood (REML) approach can be implemented to overcome this drawback, where the estimation of $\mathbf{V}_i$ is performed using iterative

algorithms, such as Expectation-Maximization (EM; Dempster et al., 1977) or Newton-Raphson. The application of these numerical methods to linear mixed effects models can be found in Laird and Ware (1982) and Lindstrom and Bates (1988), respectively.

Verbeke and Lesaffre (1997) studied the asymptotic behavior of the estimators given in equation (2.3) and showed that for normally-distributed random effects, the MLEs of the model parameters are consistent and asymptotically normal with inverse Fisher's information as the asymptotic covariance matrix. However, for non-normal random effects, even though the properties of consistency and asymptotic normality are valid, a sandwich type correction to the Fisher's information matrix is required to obtain an appropriate asymptotic covariance matrix.

A generalization of model (2.1) is the generalized linear mixed effects model

$$g\{E(Y_{ij}|\boldsymbol{\xi}_i)\} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\boldsymbol{\xi}_i,$$

where the conditional distribution of Y_{ij} given the random effects $\boldsymbol{\xi}_i$ follows a distribution $f(Y_{ij}|\boldsymbol{\xi}_i;\boldsymbol{\beta})$ from the exponential family, the repeated measurements, $Y_{i1},\ldots,Y_{in_i}$, are assumed to be independent given $\boldsymbol{\xi}_i$, and $\boldsymbol{\xi}_i = (\xi_{i1},\ldots,\xi_{iq})^{\mathrm{T}}$ is a set of random effects assumed to follow a Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_\xi$. The estimation of the parameters in these models are performed via the maximum likelihood estimation techniques. To derive the likelihood function of the parameters, the random effect $\boldsymbol{\xi}_i$ is treated as a set of unobserved variables and is integrated out as follows

$$L(\boldsymbol{\beta},\boldsymbol{\theta}_\xi;\mathbf{Y}) = \prod_{i=1}^{n}\int\prod_{j=1}^{n_i} f(Y_{ij}|\boldsymbol{\xi}_i;\boldsymbol{\beta})f(\boldsymbol{\xi}_i;\boldsymbol{\theta}_\xi)d\boldsymbol{\xi}_i,$$

where $\boldsymbol{\theta}_\xi$ is a vector containing the parameters from $\boldsymbol{\Sigma}_\xi$. The maximum likelihood estimates of all the parameters are obtained by solving the following observed score equations,

$$\begin{aligned} \mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta}_\xi|\mathbf{Y}) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{ij}\big[Y_{ij} - E\{\mu_{ij}(\boldsymbol{\xi}_i)|\mathbf{Y}_i\}\big] = 0, \\ \text{and } \mathbf{S}_{\boldsymbol{\theta}_\xi}(\boldsymbol{\beta}, \boldsymbol{\theta}_\xi|\mathbf{Y}) &= \frac{1}{2}\boldsymbol{\Sigma}_\xi^{-1}\Big\{\sum_{i=1}^{n} E(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^{\mathrm{T}}|\mathbf{Y}_i)\Big\}\boldsymbol{\Sigma}_\xi^{-1} - \frac{n}{2}\boldsymbol{\Sigma}_\xi^{-1} = 0, \end{aligned}$$

where $\mu_{ij}(\boldsymbol{\xi}_i) = E(Y_{ij}|\boldsymbol{\xi}_i)$. If the response follows a Gaussian distribution, the score equations have closed form solutions, whereas for most non-Gaussian distributions, numerical methods such as EM algorithm can be adopted. EM algorithm iterates between E-and M-steps, where the E-step evaluates the expectations in the score equations using the current values of the parameters, and the M-step solves the score equations to produce the updated parameter estimates. For higher dimensional random effects, due to computational challenges, using Monte Carlo integration is more reasonable than numerical integration methods.

The last topic we discuss in this section is generalized linear mixed models for count data. Assuming that, conditional on the random intercept ξ_i, the longitudinal count observations $Y_{i1}, \ldots, Y_{in_i}$ are independent of each other, and follows a Poisson distribution with

$$\log E(Y_{ij}|\xi_i) = \xi_i + \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \log(t_{ij}),$$

where $\{t_{ij} : i = 1, \ldots, n; j = 1 \ldots, n_i\}$ are the time points when repeated measurements were taken. Given that $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$, the conditional likelihood of the parameters are given by

$$\prod_{i=1}^{n}\binom{Y_{i\cdot}}{Y_{i1}, \ldots, Y_{in_i}}\prod_{j=1}^{n_i}\left(\frac{t_{ij}\exp^{\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}}}{\sum_{l=1}^{n_i} t_{il}\exp^{\mathbf{X}_{il}^{\mathrm{T}}\boldsymbol{\beta}}}\right)^{Y_{ij}},$$

where the contribution of the ith subject is a multinomial probability with

$$\pi_{ij} = \frac{t_{ij} \exp(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta})}{\sum_{l=1}^{n_i} t_{il} \exp(\mathbf{X}_{il}^{\mathrm{T}} \boldsymbol{\beta})}$$

representing the probability that each of the $Y_{i\cdot}$ events will belong to the jth category, for $j = 1, \ldots, n_i$. The estimates are obtained by the method of maximum likelihood.

2.2 Survival Data Analysis

Time-to-event data is collected in numerous applied fields, such as, sociology (Turnbull and Weiss, 1978), epidemiology (Lagakos et al., 1988), medicine (Avalos et al., 1993), and demography (Hyde, 1980). Examples of such data are time to appearance of tumor, heart attack, death after a transplant, failure of a machine, or recovery from a disease after treating with certain medications. In most time-to-event studies, inference becomes challenging due to *loss to follow-up* and *censoring*. Loss to follow-up occurs if information of an active participant in a study is lost after a certain period due to the sudden withdrawal of the individual from the study. If a subject under investigation does not experience the event during the pre-specified study period, we say that observation is "censored". Censoring can be left, right, or interval depending on the situation. If the event occurs after the study period ends, that individual is considered to be right censored. For example, in an animal study, after certain treatments, time to develop a disease of the mice are recorded. If a mouse relapses due to the disease during the follow-up period, the exact event-time is known. But if death does not occur during the follow-up period, the mouse is sacrificed after the specified study period ends, due to budget and time constraints, and that observation is known to be right censored. Left censoring arises when the event occurs

before the investigation was started. For example, if a person is asked when was the first time they started smoking, and the response is that sometime during high school but cannot remember the exact age, which indicates that the event occurred before the observational period.

Let $\mathcal{T}$ be a non-negative random variable from a homogeneous population denoting the time until occurrence of a specified event for an individual. Four functions, namely, survival (or reliability) function, hazard (or risk) function, probability density (or probability mass) function, and mean residual life at time t characterize the distribution of $\mathcal{T}$. Knowing either of these four functions, uniquely defines the rest of the functions. Life-time distribution can be discrete or continuous depending on the study. Usually time is considered to be continuous, but due to rounding in measurement, grouping of event-times into intervals, or simply because of integral values of life-time, $\mathcal{T}$ can take discrete values.

The survival function is defined to be the probability that an individual will experience the event after time t. In the continuous case, the survival function is defined as $\mathcal{S}(t) = P(\mathcal{T} > t)$. This is a monotone non-increasing function, and for right-censored data, equals to one at the beginning of time and approaches zero as time tends to infinity. These basic properties do not change with the change in values of any parameters of the distribution of $\mathcal{T}$. Therefore, this function is not much informative in determining underlying failure patterns, but useful when comparing two or more mortality patterns. Different techniques are required for analyzing discrete life-time distributions. In this project, we only consider the continuous case.

Next, we consider the hazard function. Let Δt be an arbitrary time interval. Hazard is considered to be the conditional probability that an individual who has survived until time t, will die (or experience the event) at the next instant of time. The hazard function is

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq \mathcal{T} < t + \Delta t | \mathcal{T} \geq t]}{\Delta t},$$

where the condition inside the probability, that is, $(\mathcal{T} \geq t)$ indicates that the individual was alive at least until time t. The expression $(t \leq \mathcal{T} < t + \Delta t)$ implies that the individual was alive at exact time t but will not survive until $(t + \Delta t)$. The limiting condition on Δt is to ensure that the person experiences the event in the next instant of time t, in other words, failure occurs within an infinitesimal time after t.

Practically, the most common shape of hazard function is increasing in nature, which is appropriate to describe the hazard associated with natural aging of an individual or wear and tear of machines, but it can be of any other shapes. For example, a hump shaped hazard function describes the failure of a patient after some kind of surgery or organ transplant, when risk is high immediately after the transplant, but starts decreasing after a certain period of time. In demography, it is a commonly known fact that the mortality of children under five years is high, it stabilizes as an infant grows up and again starts increasing after middle age. A bathtub-shaped hazard function is appropriate for this kind of data. A decreasing hazard function might arise due to some electronic devices which are likely to fail at the beginning of use but eventually the risk of failure decreases. It can be shown that hazard function can take different shapes for different parameters of a distribution. For a Weibull distribution, if the shape parameter is a positive fraction,

the hazard decreases with time. For shape parameter equal to one, Weibull has a constant hazard and when shape parameter takes a value greater than one, Weibull hazard increases with time. All these examples indicate that hazard is much useful than survival in examining the underlying failure patterns of individuals. The only practical restriction on hazard is it has to be non-negative. A related quantity in this context is the cumulative hazard $\mathcal{H}(t)$, which is important for verification of goodness-of-fit of a model. The relationship between the survival function and the cumulative hazard is $\mathcal{S}(t) = \exp[-\mathcal{H}(t)]$.

Time-to-event data can be analyzed using parametric, nonparametric, and semi-parametric techniques. Parametric inference can have several limitations when it comes to practical application. For example, exponential distribution has a constant hazard rate making it too restrictive to use in real life data. Hazard function of log-normal and log-logistic distributions are hump shaped, that is, the hazard increases up to a certain value depending on the parameters, and thereafter starts decreasing for large values of t. Hence the usefulness as life-time distributions of these models is criticized in literature, because this situation might be implausible for most data. Contrary to this, analyzing time-to-event data using distributions without assumption of any parameters is an efficient solution to this problem. Nonparametric methods not only provide the underlying empirical distribution of the population, from which the data is collected, they also help investigate if any known parametric model fits the data. For our purpose, we will discuss various nonparametric and semiparametric approaches existing in literature.

Inference methods in this review are based on the assumption that censoring is non-informative, in other words, censoring time of an individual provides no further information

of the likelihood of their survival if they would have continued in the study at a future time point. Censoring time and event-time are independent of each other and the estimators are based on right censored data. We assign $\mathcal{C}$ to the last visiting time of the individual to indicate it is a censored time. Data can be represented by a set of random variables (T, δ), where $\delta = 0$ if a subject is censored. In that case, T takes value $\mathcal{C}$. If the subject experiences the event during the follow-up period, $\delta = 1$ and T equals $\mathcal{T}$. Therefore, we can observe only $T = \min(\mathcal{T}, \mathcal{C})$. Let $t_1 < t_2 < \cdots < t_D$ be the time points when the specified event was experienced by the subjects under study. At time t_j, the number of events occurred is d_j. Let r_j be the number of individuals who are at risk at time t_j, that is, the individuals who are alive, still continuing under the study (that is, not censored before t_j), and/or experience the event at time t_j. The estimate of conditional probability, that a subject who survived just prior to t_j, will experience the event at t_j, is given by $\frac{d_j}{r_j}$.

Nonparametric Inference

The most used nonparametric estimator, known as Product-Limit (PL) estimator, introduced by Kaplan and Meier (1958) is

$$\hat{S}(t) = \begin{cases} 1, & if\ t < t_1 \\ \prod_{t_j \le t}[1 - \frac{d_j}{r_j}], & if\ t_1 \le t. \end{cases}$$

The variance of the PL estimator is given by Greenwood's formula (Greenwood, 1926)

$$\widehat{\mathrm{Var}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{r_j(r_j - d_j)}.$$

As Klein (1991) discussed, though Greenwood's variance estimator of $\hat{S}(t)$ tends to underestimate the true variance of Kaplan-Meier estimator for small to moderate sample sizes, it

comes closest to the true variance of the PL estimator and has a smaller variance in comparison to other variance estimators, except when r_j is very small. Kaplan and Meier (1958) have shown that the PL estimator is the nonparametric maximum likelihood estimator of survival function. It is also approximately unbiased, consistent, and for fixed time t, has approximate normal distribution.

Semiparametric Inference

Often researchers are interested in comparing time-to-event data for two or more groups, for which additional features of subjects might have been recorded. In this model, the response is a set of variables, that is, the observed time to event along with the event indicator. The covariates are the (treatment) groups that are of interest and other relevant explanatory (or predictor) variables that might have effect on life time. For example, demographic variables, such as age, gender, race, education, income; behavioral variables such as alcohol consumption, smoking habits, level of physical activities, diet; or physiological variables such as blood pressure, glucose level, heart rate, may be used as predictors in a regression model because these explain the response variable. These predictors might be fixed throughout the course of the study, such as gender and race, or may be time-dependent, such as age and blood pressure. The multiplicative hazard model introduced by Cox (1972), commonly known as proportional hazard model, with fixed covariates, are commonly used to model these data.

Let $\mathcal{T}$ denote the true time to event. Data is collected from n subjects and is based on the triplet $(T_i, \delta_i, \mathbf{X}_i)$, $i = 1, 2, \ldots, n$, where T_i is the observed event-time (minimum of the true event-time $\mathcal{T}_i$ and the censoring time C_i) for the ith patient, δ_i is the indicator

variable which equals to 1 if the event has occurred and 0 if the patient is right-censored, and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is the vector of fixed (time-invariant) predictors. If $h(t|\mathbf{X})$ is the hazard of an individual at time t with vector of predictors $\mathbf{X}$, then a basic model proposed by Cox (1972) is $h(t|\mathbf{X}) = h_0(t)c(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, where $h_0(t)$ is known as the baseline hazard and is treated nonparametrically. Here, $c(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ is a known function and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of parameters implying that a parametric form is assumed for the covariate effects. Hence this model is known as semiparametric. Since hazard must be positive, a common form of $c(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ is assumed to be exponential. Hence,

$$h(t|\mathbf{X}) = h_0(t)\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) = h_0(t)\exp\left(\textstyle\sum_{k=1}^{p} X_k\beta_k\right). \tag{2.4}$$

This model is known as the proportional hazard model because it can be shown that the ratio of the hazards of two individuals, with covariates $\mathbf{X}$ and $\mathbf{X}^*$, at a given time t, yields a constant equal to $\exp\left[\sum_{k=1}^{p}(X_k - X_k^*)\beta_k\right]$, and this quantity is known as the relative risk of experiencing the event for an individual with covariate $\mathbf{X}$ over an individual with covariate $\mathbf{X}^*$, at a given time t.

The response is the set of variables (T_i, δ_i), for $i = 1, 2, \ldots, n$, dependent on the predictors (or the independent variables), which include discrete (heart rate), continuous (age, blood sugar), and categorical (race, gender) variables. The model can be interpreted as follows. Suppose, we want to know the effect of a treatment on three different races, namely, Caucasian, African-American, and Hispanic, for which, two binary variables can be used. Let $X_1 = 1$, if a patient is Caucasian, 0 otherwise; and $X_2 = 1$, if a patient is African-American, and 0 otherwise. Using equation (2.4), we can interpret that the hazard for a Caucasian patient is $h_0(t)\exp(\beta_1)$, the hazard for an African-American patient is

$h_0(t)\exp(\beta_2)$, and that of a Hispanic patient is $h_0(t)$. Hence the relative risk that a Caucasian and an African-American patient will experience the event over a Hispanic patient, at a given time t, is given by $\exp(\beta_1)$ and $\exp(\beta_2)$, respectively, whereas that of a Caucasian over an African-American patient is given by $\exp(\beta_1 - \beta_2)$. Note that $h_0(t)$ need not be known when comparing the risk of two groups.

For the construction of likelihood of the parameters in the proportional hazard model, censoring is assumed to be non-informative, that is, the event-time and the censoring time for the ith subject is independent. Depending on whether ties are present among event-times or not, various likelihoods of the parameters can be constructed. When there is no tie among the event-times, we have exactly one individual having event-time t_j, that is $d_j = 1$ for all $j = 1, 2, \ldots, D$. We also define a risk set at time t_j as $R(t_j)$, which contains all the individuals who are continuing in the study just prior to t_j, out of which exactly one subject will experience the event at time t_j. The partial likelihood based on the proportional hazard model in (2.4) is given by a product of the ratios, taken over all the time points, where the numerator is the hazard of the subject who experienced the event at t_j, and the denominator consists of the information of hazard on all the subjects exposed to the event at that time point, that is, subjects belonging to the risk set $R(t_j)$. Equating partial derivatives of the log of the likelihood, with respect to the parameters, to zero, yields the efficient score equations. Using Newton-Raphson or any other iterative method produces the maximum likelihood estimates of $\boldsymbol{\beta}$. To test the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, we can adopt large sample approaches such as Wald's test, likelihood ratio test, or score test. All three test statistics asymptotically follow chi-square distribution with p degrees of freedom,

under the null hypothesis. Wald's statistic and the likelihood ratio test statistic have similar rate of convergence, and converges to the limiting chi-square distribution faster than the score statistic. In presence of ties, three different partial likelihoods can be constructed as suggested by Cox (1972), Breslow (1974) and Efron (1977). For small number of ties, the likelihood proposed by Breslow (1974) works well and is similar to Efron's likelihood. The likelihood proposed by Cox (1972) assumes a logistic hazard model for discrete event-times. When there are no ties, all three likelihoods reduce to the partial likelihood discussed above for no ties. To estimate the cumulative hazard or survival function from Cox's model, the baseline hazard must be fitted. For fixed values of $\boldsymbol{\beta}$, the complete censored-data likelihood (Johansen, 1983) can be treated as a function of $h_0(t)$ and the value $\hat{h}_0(t)$ for which the profile likelihood is maximized, can be treated as an estimate of the baseline hazard, at a given time t.

Comparison of Nonparametric and Semiparametric Methods

Despite being the most frequently used estimator of survival function, the PL estimator has its limitations. This estimator is based on the assumption of non-informative censoring. Violation of this assumption may incur bias in estimation. Also, PL estimator is not well defined for any time point larger than the biggest observed time t_{max}. If t_{max} is a true event, this estimator is well defined and the estimated survival beyond this point is zero. But if t_{max} is a censoring time, problem arises in estimation. Assuming $\hat{S}(t) = 0$ beyond t_{max} leads us to assume that all the individuals will experience the event immediately after the study is over, and this causes negative bias in estimation. Approximating $\hat{S}(t)$ by $\hat{S}(t_{max})$ for all $t > t_{max}$ implies that an individual will experience the event at infinity, and

this causes positive bias in estimation. A solution to this problem is to complete the tail by an exponential or Weibull distribution starting at t_{max}.

Another limitation of the PL estimator is that it does not allow for the analysis of the effect of covariates on survival of the individuals. The proportional hazard model can be used as a solution to this problem under semiparametric framework. When comparison between groups are of interest over the underlying pattern of hazard, Cox's model works the best. For example, at a given time point, if the mortality in the control group is twice as much as that of the treatment group, knowledge of the hazard distribution for each group becomes unnecessary, because the proportional hazard model can provide the relative risk of two groups at a given time without any assumption of the underlying hazard distributions of the two populations from which the patients in two groups were selected. But one should be careful about the proportionality assumptions before fitting Cox's model to the data. Several graphical checks exist in literature to verify the proportionality assumption in this hazard model (Andersen, 1982; Klein and Moeschberger, 2005). For example, if a binary predictor X truly fits the proportional hazard model, the plot of the difference of the logarithms of the estimated hazards (or cumulative hazards), given $X = 1$ and $X = 0$ respectively, versus time t, should yield a line equal to the corresponding coefficient of X in (2.4). If the proportionality assumption does not hold for a predictor, that variable can be stratified in a way, so that, different strata have different baseline hazard functions, but within each stratum, the assumption holds for all the predictors.

Cox's model assumes the effect of the predictor variables are time-invariant, only the baseline hazard changes as a function of time. But as mentioned earlier, covariates can

change its values during the course of the study, affecting the risk of experiencing the event of the individuals. A few methods were discussed by Klein and Moeschberger (2005). In one method, instead of using the fixed covariate vector $\mathbf{X}_i$, the covariate information for the ith individual at time t_j is considered to be $\mathbf{X}_i(t_j)$. Inference can be performed following the usual methods discussed above for Cox's model. The problem with this model is that, the coefficients of the covariates are considered to be fixed over time. Another possible solution is to use the time-varying covariate as an indicator variable where it changes value from 0 to 1 after the occurrence of an intermediate event. If proportional hazard is not satisfied for a binary predictor X_1, a time-varying predictor $X_2 = X_1.g(t)$ can be introduced, where $g(t)$ is a function of time, but often the form of $g(t)$ is unknown, making it difficult to analyze the effect of time. A piece-wise proportional hazard model (Matthews and Farewell, 1982) can be considered to find the sudden "change point" of the effect of a covariate, but neither of these methods allow for the continuous effect of time on survival.

2.3 Varying Coefficient Models (VCM)

In statistical data analysis, when the main interest is in investigating the relationship between two or more variables, we often employ parametric regression models. However, despite being popular and widely used in literature, they cannot capture the dynamic features of the data. For instance, in longitudinal studies, repeated measurements on weights of infants (response) born from HIV infected mothers are dependent on maternal vitamin A levels (predictor), where the regression coefficient of the predictor may vary as a function of time (Hoover et al., 1998). In ecological studies, it is believed that net

ecosystem exchange of CO_2 (response) varies as a nonlinear function of photosynthetically active radiation (predictor), where the regression coefficients change depending on the temperature (Kürüm et al., 2014). In both of these real-life applications, the aforementioned parametric regression models would have assumed that these coefficient functions would be constant/time-invariant and miss the dynamic structure in these studies, which would have resulted in large modeling bias and incorrect inference. To increase the flexibility of traditional regression models and to reduce modeling bias, varying coefficient models were introduced (Cleveland et al., 1991; Hastie and Tibshirani, 1993) as

$$Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U) + \varepsilon, \tag{2.5}$$

where Y is the response variable, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ is the vector of predictors with the corresponding regression coefficients $\boldsymbol{\beta}(U) = \{\beta_1(U), \beta_2(U), \ldots, \beta_p(U)\}^{\mathrm{T}}$, which are assumed to vary as functions of the scalar variable U, and ε is the random error with the conditional mean 0 and conditional variance $\sigma^2(u)$. Note that, in this section, U will be referred to as the "covariate".

The coefficient functions in model (2.5) can be estimated by three different methods, namely, local polynomials (Hoover et al., 1998), polynomial splines (Huang and Shen, 2004), and smoothing splines (Hastie and Tibshirani, 1993). Polynomial splines and smoothing splines usually require relatively larger number of parameters, whereas local polynomials, specifically, local linear techniques, can adequately approximate the regression functions locally around the neighborhood of a point u with only two parameters (Fan and Gijbels, 1996). Therefore, we choose to use local linear fitting technique for estimation of the coefficient functions in this dissertation, where instead of increasing the number of parameters

globally, we divide the domain of the covariate U in several neighborhoods, and locally solve linear regression problems in these neighborhoods.

Local polynomial regression is originally a weighted least squares fitting problem, where the weight $K_h(\cdot) = h^{-1}K(\cdot/h)$, known as kernel, is a real valued and symmetric function with a smoothing parameter (or bandwidth) h defining the size of the local neighborhood. Several kernel functions, namely, uniform, triangular, quartic, triweight, tricube, Gaussian, and cosine, are available for estimation. In our method, we use *Epanechnikov* kernel, $K(x) = \frac{3}{4}(1-u^2)$, because of its desirable asymptotic properties, such as the smallest bias and variance (Fan and Gijbels, 1996).

In this local linear estimation technique, we start by approximating the local regression function at each point u, by the local linear regression $\beta_k(u_i) \approx \beta_k(u) + \beta_k'(u)(u_i - u) \equiv a_k + b_k(u_i - u)$, using Taylor's expansion, for all $k = 1, 2, \ldots, p$, and for u_i in the neighborhood of u, where β_k' denotes the first order derivative of β_k. This is a special case of local polynomials, where the degree of the local polynomials are one. We minimize the least squares function

$$\sum_{i=1}^{n} \left[Y_i - \left\{ \mathbf{X}_i^{\mathrm{T}}\mathbf{a} + \mathbf{X}_i^{\mathrm{T}}\mathbf{b}(U_i - u) \right\} \right]^2 K_h(U_i - u)$$

with respect to the local parameters $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$ to estimate these coefficient functions. The estimators which minimize this function are given by $\hat{\boldsymbol{\beta}} = \hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_p)^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}}' = \hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_p)^{\mathrm{T}}$. $\hat{\boldsymbol{\beta}}(u)$ is the linear estimator of $\boldsymbol{\beta}(u)$ and is asymptotically normally distributed according to the results presented by Zhang and Lee (2000).

Theorem 1. Under the conditions provided by Zhang and Lee (2000), we have

$$\mathrm{Cov}^{-1/2}\{\hat{\boldsymbol{\beta}}(u)\}\left[\hat{\boldsymbol{\beta}}(u)-\boldsymbol{\beta}(u)-\mathrm{bias}\{\hat{\boldsymbol{\beta}}(u)\}\right] \xrightarrow{\mathcal{D}} N(\mathbf{0},\mathbf{I}_p),$$

where

$$\mathrm{bias}\{\hat{\boldsymbol{\beta}}(u)\}=\frac{1}{2}\mu_2\boldsymbol{\beta}''(u)h^2, \text{ and } \mathrm{Cov}\{\hat{\boldsymbol{\beta}}(u)\}=\left\{nhf(u)E(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U=u)\right\}^{-1}\nu_0\sigma^2(u),$$

$\mu_\kappa=\int u^\kappa K(u)du$, $\nu_\kappa=\int u^\kappa K^2(u)du$, $f(u)$ is the marginal density of U, $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^{\mathrm{T}}$, and $\mathbf{I}_p$ is the identity matrix of order p.

The choice of bandwidth plays a crucial role in the estimation of the coefficient functions. When a coefficient function is simple and homogeneous in nature, a large and constant bandwidth is sufficient. However, if the function is complex, that is, with irregular crests and troughs, a smaller bandwidth becomes essential to accurately approximate and capture the trends in the coefficient function. A smaller bandwidth helps to reduce the modeling bias by assigning more weights to the closest neighboring observations around u, but due to the smaller size of the neighborhood, only a few observations are considered for the estimation, which may increase the variance of the estimator. On the other hand, a larger bandwidth implies that the local weights are distributed to more number of data points. This helps controlling the variance, but may increase the modeling bias due to inaccurate approximations led by the inclusion of observations which are far away from u. Hence, we need to find an optimal bandwidth for this bias-variance trade-off. The cross-validation technique is widely used for the selection of the bandwidth. In this technique, the ith observation is removed to fit the model to the data, for all $i=1,2,\ldots,n$, and the corresponding residual sum of squares (RSE) is calculated in each step. An optimal

bandwidth is the one which minimizes the cross-validation score $CV(h) = \sum_i (Y_i - \hat{Y}_{-i})^2$, where $\hat{Y}_{-i}$ is the fitted value with ith observation excluded. This method assumes that all coefficient functions in the model (2.5) have the same level of smoothness. Zhang and Lee (2000) proposed a variable bandwidth selection procedure for estimating complicated coefficient functions having different degrees of smoothness.

A generalization of model (2.5) to a generalized linear regression setting was proposed by Cai et al. (2000). They employed a local maximum likelihood estimation (MLE) procedure to estimate the varying coefficients in the model

$$g\{m(u, \mathbf{x})\} = \sum_{k=1}^{p} \beta_k(u) x_k,$$

where $m(u, \mathbf{x})$ is the conditional mean regression function of the response Y given $\mathbf{X} = \mathbf{x}$ and $U = u$. For each given point u, the coefficient function for the ith individual, $\beta_k(u_i)$ is approximated by a local linear regression $\beta_k(u_i) \approx a_k + b_k(u_i - u)$ for u_i in the neighborhood of u, where $a_k = \beta_k(u)$ and $b_k = \beta_k'(u)$. With $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$ as the parameters, the local likelihood function,

$$\ell(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(g^{-1}\left[\sum_{k=1}^{p} \Big\{ a_k + b_k(U_i - u) \Big\} X_{ik} \right], Y_i \right) K_h(U_i - u)$$

is maximized to obtain $\hat{\boldsymbol{\beta}} = \hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_p)^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}}' = \hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_p)^{\mathrm{T}}$, the MLE of $\mathbf{a}$ and $\mathbf{b}$, respectively. Due to the generalized linear relationship between Y and $\mathbf{X}$, it is not possible to obtain closed form solutions for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ for non-Gaussian distributions. An iterative maximum likelihood technique using Newton-Raphson algorithm can be used to estimate the parameters, but that is computationally expensive due to the use of cross-validation as the selection procedure of the bandwidth, the presence of many predictors,

and most importantly because the likelihood is to be maximized for numerous values of u. Hence, Cai et al. (2000) proposed a one-step estimator of the parameters using Newton-Raphson algorithm. Let $\hat{\boldsymbol{\beta}}_0$ be a given initial estimator of $\boldsymbol{\beta}$, and if $\ell'(\boldsymbol{\beta})$ and $\ell''(\boldsymbol{\beta})$ be the gradient and hessian of the likelihood $\ell(\boldsymbol{\beta}) = \ell(\mathbf{a}, \mathbf{b})$ respectively, then the proposed one-step estimator is given by

$$\hat{\boldsymbol{\beta}}_{OS} = \hat{\boldsymbol{\beta}}_0 - \{\ell''(\hat{\boldsymbol{\beta}}_0)\}^{-1}\ell'(\hat{\boldsymbol{\beta}}_0).$$

Cai et al. (2000) derived the following theorem to show the asymptotic normality of the estimators obtained using the iterative likelihood algorithm.

Theorem 2. Under the conditions provided by Cai et al. (2000), when $h = h_n \to 0$, and $nh \to \infty$ as $n \to \infty$,

$$\sqrt{nh}\left[\mathbf{H}\Big\{\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u)\Big\} - \frac{h^2}{2(\mu_2 - \mu_1^2)}\begin{Bmatrix}(\mu_2^2 - \mu_1\mu_3)\boldsymbol{\beta}''(u)\\ \\ (\mu_3 - \mu_1\mu_2)\boldsymbol{\beta}''(u)\end{Bmatrix} + o_p(h^2)\right] \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^{-1})$$

Furthermore, for a symmetric kernel function $K(\cdot)$,

$$\sqrt{nh}\left\{\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u) - \frac{h^2\mu_2}{2}\boldsymbol{\beta}''(u) + o_p(h^2)\right\} \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \boldsymbol{\Sigma}(u)\},$$

where $\mu_\kappa = \int u^\kappa K(u)du$, $\nu_\kappa = \int u^\kappa K^2(u)du$, $\mathbf{H} = \text{diag}(1, h) \otimes \mathbf{I}_p$, $\otimes$ is the Kronecker product, $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\boldsymbol{\Sigma}(u) = \nu_0\boldsymbol{\Gamma}^{-1}(u)/f(u)$ with $f(u)$ as the marginal density of U, $\boldsymbol{\Gamma}(u) = E\big[\rho(U, \mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}|U = u\big]$, $\rho(u, \mathbf{x}) = [g_1 m(u, \mathbf{x})]^2\text{Var}(Y|U = u, \mathbf{X} = \mathbf{x})$, $g_1(\cdot) = g_0^{(1)}(\cdot)/g^{(1)}(\cdot)$, $g_0(\cdot)$ is the canonical link,

$$\boldsymbol{\Delta} = f_U(u)\begin{pmatrix}1 & \mu_1\\ \\ \mu_1 & \mu_2\end{pmatrix} \otimes \boldsymbol{\Gamma}(u), \text{ and } \boldsymbol{\Lambda} = f_U(u)\begin{pmatrix}\nu_0 & \nu_1\\ \\ \nu_1 & \nu_2\end{pmatrix} \otimes \boldsymbol{\Gamma}(u).$$

They have also proved under the conditions of Theorem 2, if the initial estimator $\hat{\boldsymbol{\beta}}_0$ satisfies $\mathbf{H}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) = O_p\big\{h^2 + (nh)^{-1/2}\big\}$, the one-step local MLE $\hat{\boldsymbol{\beta}}_{OS}$ has the same asymptotic

distribution as the fully iterative MLE $\hat{\boldsymbol{\beta}}$, and hence using the one-step estimator increases the computational efficiency significantly, without sacrificing asymptotic performance. They have also proposed a nonparametric maximum likelihood ratio test as a test of significance of the parameters, as well as to investigate if the coefficient functions are really varying. The asymptotic normality of the null distribution of their test statistic was established by a conditional bootstrap method. Please refer to Cai et al. (2000) for further details.

The extension of varying coefficients to a nonlinear setting was proposed by Kürüm et al. (2014). In their model, the relationship between the response for the ith subject, that is, Y_i, and the corresponding predictors $\mathbf{X}_i$ is allowed to be a nonlinear function as follows

$$Y_i = f\{\mathbf{X}_i, \boldsymbol{\beta}(U_i)\} + \epsilon_i, \tag{2.6}$$

where $f(.)$ is a pre-specified function. Estimation under this framework has challenges. First, although pre-specified, $f(.)$ can be nonlinear; therefore, closed form solutions for the local parameters may not exist. Second, using Newton-Raphson algorithm for estimation may lead to a Hessian matrix that is not positive definite. Hence, in order to overcome these problems, Kürüm et al. (2014) proposed an iterative local linear search algorithm, where they minimized the local least squares,

$$\ell(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \Big[Y_i - f\big\{\mathbf{X}_i, \mathbf{a} + \mathbf{b}(U_i - u)\big\} \Big]^2 K_h(U_i - u)$$

with respect to the parameters $\mathbf{a}$ and $\mathbf{b}$, to obtain the local estimators $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, respectively. During the course of the algorithm, they updated the estimates as follows

$$\begin{pmatrix} \mathbf{a}^{(\kappa+1)} \\ \\ \mathbf{b}^{(\kappa+1)} \end{pmatrix} = (\mathbf{F}_\kappa^{\mathrm{T}} \mathbf{W} \mathbf{F}_\kappa)^{-1} \mathbf{F}_\kappa^{\mathrm{T}} \mathbf{W} \mathcal{Y}_\kappa,$$

where $\mathbf{a}^{(\kappa+1)}$ denotes the $(\kappa+1)$th iteration for estimating $\mathbf{a}$, $\mathbf{F} = \mathbf{F}_\kappa$ is an $n \times 2p$ matrix with the ith row as

$$\left[f'^{\mathrm{T}}\{\mathbf{X}_i, \mathbf{a}^{(\kappa)} + \mathbf{b}^{(\kappa)}(U_i - u)\},\ (U_i - u)f'^{\mathrm{T}}\{\mathbf{X}_i, \mathbf{a}^{(\kappa)} + \mathbf{b}^{(\kappa)}(U_i - u)\}\right],$$

$\mathbf{W} = \mathrm{diag}\{K_h(U_1 - u), \ldots, K_h(U_n - u)\}$, and $\mathcal{Y}_\kappa = (Y_{1,\kappa}, \ldots, Y_{n,\kappa})^{\mathrm{T}}$ with

$$Y_{i,\kappa} = Y_i - f\{\mathbf{X}_i, \mathbf{a}^{(\kappa)} + \mathbf{b}^{(\kappa)}(U_i - u)\} + \{\mathbf{a}^{(\kappa)} + \mathbf{b}^{(\kappa)}(U_i - u)\}^{\mathrm{T}} f'\{\mathbf{X}_i, \mathbf{a}^{(\kappa)} + \mathbf{b}^{(\kappa)}(U_i - u)\}.$$

The solution of this iterative linear search algorithm satisfies $\ell(\mathbf{a}, \mathbf{b}) = 0$. Kürüm et al. (2014) proved the following theorems to present the asymptotic normality of these estimators and to derive the consistent estimator of their asymptotic variances. Let $\boldsymbol{\theta}(u) = (a_1, \ldots, a_p, b_1, \ldots, b_p)^{\mathrm{T}}$, $\hat{\boldsymbol{\theta}}(u) = (\hat{\mathbf{a}}^{\mathrm{T}}, \hat{\mathbf{b}}^{\mathrm{T}})^{\mathrm{T}}$, $c(u)$ as the marginal density of U,

$$\boldsymbol{\Gamma}_1(u) = E\left(f'\{\mathbf{X}; \boldsymbol{\beta}(u)\}\left[f'\{\mathbf{X}; \boldsymbol{\beta}(u)\}\right]^{\mathrm{T}} \middle| U = u\right)_{p \times p},$$

and

$$\boldsymbol{\Gamma}_2(u) = E\left(\sigma^2(u, \mathbf{X}) f'\{\mathbf{X}; \boldsymbol{\beta}(u)\}\left[f'\{\mathbf{X}; \boldsymbol{\beta}(u)\}\right]^{\mathrm{T}} \middle| U = u\right)_{p \times p}.$$

Theorem 3(a). Under the regularity conditions by Kürüm et al. (2014), as $n \to \infty$,

$$\sqrt{nh}\left[\mathbf{H}\{\hat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}(u)\} - \frac{h^2}{2(\mu_2 - \mu_1^2)} \begin{Bmatrix} (\mu_2^2 - \mu_1\mu_3)\boldsymbol{\beta}''(u) \\ (\mu_3 - \mu_1\mu_2)\boldsymbol{\beta}''(u) \end{Bmatrix} + o_p(h^2)\right] \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^{-1}),$$

where

$$\boldsymbol{\Delta} = c(u) \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes \boldsymbol{\Gamma}_1(u) \text{ and } \boldsymbol{\Lambda} = c(u) \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \boldsymbol{\Gamma}_2(u).$$

Furthermore, for symmetric $K(\cdot)$, it can be shown that, as $n \to \infty$,

$$\sqrt{nh}\left\{\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u) - \frac{h^2\mu_2}{2}\boldsymbol{\beta}''(u) + o_p(h^2)\right\} \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \boldsymbol{\Sigma}(u)\},$$

where $\boldsymbol{\Sigma}(u) = \nu_0\boldsymbol{\Gamma}_1^{-1}(u)\boldsymbol{\Gamma}_2(u)\boldsymbol{\Gamma}_1^{-1}(u)/c(u)$.

Theorem 3(b). Under the regularity conditions (Kürüm et al., 2014), as $n \to \infty$,

$$\mathbf{H}(\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathcal{Q}\mathbf{W}\mathbf{F}(\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{H} \xrightarrow{\mathcal{P}} \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^{-1},$$

where $\mathcal{Q} = \mathrm{diag}(e_1^2, \ldots, e_n^2)$ with $e_i = Y_i - f\{\mathbf{X}_i, \hat{\boldsymbol{\beta}}(U_i)\}$. If the errors are assumed to be normal, their algorithm is same as Fisher's scoring algorithm, and hence shares its convergence properties. The consistent estimator of asymptotic variance of the one-step estimator is given by

$$\widehat{\mathrm{Cov}}\{\hat{\boldsymbol{\theta}}(u)\} = (\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathcal{Q}\mathbf{W}\mathbf{F}(\mathbf{F}^{\mathrm{T}}\mathbf{W}\mathbf{F})^{-1}.$$

It is shown that this iterative algorithm works well only when all of the coefficient functions have same degree of smoothness. For the coefficient functions with different degrees of smoothness, they have proposed a two-step local linear estimator and used a bootstrap confidence interval for inferences. Though the two-step algorithm can increase computational cost, it gains significant efficiency in the estimation. In addition to these estimators, Kürüm et al. (2014) suggested a generalized F-test for testing the significance and the varying nature of the coefficient functions, where they proved that the null distribution of the generalized likelihood ratio test statistic is chi-square.

Time-varying Coefficient Models

As we described in Section 2.1, repeated measurements of a response, collected longitudinally from a set of subjects, result in correlated data within each subject, for which the response-predictor relationships may change over time. Widely used parametric models specified in Section 2.1 do not allow the coefficients to vary as functions of time and ignoring the effect of time may induce large modeling bias. Considering time as the covariate in a VCM, Hoover et al. (1998) proposed the following model,

$$Y_i(t) = \mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t) + \varepsilon_i(t),$$

where $Y_i(t)$ and $\mathbf{X}_i(t) = \{1, X_{i1}(t), \ldots, X_{ip}(t)\}^{\mathrm{T}}$ denote the response and the vector of predictors at time t, respectively, and the measurement times are denoted as $t = t_{ij}$, for $i = 1, \ldots, n$ and $j = 1, 2, \ldots, n_i$. The coefficient functions $\boldsymbol{\beta}(t) = \{\beta_0(t), \beta_1(t), \ldots, \beta_p(t)\}^{\mathrm{T}}$ are allowed to be smooth nonparametric functions and $\varepsilon_i(t)$ is a zero mean stochastic process. In this model, the subjects were assumed to be independent, but within subject correlation is present over time. Hoover et al. (1998) derived two nonparametric estimators, namely, smoothing spline and a locally weighted polynomial. Let us briefly describe the smoothing spline method here. This method involves minimizing the objective function

$$J(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_i(t_{ij}) - \left\{\sum_{k=0}^{p} X_{ik}(t_{ij})\beta_k(t_{ij})\right\}\right]^2 + \sum_{k=0}^{p}\lambda_k \int \{\beta_k''(t)\}^2\, dt,$$

where the first term is a measure of the model bias and the second term penalizes the roughness of the coefficient functions through the positive-valued smoothing parameters $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$. Large λ_k gives excessive penalty resulting in oversmoothed coefficients and small λ_k produces undersmoothed coefficients. Therefore, the choice of $\boldsymbol{\lambda}$ becomes

very important in practice. For a single predictor, a scatter plot of the data may help to subjectively decide a value for $\boldsymbol{\lambda}$, but for more than one predictors, an automated technique is required. Hence, Hoover et al. (1998) proposed a cross-validation criterion to select the smoothing parameter $\boldsymbol{\lambda}$. In particular, they suggested a cross-validation technique where the entire set of observations for a subject is left out at a time rather than a single observation, since the latter approach is inappropriate when there is intra-subject dependence.

Hoover et al. (1998) also derived kernel estimators for the coefficient functions, which is a special case of the local polynomials, where the local regression function is approximated by a polynomial of degree zero, and established their asymptotic properties. In addition, they proved that smoothing splines become advantageous over local polynomials, due to the presence of multiple smoothing parameters.

Time-varying coefficient models can be extended to generalized linear settings. Kürüm et al. (2016) suggested these models, which are the extensions of generalized varying coefficient model proposed by Cai et al. (2000) to longitudinal settings. The outcome $Y_i(t)$ in their model was a binary time-dependent response, which was assumed to have an underlying normal latent variable $W_i(t) = \mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t) + \varepsilon_i(t)$ and hence, modeled through the probit link

$$P\{Y_i(t) = 1|\mathbf{X}_i(t)\} = \Phi\left\{\frac{\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t)}{\sigma(t)}\right\},$$

where $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^{\mathrm{T}}$ is the vector of predictors with corresponding vector of time-varying coefficients $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_p(t)\}^{\mathrm{T}}$, and the random error $\boldsymbol{\varepsilon}_i(t)$ follows a normal distribution with zero mean and variance $\sigma^2(t)$. With t_{ij} as the jth measurement

time for the ith subject, for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, the authors maximized the following local log-likelihood

$$\ell(\mathbf{a}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ell\left(g^{-1}\left[\sum_{k=1}^{p} \left\{ a_k + b_k(t_{ij} - t_0) \right\} X_{ik}(t_{ij}) \right], Y_i(t_{ij}) \right) K_h(t_{ij} - t)$$

with respect to $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$, where (a_k, b_k) are such that $\beta_k(t_{ij}) \approx \beta_k(t_0) + \beta_k'(t_0)(t_{ij} - t_0) \equiv a_k + b_k(t_{ij} - t_0)$, t_{ij} in the neighborhood of size h around t_0, h is the bandwidth of the kernel function, $g(\cdot)$ is the link function, and N is the total number of observations. They proposed an iterative algorithm where the parameters are updated as follows:

$$\begin{pmatrix} \mathbf{a}^{(\kappa+1)} \\ \\ \mathbf{b}^{(\kappa+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(\kappa)} \\ \\ \mathbf{b}^{(\kappa)} \end{pmatrix} - \{\ell''(\mathbf{a}, \mathbf{b})\}^{-1} \ell'(\mathbf{a}, \mathbf{b}),$$

where $\{\mathbf{a}^{(\kappa)}, \mathbf{b}^{(\kappa)}\}^{\mathrm{T}}$ denotes the κth iteration for estimating $(\mathbf{a}, \mathbf{b})^{\mathrm{T}}$, and the solution of this iterative linear algorithm satisfies $\ell(\mathbf{a}, \mathbf{b}) = 0$. Kürüm et al. (2016) also studied the asymptotic behavior of these estimators and showed that they are asymptotically normally distributed. Please refer to Kürüm et al. (2016) for further details on this method.

Time-varying coefficients can be used for modeling survival data as well. In survival analysis, the response for the ith subject is a set of variables (T_i, δ_i). Here, $T_i = \min(\mathcal{T}_i, \mathcal{C}_i)$ is the observed event-time outcome with the event indicator δ_i, where $\delta_i = 1$ if $T_i = \mathcal{T}_i$ (true event-time), and $\delta_i = 0$ if $T_i = \mathcal{C}_i$ (censoring time). The survival of an individual may be affected by a set of predictors $\mathbf{X}(t) = \{X_1(t), \ldots, X_p(t)\}^{\mathrm{T}}$, which may include time-invariant predictors as well. Note that, the time-varying predictors are exogenous. The most commonly used model for survival data is Cox's proportional hazard model, where

Fan et al. (2006) introduced time-varying coefficients as follows,

$$h(t|\mathbf{X}, U) = h_0(t) \exp\left[\beta_0\{U(t)\} + \mathbf{X}^{\mathrm{T}}(t)\boldsymbol{\beta}\{U(t)\}\right], \tag{2.7}$$

where $\beta_0\{U(t)\}$ and $\boldsymbol{\beta}\{U(t)\} = [\beta_1\{U(t)\}, \ldots, \beta_p\{U(t)\}]^{\mathrm{T}}$ are unknown coefficient functions depending on time t through an exposure variable $U(\cdot)$. Note that, when $U(t) = t$, model (2.7) becomes time-dependent Cox's hazard model, where the proportionality assumption no longer holds, unless $\boldsymbol{\beta}(t)$ is time invariant. However, model (2.7) can still be used to investigate the extent to which the predictors $\mathbf{X}(t)$ interact nonlinearly with the exposure variable $U(t)$. It is notable that the term $\beta_0\{U(t)\}$ is not incorporated with the predictors $\mathbf{X}(t)$ because the local intercept for $\beta_0(\cdot)$ cancels out in the local partial likelihood in (2.8) leading to a different estimator rule for β_0.

The estimation for the model (2.7) is performed via maximizing the partial likelihood

$$L\{\beta_0(\cdot), \boldsymbol{\beta}(\cdot)\} = \prod_{i=1}^{n} \left[\frac{\exp\left\{\beta_0(U_i) + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}(U_i)\right\}}{\sum_{j\epsilon R(t_i)} \exp\left\{\beta_0(U_j) + \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}(U_j)\right\}}\right]^{\delta_i},$$

where $R(t) = \{i : T_i \geq t\}$ is the risk set at time t. By Taylor's expansion, the local regression functions at u_i, in the neighborhood of u of size h, are approximated by $\beta_{1k}(u_i) \approx \beta_{1k}(u) + \beta'_{1k}(u)(u_i - u) \equiv a_{1k} + b_{1k}(u_i - u)$, for all $k = 1, 2, \ldots, p$, and $\beta_0(u_i) \approx \beta_0(u) + \beta'_0(u)(u_i - u) \equiv a_0 + b_0(u_i - u)$. Substituting the approximations in the above likelihood, the partial log-likelihood for $\mathbf{a}_1 = (a_{11}, \ldots, a_{1p})^{\mathrm{T}}$, $\mathbf{b}_1 = (b_{11}, \ldots, b_{1p})^{\mathrm{T}}$ and b_0 can be given by

$$\ell(\mathbf{a}_1, \mathbf{b}_1, b_0) = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)\delta_i \times \Bigg(\mathbf{X}_i^{\mathrm{T}}\mathbf{a}_1 + \mathbf{X}_i^{\mathrm{T}}\mathbf{b}_1(U_i - u) + b_0(U_i - u) - \log\left[\sum_{j\epsilon R(t_i)} \exp\Big\{\mathbf{X}_j^{\mathrm{T}}\mathbf{a}_1 + \mathbf{X}_j^{\mathrm{T}}\mathbf{b}_1(U_j - u) + b_0(U_j - u)\Big\}K_h(U_j - u)\right]\Bigg), \tag{2.8}$$

where $K_h(\cdot)$ is a the kernel function with bandwidth h. Equation (2.8) can be numerically solved to find estimates of the parameters using the Newton-Raphson or Fisher's scoring algorithms. However, these algorithms are computationally costly and in some applications, due to very few number of data points around u, the local partial-likelihood estimators might not exist. Hence, Fan et al. (2006) proposed a one-step local partial likelihood estimator

$$\hat{\boldsymbol{\zeta}}_{OS} = \hat{\boldsymbol{\zeta}}_0 - \{\ell''(\hat{\boldsymbol{\zeta}}_0)\}^{-1}\ell'(\hat{\boldsymbol{\zeta}}_0),$$

where $\hat{\boldsymbol{\zeta}}_{OS}$ is the one-step estimator and $\hat{\boldsymbol{\zeta}}_0$ is a given initial value of $\boldsymbol{\zeta} = \{\boldsymbol{\beta}_1^{\mathrm{T}}, (\boldsymbol{\beta}_1')^{\mathrm{T}}, \beta_0'\}^{\mathrm{T}}$, respectively.

The asymptotic normality of the maximum partial likelihood estimator can be given by the Theorem 4. Let $\hat{\boldsymbol{\zeta}}(u)$ be the maximum partial likelihood estimator of the vector containing true coefficient functions $\boldsymbol{\zeta}(u) = \{\boldsymbol{\beta}_1^{\mathrm{T}}(u), \boldsymbol{\beta}_1'(u)^{\mathrm{T}}, \beta_0'(u)\}^{\mathrm{T}}$.

Theorem 4. Under the conditions provided by Fan et al. (2006),

$$\sqrt{nh}\big[\mathbf{H}\big\{\hat{\boldsymbol{\zeta}}(u) - \boldsymbol{\zeta}(u)\big\} - \frac{1}{2}h^2\boldsymbol{e}_p\boldsymbol{\zeta}''(u)\mu_2\big] \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \boldsymbol{\Sigma}(\tau, u)\},$$

where $\boldsymbol{\Sigma}(\tau, u)$ is the covariance matrix, τ is a finite time point up to which the data points are used, $\mathbf{H}$ is a $(2p+1)$ order diagonal matrix, with the first p diagonal elements as 1 and the remaining $(p+1)$ elements as h, and $\boldsymbol{e}_p$ is another diagonal matrix of order $(2p+1)$ with the first p diagonal elements as 1 and the rest $(p+1)$ elements as 0. The authors proved that under the conditions of Theorem 4, if $\mathbf{H}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) = O_p\{h^2 + (nh)^{-1/2}\}$, then the one-step estimator $\hat{\boldsymbol{\zeta}}_{OS}$ has the same asymptotic distribution as the maximum local partial likelihood estimator $\hat{\boldsymbol{\zeta}}$, which means that the one-step estimation technique is as efficient as the fully iterative algorithm. In addition, Fan et al. (2006) derived consistent estimators of bias and variance of $\hat{\boldsymbol{\zeta}}$.

The authors proposed a consistent estimator of baseline hazard $h_0(t)$ using the kernel method. The expression for baseline hazard is given as

$$\hat{h}_0(t) = \int K_h(t-s) d\hat{\mathscr{H}}_0(s),$$

where $K_h(\cdot)$ is a the kernel function with bandwidth h and $\hat{\mathscr{H}}_0(t)$ is an estimate of cumulative hazard function

$$\hat{\mathscr{H}}_0(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{d\mathscr{N}_i(s)}{\frac{1}{n}\sum_{j=1}^{n} I(T_j \geq s)\exp\{\hat{\beta}_0(U_j) + \mathbf{X}_j^{\mathrm{T}}(s)\hat{\boldsymbol{\beta}}(U_j)\}}$$

with $\mathscr{N}_i(s) = I(\mathscr{T}_i \leq s, \delta_i = 1)$.

2.4 Joint Models

A number of models have been developed for the joint modeling of longitudinal and survival outcomes. A commonly used modeling framework is the shared-parameter models. In this approach, the longitudinal and the time-to-event outcomes are assumed to be conditionally independent given a set of underlying latent variables shared by both submodels, which in most cases, are included in the form of random effects. Additionally, these shared parameters or the random effects account for the within subject correlation of the longitudinal repeated measurements. Note that the longitudinal measurements are taken from the subjects intermittently and these observations might contain measurement errors. Therefore, the shared-parameter models focus on modeling the "true" unobserved longitudinal process to have complete information on the history of the subjects until the observed event-time. Tsiatis et al. (1995), and Dafni and Tsiatis (1998) proposed a two-stage approach, where a separate model is fit to each outcome, to estimate the parameters in the

shared-parameter models. At the first stage of their estimation procedure, empirical Bayes estimate of the longitudinal outcome was computed at each event-time by using growth curve models with random effects. In the second stage, a Cox model, where the estimated longitudinal values were included as a predictor, was fit via a partial likelihood. However, Wulfsohn and Tsiatis (1997), Tsiatis and Davidian (2004), and Sweeting and Thompson (2011) argued that estimation methods, where both outcomes are analyzed jointly, yields more efficient and accurate estimates than the two-stage approaches.

We describe an improved joint likelihood-based method (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2004) in this section. Let $Y_i(t)$ and $m_i(t)$ be the respective observed and "true" longitudinal response for the ith subject at time $t = t_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. To account for the subject-specific time dependence, the longitudinal submodel at time t for the ith subject is given by a linear mixed effects model

$$\begin{aligned} Y_i(t) &= m_i(t) + \varepsilon_i(t), \\ m_i(t) &= \mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}(t)\boldsymbol{\xi}_i, \\ \boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\xi), &\quad \text{and} \quad \varepsilon_i(t) \overset{iid}{\sim} N(0, \sigma^2), \end{aligned} \tag{2.9}$$

where $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^{\mathrm{T}}$ is the design matrix for the fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ at time t, and $\mathbf{Z}_i(t) = \{Z_{i1}(t), \ldots, Z_{iq}(t)\}^{\mathrm{T}}$ is the design matrix for the subject-specific random effects $\boldsymbol{\xi}_i = (\xi_1, \ldots, \xi_q)^{\mathrm{T}}$. The random effects $\boldsymbol{\xi}_i$ and the measurement errors ε_i are assumed to be independent of each other.

Let $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ be the history of the true unobserved longitudinal process $m_i(t)$ up to time point t, $T_i = \min(\mathcal{T}_i, \mathcal{C}_i)$ be the observed event outcome of the ith individual, with an indicator δ_i, where $\delta_i = 1$ if $T_i = \mathcal{T}_i$ (event-time), and $\delta_i = 0$ if

$T_i = C_i$ (censoring time). Then the survival submodel for the ith subject is given by the proportional hazard model

$$h_i\{t|\mathcal{M}_i(t), \mathbf{W}_i\} = h_0(t)\exp\{\gamma m_i(t) + \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\eta}\}, \tag{2.10}$$

where $h_0(t)$ is the baseline hazard, $\mathbf{W}_i$ is the vector of baseline exposure variables (such as gender and race), $\boldsymbol{\eta}$ is the regression parameter vector corresponding to the baseline variables, and γ is the measure of extent to which the true longitudinal process has an impact on the event-time. Model (2.10) implies that the relative risk of two subjects i and i^* for an event at time t depends only on the current value of their time-dependent marker $m_i(t)$ and $m_{i^*}(t)$. However, this is not true for the survival function

$$\mathcal{S}_i(t|\mathcal{M}_i(t), \mathbf{W}_i) = P\{\mathcal{T}_i > t|\mathcal{M}_i(t), \mathbf{W}_i\} = \exp\left[-\int_0^t h_0(s)\exp\left\{\gamma m_i(s) + \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\eta}\right\}ds\right],$$

because it depends on the entire history of that patient until time t. This feature is one of the important things to be considered as the survival function is a part of the joint likelihood. To specify model (2.10) fully, discussion of choice of the baseline hazard $h_0(t)$ is important. In standard survival analysis, this term is left unspecified to avoid bias induced by misspecification of its distribution, but within joint modeling framework, this choice leads to underestimation of the standard errors of the model parameters (Hsieh et al., 2006). Therefore, "flexible parametric" models, namely, piece-wise constant and regression spline approaches can be adopted (Rizopoulos, 2012).

Based on the assumption that the vector of the time-independent random effects $\boldsymbol{\xi}_i$ underlies both longitudinal and survival processes, and thus accounts for the association between these outcomes as well as the within subject correlation of the repeated measurements, the observed (incomplete) likelihood $f(\mathbf{Y}_i, T_i, \delta_i; \boldsymbol{\theta})$ is obtained by integrating out

the unknown random effects $\boldsymbol{\xi}_i$ from the complete data likelihood as follows

$$\begin{aligned} f(\mathbf{Y}_i, T_i, \delta_i; \boldsymbol{\theta}) &= \int f(\mathbf{Y}_i, T_i, \delta_i, \boldsymbol{\xi}_i; \boldsymbol{\theta}) d\boldsymbol{\xi}_i \\ &= \int f(\mathbf{Y}_i, T_i, \delta_i | \boldsymbol{\xi}_i; \boldsymbol{\theta}) f(\boldsymbol{\xi}_i; \boldsymbol{\theta}) d\boldsymbol{\xi}_i \\ &= \int f(\mathbf{Y}_i | \boldsymbol{\xi}_i; \boldsymbol{\theta}) f(\boldsymbol{\xi}_i; \boldsymbol{\theta}) f(T_i, \delta_i | \boldsymbol{\xi}_i; \boldsymbol{\theta}) d\boldsymbol{\xi}_i \\ &= \int \left[\prod_{j=1}^{n_i} f\big\{ Y_i(t_{ij}) | \boldsymbol{\xi}_i; \boldsymbol{\theta} \big\} \right] f(\boldsymbol{\xi}_i; \boldsymbol{\theta}) f(T_i, \delta_i | \boldsymbol{\xi}_i; \boldsymbol{\theta}) d\boldsymbol{\xi}_i, \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_s^{\mathrm{T}}, \boldsymbol{\theta}_\xi^{\mathrm{T}})^{\mathrm{T}}$ is the full parameter vector, with $\boldsymbol{\theta}_y$, $\boldsymbol{\theta}_s$, and $\boldsymbol{\theta}_\xi$ denoting the parameters from the longitudinal submodel, time-to-event submodel, and the elements from the covariance matrix of the subject-specific unique random-effects, respectively. Additionally, it is assumed that given the observed history, censoring (loss to follow-up) and visiting (time of collection of longitudinal measurements) are independent of the true event-times and future longitudinal responses. These assumptions practically imply that even though the time of visit or withdrawal of a subject from a study depends on their past history, it does not depend on any latent characteristics related to the prognosis.

The parameters can be estimated by maximizing the observed log-likelihood function $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{Y}_i, T_i, \delta_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, using standard maximizing techniques such as Expectation-Maximization (EM) (Dempster et al., 1977) or Newton-Raphson (NR) algorithm. Though the EM algorithm has slow convergence rate near the maximum, the score function corresponding to the log-likelihood is a key factor required for both EM and NR algorithms. The score function can be written as

$$\mathbf{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int A(\boldsymbol{\theta}, \boldsymbol{\xi}_i) f(\boldsymbol{\xi}_i | \mathbf{Y}_i, T_i, \delta_i; \boldsymbol{\theta}) d\boldsymbol{\xi}_i, \tag{2.11}$$

where $A(\boldsymbol{\theta}, \boldsymbol{\xi}_i) = \partial\{\log f(\mathbf{Y}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta}) + \log f(\boldsymbol{\xi}_i;\boldsymbol{\theta}) + \log f(T_i, \delta_i|\boldsymbol{\xi}_i;\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}^{\mathrm{T}}$ denotes the complete data score vector. Rizopoulos et al. (2009) observed that if the score equations in (2.11) are solved with respect to $\boldsymbol{\theta}$, with $f(\boldsymbol{\xi}_i|\mathbf{Y}_i, T_i, \delta_i;\boldsymbol{\theta})$ fixed at the previous iterated value of $\boldsymbol{\theta}$, then this corresponds to an EM algorithm, whereas if the score equations are solved with respect to $\boldsymbol{\theta}$, considering $f(\boldsymbol{\xi}_i|\mathbf{Y}_i, T_i, \delta_i;\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, then this corresponds to direct maximization of the likelihood. The EM algorithm involves approximation of integrals in order to estimate the random effects. Higher number of random effects increases the dimension of the integration in the E-step. Rizopoulos et al. (2009) proposed fully exponential Laplace approximations to obtain the integrals to increase computational efficiency in joint models. However, they observed that in the low-dimensional scenario, the Gauss-Hermite quadrature rule was faster than the Laplace method and resulted in equivalent estimates. A package named `JM`, developed by Rizopoulos (2010), can be found in R software repository, which uses the methods described above for the joint analysis of longitudinal and time-to-event data.

Although the idea behind general formulation of joint likelihood and estimation techniques mentioned above are similar, alternative models have been postulated by several authors in joint models literature. One such joint modeling technique was proposed by De Gruttola and Tu (1994). They used the same set of random effects to model both longitudinal and survival responses as follows

$$Y_i(t) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\xi}_i + \varepsilon_i(t), \tag{2.12}$$

$$\mathcal{T}_i = \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\zeta} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\xi}_i + r_i, \tag{2.13}$$

where we observe that the longitudinal submodel in (2.12) is similar to (2.9), but the survival submodel (2.13) is different than (2.10). Here, the true event-time (or a one-to-one transformation of this variable), denoted by $\mathcal{T}_i$, is linked to the fixed effects $\boldsymbol{\zeta}$ through a design vector $\mathbf{W}_i$, the random effects $\boldsymbol{\xi}_i$ are connected through a set of unknown parameters $\boldsymbol{\lambda}$, and r_i is a zero-mean normal residual with variance σ_r^2. Assuming that the first measurement time t_{i1} is fixed at zero for all subjects, and $\delta_i = I(\mathcal{T}_i < \mathcal{C}_i)$ with $\mathcal{T}_i$ as the true event-time and $\mathcal{C}_i$ as the censoring time, the joint likelihood for the ith subject is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^{N^o} \log \left[\int \phi\{\mathbf{Y}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta}\}\phi(\boldsymbol{\xi}_i;\boldsymbol{\theta})\phi(\mathcal{T}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta})d\boldsymbol{\xi}_i \right] \\ &\quad + \sum_{i=1}^{N^c} \log \left[\int \phi\{\mathbf{Y}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta}\}\phi(\boldsymbol{\xi}_i;\boldsymbol{\theta})\{1-\Phi(\mathcal{C}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta})\}d\boldsymbol{\xi}_i \right], \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\zeta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}}, \sigma^2, \sigma_r^2)^{\mathrm{T}}$ is the full parameter vector, N^o and N^c are the respective numbers subjects having true and censored event-times, and $\Phi(\mathcal{C}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta}) = \int_{-\infty}^{\mathcal{C}_i} \phi(\mathcal{T}_i|\boldsymbol{\xi}_i;\boldsymbol{\theta})d\mathcal{T}_i$ with $\phi(\cdot)$ and $\Phi(\cdot)$ as the probability density and cumulative distribution functions of a standard normal distribution, respectively. The authors used Expectation-Maximization (EM) algorithm to maximize $L(\boldsymbol{\theta})$ to estimate the parameters.

Henderson et al. (2000) took a similar approach as De Gruttola and Tu (1994), but they used a bivariate latent random process to link the two outcomes instead of random effects as shared parameters. Assuming that n subjects were followed over the time interval $[0, \tau)$, the longitudinal submodel for the ith subject is given as

$$Y_i(t) = \mathbf{X}_{1i}^{\mathrm{T}}(t)\boldsymbol{\beta}_1 + W_{1i}(t) + \varepsilon_i(t),$$

and the survival submodel is given by

$$h_i(t) = h_0(t)\mathcal{R}_i(t)\exp\{\mathbf{X}_{2i}^{\mathrm{T}}(t)\boldsymbol{\beta}_2 + W_{2i}(t)\},$$

where $\mathbf{W}_i(t) = \{W_{1i}(t), W_{2i}(t)\}$ is an unobserved latent zero-mean bivariate Gaussian process, $\{\mathcal{R}_i(t) : 0 \leq t \leq \tau\}$ indicates whether a subject is at risk at time t, and the design matrices $\mathbf{X}_{1i}(t)$ and $\mathbf{X}_{2i}(t)$ may include common predictors. This model can be considered as an extension of a number of specific joint models, for example, specifying $W_{1i}(t) = \xi_{i1} + \xi_{i2}t$ and $W_{2i}(t) = \gamma W_{1i}(t)$, where $(\xi_{i1}, \xi_{i2}) \sim N(\mathbf{0}, \mathbf{\Sigma}_\xi)$ are the subject-specific random effects, would reduce this model to the joint models proposed by Tsiatis et al. (1995), Faucett and Thomas (1996), and Wulfsohn and Tsiatis (1997). Additionally, in absence of significant association between the two outcomes, this model reduces to two separate marginal models, a normal linear model with correlated errors for the repeated measurements, and a proportional hazard model for the event-time.

The joint modeling approaches mentioned above assume that the shared parameters are normally distributed. An alternative approach, known as conditional score method, was proposed by Tsiatis and Davidian (2001), where no distributional assumptions were imposed on the random effects. The idea behind their method is to treat the random effects as nuisance parameters and condition on appropriate sufficient statistic of the random effects to obtain unbiased estimating equations of the coefficients in the hazard model. This method produces consistent and asymptotically normal estimates of the parameters. In one of their later papers (Tsiatis and Davidian, 2004), they formally discussed the assumptions behind the joint models, and presented a comparison between the joint likelihood-based versus the conditional score-based approaches. The authors have shown that the likelihood-based approaches may yield the most precise inferences, but can be computationally demanding, whereas the conditional score approach is easier to compute, but loses efficiency in compar-

ison to the likelihood-based approach, because it does not exploit the full information in the longitudinal data.

To test the significance of the model parameters, standard asymptotic tests are available, for example, likelihood ratio test, score test, and Wald test. All of these test statistics follow chi-square distribution with degrees of freedom equal to the number of parameters being tested. In large sample, these test statistics are low-order Taylor series expansion of each other, and are asymptotically equivalent. But for finite sample size, though the likelihood ratio test is computationally more demanding, it is considered to be more reliable than the other two.

In longitudinal and time-to-event studies, the predictors are often time-dependent and the regression coefficients may change over time. Additionally, the association between the two outcomes may be time-varying. Ordinary joint models such as the aforementioned random effects models proposed for joint modeling, cannot capture this dynamic structure of the data. To increase the flexibility of joint models, recent developments in joint modeling literature allow the associations to vary as flexible functions of time. We refer this class of joint models as dynamic joint models in this dissertation. Dynamic joint models is a fairly new research area and some of the earlier works include Song and Wang (2008) and Ye et al. (2015). Song and Wang (2008) proposed two methods for estimating time-varying associations between multiple longitudinal processes and the survival outcome. In their model, let $X_{ik}(u)$ be the kth unobserved longitudinal process for the ith subject at time u, where $i = 1, \ldots, n$, $k = 1, \ldots, p$, and Y_{ikj} be the observed measurement at time t_{ikj}, for

$j = 1, \ldots, n_k$. The longitudinal processes are modeled using the linear mixed effect model

$$\begin{aligned} X_{ik}(u) &= \boldsymbol{\xi}_{ik}^{\mathrm{T}} \mathbf{f}_k(u), \\ Y_{ikj} &= X_{ik}(t_{ikj}) + \varepsilon_{ikj}, \end{aligned}$$

where $\mathbf{f}_k(u) = \{f_{k1}(u), \ldots, f_{kq_k}(u)\}^{\mathrm{T}}$ is a vector of known functions of u and $\boldsymbol{\xi}_{ik} = \{\xi_{ik1}, \ldots, \xi_{ikq_k}\}^{\mathrm{T}}$ is a vector of corresponding random effects. Note that $\mathbf{f}_k(\cdot)$ and $\boldsymbol{\xi}_{ik}$ can be different for different k, which allows flexible modeling of the time trajectory of each covariate via polynomial or spline. The random effects $\boldsymbol{\xi}_{ik}$ may be correlated across k and no distributional assumption was placed on $\boldsymbol{\xi}_i = (\boldsymbol{\xi}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_{ip}^{\mathrm{T}})^{\mathrm{T}}$. For time-independent covariates, ξ_{ik} is a scalar and $\mathbf{f}_k(u) = 1$. The errors ε_{ikj} are assumed to be normally distributed with mean zero and variance σ_{kk}^2 that may reflect both biological variation and measurement error.

To describe the survival submodel, they first defined the true event-time as $\mathcal{T}_i$ and the censoring time as $\mathcal{C}_i$. The observed event-time outcome was defined as $T_i = \min(\mathcal{T}_i, \mathcal{C}_i)$ with an indicator $\delta_i = I(\mathcal{T}_i \leq \mathcal{C}_i)$. They proposed a time-varying coefficient hazard model to describe the relationship between the hazard of failure and the longitudinal processes

$$\begin{aligned} h_i\{u|\mathbf{X}_i(u)\} &= \lim_{du \to 0} du^{-1} P\{u \leq \mathcal{T}_i < u + du | \mathcal{T}_i \geq u; \boldsymbol{\xi}_i, \mathcal{C}_i, t_i(u), \varepsilon_i(u)\} \\ &= h_0(u) \exp\{\boldsymbol{\beta}(u)^{\mathrm{T}} \mathbf{X}_i(u)\}, \end{aligned}$$

where the baseline hazard $h_0(u)$ was left fully unspecified, $t_i(u) = (t_{ikj} < u; k = 1, \ldots, p)$ denoted the observation times before u, $\varepsilon_i(u) = \{\varepsilon_{ikj} : t_{ikj} < u, k = 1, \ldots, p, j = 1, \ldots, n_{ik}\}$, and the estimation was focused on the time-varying coefficients $\boldsymbol{\beta}(u) = \{\beta_1(u), \ldots, \beta_p(u)\}^{\mathrm{T}}$.

Song and Wang (2008) proposed two approaches for the estimation of the time-varying regression parameters in the survival submodel, namely, local corrected score esti-

mator and local conditional score estimator. For both methods, they first locally approximated the coefficients as $\boldsymbol{\beta}(u) \approx \boldsymbol{\beta}(t) + \boldsymbol{\beta}'(t)(u-t)$ where u was in the neighborhood of t. Let $\mathbf{b} = (\mathbf{b}_0^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \boldsymbol{\beta}'(t)^{\mathrm{T}}\}^{\mathrm{T}}$ be the local estimator. The corrected score estimator was given as

$$\begin{aligned}\widehat{\mathbf{U}}_{CR}(\mathbf{b}) &= (n\mathbf{H})^{-1}\sum_{i=1}^{n}\int_0^L K_h(u-t)\times\Bigg\{\widehat{\mathbf{X}}_i(u,u-t) \\ &\quad + \boldsymbol{\Sigma}_i(u,u-t)\mathbf{b} - \frac{\widehat{\mathbf{G}}_{CR,1}(\mathbf{b},u,u-t)}{\widehat{\mathbf{G}}_{CR,0}(\mathbf{b},u,u-t)}\Bigg\}dN_i(u) = \mathbf{0}, \qquad (2.14)\end{aligned}$$

where L is a fixed time, $\mathbf{H} = \mathrm{diag}(\mathbf{I}_p, h\mathbf{I}_p)$ with $\mathbf{I}_p$ is a p-dimensional identity matrix, h is the bandwidth of the kernel function $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $\widehat{\mathbf{G}}_{CR,r}(\mathbf{b},u,u-t) = n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{G}}_{CR,r_i}(\mathbf{b},u,u-t)$ with

$\widehat{\mathbf{G}}_{CR,r_i}(\mathbf{b},u,u-t) = \mathbf{W}_i(u)\widehat{\mathbf{X}}_i^{\otimes r}(u,u-t)\times\exp\{\mathbf{b}^{\mathrm{T}}\widehat{\mathbf{X}}_i(u,u-t) - \frac{1}{2}\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_i(u,u-t)\mathbf{b}\}$,

$\mathbf{c}^{\otimes r} = 1, \mathbf{c}, \mathbf{c}\mathbf{c}^{\mathrm{T}},$ for $r = 0, 1, 2,$ respectively,

$\boldsymbol{\Sigma}_i(u,u-t) = \mathrm{Var}\{\widehat{X}_i(u,u-t)|\boldsymbol{\xi}_i, t_i(u)\}$,

$\widehat{\mathbf{X}}_i(u,u-t) = (1,u-t)^{\mathrm{T}}\otimes\widehat{\mathbf{X}}_i(u)$,

$\widehat{\mathbf{X}}_i(u) = \{\widehat{X}_{i1}(u),\ldots,\widehat{X}_{ip}(u)\}^{\mathrm{T}}$ is the ordinary least square estimator of $\mathbf{X}_i(u)$,

$\mathbf{W}_i(u) = I(T_i \geq u, n_{ik}(u) \geq q_k)$,

$N_i(u) = I(T_i \leq u, \delta_i = 1, n_{ik} \geq q_k)$.

The authors have shown that the bias, which arises in a local estimating equation due to using the ordinary least squares estimator $\widehat{\mathbf{X}}_i(u)$, can be removed by the "corrected" score function mention in (2.14) and these estimators are consistent and asymptotically normally distributed.

The second approach that Song and Wang (2008) proposed was the conditional score estimator, which "conditions away" the nuisance random effects based on its sufficient statistic. Given $\boldsymbol{\xi}_i$, $t_i(u)$, and $\mathbf{W}_i(u) = 1$, the conditional distribution of $dN_i(u)$ is given by a Bernoulli distribution having the success probability locally approximated as $h_0(u)du\exp\{\mathbf{b}^{\mathrm{T}}\mathbf{X}_i(u, u-t)\}$. It can be shown that the sufficient statistic for $\boldsymbol{\xi}_i$ is $\mathbf{S}_i(\mathbf{b}, u, u-t) = \widehat{\mathbf{X}}_i(u, u-t) + \boldsymbol{\Sigma}_i(u, u-t)\mathbf{b}\ dN_i(u)$. Conditional on this sufficient statistics, the local conditional score estimating equation can be presented as

$$\begin{aligned}\widehat{\mathbf{U}}_{CD}(\mathbf{b}) &= (n\mathbf{H})^{-1}\sum_{i=1}^{n}\int_0^L K_h(u-t) \\ &\quad\times\left\{\mathbf{S}_i(\mathbf{b}, u, u-t) - \frac{\widehat{\mathbf{G}}_{CD,1}(\mathbf{b}, u, u-t)}{\widehat{\mathbf{G}}_{CD,0}(\mathbf{b}, u, u-t)}\right\}dN_i(u) = \mathbf{0}, \qquad (2.15)\end{aligned}$$

where $\widehat{\mathbf{G}}_{CD,r}(\mathbf{b}, u, u-t) = n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{G}}_{CD,r_i}(\mathbf{b}, u, u-t)$ with

$$\widehat{\mathbf{G}}_{CD,r_i}(\mathbf{b}, u, u-t) = \mathbf{W}_i(u)\mathbf{S}_i^{\otimes r}(\mathbf{b}, u, u-t) \times \exp\{\mathbf{b}^{\mathrm{T}}\mathbf{S}_i(\mathbf{b}, u, u-t) - \frac{1}{2}\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}_i(u, u-t)\mathbf{b}\}.$$

The authors have shown that although the corrected score estimator is asymptotically equivalent to the conditional score estimator, the latter outperforms the former for finite samples, especially in presence of relatively large measurement error.

The primary focus of the model proposed by Song and Wang (2008) was on the hazard submodel with time-varying coefficients, where these coefficients quantify the time-dependent effects of the corresponding longitudinal predictors on hazard of the subjects. In their method, these longitudinal predictors were limited to be continuous variables only. Ye et al. (2015) on the other hand, postulated a dynamic joint model for a single longitudinal outcome, which belongs to the canonical exponential family and thus, the values of this variable are allowed to be both discrete and continuous. They assumed that the longitudinal

trajectory is driven by a latent Gaussian process $X_i(t)$, which is essentially the canonical parameter of the corresponding exponential family, and modelled this latent process by employing functional principal component analyses (FPCA) as follows,

$$X_i(t) = \mu(t) + \boldsymbol{\psi}(t)^{\mathrm{T}}\boldsymbol{\xi}_i,$$

where $\mu(t)$ is the mean longitudinal trend, $\boldsymbol{\psi}(t)$ is a vector of orthonormal functions (known as eigenfunctions in FPCA), and $\boldsymbol{\xi}_i$ is the vector of principal component scores following zero-mean normal distribution with a diagonal matrix $\boldsymbol{\Sigma}_\xi$ as covariance matrix, the diagonal elements being eigenvalues. The authors also assumed that the time-to-event outcome depends on the longitudinal outcome only through this latent process. They used penalized B-Splines to estimate $\mu(t)$, $\boldsymbol{\psi}(t)$, and the time-dependent association, which were expressed as linear combinations of the eigenfunctions.

The joint modelling approaches proposed by Song and Wang (2008) and Ye et al. (2015) did not allow to investigate the effects of exploratory variables on the longitudinal processes. Andrinopoulou et al. (2018) proposed a dynamic joint model which accommodated exploratory variables in both longitudinal and hazard submodels. The longitudinal submodel in their joint modeling framework is same as the model specified in (2.9), but their survival component includes a flexible function to model the association between the two outcomes. The survival submodel is given by the hazard function

$$h_i\{t|\mathcal{M}_i(t), \mathbf{W}_i\} = h_0(t)\exp\left[\mathbf{W}_i^{\mathrm{T}}\boldsymbol{\eta} + f\{\gamma(t), \mathcal{M}_i(t)\}\right],$$

where $\mathbf{W}_i$ is a vector of baseline predictors with a corresponding vector of regression coefficients $\boldsymbol{\eta}$, and $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to time t. The function $f\{\gamma(t), \mathcal{M}_i(t)\}$ specifies which features of the

longitudinal submodel are included in the relative risk model. They suggested some forms of the function $f(\cdot)$, namely, $f\{\gamma(t), \mathcal{M}_i(t)\} = \gamma(t)m_i(t)$ and $f\{\gamma(t), \mathcal{M}_i(t)\} = \gamma(t)\int_0^t m_i(s)ds$. For modeling the time-varying association $\gamma(t)$ and the log of the baseline $h_0(t)$, they suggested using the P-spline approach to avoid misspecification of number and location of knots in commonly used smoothing techniques, such as, B-spline. The idea behind P-spline method is that it gains sufficient smoothness by using a relatively higher number of equally spaced knots, and it uses a penalty term to avoid over-fitting. The spline models for the association and the log-baseline hazard are

$$\gamma(t) = \sum_{\ell=1}^{L} \alpha_\ell \mathcal{B}_\ell(t) \quad \text{and} \quad \log\{h_0(t)\} = \sum_{u=1}^{U} \lambda_{u,h_0} \mathcal{B}_u(t),$$

respectively, where α_ℓ is a set of parameters which capture the strength of association between the longitudinal and the event-time outcomes, $\mathcal{B}_\ell(t)$ is the ℓth basis function of a B-spline, λ_{u,h_0} are the spline coefficients for the baseline hazard, and $\mathcal{B}_u(t)$ is the uth basis function of a B-spline. They employed Markov Chain Monte Carlo method to estimate the spline parameters, and the inference is based on the posterior of the joint model. Under the assumptions of the shared-parameter models, that is, conditioned on the random effects, the within subject observations, as well as the two outcomes become independent, Andrinopoulou et al. (2018) postulated the following posterior density,

$$f(\boldsymbol{\theta}, \boldsymbol{\xi}_i | \mathbf{Y}_i, T_i, \delta_i) \propto \prod_{j=1}^{n_i} f(Y_{ij} | \boldsymbol{\xi}_i, \boldsymbol{\theta}_y) f\{T_i, \delta_i | m_i(T_i), \boldsymbol{\theta}_s\} f(\boldsymbol{\xi}_i | \boldsymbol{\theta}_y) f(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_s^{\mathrm{T}})^{\mathrm{T}}$ is the parameter vector for the longitudinal and the survival outcomes, respectively. The smoothness of $\gamma(t)$ and $h_0(t)$ are controlled by the prior distributions of the set of spline coefficients $\{\alpha_\ell : \ell = 1, \ldots, L\}$ and $\{\lambda_{u,h_0} : u = 1, \ldots, U\}$, more specifically, by

the hyper-priors of the variances of these spline coefficients. For more details on the choice of prior, hyper-prior, and penalty matrices, please refer to Andrinopoulou et al. (2018). They calculated AUC (area under the receiver operative characteristic curve) and PE (prediction error) in order to measure the performance of their model. AUC can discriminate between patients who will experience the event versus who will not. PE was computed to measure the accuracy of their model. They have shown through simulation studies that their dynamic joint model has higher AUC and lower PE values in comparison to a constant coefficient joint model.

Piulachs et al. (2021) extended Bayesian dynamic joint models proposed by Andrinopoulou et al. (2018) to a generalized setting where the longitudinal response was zero-inflated count data. They used a hierarchical negative-binomial model to fit the longitudinal process. The time-to-event outcome was left-truncated, and they allowed time-varying exogenous predictors in their hazard submodel. The time-varying association between the longitudinal and the event-time outcomes was specified by penalized B-splines. For more information on dynamic joint-models, please see Suresh et al. (2017), Li and Luo (2017), Barrett and Su (2017), Hong et al. (2021), and Martins (2021).

Chapter 3

Time-Varying Joint Models for Longitudinal and Time-to-Event Outcomes

In this chapter, we present our time-varying joint modeling (TV-JM) approach. The model is specified in Section 3.1, the estimation procedure is described in Section 3.2. In Section 3.3, we discuss practical issues related to the application of our models. The simulation study and the application on WIHS data are presented in Sections 3.4 and 3.5, respectively.

3.1 Model Specification:

The joint modeling of longitudinal and survival outcomes involves submodels for each outcome. Let us start with the longitudinal outcome. In practice, the longitudinal

response is collected sparsely over time and it may contain measurement error and biological variability, and thus, the true longitudinal process for a subject may not be observable. Therefore, we denote $Y_i(t)$ as the observed longitudinal response and $m_i(t)$ as the true longitudinal process for the ith subject at time $t = t_{ij}$ with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. For the longitudinal outcome, we propose the following time-varying coefficient model with the subject-level random effect ξ_i,

$$\begin{aligned} Y_i(t) &= m_i(t) + \varepsilon_i(t), \\ m_i(t) &= \mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t) + \xi_i, \end{aligned}$$

where $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^{\mathrm{T}}$ is the vector of predictors with the corresponding coefficient vector $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_p(t)\}^{\mathrm{T}}$. The subject-specific random effect ξ_i and the error term $\varepsilon_i(t)$ follow a zero-mean normal distribution with variances σ_ξ^2 and $\sigma^2(t)$, respectively. Note that ξ_i and $\varepsilon_i(t)$ are assumed to be independent of each other.

In the survival submodel of the joint model, for the ith subject, let $\mathbf{W}_i(t) = \{W_{i1}(t), \ldots, W_{iq}(t)\}^{\mathrm{T}}$ be the vector of exogenous time-varying risk factors with the corresponding coefficient vector $\boldsymbol{\eta}(t) = \{\eta_1(t), \ldots, \eta_q(t)\}^{\mathrm{T}}$ and $T_i = \min(\mathcal{T}_i, C_i)$ be the observed event-time outcome with the event indicator δ_i, where $\delta_i = 1$ if $T_i = \mathcal{T}_i$ (true event-time), and $\delta_i = 0$ if $T_i = C_i$ (censoring time). The hazard and survival functions are defined as follows

$$h_i\{t|\mathcal{M}_i(t), \mathbf{W}_i(t)\} = h_0(t)\exp\left\{m_i(t)\gamma(t) + \mathbf{W}_i^{\mathrm{T}}(t)\boldsymbol{\eta}(t)\right\} \text{ and} \tag{3.1}$$

$$\begin{aligned} \mathcal{S}_i\{t \mid \mathcal{M}_i(t), \mathbf{W}_i(t)\} &= Pr\{T_i > t \mid \mathcal{M}_i(t), \mathbf{W}_i(t)\} \\ &= \exp\left[-\int_0^t h_i\{u|\mathcal{M}_i(u), \mathbf{W}_i(u)\}du\right], \end{aligned} \tag{3.2}$$

respectively, with $h_0(t)$ as the baseline hazard function, $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ as the history of the true unobserved longitudinal process $m_i(t)$ up to time point t, $\gamma(t)$ denoting the time-varying regression coefficient that quantifies the effect of the true longitudinal process on the risk of an event. Note that although we denote the predictors in both submodels as $\mathbf{X}_i(t)$ and $\mathbf{W}_i(t)$, they can include time-invariant (baseline) predictors as well. The definitions of both hazard and survival functions indicate that the failure time of a subject depends on their longitudinal outcome, that is, the CD4 cell percentage in the WIHS data (Section 3.5). The hazard function in (3.1) shows that the relative risk of a person experiencing the event at any time t depends on the value of their CD4 cell percentage at that time point only, whereas (3.2) implies that the survival of that individual is dependent on their entire CD4 cell percentage history up to time t.

To define the joint distribution of the longitudinal and survival outcomes, it is assumed that the random effect ξ_i underlies both outcomes. More specifically, the random effect is considered to account for the association between the two responses and the within-subject correlation of the longitudinal outcome. Therefore, given the random effect ξ_i, not only the two types of responses become independent of each other, all the longitudinal observations for a given individual become independent as well. Therefore, the joint distribution for the longitudinal and survival outcomes is

$$f(\mathbf{Y}_i, T_i, \delta_i, \xi_i; \boldsymbol{\theta}) = f(\mathbf{Y}_i|\xi_i; \boldsymbol{\theta}) f(\xi_i; \boldsymbol{\theta}) f(T_i, \delta_i|\xi_i; \boldsymbol{\theta}), \tag{3.3}$$

where $\mathbf{Y}_i = \{Y_i(t_{i1}), \ldots, Y_i(t_{in_i})\}^{\mathrm{T}}$ denotes the longitudinal outcome for the ith subject, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_y^{\mathrm{T}}(t), \boldsymbol{\theta}_s^{\mathrm{T}}(t), \theta_\xi\}^{\mathrm{T}}$ is the full parameter vector with $\boldsymbol{\theta}_y(t) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \sigma^2(t)\}^{\mathrm{T}}$, $\boldsymbol{\theta}_s(t) = \{\boldsymbol{\theta}_{h_0}^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}\}^{\mathrm{T}}$, and $\theta_\xi = \sigma_\xi^2$ denoting the parameters for the longitudinal outcome, the

survival outcome, and the variance of the subject-level random effect, respectively, and $\boldsymbol{\theta}_{h_0}^{\mathrm{T}}$ denoting the vector of parameters in the baseline hazard function $h_0(\cdot)$. The joint density for the longitudinal outcome and the random effects in (3.3) is defined as

$$\begin{aligned} f(\mathbf{Y}_i|\xi_i;\boldsymbol{\theta})f(\xi_i;\boldsymbol{\theta}) &= \Big[\prod_{j=1}^{n_i} f\{Y_i(t_{ij})|\xi_i;\boldsymbol{\theta}_y(t_{ij})\}\Big] f(\xi_i;\boldsymbol{\theta}) \\ &= \Big[\prod_{j=1}^{n_i}\{2\pi\sigma^2(t_{ij})\}^{-1/2}\Big] \exp\Big[-\sum_{j=1}^{n_i}\frac{\big\{Y_i(t_{ij})-\mathbf{X}_i^{\mathrm{T}}(t_{ij})\boldsymbol{\beta}(t_{ij})-\xi_i\big\}^2}{2\sigma^2(t_{ij})}\Big] \\ &\quad \times(2\pi\sigma_\xi^2)^{-1/2}\exp\left\{-\xi_i^2/(2\sigma_\xi^2)\right\}. \end{aligned}$$

Furthermore, the density for the observed event-time T_i given the random effect ξ_i in (3.3) is given by

$$\begin{aligned} f(T_i,\delta_i|\xi_i;\boldsymbol{\theta}) &= h_i\{T_i|\mathcal{M}_i(T_i),\mathbf{W}_i(T_i);\boldsymbol{\theta}\}^{\delta_i}\mathcal{S}_i\{T_i|\mathcal{M}_i(T_i),\mathbf{W}_i(T_i);\boldsymbol{\theta}\} \\ &= \Big[h_0(T_i)\exp\Big\{m_i(T_i)\gamma(T_i)+\mathbf{W}_i^{\mathrm{T}}(T_i)\boldsymbol{\eta}(T_i)\Big\}\Big]^{\delta_i} \\ &\quad \times\exp\Big[-\int_0^{T_i} h_i\{u|\mathcal{M}_i(u),\mathbf{W}_i(u)\}du\Big]. \end{aligned}$$

3.2 Estimation and Inference

In this section, we propose an estimation procedure based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), in which, the subject-level random effects are treated as missing data. The proposed EM algorithm iterates between two-steps: in the E-step, we estimate the random effects $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)^{\mathrm{T}}$ and in the M-step, we maximize the conditional expectation of the complete likelihood to estimate $\boldsymbol{\theta} = \{\boldsymbol{\theta}_y^{\mathrm{T}}(t),\boldsymbol{\theta}_s^{\mathrm{T}}(t),\theta_\xi\}^{\mathrm{T}}$. The complete joint log-likelihood $\ell(\boldsymbol{\xi},\boldsymbol{\theta})$ is defined by

$$\ell(\boldsymbol{\xi},\boldsymbol{\theta}) = \sum_{i=1}^{n}\ell_i(\xi_i,\boldsymbol{\theta}) = \sum_{i=1}^{n}\log L_i(\xi_i,\boldsymbol{\theta}) = \sum_{i=1}^{n}\log f(\mathbf{Y}_i,T_i,\delta_i,\xi_i;\boldsymbol{\theta}),$$

where $L_i(\xi_i, \boldsymbol{\theta})$ denotes the likelihood contribution of the ith subject and the log-likelihood $\ell_i(\xi_i, \boldsymbol{\theta})$ is given by

$$\begin{aligned}
\ell_i(\xi_i, \boldsymbol{\theta}) &= \log f(\mathbf{Y}_i, T_i, \delta_i, \xi_i; \boldsymbol{\theta}) \\
&= \log f(\mathbf{Y}_i|\xi_i; \boldsymbol{\theta}) + \log f(\xi_i; \boldsymbol{\theta}) + \log f(T_i, \delta_i|\xi_i; \boldsymbol{\theta}) \\
&= -\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{\left\{Y_i(t_{ij}) - m_i(t_{ij})\right\}^2}{\sigma^2(t_{ij})} + \log\{2\pi\sigma^2(t_{ij})\}\right] - \frac{\xi_i^2}{2\sigma_\xi^2} - \frac{1}{2}\log(2\pi\sigma_\xi^2) \\
&\quad +\delta_i\left\{\log h_0(T_i) + m_i(T_i)\gamma(T_i) + \mathbf{W}_i^{\mathrm{T}}(T_i)\boldsymbol{\eta}(T_i)\right\} \\
&\quad -\int_0^{T_i} h_0(u)\exp\left\{m_i(u)\gamma(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}(u)\right\} du
\end{aligned}$$

with $m_i(\cdot) = \mathbf{X}_i^{\mathrm{T}}(\cdot)\boldsymbol{\beta}(\cdot) + \xi_i$. The incomplete log-likelihood is defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\mathcal{L}_i(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\left\{\int L_i(\xi_i, \boldsymbol{\theta})d\xi_i\right\}. \tag{3.4}$$

The estimation steps of the proposed approximate EM algorithm are described as follows.

1. The initial values for all the parameters $\boldsymbol{\theta}^0 = \{\boldsymbol{\theta}_y^{0\mathrm{T}}(t), \boldsymbol{\theta}_s^{0\mathrm{T}}(t), \theta_\xi^0\}^{\mathrm{T}}$ are selected by fitting separate models to each outcome. A linear mixed effects (LME) and Cox models are used to model the longitudinal and survival outcomes, respectively.

2. (E-step) In the rth iteration, the estimates of the posterior mean and variance of the subject-level random effects ξ_i are obtained via Gauss-Hermite quadrature approximations, leading to the approximated conditional expectation of the complete log-likelihood.

3. (M-step) The incomplete log-likelihood is maximized to obtain closed form solutions for the current estimate of θ_ξ ($\theta_\xi = \sigma_\xi^2$). The approximated expected complete log-likelihood is maximized to obtain the rest of the current estimates $\boldsymbol{\theta}^{\setminus\sigma} =$

$\{\boldsymbol{\theta}_y^{\mathrm{T}}(t), \boldsymbol{\theta}_s^{\mathrm{T}}(t)\}^{\mathrm{T}}$ via a Newton-Raphson algorithm. Note that for the time-varying parameters in $\boldsymbol{\theta}^{\setminus\sigma}$, that is, $\{\boldsymbol{\beta}(t)^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}\}^{\mathrm{T}}$, we employ local linear fitting techniques (Fan and Gijbels, 1996).

4. The algorithm iterates between steps 2-3 until the difference between two consecutive incomplete log-likelihood values are less than a predefined tolerance level ϵ.

3.2.1 E-step and the Gauss-Hermite Quadrature Approximation

The posterior mean and the variance of the random effect which are denoted by ξ_{i0} and υ_{i0}, respectively, are defined as

$$\xi_{i0} = \frac{\int \xi_i L_i(\xi_i, \boldsymbol{\theta}) d\xi_i}{\int L_i(\xi_i, \boldsymbol{\theta}) d\xi_i} \text{ and } \upsilon_{i0} = \frac{\int (\xi_i - \xi_{i0})^2 L_i(\xi_i, \boldsymbol{\theta}) d\xi_i}{\int L_i(\xi_i, \boldsymbol{\theta}) d\xi_i}. \tag{3.5}$$

For approximating the integrals in (3.5), numerical integration methods such as Gauss-Hermite quadrature or Laplace approximation should be adopted. As the Laplace approximation is an asymptotic method and requires large number of observations within each individual, under our circumstances, that is, with only one random effect (integration of order one) and a small number of observations within each subject, the Gauss-Hermite quadrature performs better than the Laplace approximation (Rizopoulos et al., 2009). A brief description of the Gauss-Hermite quadrature method and its application in our model is presented in Appendix A.

Once the posterior mean and the variance of the random effect ξ_i are estimated, we can proceed to approximating the conditional expectation of the complete joint log-likelihood in the E-step, that is, $\sum_{i=1}^{n} E\left[\ell_i\{\xi_i, \boldsymbol{\theta}\} \middle| \mathbf{Y}_i, T_i, \delta_i, \mathbf{X}_i(t), \mathbf{W}_i(t), \boldsymbol{\theta}^*\right]$, where $\boldsymbol{\theta}^* = \{\boldsymbol{\theta}_y^{*\mathrm{T}}(t), \boldsymbol{\theta}_s^{*\mathrm{T}}(t), \theta_\xi^*\}^{\mathrm{T}}$ denotes the current parameter estimates with $\boldsymbol{\theta}_y^*(t) = \{\boldsymbol{\beta}^{*\mathrm{T}}(t), \sigma^{2*}(t)\}^{\mathrm{T}}$,

$\boldsymbol{\theta}_s^*(t) = \{\boldsymbol{\theta}_{h_0}^{*\mathrm{T}}, \gamma^*(t), \boldsymbol{\eta}^*(t)^{\mathrm{T}}\}^{\mathrm{T}}$, and $\theta_\xi^* = \sigma_\xi^{2*}$. Since the closed form of the conditional expectation $\sum_{i=1}^n E\left[\ell_i\{\xi_i, \boldsymbol{\theta}\}\middle|\mathbf{Y}_i, T_i, \delta_i, \mathbf{X}_i(t), \mathbf{W}_i(t), \boldsymbol{\theta}^*\right]$ is challenging to derive, a second-degree Taylor's expansion around ξ_{i0}^* is employed to approximate the expected log-likelihood

$$\sum_{i=1}^n \left[\ell_i(\xi_{i0}^*, \boldsymbol{\theta}^*) + \ell_i'\{\xi_{i0}^*, \boldsymbol{\theta}^*\}E(\xi_i - \xi_{i0}^*) - \frac{1}{2}E\big\{\Sigma_i^*(\xi_i - \xi_{i0}^*)^2\big\}\right],$$

where ξ_{i0}^* denotes the estimated posterior mean of ξ_i based on the current parameter estimates, $\ell_i'(\xi_{i0}^*, \boldsymbol{\theta}^*) = \partial\ell_i(\xi_i, \boldsymbol{\theta})/\partial\xi_i|_{\xi_i=\xi_{i0}^*, \boldsymbol{\theta}=\boldsymbol{\theta}^*}$, and $\Sigma_i^* = -\partial^2\ell_i(\xi_i, \boldsymbol{\theta})/\partial\xi_i^2|_{\xi_i=\xi_{i0}^*, \boldsymbol{\theta}=\boldsymbol{\theta}^*}$. Note that $E(\xi_i - \xi_{i0}^*) = 0$ and $\Sigma_i^*(\xi_i - \xi_{i0}^*)^2$ follows a Chi-Square distribution with degrees of freedom 1 since the posterior variance $\upsilon_{i0} = \Sigma_i^{*-1}$. Therefore, the expected log-likelihood can be approximated as follows,

$$\begin{aligned}
\sum_{i=1}^n E\left[\ell_i\{\xi_i, \boldsymbol{\theta}\}\right] &\approx \sum_{i=1}^n \ell_i(\xi_{i0}^*, \boldsymbol{\theta}^*) - \frac{n}{2} \qquad (3.6)\\
&= \sum_{i=1}^n \Bigg(-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{\big\{Y_i(t_{ij}) - m_i^*(t_{ij})\big\}^2}{\sigma^{2*}(t_{ij})} + \log\{2\pi\sigma^{2*}(t_{ij})\}\right] - \frac{\xi_{i0}^{*2}}{2\sigma_\xi^{2*}}\\
&\quad -\frac{1}{2}\log(2\pi\sigma_\xi^{2*}) + \delta_i\left\{\log h_0^*(T_i) + m_i^*(T_i)\gamma^*(T_i) + \mathbf{W}_i^{\mathrm{T}}(T_i)\boldsymbol{\eta}^*(T_i)\right\}\\
&\quad -\int_0^{T_i} h_0^*(u)\exp\left\{m_i^*(u)\gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u)\right\}du\Bigg) - \frac{n}{2},
\end{aligned}$$

where $m_i^*(\cdot) = \mathbf{X}_i^{\mathrm{T}}(\cdot)\boldsymbol{\beta}^*(\cdot) + \xi_{i0}^*$.

3.2.2 M-step

In this step, for estimation of the random effect variance σ_ξ^2, we directly maximize the incomplete log-likelihood $\ell(\boldsymbol{\theta})$ in (3.4) and set the following score function to zero

$$V(\sigma_\xi^2) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\sigma_\xi^2} = \sum_{i=1}^n \frac{\partial}{\partial\sigma_\xi^2}\log\left\{\int L_i(\xi_i, \boldsymbol{\theta})d\xi_i\right\} = \sum_{i=1}^n \int\left(\frac{\xi_i^2}{2\sigma_\xi^4} - \frac{1}{2\sigma_\xi^2}\right)\mathcal{F}(\xi_i)d\xi_i,$$

where $\mathcal{F}(\xi_i) = L_i(\xi_i, \boldsymbol{\theta}) / \int L_i(\xi_i, \boldsymbol{\theta}) d\xi_i$ is the posterior density of ξ_i. This leads to the following estimate of σ_ξ^2 at the current iteration

$$\widehat{\sigma}_\xi^{2*} = n^{-1} \sum_{i=1}^{n} \left\{ (\xi_{i0}^*)^2 + \upsilon_{i0}^* \right\},$$

where ξ_{i0}^* and υ_{i0}^* are the estimates of the posterior mean and variance, respectively, at the current stage of the EM-iteration. For inference, the likelihood-based standard error (SE) for $\widehat{\sigma}_\xi^2$ is equal to the square root of $\{-\mathcal{H}\ell_\xi(\widehat{\sigma}_\xi^2)\}^{-1}$, where $\mathcal{H}\ell_\xi(\sigma_\xi^2) = \partial^2 \ell(\boldsymbol{\theta}) / (\partial \sigma_\xi^2)^2$ is the hessian and $\widehat{\sigma}_\xi^2$ is the estimate of the random effect variance at the last EM iteration.

The rest of the parameters, that is, $\boldsymbol{\theta}_y(t) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \sigma^2(t)\}^{\mathrm{T}}$ and $\boldsymbol{\theta}_s(t) = \{\boldsymbol{\theta}_{h_0}^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}\}^{\mathrm{T}}$ do not have closed form solutions; therefore, cannot be estimated via maximizing the incomplete likelihood. We first focus on estimating the time-varying coefficient functions $\boldsymbol{\alpha}(t) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}\}^{\mathrm{T}}$ by employing the local linear regression method (Fan and Gijbels, 1996). We locally approximate the regression coefficient functions in a neighborhood of a fixed point t_0 using Taylor's approximation,

$$\boldsymbol{\alpha}(t) \approx \boldsymbol{\alpha}(t_0) + \boldsymbol{\alpha}'(t_0)(t - t_0) \equiv \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1(t - t_0),$$

where $\boldsymbol{\alpha}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \gamma_0, \boldsymbol{\eta}_0^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \gamma_1, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^{\mathrm{T}}$, $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0q})^{\mathrm{T}}$, and $\boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{1q})^{\mathrm{T}}$. We maximize the following approximated expected local log-likelihood

$$\sum_{i=1}^{n}\Bigg\{-\frac{1}{2}\sum_{j=1}^{n_i}\Bigg[\frac{\{Y_i(t_{ij})-m_{il}^*(t_{ij})\}^2}{\sigma^{2*}(t_{ij})}+\log\{2\pi\sigma^{2*}(t_{ij})\}+\frac{\xi_{i0}^{*2}}{\sigma_\xi^{2*}} \tag{3.7}$$
$$+\log(2\pi\sigma_\xi^{2*})\Bigg]K_{h_1}(t_{ij}-t_0)$$
$$+\Bigg(\delta_i\Big[\log h_0^*(T_i)+m_{il}^*(T_i)\{\gamma_0^*+\gamma_1^*(T_i-t_0)\}+\mathbf{W}_i^{\mathrm{T}}(T_i)\{\boldsymbol{\eta}_0^*+\boldsymbol{\eta}_1^*(T_i-t_0)\}\Big]$$
$$-\int_0^{T_i}h_0^*(u)\exp\Big[m_{il}^*(u)\{\gamma_0^*+\gamma_1^*(u-t_0)\}+\mathbf{W}_i^{\mathrm{T}}(u)\{\boldsymbol{\eta}_0^*+\boldsymbol{\eta}_1^*(u-t_0)\}\Big]du\Bigg)$$
$$\times\ K_{h_2}(T_i-t_0)\Bigg\}-\frac{n}{2},$$

with respect to $(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)$, where for any c, the estimates at the current EM-iteration is denoted by c^*, $m_{il}^*(z)=\mathbf{X}_i^{\mathrm{T}}(z)\{\boldsymbol{\beta}_0^*+\boldsymbol{\beta}_1^*(z-t_0)\}+\xi_{i0}^*$, and $K_{h_1}(\cdot)$ and $K_{h_2}(\cdot)$ are kernel functions for the longitudinal and the survival components respectively, with $K_h(\cdot)=h^{-1}K(\cdot/h)$ and bandwidth h. We use a Newton-Raphson (NR) algorithm to maximize (3.7) and obtain $(\hat{\boldsymbol{\alpha}}_0^{\mathrm{T}},\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}})^{\mathrm{T}}$. Let $\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\}$ be the estimate of $(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)$ at the current iteration of the NR algorithm and we update $(\boldsymbol{\alpha}_0^{\mathrm{T}},\boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$ according to

$$\begin{Bmatrix}\boldsymbol{\alpha}_0^{(it+1)}\\ \boldsymbol{\alpha}_1^{(it+1)}\end{Bmatrix}=\begin{Bmatrix}\boldsymbol{\alpha}_0^{(it)}\\ \boldsymbol{\alpha}_1^{(it)}\end{Bmatrix}-[\ell''\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\}]^{-1}\ell'\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\},$$

where $\ell'\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\}$ and $\ell''\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\}$ are the score function and Hessian of the approximated expected local log-likelihood (3.7) with respect to $(\boldsymbol{\alpha}_0^{\mathrm{T}},\boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$, respectively, evaluated at the current estimates $\{\boldsymbol{\alpha}_0^{(it)},\boldsymbol{\alpha}_1^{(it)}\}$. The likelihood-based SEs at time point t_0 are obtained via the square root of the diagonal elements of $\{-\mathcal{H}(t_0)\}^{-1}$, where $\mathcal{H}(t_0)=\ell''(\widehat{\boldsymbol{\alpha}}_0,\widehat{\boldsymbol{\alpha}}_1)$ is the hessian matrix evaluated at $(\widehat{\boldsymbol{\alpha}}_0^{\mathrm{T}},\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}})^{\mathrm{T}}$ which are the estimates of the local parameters at the last iteration of the EM algorithm.

Next, we estimate the time-invariant parameter $\boldsymbol{\theta}_{h_0}$ using a Newton-Raphson algorithm, in which, we maximize the approximated expected log-likelihood in (3.6). The updated estimator is obtained by $\boldsymbol{\theta}_{h_0}^{(r+1)} = \boldsymbol{\theta}_{h_0}^{(r)} - \{\mathcal{H}_{h_0}^{(r)}\}^{-1}\mathbf{V}_{h_0}^{(r)}$, where r is the current iteration of the Newton-Raphson algorithm, $\mathbf{V}_{h_0}^{(r)}$ and $\mathcal{H}_{h_0}^{(r)}$ are the score function and the Hessian matrix of the approximated expected log-likelihood (3.6) with respect to $\boldsymbol{\theta}_{h_0}$, respectively, evaluated at the current estimates $\boldsymbol{\theta}_{h_0}^{(r)}$. The likelihood-based SEs for $\hat{\boldsymbol{\theta}}_{h_0}$ are equal to the square root of the diagonal elements of negative $\mathcal{H}_{h_0}^{-1}$, where $\mathcal{H}_{h_0}$ contains hessian values from the last EM iteration.

Note that, the likelihood-based SEs of the estimators are expected to be biased in estimating the true SEs since the variability in the estimation of the random effects are not taken into account in the EM algorithm (Hsieh et al., 2006; Kass and Steffey, 1989). Therefore, we examine the extent of this bias in the likelihood-based SEs in the simulation studies. Furthermore, we propose bootstrap estimates of SEs and investigate their performance in simulation studies.

Once the regression coefficients and the baseline hazard parameters are estimated, we conclude the M-step of the current EM-iteration by using the kernel estimator to obtain $\widehat{\sigma}^2(\cdot)$ at the fixed point t_0

$$\widehat{\sigma}^{2*}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} e_i^{*2}(t_{ij})K_{h_1}(t_{ij}-t_0)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i} K_{h_1}(t_{ij}-t_0)}, \tag{3.8}$$

where $e_i^*(t_{ij}) = Y_i(t_{ij}) - \{\mathbf{X}_i^{\mathrm{T}}(t_{ij})\boldsymbol{\beta}^*(t_{ij}) + \xi_{i0}^*\}$ is the residual for the jth longitudinal measurement of the ith subject at the current EM-iteration with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. The explicit expressions for all the derivatives of the expected (local) log-likelihood and the incomplete log-likelihood along with further details of the M-step of the EM-algorithm are

provided in Appendix B, where the survival integrations in each of those equations were approximated using Simpson's one third rule.

In summary, at the M-step, we proposed the following steps to obtain the estimates of $\boldsymbol{\theta}(t) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \sigma^2(t), \boldsymbol{\theta}_{h_0}^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}, \sigma_\xi^2\}^{\mathrm{T}}$ at the current iteration.

1. We estimated the random effect variance σ_ξ^2 by maximizing the incomplete log-likelihood $\ell(\boldsymbol{\theta})$ specified in (3.4).

2. We employed NR algorithm to maximize the local log-likelihood in (3.7) to estimate the local parameters $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{\beta}_0^{\mathrm{T}}, \gamma_0, \boldsymbol{\eta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \gamma_1, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$.

3. The baseline hazard function $h_0(t)$ was estimated via the restricted cubic spline approach. In particular, we maximized the global log-likelihood in (3.6) with respect to the spline coefficients $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_K\}^{\mathrm{T}}$ (defined in Section 3.3) using NR optimization.

4. We estimated the error variance $\sigma^2(t)$ using the kernel estimator presented in (3.8).

3.3 Practical Issues

For the practical application of our approach, we need to address four important issues. The first one is the choice of the baseline hazard function $h_0(t)$. In our modeling framework, we opted for a restricted cubic spline as it presents a flexible model for the baseline hazard. In addition, this approach provides a smooth approximation of the function with a small number of knots. The baseline hazard function in this approach is given by

$$h_0(t) = \exp\left\{\sum_{\kappa=1}^{K-2} \varphi_\kappa \omega_\kappa(t) + \varphi_{K-1} t + \varphi_K\right\},$$

where $\omega_\kappa(t) = (t-\vartheta_\kappa)_+^3 - \frac{(t-\vartheta_{K-1})_+^3(\vartheta_K-\vartheta_\kappa)}{(\vartheta_K-\vartheta_{K-1})} + \frac{(t-\vartheta_K)_+^3(\vartheta_{K-1}-\vartheta_\kappa)}{(\vartheta_K-\vartheta_{K-1})}$, for $\kappa = 1,\ldots,(K-2)$, and $(z)_+ = \max(0,z)$. The knots $\vartheta_1,\ldots,\vartheta_K$ are time points which satisfy $0 < \vartheta_1 < \cdots < \vartheta_K < \max(T_i)$. The spline coefficients $\boldsymbol{\varphi} = \{\varphi_1,\ldots,\varphi_K\}^{\mathrm{T}}$ construct the baseline hazard parameter $\boldsymbol{\theta}_{h_0}$ in this case. To balance between the bias and the variance, choosing the optimal number of knots is crucial. Keeping the total number of parameters in the survival sub-component of the joint model between $1/20^{\text{th}}$ and $1/10^{\text{th}}$ of the number of true events provides satisfactory estimates (Rizopoulos, 2012). To allow enough data-points within each interval and to assign more flexibility in the region of greater density, the knots can be placed on equispaced quantiles of the true event-times (Harrell Jr., 2001; Rizopoulos, 2012).

The second practical issue is the choice of bandwidth for the kernel functions. We recommend a form of cross-validation proposed by Fan and Zhang (2008) and Kürüm et al. (2018) to choose h_1 and h_2 simultaneously. In this approach, we leave out a single subject at a time rather than a single observation as the latter approach is inappropriate when there is within-subject dependence (Hoover et al., 1998). After removing the ith subject, we estimate $\boldsymbol{\theta}(\cdot)$ based on the remaining subjects. After doing this for each subject, we calculate the cross-validation score as follows

$$CV(h_1,h_2) = -\sum_{i=1}^{n} \mathcal{L}_i\{\widehat{\boldsymbol{\theta}}_{(-i)}(t)\},$$

where $\mathcal{L}_i\{\widehat{\boldsymbol{\theta}}_{(-i)}(t)\}$ is the observed (incomplete) data log-likelihood specified in (3.4) evaluated at $\widehat{\boldsymbol{\theta}}_{(-i)}(t)$ which is the leave-i-estimate of $\boldsymbol{\theta}(t)$. The pair (h_1,h_2) with the minimum $CV(h_1,h_2)$ is chosen as the optimum bandwidth combination.

The third issue is choosing the kernel function. We used *Epanechnikov kernel* (Epanechnikov, 1969), given by $K(t) = 0.75(1-t^2)$ with $|t| \leq 1$, for both longitudinal and

survival components because it minimizes the asymptotic mean-squared error of the local linear estimators (Fan and Gijbels, 1996).

The final practical issue is choosing the initial values of the parameters. The estimates obtained by fitting an ordinary LME model were chosen as initial values for $\boldsymbol{\beta}(t)$ at all time points t and the random effect variance σ^2_ξ. Similarly, the initial estimates at all time points for the coefficients $\gamma(t)$ and $\boldsymbol{\eta}(t)$ in the survival submodel were obtained by fitting an ordinary Cox model using $Y_i(t_{in_i})$ and $\mathbf{W}_i(t_{in_i})$ as the respective predictors. For the initial estimate of the vector of the baseline hazard parameter $\boldsymbol{\theta}_{h_0}$, we used a K dimensional vector having $\varphi_\kappa = 0$ for $\kappa \in \{1, \ldots, (K-1)\}$ and $\varphi_K = \log\left\{\sum_{i=1}^n T_i \Big/ \sum_{i=1}^n \delta_i\right\}$ where K is the number of knots used in the restricted cubic spline function. The knots were placed at percentiles of the uncensored event-times. We used the sample variance of the residuals obtained from fitting the LME as the initial estimates for the error variance at each grid-time. Using these initial values, we first estimated the posterior mean and variance of the random effects. Then, we used these values to maximize the expected local log-likelihood in (3.7) at all the grids.

3.4 Simulation Studies

We studied the performance of the proposed TV-JM via two simulation studies. For each study, the reported results were based on 150 data sets and each data set consisted of $n = 300$ subjects. For the ith subject, we generated $m = 30$ irregular longitudinal observation times from a standard uniform distribution. In the first simulation set-up, for each measurement time t, we generated a time-varying predictor $X_i(t)$ from a

standard normal distribution for the longitudinal submodel, and a time-invariant predictor W_i from a zero-mean normal distribution with variance 3 for the survival submodel. In the second set-up, we used the same time-varying predictor $X_i(t)$ in both submodels to assess the performance of our estimation when time-varying exogenous predictors are included in the survival submodel. The time-varying parameters in (3.3) were defined as $\beta_0(t) = 0.5\sin(3\pi t)$, $\beta_1(t) = 0.5\cos(3\pi t)$, $\gamma(t) = 0.5\cos(2\pi t)$, and $\eta(t) = \sin(\pi t) - 0.5$. The random effect ξ_i for the ith subject was simulated from a normal distribution with zero mean and $\sigma_\xi^2 = 1.5$.

The longitudinal response $Y_i(t)$ was generated from a normal distribution with mean $\{\beta_0(t) + X_i(t)\beta_1(t) + \xi_i\}$ and variance $\sigma^2(t) = 0.5 + \sin^2(1.5\pi t)$. The true event-time $\mathcal{T}_i$ was simulated using inverse probability integral transformation (Bender et al., 2005) with a Weibull baseline hazard $h_0(t) = \lambda t^{\lambda-1}$ where the shape parameter was $\lambda = 1.5$. In order to restrict our analysis time window between 0 and 1, we specified the censoring time as $\mathcal{C}_i = \min(1, u_i)$ where u_i is a random sample from an exponential distribution having mean 0.9. The observed event-time was defined as $T_i = \min(\mathcal{T}_i, \mathcal{C}_i)$ with $\delta_i = 1$ if $\mathcal{T}_i < \mathcal{C}_i$ and 0 otherwise. For both simulation set-ups, this led to an average censoring rate of approximately 55% with approximately 47% and 62% as the respective minimum and maximum. In practice, since longitudinal information after the observed event is usually no longer available due to death or dropout, we deleted repeated measurements after the observed event-time for all subjects. This led to an average of 13 observations per subject in both set-ups.

The local parameters were estimated at an equidistant set of grid points $\{t_r : r = 1, 2, \ldots, n_{grid}\}$ between 0 and 1 with $n_{grid} = 200$. In our simulation studies, we generated several pilot data sets, and used a cross-validation bandwidth selector to get an overall picture about the optimal bandwidths. To save computing time, we fixed the bandwidths to be close to the optimal ones from the pilot simulation data sets. Specifically, we set the bandwidths to $h_1 = 0.025$ and $h_2 = 0.39$ for the first scenario (time-invariant survival predictor) and $h_1 = 0.02$ and $h_2 = 0.42$ for the second scenario (time-varying exogenous survival predictor). For the estimation of the baseline hazard function via a restricted cubic spline, we placed four knots at the 5th, 35th, 65th, and 95th percentiles of the uncensored event-times.

For conducting inference on the model parameters, we calculated standard errors (SEs) using both likelihood-based and bootstrap-based approaches. The bias and the SEs from the two scenarios are presented in Tables 3.1 and 3.2, respectively. The "true" standard deviations of the parameters were assumed to be the sample standard deviation of their estimates obtained from 150 data sets (denoted as SD). The sample average and the sample standard deviation of 150 estimated likelihood-based SEs are denoted by SE and SD_{SE}, respectively. In addition, Boot_{SE} and $\text{Boot}_{\text{SD}_{\text{SE}}}$ denote the bootstrap-based SEs and their sample standard deviations using 100 data sets, each data set having 50 bootstrap samples. For both time-invariant and time-varying cases, we observe that the estimation bias of all the coefficients and the random effect variance is relatively small and is less than the corresponding SD, indicating that our proposed model performs well under both cases. However, we observe that the likelihood-based SEs underestimate the true SD values

(the difference between SD and SE is larger than the twice SD_{SE}), whereas the bootstrap-based standard errors (Boot_{SE}) cover the true SDs reasonably well within twice $\text{Boot}_{\text{SD}_{\text{SE}}}$. Based on these simulation results, we suggest that using bootstrap estimates of SEs are more suitable in practice and hence, we applied bootstrap SEs to form confidence intervals for TV-JM estimates in the WIHS data analysis. The plots of the average estimates of the regression coefficients, error variance, and the baseline hazard, along with their 95% bootstrap confidence intervals for the two scenarios are presented in Figures 3.1 and 3.2, respectively. We observe that the average estimates at each time point are close to the true values and the true values are covered by the confidence bands. Thus, estimation under TV-JM captures the dynamic trends of all the parameters efficiently and accurately.

Parameter	Time	Bias	SD	SE (SD_{SE})	Boot_{SE} ($\text{Boot}_{\text{SD}_{\text{SE}}}$)
$\beta_0(t)$	0.25	-0.010	0.114	0.015 (0.001)	0.105 (0.012)
	0.50	-0.042	0.137	0.017 (0.001)	0.117 (0.017)
	0.75	-0.052	0.138	0.019 (0.002)	0.125 (0.022)
$\beta_1(t)$	0.25	0.004	0.075	0.015 (0.001)	0.075 (0.010)
	0.50	-0.009	0.084	0.017 (0.002)	0.084 (0.014)
	0.75	0.009	0.106	0.019 (0.003)	0.096 (0.022)
$\gamma(t)$	0.25	-0.059	0.167	0.066 (0.028)	0.161 (0.070)
	0.50	-0.070	0.173	0.066 (0.022)	0.206 (0.062)
	0.75	-0.069	0.141	0.111 (0.024)	0.164 (0.034)
$\eta(t)$	0.25	0.074	0.084	0.061 (0.009)	0.094 (0.018)
	0.50	0.086	0.111	0.069 (0.010)	0.111 (0.022)
	0.75	0.102	0.116	0.092 (0.018)	0.137 (0.031)
σ_ξ^2	—	0.021	0.133	0.118 (0.011)	0.131 (0.019)

Table 3.1: Results for the simulation set-up with time-invariant survival predictor (averaged over 150 data sets). Given are bias, standard deviation (SD), likelihood-based standard errors (SE), and bootstrap SE (Boot_{SE}). Given in parentheses (SD_{SE} and $\text{Boot}_{\text{SD}_{\text{SE}}}$) are standard deviations of the corresponding quantities.

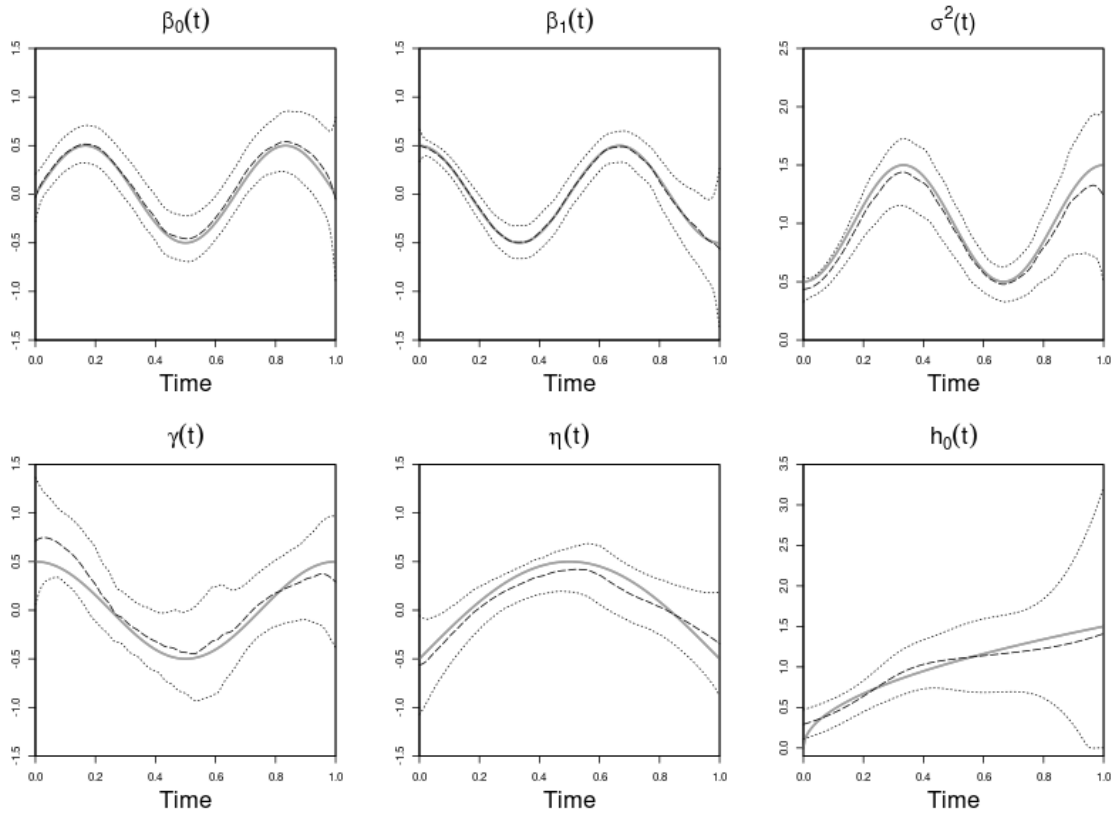


Figure 3.1: Estimated parameters from the simulation scenario with the time-invariant survival predictor. The solid and the dashed curves represent the true and the average estimated values of the parameters, respectively. The dotted lines are 95% bootstrap confidence bands.

Parameter	Time	Bias	SD	SE (SD_{SE})	$Boot_{SE}$ ($Boot_{SD_{SE}}$)
$\beta_0(t)$	0.25	-0.030	0.111	0.015 (0.001)	0.114 (0.016)
	0.50	-0.038	0.133	0.017 (0.002)	0.125 (0.016)
	0.75	-0.045	0.131	0.019 (0.002)	0.138 (0.024)
$\beta_1(t)$	0.25	0.006	0.086	0.015 (0.001)	0.085 (0.014)
	0.50	-0.010	0.096	0.017 (0.002)	0.096 (0.018)
	0.75	0.010	0.120	0.019 (0.003)	0.114 (0.027)
$\gamma(t)$	0.25	-0.068	0.233	0.062 (0.022)	0.225 (0.067)
	0.50	-0.080	0.178	0.072 (0.092)	0.218 (0.068)
	0.75	-0.040	0.176	0.093 (0.030)	0.213 (0.055)
$\eta(t)$	0.25	-0.063	0.227	0.119 (0.014)	0.263 (0.043)
	0.50	0.038	0.169	0.107 (0.040)	0.213 (0.034)
	0.75	0.086	0.185	0.145 (0.019)	0.235 (0.037)
σ_ξ^2	—	0.022	0.138	0.118 (0.011)	0.131 (0.020)

Table 3.2: Results for the simulation set-up with the time-varying exogenous survival predictor (averaged over 150 data sets). Given are bias, standard deviation (SD), likelihood-based standard errors (SE), and bootstrap SE ($Boot_{SE}$). Given in parentheses (SD_{SE} and $Boot_{SD_{SE}}$) are standard deviations of the corresponding quantities.

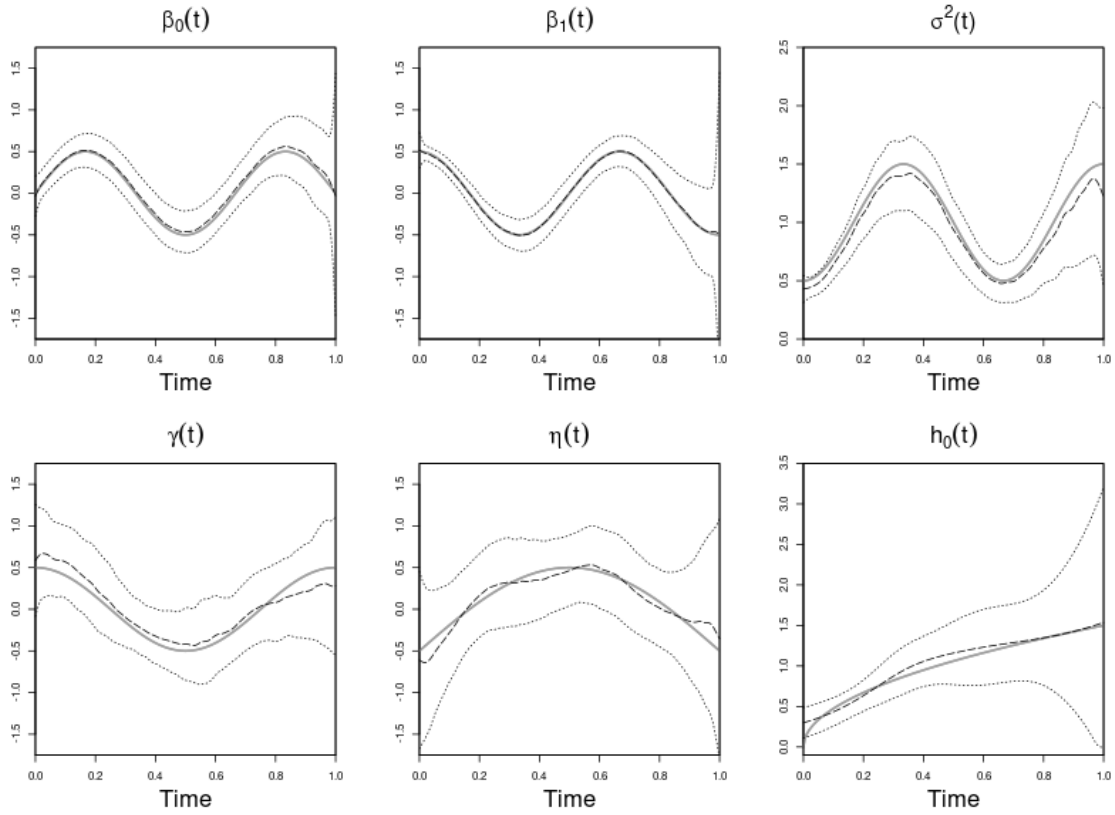


Figure 3.2: Estimated parameters from the simulation scenario with the time-varying exogenous survival predictor. The solid and the dashed curves represent the true and the estimated mean values of the parameters, respectively. The dotted lines are 95% bootstrap confidence bands.

3.5 Application to the Women's Interagency HIV Study

We utilized our proposed time-varying joint model for the analysis of a subset of Women's Interagency HIV Study (WIHS) data. The data consists of follow-up information between 1994 and 2015 on 901 HIV positive and 1280 HIV negative women recruited from HIV testing sites across ten cities in the U.S.. Participants were scheduled to have semiannual interviews where they received physical and oral examinations, gave blood, urine, and

gynaecological specimens, and also answered a series of questions about their daily activities such as sexual behaviours, tobacco, drug, and alcohol use.

An important goal of HIV/AIDS research is to investigate the response-predictor relationships, that is, to monitor how potential exploratory risk factors affect the progression of the human immunodeficiency virus through the trend of CD4 cell percentage and time to death of the patients. Additionally, since the presence of association between these two outcomes is highly likely, it is crucial to inspect this response-response relationship as well. Furthermore, as the data is longitudinal, both relationships may change over years since the time of seroconversion. Therefore, we limited our analysis to the 901 HIV positive subjects. In this subset, the age of the subjects varied between 19 and 73 during recruitment, 55.5%, 19.5%, 18.3%, and 6.7% of the subjects self-identified as African-American non-Hispanic origin, Caucasian non-Hispanic origin, Latina or Hispanic, and others/mixed race, respectively, and 86% of the participants were heterosexual. As many participants missed some of their scheduled visits, the number of measurements and measurement times vary from subject to subject. The minimum, maximum, and average number of longitudinal observations per subject were 1, 34, and 15.8, respectively. Out of the 901 subjects, 447 died during the study which lead to a censoring rate of 50.4%. Based on previous HIV literature (Zeger and Diggle, 1994; Kürüm et al., 2015), we chose a set of predictors as follows. For the CD4 percentage, current smoking behavior, current HPV infection status, CES-D depression score, history of AIDS diagnosis, baseline CD4 percentage, sexual-orientation, and race of the subjects were selected as predictors. We chose sexual orientation and race for the hazard submodel. For the ith subject, the true longitudinal trajectory of the percentage of CD4

cells was modelled as

$$m_i(t) = \mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\zeta}_1(t) + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\zeta}_2(t) + \xi_i,$$

where $\mathbf{X}_i(t) = \{1, X_{i1}(t), \ldots, X_{i4}(t)\}^{\mathrm{T}}$ was the vector of time-varying predictors with

$$\begin{aligned} X_{i1}(t) &= \text{indicator of smoking behavior at time } t, \\ X_{i2}(t) &= \text{indicator of HPV infection status at time } t, \\ X_{i3}(t) &= \text{CES-D depression score at time } t, \text{ and} \\ X_{i4}(t) &= \text{indicator of AIDS diagnosis at or before time } t, \end{aligned}$$

and corresponding vector of coefficients $\boldsymbol{\zeta}_1(t) = \{\beta_0(t), \beta_1(t), \ldots, \beta_4(t)\}^{\mathrm{T}}$. Here, ξ_i was the unknown underlying random intercept. The vector of time-invariant or baseline predictors was $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i5})^{\mathrm{T}}$, where

$$\begin{aligned} Z_{i1} &= \text{baseline CD4 percentage (CD4 at the time of HIV detection)}, \\ Z_{i2} &= \text{binary variable denoting if a subject was heterosexual}, \end{aligned}$$

and three indicator variables for race, Z_{i3} = African-American non-Hispanic, Z_{i4} = Hispanic, and Z_{i5} = Caucasian non-Hispanic, with corresponding vector of coefficients $\boldsymbol{\zeta}_2(t) = \{\beta_5(t), \ldots, \beta_9(t)\}^{\mathrm{T}}$. The hazard submodel was

$$h_i\{t|\mathcal{M}_i(t), \mathbf{W}_i\} = h_0(t) \exp\Big\{m_i(t)\gamma(t) + \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\eta}(t)\Big\},$$

where $\mathbf{W}_i = (W_{i1}, \ldots, \mathbf{W}_{i4})^{\mathrm{T}}$ was a subset of the baseline predictors $\mathbf{Z}_i$ which included sexual-orientation (W_{i1}) and three indicator variables of race ($W_{i2} = Z_{i3}, W_{i3} = Z_{i4}, W_{i4} = Z_{i5}$), respectively, with the vector of coefficients $\boldsymbol{\eta}(t) = \{\eta_1(t), \ldots, \eta_4(t)\}^{\mathrm{T}}$. Considering the number of years since the first detection of HIV as our underlying time-scale, we estimated

the time-varying parameters at 200 equally spaced grid points in the interval [0, 18.5] years with optimal bandwidth selected as ($h_1 = 8$, $h_2 = 12$) using a cross-validation bandwidth selector. For the estimation of the baseline hazard function, we used a restricted cubic spline function with $K = 5$ knots placed at 5th, 27.5th, 50th, 72.5th, 95th percentiles of the true event-times.

The estimate of the random effect variance from the WIHS data analysis is $\widehat{\sigma}^2_\xi = 70.66$ with 95% bootstrap confidence interval (55.15, 86.17). The rest of the parameters are time-varying functions and therefore, are presented through Figures 3.3 and 3.4. Figures 3.3 (a)-(i) represent the estimated parameters from the longitudinal submodel, along with their 95% bootstrap confidence bands. Following are the interpretations of the results.

(a) The intercept is significantly negative for the first 5.5 years since HIV diagnosis. It shows an increasing trend for the first 12 years, stabilizes in between 12th and 15th years, and declines afterwards.

(b) Although smoking shows a positive effect on CD4 percentage for the first 4.5 years, it is declining in nature throughout the entire study window, and starts showing a significantly negative impact on CD4 percentage after 8 years of seroconversion.

(c) Co-infection with HPV has significantly negative impact on CD4 for most of the study window starting after the first year since HIV diagnosis.

(d) The estimated coefficient function of CES-D depression scores indicates that it has a significantly negative effect on CD4 percentage between 3 and 13.5 years after the first HIV diagnosis.

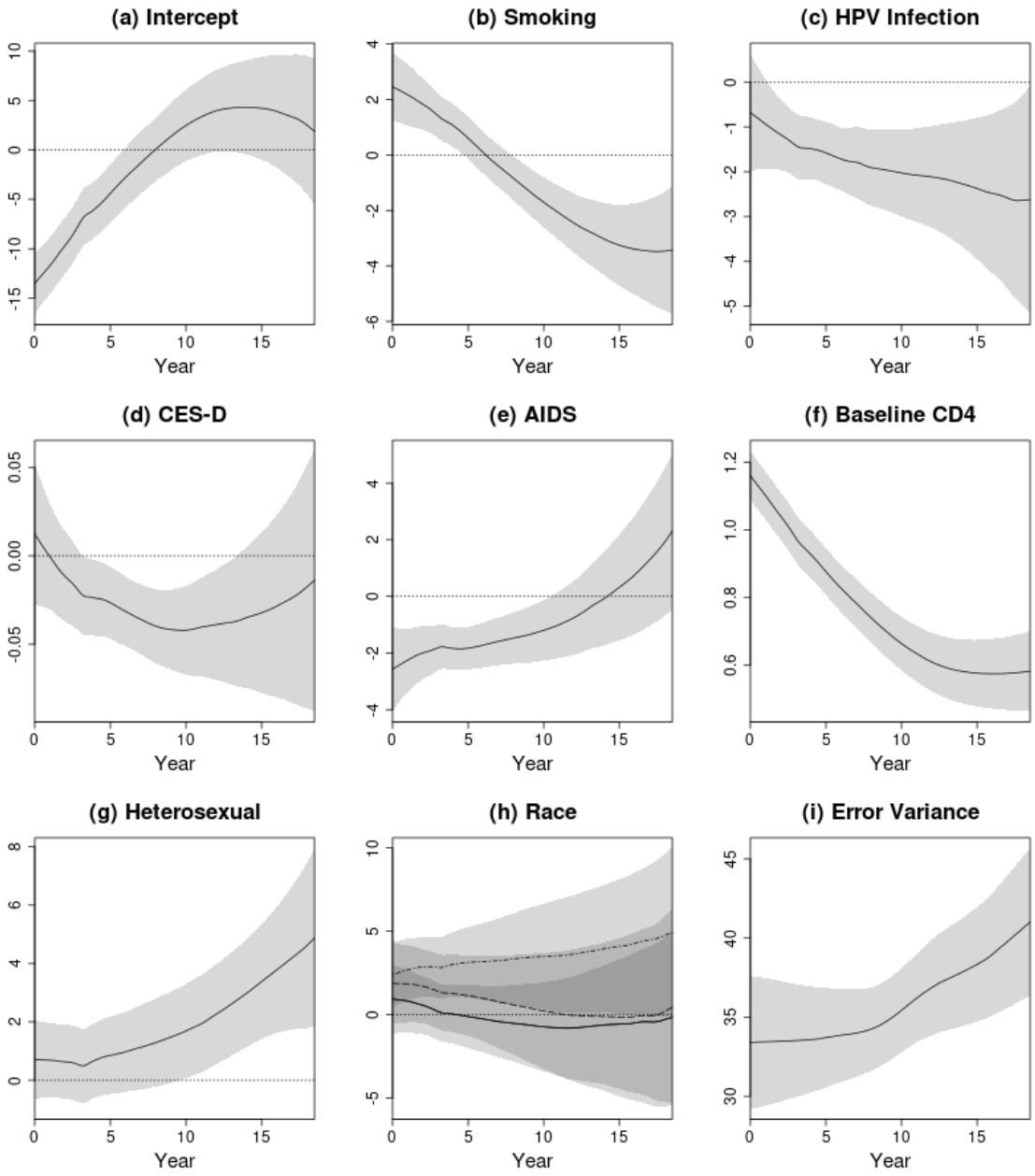


Figure 3.3: Estimated parameters (solid, dashed, or dotted-dashed curves) of the longitudinal submodel from the WIHS data analysis along with their 95% bootstrap confidence bands (shaded region).

(e) Subjects who were diagnosed with AIDS had significantly lower CD4 percentage than the subjects who did not have AIDS for the first 10.5 years. This difference is decreasing with time, and vanishes after 10.5 years of first HIV detection.

(f) Baseline CD4 has significantly positive impact on CD4 percentage for the entire study period, though this effect is declining until 13 years after the first HIV diagnosis, and it stabilizes after that.

(g) We found that after 9.5 years since seroconversion, heterosexual subjects start having higher CD4 percentage in comparison to that of subjects with other sexual orientations, and this difference increases with time.

(h) In this figure, the solid, dashed, and dotted-dashed curves represent the estimated coefficients of the subjects from African-American non-Hispanic, Hispanic, and Caucasian non-Hispanic origins, respectively. Although we observe that the CD4 trend of the Caucasian subjects diverges from the trends of the other two races with time, the difference among the three groups were not significant at any time within the study period.

(i) We observe that the error variance increases over time.

Figures 3.4 (a)-(d) display the estimated parameters along with their 95% bootstrap confidence bands from the survival submodel. We interpret the results from these figures below.

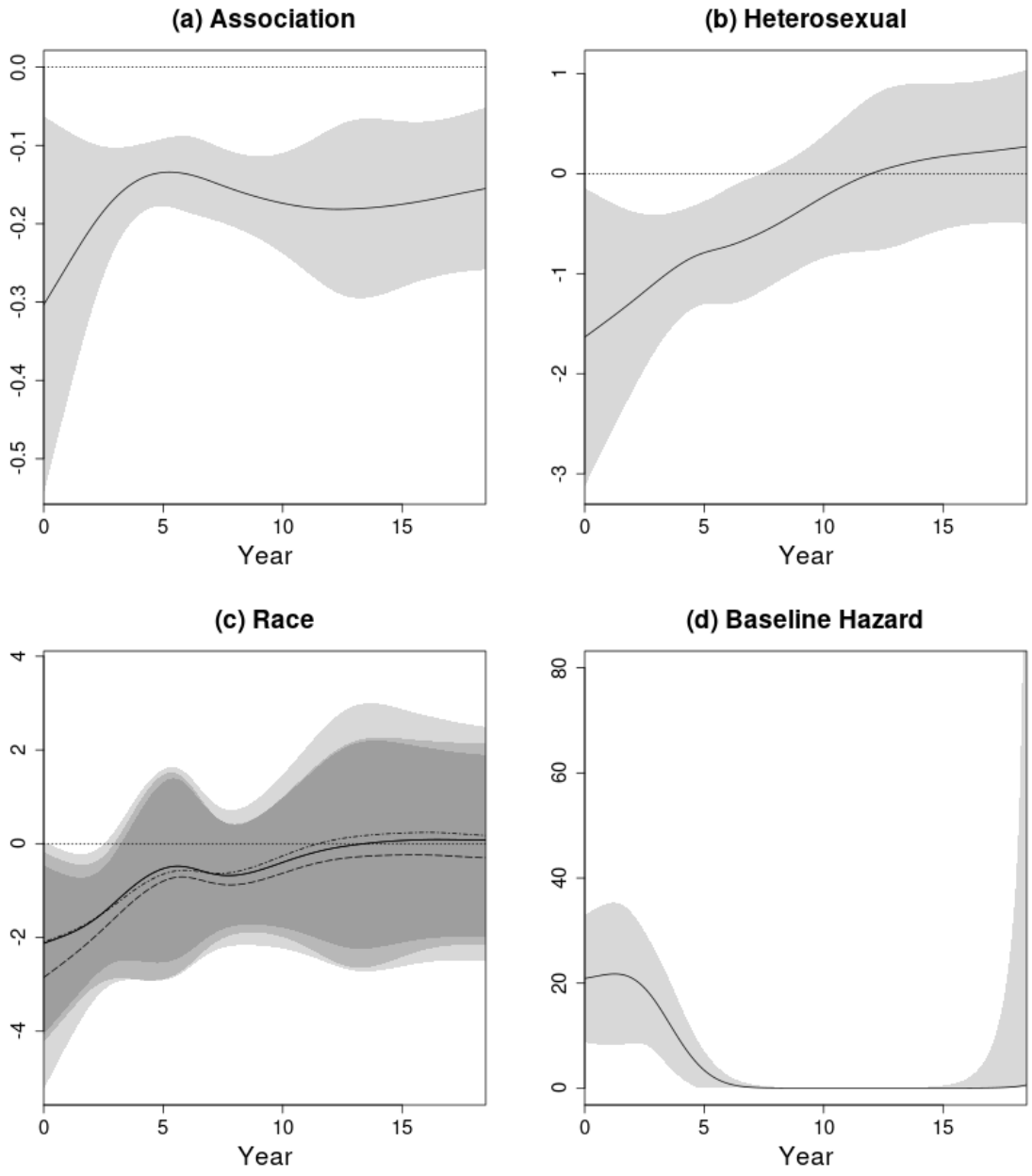


Figure 3.4: Estimated parameters (solid, dashed, or dotted-dashed curves) of the survival submodel from the WIHS data analysis along with their 95% bootstrap confidence bands (shaded region).

(a) The association between CD4 percentage and time to death is significantly negative for the entire 18.5 years follow-up time. This implies that a declining CD4 percentage increases the risk of death in a patient for the entire study window.

(b) Sexual orientation has significant effect on time to death for the first 7 years, where the heterosexual subjects are at slightly less risk of death in comparison to the other subjects. Note that, in a Cox hazard model, the negative value of a coefficient indicates smaller risk of death with higher value of the corresponding predictor.

(c) This figure displays the estimated coefficients of all three races from the survival submodel, where the solid, dashed, and dotted-dashed curves represent the estimated effects of the subjects from African-American non-Hispanic, Hispanic, and Caucasian non-Hispanic origins, respectively. We observe from this plot that the trend of these three races are approximately the same, but they have slightly smaller risk of death during the first 3 years since seroconversion in comparison to the subjects who belong to the other/mixed race category.

(d) The baseline hazard function indicates that at the beginning of the study window, that is, when a subject was just found to be HIV positive, has very high risk of death. This risk decreases and eventually becomes very close to zero between 7 and 15 years.

To summarize, in this chapter, we proposed our novel time-varying joint model (TV-JM) (Section 3.1). We demonstrated the estimation technique via an EM algorithm (Section 3.2), where in the E-step, the random effects were estimated (Section 3.2.1) and in the M-step, we applied local linear regression techniques to estimate the time-varying

coefficients (Section 3.2.2). The practical issues related to the estimation were discussed (Section 3.3) and extensive simulation studies were conducted to illustrate the finite sample performance of the TV-JM estimators (Section 3.4). Finally, we applied TV-JM for the analysis of the WIHS data (Section 3.5). Application of this methodology allows researchers to examine the time-varying association between CD4 percentage and time to death of HIV positive patients while also uncovering the complex dynamic effects of various sociodemographic, behavioral, and pathological exploratory factors on these outcomes. This may lead to a better understanding of HIV progression at various stages of life and therefore, leading to insightful contributions to HIV/AIDS research. For example, several intervention programs can be established for the HIV positive individuals in order to change their smoking behavior or control their depressive phases, which might help to keep CD4 percentage from dropping below the critical level. This in turn might keep their immunity against co-infections intact and despite having HIV, might help them to lead a healthier and longer life.

Chapter 4

Analysis of the Purpose in Life Data: A Novel Application of Generalized Time-Varying Coefficient Models

4.1 Background

Purpose in life has been established as a major psychological resource linked with numerous health benefits, adaptive aging, and longevity. Individuals with a strong sense of purpose in life are likely to have more motivation, future-oriented thoughts, higher self-control, and they are likely to strive more for life-goals even under adverse environments and traumatic experiences. The formation of purposefulness is an integral component of

identity development in early life, particularly during adolescence. Research suggests that early-life socialization processes shaped by advantaged or disadvantaged origins can either benefit or obstruct purposeful thinking about future life pursuits. The processes of identity development and socialization are shaped by socioeconomic status (SES). Despite a growing body of research on adult SES and purpose in life, we know only a little about the extent to which early-life SES contributes to social stratification of purpose in midlife and beyond. As one advances from midlife to old age, sense of purpose tends to diminish, possibly due to loss of social, familial, and physical roles through life transitions, including retirement, widowhood, and health problems of oneself and close family members.

Studies that employ longitudinal data only examined a short time span (less than 10 years) and investigated changes in purpose across multiple observations rather than age. Therefore, additional research is needed to understand the trajectory of purposeful thought at different stages of life, particularly from midlife to old age when psychological well-being is expected to decline. How sense of purpose changes in later life for the upwardly mobile group is an open question which we investigated in this project. *We employed the cutting-edge statistical method, generalized varying-coefficient model (GVCM), which has never been used to analyze data on healthy aging.* Since these models do not pre-specify the response-predictor relationships, its application helped us to uncover the complex age-varying effects of social mobility, demographic status, and major life events on purpose in life. More specifically, with longitudinal data from the Midlife in the United States (MIDUS) study, we used these models to investigate (1) the age trajectory of purpose during the second half of the life span (ages 40-80), (2) the role of major life events (such as retirement and

illness) in explaining the later-life purpose, and (3) how the trajectory of purpose in life differs across social mobility groups (measured by childhood and adult SES) after midlife. However, the MIDUS data contains a significant amount of missing values; therefore, we first investigated the performance of GVCM under missingness mechanisms to ensure that it leads to accurate and efficient estimates.

This chapter is divided into the following sections. Section 4.2 presents a description of the MIDUS data. Section 4.3 describes the model we used for the analysis. A brief literature on missing data mechanisms along with how they are handled under the GVCM framework are discussed in Section 4.4. A simulation study to demonstrate the performance of GVCM under missing at random mechanism is presented in Section 4.5. Finally, in Section 4.6, the analysis results of the MIDUS data are shown.

4.2 Midlife in the United States (MIDUS) Data

The MIDUS data was collected in three waves M1 (1995–1996), M2 (2004–2005), and M3 (2013–2014). The study targeted non-institutionalized, English-speaking adults aged between 25 and 74 (during recruitment) in the United States and it consisted of a two-stage survey: a telephone interview and a self-administered questionnaire. There were 6325 subjects who responded for the survey. Since our goal was to explore the purpose in life during midlife to older ages only, we considered a subset of the data consisting of 5559 individuals between ages 40 and 80. In our analysis, an observation was considered incomplete when the outcome (purpose in life) or at least one of the predictors was missing at that measurement time, but we did not completely ignore that individual if information

on all other waves were available. After excluding these missing observations, we had 4656 subjects in the final analytic sample.

Participants completed measures of purpose in life (PIL) at M1, M2, and M3, where they were asked to respond to a three-item purpose sub-scale of Ryff's Psychological Well-Being measure (Ryff and Keyes, 1995), that is, how much they agree with the following three questions on a scale from 1 (strongly disagree) to 7 (strongly agree): "Some people wander aimlessly through life, but I am not one of them," "I live life one day at a time and do not really think about the future," and "I sometimes feel as if I have done all there is to do in life". In order to investigate the role of major later-life events on the PIL score, we formed six binary variables to capture social, familial, or physical role-related transitions or events. In particular, retirement was coded based on if respondents retired from their employment; widowed includes respondents whose last marriage ended with the death of their spouse and who did not remarry, and self-illness indicates whether respondents have ever been diagnosed with cancer, stroke, or heart problems. We also created three binary variables related to worsening health of family members (parents, spouse/partner, and children) based on questions regarding whether these family members had a chronic disease or disability in the past 12 months. To explore the effect of social mobility on the PIL score, we constructed the following five groups based on social class memberships during childhood and adulthood: stable low (low in childhood/low in adulthood, n=596), downward mobility (high/middle, high/low, or middle/low, n=1281), stable middle (middle/middle, n=554), upward mobility (low/middle, low/high, or middle/high, n=1,444), and stable high (high/high, n=781). In our analysis, we controlled for age (the underlying timescale), gender, race (Caucasian vs.

others), and attrition status. While men (48%) and women (52%) were almost evenly distributed, the majority of the sample (94%) was Caucasian. Out of 4,656 respondents in the final sample, 61% remained in the study throughout all three waves, whereas 39% died or were lost to follow-up (LFU) following M1 or M2. We created a categorical variable that reflects five different patterns of attrition: (i) participated in all three waves (n=2834), (ii) participated in M1 and subsequently LFU (n=414), (iii) participated in M1 and then died (n=379), (iv) participated in M1, M2, and then LFU (n=629), and (v) participated in M1, M2, and then died (n=400).

4.3 Generalized Varying-Coefficient Models

As the response, purpose in life (PIL) score, was composed of discrete measures, we assumed it follows a Poisson distribution; and therefore, employed a generalized time-varying coefficient model (Cai et al., 2000) to analyze this data set. In this section, we have briefly described the GVCM model, its likelihood construction, estimation procedure, inference on the parameters, and practical issues.

Let $Y_i(t_{ij})$ be the PIL score of the ith subject measured at jth time point t_{ij}, for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and the time-varying predictors at any time t are defined as $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^{\mathrm{T}}$. Since GVCM accounts for the within-subject correlation under independent working correlation structure (discussed in Section 4.4), we treated all $N = \sum_{i=1}^{n} n_i$ observations independently. The GVCM model is expressed as

$$E\{Y_i(t)|\mathbf{X}_i(t)\} = g^{-1}\{\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t)\},$$

where $g(\cdot)$ is the canonical link and $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_p(t)\}^{\mathrm{T}}$ is the vector of coefficient functions quantifying the time-varying effect of the predictors on the mean PIL scores. Since the response is assumed to follow a Poisson distribution, we used a log-link to describe the relationship between the response and the predictors. Hence, the model can be rewritten as $E\{Y_i(t)|\mathbf{X}_i(t)\} = \exp\{\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t)\}$ and the log-likelihood under this GVCM model is given by

$$\ell(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[-\exp\left\{\mathbf{X}_i^{\mathrm{T}}(t_{ij})\boldsymbol{\beta}(t_{ij})\right\} + Y_i(t_{ij})\left\{\mathbf{X}_i^{\mathrm{T}}(t_{ij})\boldsymbol{\beta}(t_{ij})\right\} - \log\{Y_i(t_{ij})!\}\Big].$$

For each given time t_0, the coefficient function for the (i,j)th measurement, $\beta_\kappa(t_{ij})$ was approximated by a local linear regression $\beta_\kappa(t_{ij}) \approx a_\kappa + b_\kappa(t_{ij} - t_0)$ for t_{ij} in the neighborhood of t_0, where $a_\kappa = \beta_\kappa(t_0)$ and $b_\kappa = \beta'_\kappa(t_0)$, for $\kappa = 1, \ldots, p$. With $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$ as the parameters, the local log-likelihood function

$$\ell(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[Y_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})\left\{\mathbf{a} + \mathbf{b}(t_{ij} - t_0)\right\} - Z_i(t_{ij})\Big]K_h(t_{ij} - t_0) \tag{4.1}$$

is maximized to obtain $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{a}} = (\widehat{a}_1, \ldots, \widehat{a}_p)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}' = \widehat{\mathbf{b}} = (\widehat{b}_1, \ldots, \widehat{b}_p)^{\mathrm{T}}$, the maximum likelihood estimates of $\mathbf{a}$ and $\mathbf{b}$, respectively, where $Z_i(t) = \exp\left[\mathbf{X}_i^{\mathrm{T}}(t)\left\{\mathbf{a} + \mathbf{b}(t - t_0)\right\}\right]$ and $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a kernel function with bandwidth h. Since no closed-form solution to the likelihood in (4.1) is available, an iterative maximum likelihood technique via the Newton-Raphson (NR) algorithm is employed to estimate the parameters as

$$\begin{Bmatrix} \mathbf{a}^{(r+1)} \\ \\ \mathbf{b}^{(r+1)} \end{Bmatrix} = \begin{Bmatrix} \mathbf{a}^{(r)} \\ \\ \mathbf{b}^{(r)} \end{Bmatrix} - [\ell''\{\mathbf{a}^{(r)}, \mathbf{b}^{(r)}\}]^{-1}\ell'\{\mathbf{a}^{(r)}, \mathbf{b}^{(r)}\},$$

where $\{\mathbf{a}^{(r)}, \mathbf{b}^{(r)}\}$ is the estimate of $(\mathbf{a}, \mathbf{b})$ at the rth NR iteration, and $\ell'\{\mathbf{a}^{(r)}, \mathbf{b}^{(r)}\}$ and $\ell''\{\mathbf{a}^{(r)}, \mathbf{b}^{(r)}\}$ are the respective first and second order derivatives of the local log-likelihood

in (4.1) evaluated at the rth NR iteration. The first order derivatives with respect to the local parameters are

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{a}}\ell(\mathbf{a},\mathbf{b}) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[\{Y_i(t_{ij}) - Z_i(t_{ij})\}\mathbf{X}_i(t_{ij})\Big]K_h(t_{ij}-t_0) \text{ and} \\
\frac{\partial}{\partial \mathbf{b}}\ell(\mathbf{a},\mathbf{b}) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big[\{Y_i(t_{ij}) - Z_i(t_{ij})\}\mathbf{X}_i(t_{ij})(t_{ij}-t_0)\Big]K_h(t_{ij}-t_0),
\end{aligned}$$

and the second order derivatives of the local likelihood with respect to the local parameters are given by

$$\begin{aligned}
\frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{a}^{\mathrm{T}}}\ell(\mathbf{a},\mathbf{b}) &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big\{Z_i(t_{ij})\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})\Big\}K_h(t_{ij}-t_0), \\
\frac{\partial^2}{\partial \mathbf{b}\partial \mathbf{b}^{\mathrm{T}}}\ell(\mathbf{a},\mathbf{b}) &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big\{Z_i(t_{ij})\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})(t_{ij}-t_0)^2\Big\}K_h(t_{ij}-t_0), \text{ and} \\
\frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{b}^{\mathrm{T}}}\ell(\mathbf{a},\mathbf{b}) &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big\{Z_i(t_{ij})\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})(t_{ij}-t_0)\Big\}K_h(t_{ij}-t_0).
\end{aligned}$$

For inference on the parameters, we first need to estimate the variances of the estimators. For this purpose, we employed sandwich estimators, which were proposed by Carroll et al. (1998) under the local estimating equations, and were shown to provide consistent estimates (Huber et al., 1967). Following the notations presented in Cai et al. (2000), the covariance matrix at time t_0 was estimated by

$$\widehat{\mathbf{\Sigma}}(t_0) = \widehat{\mathbf{\Gamma}}(t_0)^{-1}\widehat{\mathbf{\Lambda}}(t_0)\widehat{\mathbf{\Gamma}}(t_0)^{-1},$$

where

$$
\begin{aligned}
\widehat{\mathbf{\Gamma}}(t_0) &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[q_2\big\{S_{ij}, Y_i(t_{ij})\big\}\mathbf{M}_i(t_{ij})\mathbf{M}_i^{\mathrm{T}}(t_{ij})\right]K_h(t_{ij}-t_0),\\
\widehat{\mathbf{\Lambda}}(t_0) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[q_1^2\big\{S_{ij}, Y_i(t_{ij})\big\}\mathbf{M}_i(t_{ij})\mathbf{M}_i^{\mathrm{T}}(t_{ij})\right]K_h^2(t_{ij}-t_0),\\
\mathbf{M}_i(t_{ij}) &= \begin{bmatrix} \mathbf{X}_i(t_{ij}) \\ \\ \mathbf{X}_i(t_{ij})(t_{ij}-t_0)\end{bmatrix},\\
q_1\big\{S_{ij}, Y_i(t_{ij})\big\} &= \frac{\partial}{\partial S_{ij}}\ell(\mathbf{a},\mathbf{b}) = Y_i(t_{ij}) - Z_i(t_{ij}),\\
q_2\Big\{S_{ij}, Y_i(t_{ij})\Big\} &= \frac{\partial^2}{\partial S_{ij}^2}\ell(\mathbf{a},\mathbf{b}) = -Z_i(t_{ij}),
\end{aligned}
$$

with $S_{ij} = \mathbf{X}_i^{\mathrm{T}}(t_{ij})\left\{\mathbf{a} + \mathbf{b}(t_{ij}-t_0)\right\}$ and $Z_i(t_{ij}) = \exp(S_{ij})$.

One practical issue for GVCM is to choose an optimal bandwidth h, which is crucial for the bias-variance trade-off. We followed a leave-one-subject-out cross-validation technique (Hoover et al., 1998) to find the optimum bandwidth. The cross-validation score is given as

$$
CV(h) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Big\{Y_i(t_{ij}) - \widehat{Y}_{-i}(t_{ij})\Big\}^2,
$$

where $Y_i(t)$ denotes the observed value of the response for the ith subject at time t and $\widehat{Y}_{-i}(t)$ is the fitted value of this response with subject i excluded. The cross-validation score was calculated for a set of different values of h and the value that lead to the minimum score was chosen to be the optimal bandwidth.

4.4 Missing Data Mechanisms

In the MIDUS data set, we considered 5559 respondents between ages 40 and 80 who had information on 16048 repeated measurements. Among these participants, 3531 (63.5%) provided complete information at the first wave, 3043 (54.7%) at the second wave, and 2077 (37.4%) at the third wave, making only 8651 out of 16048 (54%) measurements complete (or 46% incomplete). In longitudinal experiments, missing data of this kind is very common due to dropouts or irregular visits of the subjects. Understanding and handling the missing data mechanism with the appropriate statistical method is crucial for obtaining accurate and efficient estimates. In this section, we briefly discuss various missing data mechanisms, the challenges they create in statistical analysis, and how they are handled by GVCM.

There are three types of missing data mechanisms, namely, *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). When the underlying reason for missing data is independent of the response, the mechanism is known as MCAR. MAR and MNAR scenarios arise when missing data are dependent on observed and missing (or unobserved) outcomes, respectively. MCAR and MAR are usually identified as ignorable missingness, whereas MNAR is known as non-ignorable and it requires to be modelled appropriately during the data analysis. When the underlying missingness mechanism is assumed to be ignorable, that is, MCAR or MAR, it is common to analyze the data using the subjects with complete information only (complete-case analysis). In addition to being computationally fairly straightforward to implement, under the complete-case scenario, the maximum likelihood estimation employed for normally-distributed outcomes

is shown to produce consistent estimates (Verbeke and Molenberghs, 1997). However, for non-normal outcomes, as no unified likelihood-based approach exists, generalized estimating equations (GEE) are employed. The advantages of GEE approach are that it can handle unbalanced designs and it takes into account the within-subject dependence via specifying a working correlation matrix, which approximates the true underlying correlation matrix for the response variable and is an important component of the estimation. However, the GEE approach produces consistent estimators only when the underlying missingness mechanism is assumed to be MCAR. When the missingness mechanism is MAR, the consistency of the GEE estimators depends on the specification of the working correlation matrix, more specifically, it either has to closely follow the true structure or working independence structure should be used (Fitzmaurice et al., 2008). Therefore, due to convenience, for data with MAR mechanism, Salazar et al. (2016) suggested assigning the independent correlation structure to the working correlation matrix instead of searching for the true correlation structure. Lin and Carroll (2000) showed that the efficiency of GEE estimators is still valid under the working independence assumption, in particular, the authors showed that independent correlation structure provides the smallest variance under local estimation without affecting the bias of the estimators. Additionally, Kauermann (2000) proposed a locally weighted version of GEE with an independent working correlation matrix, which is referred as weighted generalized estimating equations (WGEE), to model an ordinal longitudinal response. They proved that the desirable properties of GEE estimators, that is, efficiency and consistency are still valid under the WGEE approach.

As indicated in Section 4.3, our analysis of purpose in life score is performed via the generalized varying-coefficient models (GVCM), where our estimation is based on the local linear techniques. The local linear estimation techniques are in the same spirit as the WGEE approach since they also involve a weight matrix, which is computed using a kernel function (Carroll et al., 1998), and in addition, we also assume the working independence correlation structure. In order to demonstrate that the benefits of using WGEE extend to the estimation under GVCM for data sets with MAR missingness mechanism, we conducted simulation studies. The results are presented in Section 4.5 and they are consistent with the literature described above.

4.5 Simulation Studies

For validation of our analysis method for the MIDUS data, that is, to demonstrate that GVCM leads to accurate and efficient estimates under MAR, we performed simulation studies with two different cases: time-invariant categorical and time-varying continuous predictors. For each of these cases, we provided comparison between results obtained from complete data (that is, no missing observations) and MAR scenarios. We generated 400 data sets for each of the simulation set-ups and presented the bias and the standard errors of the estimates in Tables 4.1 and 4.2. The mean estimates along with their confidence bands obtained using sandwich standard errors (SEs) (referred as sandwich confidence bands from here onward), and the true coefficient functions are presented in Figures 4.1 and 4.2.

4.5.1 Categorical Predictors

We designed this case with time-invariant categorical predictors in order to make our simulation comparable to the MIDUS data. For each of the 400 data sets, we considered $n = 400$ subjects. Since we had a large number of participants with only less than or equal to three longitudinal time points in the MIDUS data, we generated $n_i = 3$ irregular measurement times from the standard uniform distribution for $i = 1, \ldots, n$. The first predictor X_{i1} was generated from a Bernoulli distribution with probability 0.55. The second predictor was generated from a multinomial distribution with the vector of success probabilities (0.33, 0.33, 0.34). Due to this predictor variable having three categories, we introduced two indicator variables X_{i2} (which equals to 1 if the random sample from the above distribution was (0, 1, 0) and equals to 0 otherwise) and X_{i3} (equals to 1 if the sample was (0, 0, 1), and 0 otherwise), respectively. The response variable $Y_i(t)$ was generated from a GVCM

$$E\{Y_i(t)|\mathbf{X}_i\} = \exp\{\beta_0(t) + \beta_1(t)X_{i1} + \beta_2(t)X_{i2} + \beta_3(t)X_{i3}\}, \tag{4.2}$$

where $\mathbf{X}_i = \{1, X_{i1}, X_{i2}, X_{i3}\}^{\mathrm{T}}$, $\beta_0(t) = 0.5 + \sin(2.25t)$, $\beta_1(t) = 0.5 + 2t(1-t)$, $\beta_2(t) = \cos(1.5\pi t)$, and $\beta_3(t) = -\sin(\pi t)^2$. We generated the Poisson-distributed response using Gaussian copulas in order to impose the appropriate association among the repeated measurements. In particular, we simulated the response as follows:

(i) We constructed the correlation matrix of order $n_i = 3$ for the ith individual by applying the correlation function $\rho(t, u) = 2^{-|t-u|}$, where t and u are two different time points.

(ii) The corresponding Cholesky roots were calculated to impose the correlation structure on a random sample from an n_i-variate $N(\mathbf{0}, \mathbf{I})$ distribution resulting in a vector, say, $\{W_i(t_{i1}), \ldots, W_i(t_{in_i})\}^{\mathrm{T}}$, where $\mathbf{I}$ is an n_i-dimensional identity matrix.

(iii) We applied probability integral transformation (PIT) to obtain

$$\mathbf{U} = \{U_i(t_{i1}), \ldots, U_i(t_{in_i})\}^{\mathrm{T}} = \left[\Phi^{-1}\{W_i(t_{i1})\}, \ldots, \Phi^{-1}\{W_i(t_{in_i})\}\right]^{\mathrm{T}}$$

with each element of $\mathbf{U}$ following standard uniform distribution and $\Phi^{-1}(\cdot)$ was the inverse cumulative distribution function (CDF) of a standard normal distribution.

(iv) We applied inverse PIT on $\mathbf{U}$ to generate the response

$$\{Y_i(t_{i1}), \ldots, Y_i(t_{in_i})\}^{\mathrm{T}} = \left[F^{-1}\{U_i(t_{i1})\}, \ldots, F^{-1}\{U_i(t_{in_i})\}\right]^{\mathrm{T}},$$

where $F^{-1}(\cdot)$ is the inverse CDF corresponding to the Poisson distribution with mean $E\{Y_i(t)|\mathbf{X}_i\}$ specified in (4.2) for each $i = 1, 2, \ldots, n$.

We performed the following steps on the complete data set to obtain the data with MAR mechanism, which resulted in approximately 40% missingness.

(a) For each of the (i, j)th observation, where $i = 1 \ldots, n$ and $j = 1, \ldots, n_i$, we randomly generated an observation b from a Bernoulli distribution with success probability $p_{ij} = \frac{\exp\{0.2X_{i1}+0.8X_{i2}+0.1X_{i3}\}}{1+\exp\{0.2X_{i1}+0.8X_{i2}+0.1X_{i3}\}}$.

(b) If $b = 0$, we considered the corresponding measurement as a missing value and deleted it from the data set, otherwise we retained all information on that observation.

For each of the above scenarios, the coefficient functions were locally estimated on equally spaced 200 grid points within the time interval [0, 1]. We used Epanechnikov kernel, given

by $K(t) = 0.75(1 - t^2)$ with $|t| \leq 1$, because it is shown to minimize the asymptotic mean-squared error of the local linear estimators (Fan and Gijbels, 1996). Estimates obtained by fitting a generalized linear model were treated as initial values of the regression coefficient functions in the Newton-Raphson (NR) iteration step during maximization of the local log-likelihood. For comparison purposes, we used the same bandwidth h for both complete (no missing observations) and MAR scenarios. To save time, we generated some pilot data sets and performed cross-validation on the complete data sets. We chose a value of h which was close to the optimal value obtained by fitting the pilot data. Specifically, we used $h = 0.075$ to fit all 400 data sets under the categorical predictor case.

The bias and the standard errors at three time points, namely, 0.3, 0.5, and 0.7, are presented in Table 4.1. Similar to the results in Figure 4.1, we observe that the bias is relatively small. For each time point, the standard deviation of the estimates across the 400 data sets are assumed to be the "true" standard deviation (denoted by SD) of the estimator at that time. A sandwich variance estimator (Carroll et al., 1998), presented in Section 4.3, is used for the estimation of the standard errors. The sample average and the sample standard deviation of the 400 estimated standard errors, denoted by SE and SD_{SE}, respectively, in Table 4.1, summarize the performance of the sandwich estimator under complete and MAR scenarios. From Table 4.1, we observe that the SDs as well as the SEs are slightly larger in MAR cases than that in the complete data case, which is expected due to 40% less observations in the MAR scenario. However, since all the SDs are within the interval $\text{SE} \pm 2\text{SD}_{\text{SE}}$, we can conclude that our estimators perform well under both complete

and MAR scenarios. Another point to note is that the SEs slightly underestimate the SDs, which is expected for sandwich estimators (Kürüm et al., 2016).

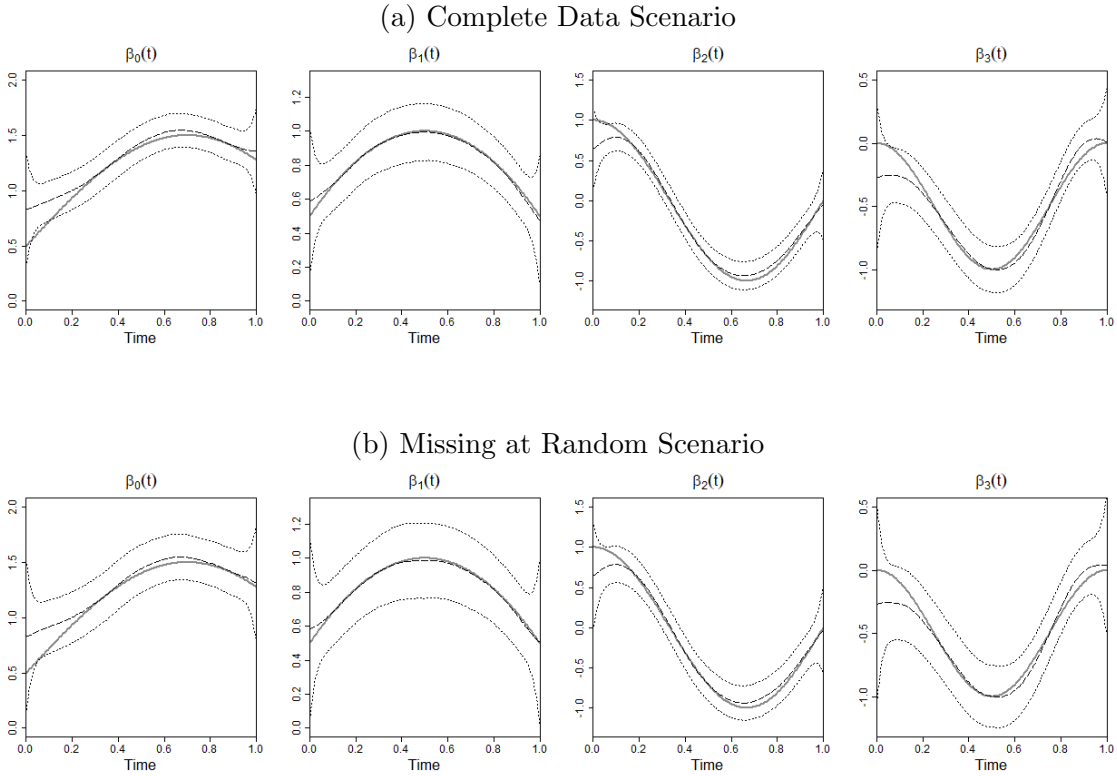


Figure 4.1: Results from our simulation study with time-invariant categorical predictors. Each row shows four plots for a given scenario. Each plot includes the true function (solid), the estimated varying-coefficient function (dashed), and the pointwise 95% sandwich confidence band (dotted).

Scenario	**Time**	Bias	SD	SE (SD_{SE})	Bias	SD	SE (SD_{SE})
		$\boldsymbol{\beta}_0(t)$			$\boldsymbol{\beta}_1(t)$		
Complete	0.3	-0.006	0.093	0.084 (0.011)	-0.001	0.092	0.081 (0.009)
	0.5	-0.035	0.090	0.080 (0.011)	0.006	0.087	0.084 (0.009)
	0.7	-0.043	0.097	0.076 (0.011)	0.004	0.096	0.079 (0.009)
MAR	0.3	-0.006	0.123	0.112 (0.018)	-0.001	0.122	0.104 (0.014)
	0.5	-0.041	0.125	0.108 (0.018)	0.016	0.111	0.110 (0.015)
	0.7	-0.045	0.125	0.103 (0.020)	0.008	0.121	0.106 (0.016)
		$\boldsymbol{\beta}_2(t)$			$\boldsymbol{\beta}_3(t)$		
Complete	0.3	-0.029	0.094	0.080 (0.008)	-0.015	0.102	0.098 (0.011)
	0.5	-0.001	0.098	0.084 (0.010)	-0.003	0.099	0.092 (0.012)
	0.7	-0.073	0.094	0.088 (0.010)	0.035	0.108	0.082 (0.009)
MAR	0.3	-0.031	0.114	0.101 (0.015)	-0.010	0.134	0.130 (0.020)
	0.5	-0.004	0.130	0.104 (0.016)	-0.002	0.137	0.123 (0.020)
	0.7	-0.074	0.121	0.109 (0.015)	0.040	0.149	0.110 (0.019)

Table 4.1: Bias and standard errors for the time-invariant categorical predictors in complete case and MAR scenarios.

4.5.2 Continuous Predictors

In this case, we considered two time-varying continuous predictors. For the ith subject, we randomly generated $n_i = 5$ irregular longitudinal time points from a standard uniform distribution, where $i = 1, \ldots, n$, with $n = 200$. For each time point, two predictors $X_{i1}(t)$ and $X_{i2}(t)$ were simulated independently from two Gaussian distributions with respective means 1 and 0, and standard deviations 0.25 and 0.1. The response variable $Y_i(t)$ was generated from a GVCM

$$E\{Y_i(t)|\mathbf{X}_i(t)\} = \exp\left\{\beta_0(t) + \beta_1(t)X_{i1}(t) + \beta_2(t)X_{i2}\right\}, \tag{4.3}$$

where $\mathbf{X}_i(t) = \{1, X_{i1}(t), X_{i2}(t)\}^{\mathrm{T}}$, $\beta_0(t) = 1 + \sin(2\pi t)$, $\beta_1(t) = 1 - \sin(2\pi t)$, and $\beta_2(t) = \cos(2\pi t)$. We applied the same correlation function and Gaussian copula technique discussed in Section 4.5.1 to generate the longitudinal outcome $Y_i(t)$ from a Poisson distribu-

tion having mean (4.3). The following steps performed on the complete data resulted in approximately 40% missingness.

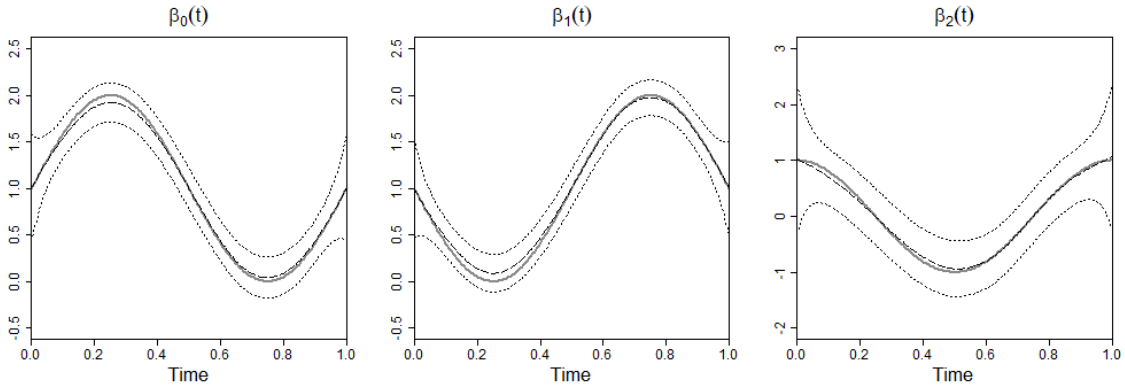
(i) For the jth observation of the ith subject, $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, we randomly generated an observation b from a Bernoulli distribution with success probability $p_{ij} = \frac{\exp\{0.4X_{i1}(t_{ij})+0.5X_{i2}(t_{ij})\}}{1+\exp\{0.4X_{i1}(t_{ij})+0.5X_{i2}(t_{ij})\}}$.

(ii) If $b = 0$, we considered the corresponding measurement as a missing value and deleted it from the data set, otherwise we retained all information on that measurement time.

Similar to the previous simulation set-up, we estimated the coefficients on 200 equally spaced grids within the time interval [0,1] and used Epanechnikov kernel. The bandwidth selected for fitting the models under both complete and MAR scenarios was $h = 0.12$.

The bias and the standard errors of the estimates are calculated in the same fashion described in Section 4.5.1 and presented in Table 4.2. Both Figure 4.2 and Table 4.2 indicate that the bias is relatively small and the shape of the regression functions are captured accurately. Similar to the categorical predictor case (Section 4.5.1), we observe that SDs are underestimated by the sandwich SEs, and are slightly higher in the MAR scenario in comparison to the complete data scenario. However, since the interval $\text{SE}\pm 2\text{SD}_{\text{SE}}$ covers the true SDs and the estimates have relatively small bias, we can conclude that GVCM performs well under both complete and MAR scenarios with continuous time-varying predictors.

(a) Complete Data Scenario



(b) Missing at Random Scenario

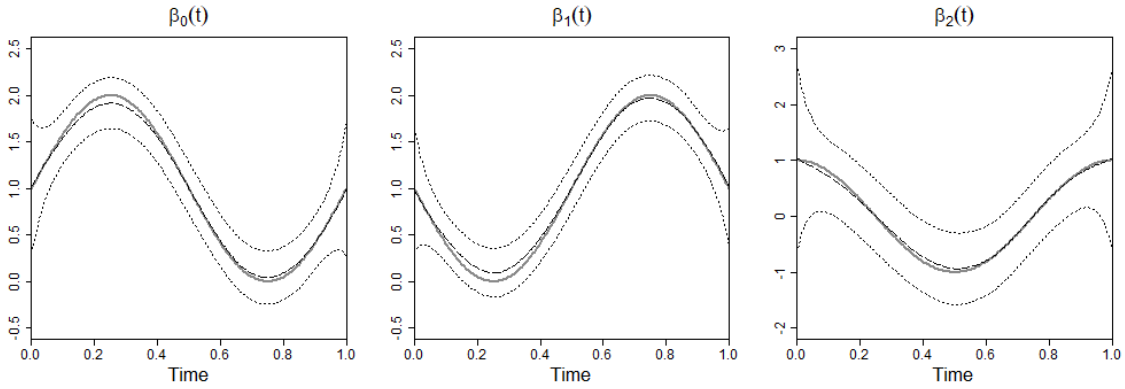


Figure 4.2: Results from our simulation study with time-varying continuous predictors. Each row shows three plots for a given scenario. Each plot includes the true function (solid), the estimated varying-coefficient function (dashed), and the pointwise 95% confidence band (dotted).

Case	**Time**	Bias	SD	SE (SD_{SE})	Bias	SD	SE (SD_{SE})
		$\boldsymbol{\beta}_0(t)$			$\boldsymbol{\beta}_1(t)$		
	0.3	0.069	0.121	0.105 (0.013)	-0.087	0.112	0.101 (0.012)
Complete	0.5	-0.001	0.125	0.109 (0.013)	-0.006	0.111	0.099 (0.013)
	0.7	-0.036	0.119	0.110 (0.014)	0.018	0.100	0.096 (0.014)
	0.3	0.076	0.155	0.137 (0.021)	-0.091	0.146	0.130 (0.020)
MAR	0.5	0.000	0.157	0.141 (0.020)	-0.007	0.140	0.128 (0.020)
	0.7	-0.036	0.158	0.143 (0.021)	0.023	0.135	0.123 (0.021)
		$\beta_2(t)$					
	0.3	-0.034	0.274	0.253 (0.031)			
Complete	0.5	-0.063	0.264	0.250 (0.029)			
	0.7	0.008	0.261	0.238 (0.030)			
	0.3	-0.049	0.351	0.326 (0.050)			
MAR	0.5	-0.062	0.329	0.320 (0.048)			
	0.7	-0.001	0.329	0.304 (0.049)			

Table 4.2: Bias and standard error for the time-varying continuous predictors in complete and MAR scenarios.

4.6 Analysis of the MIDUS Data

In this section, we present the results of the analysis of the Midlife in the United States (MIDUS) data (detailed description given in Section 4.2). The data consists of purpose in life (PIL) scores and age of the participants from three measurement times along with their social mobility groups (stable low, stable high, stable middle, downward, and upward), race, gender, attrition status, and six major later-life events: retirement, widowhood, chronic self-illness, and illness of children, spouse/partner, and parents. There was about 46% missing observations in the data. In Section 4.5, we have demonstrated that GVCM can produce accurate and efficient estimates under the MAR scenario. Under the assumption that the missingness in the MIDUS data is MAR, we employed GVCM on 4656 respondents to explore: (1) how purposefulness changes in midlife through old age

(specifically, between ages 40 and 80), (2) the role of later-life events in change in purpose over time, and (3) the disparity in purpose across different social mobility groups. The local estimation was done at 200 equally spaced grids between ages [40, 80] and the optimal bandwidth was 10.5.

In our analysis, we fitted the following GVCM assuming the PIL scores follow a Poisson distribution

$$E\{Y_i(t)|\mathbf{X}_i(t), \mathbf{W}_i\} = \exp\{\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}(t) + \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\alpha}(t)\},$$

where $\mathbf{X}_i^{\mathrm{T}}(t) = \{1, X_{i1}(t), \ldots, X_{i10}(t)\}$ and $\mathbf{W}_i^{\mathrm{T}} = (W_{i1}, \ldots, W_{i6})$ with corresponding regression coefficients $\boldsymbol{\beta}(t) = \{\beta_0(t), \beta_1(t), \ldots, \beta_{10}(t)\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}(t) = \{\alpha_1(t), \ldots, \alpha_6(t)\}^{\mathrm{T}}$, respectively. The vector of predictors $\mathbf{X}_i(t)$ consisted of binary variables for each social mobility group (stable low as the reference group) and six major life events. The predictors included in $\mathbf{W}_i$ were the control variables, that is, gender (male as the reference group), race (Caucasian as the reference group) and attrition status (binary variable for each group, that is, LFU after M1, died after M1, LFU after M2, and died after M2; participation in all three waves were used as the reference group). The overall reference group in our analysis was the group with all categorical predictors set to zero, that is, Caucasian men from stable low income group who participated in all three waves, who were not retired or widowed, and did not have any chronic illness of self and close family-members. To investigate our first goal, that is, to understand the age trajectory of PIL from midlife and beyond, we

present Figure 4.3. This figure represents the estimated average PIL trend across ages of the reference group. We observe that the PIL increases until age 57 and then decreases sharply between ages 57 and 80.

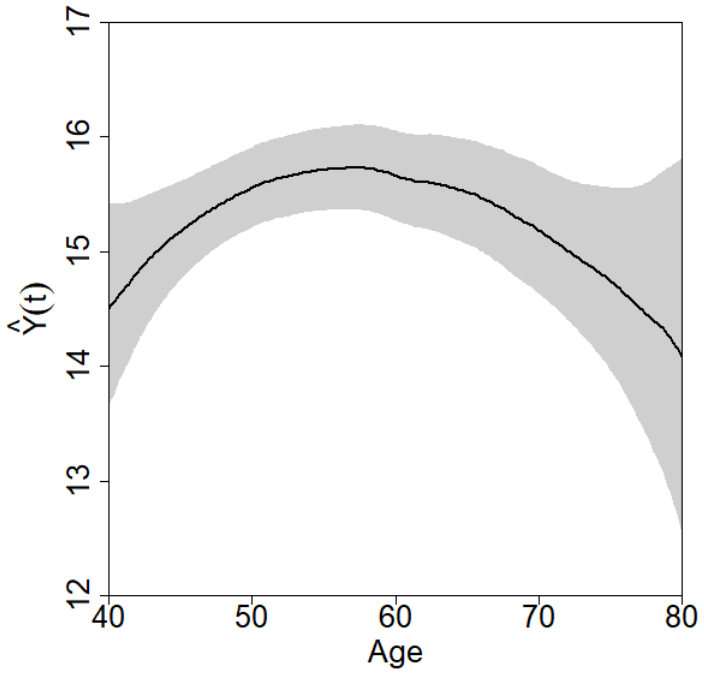


Figure 4.3: Estimated average trajectory of PIL scores as a function of age. The solid curve displays the mean PIL scores across ages for the reference group and the shaded region represents the 95% sandwich confidence band.

Figures 4.4 (a) and (b) present the estimated exponentiated coefficients along with the 95% confidence bands of the control variables gender and race, respectively. In these figures, as the confidence bands for both gender and race cover the straight line going through one, we conclude that these factors have no significant effect on PIL scores at any time point.

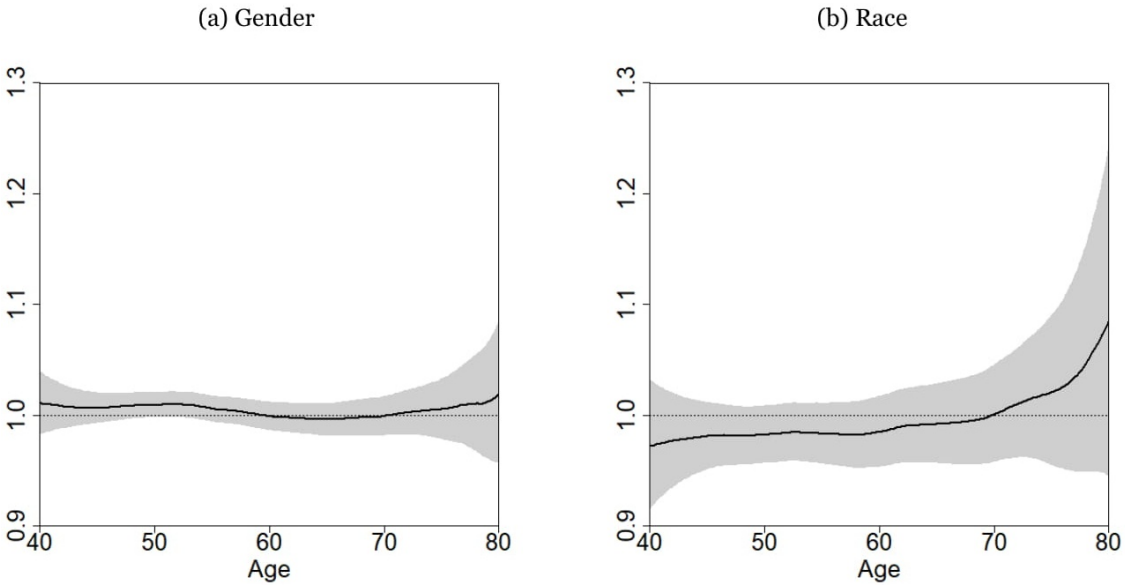


Figure 4.4: Estimated time-varying coefficients (solid) along with pointwise 95% sandwich confidence bands (shaded region) for the control variables gender and race.

To investigate our second goal, that is, how major later-life events influence the sense of purpose, we display the estimated coefficients of the six life events in Figures 4.5 (a)-(f). Figure 4.5 (d) shows a slight negative impact of serious self-illness during early midlife (between ages 44 to 50), and also between ages 70 and 75. Figure 4.5 (e) indicates that widowhood has a significantly negative impact on PIL scores between ages 57 and 80. According to Figures 4.5 (a)-(c), and (f), the rest of the life events do not have any effect on the purposeful thoughts during midlife to older adulthood.

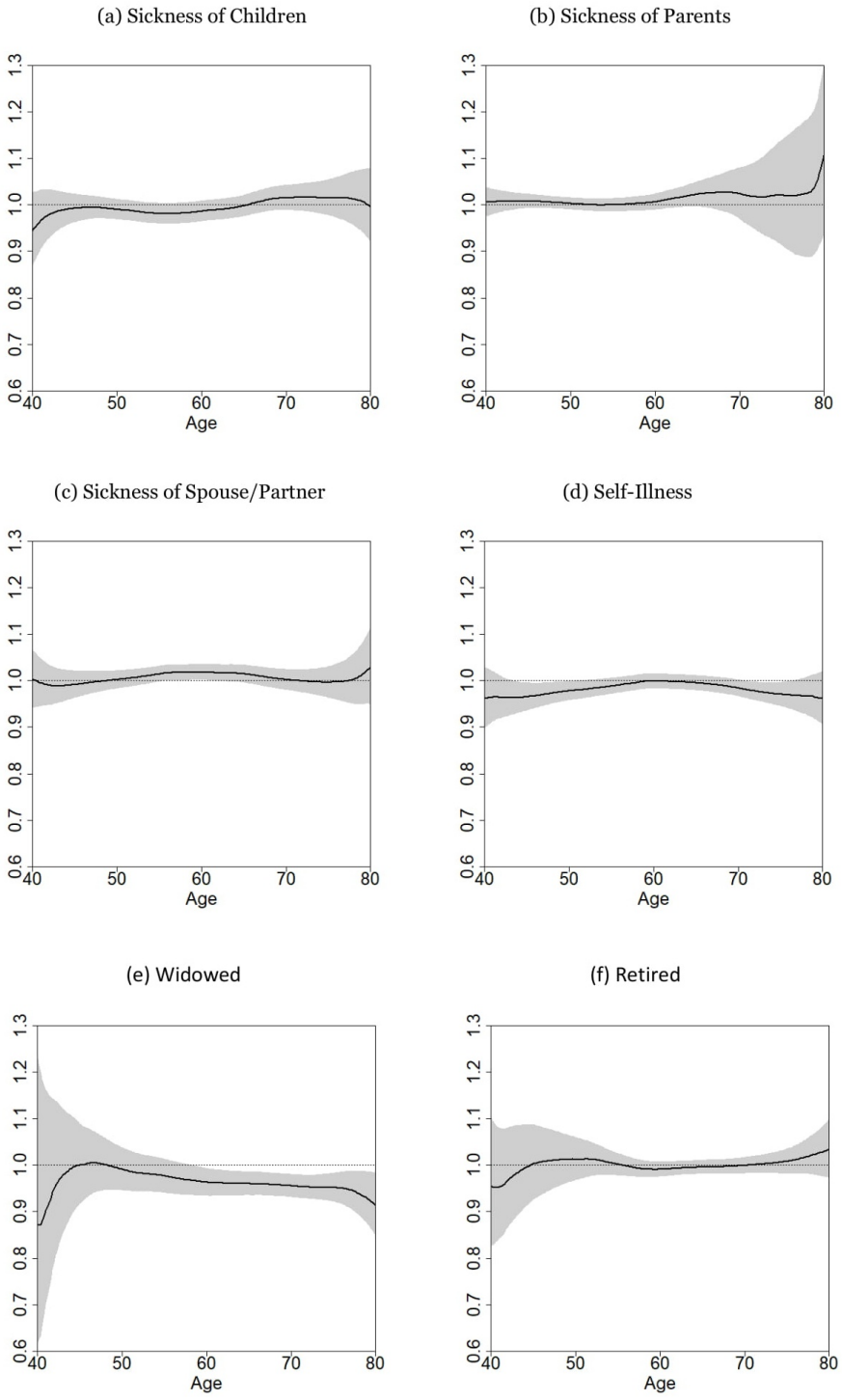


Figure 4.5: Estimated coefficient functions (solid) of later-life events, that is, sickness of self and close family members, widowhood, and retirement, along with their corresponding 95% confidence bands (shaded region).

To explore the effect of social mobility on PIL scores, that is, our third goal, we show the estimated time-varying coefficients in Figures 4.6 (a)-(d). The reference group consisted of the consistently socioeconomically disadvantaged participants, that is, individuals in the stable low social class. Since the straight line going through one was not covered by the confidence bands of the stable high, upward, stable middle, and downward groups, we can conclude from these figures that all of these social mobility groups have significantly higher PIL than the stable low socioeconomic class.

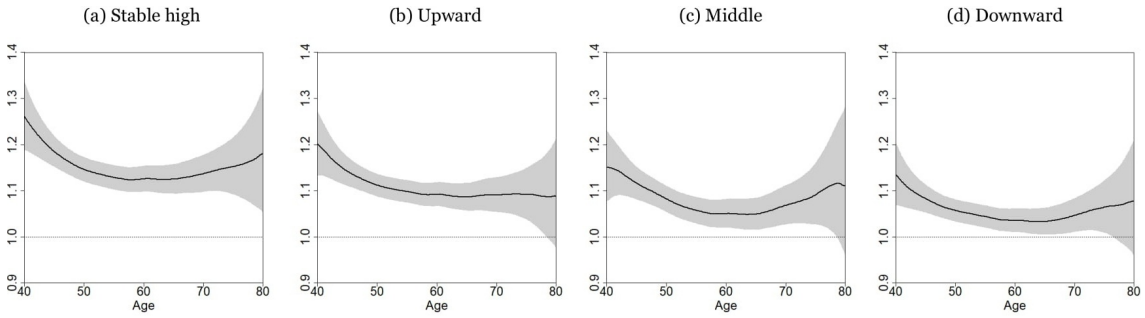


Figure 4.6: Estimated coefficients of the social mobility groups (solid) and the 95% confidence bands (shaded regions).

In order to further explore the relationships among the social mobility groups, we display the fitted PIL values for each of the five groups in Figures 4.7 (a) and (b). From Figure 4.7 (a), which includes the fitted PIL scores along with their sandwich confidence bands, we observe that subjects in the stable low group, that is, subjects who were born in the lower income group and remained in the same for their entire adulthood, have significantly lower purposefulness than the downward mobile group between ages 40 and 55. Additionally, their sense of purpose is significantly smaller than the rest of the three groups (stable middle, upward, and stable high) for almost the entire study window, specifically

between ages 40 and 75. Participants who consistently belonged to the advantaged or high social class (stable high group) display the highest sense of purpose throughout midlife to older adulthood. Their average PIL scores are significantly higher than the upwardly mobile groups between ages 42 and 65, these are higher than the stable middle group between ages 40 and 72, and the scores are larger than the downward mobile group between ages 40 and 76. The stable middle social class does not have significantly different purposefulness than the downward mobile group, but has slightly significant lower average scores than the upward group between ages 51 and 61. The downward group has much smaller sense of purpose than the upwardly mobile social class between ages 40 and 69. To understand the hierarchy and detect age-specific gradients among the social mobility groups clearly, we present Figure 4.7 (b) with only the fitted PIL scores. In this figure, we observe that the trend of PIL scores for the consistently disadvantaged group (stable low) increases sharply between ages 40 and 60, but declines after age 60. Contrary to this observation, all other groups show an overall declining trend in PIL scores throughout entire middle and old ages. The upward moving group has average PIL scores closer to the stable high group throughout midlife and early old ages, but the average score declines sharply after age 70. The stable middle group has mean PIL scores less than the upward group and higher than the downward mobile group. In summary, even if the sense of purpose significantly differs among the social mobility groups during midlife through early older ages, the disparity eventually diminishes with time, and the overlapping confidence bands in Figure 4.7 (a) indicate no difference after age 75.

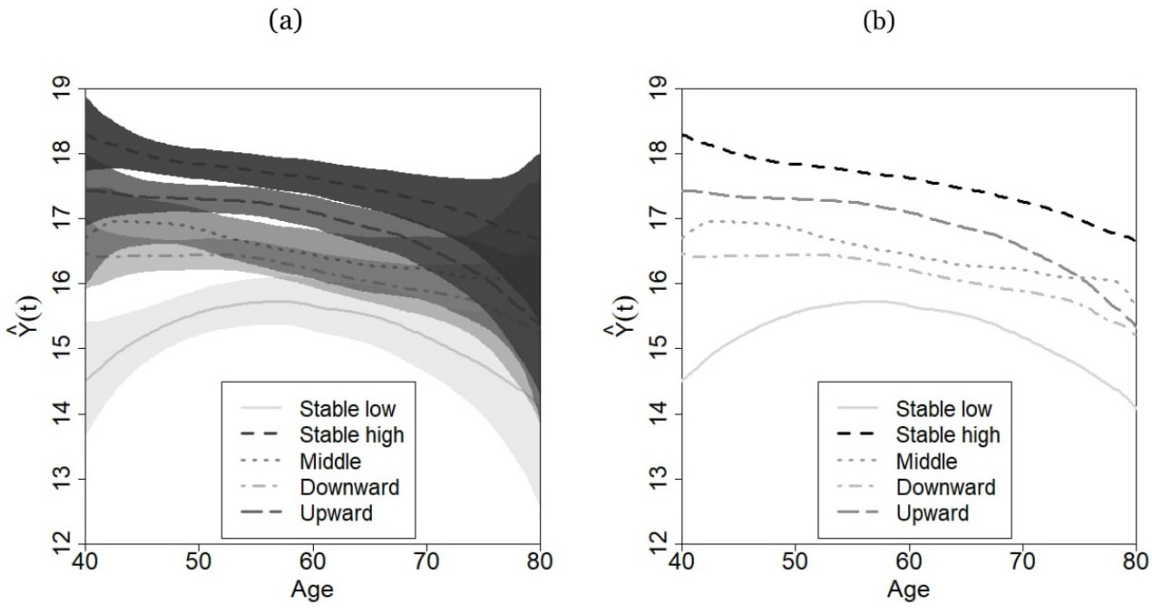


Figure 4.7: (a) Estimated PIL scores of all social mobility groups with their 95% confidence bands (shaded regions) showing the significance in difference among the groups. (b) Estimated PIL scores of all social mobility groups to display the hierarchy among the groups.

To summarize, in this chapter, we demonstrated that generalized varying coefficient models (Section 4.3) can lead to accurate and efficient estimates in the presence of MAR mechanism (Section 4.4, Section 4.5) and applied this method for the analysis of the MIDUS data (Section 4.6). Application of this methodology advances the literature in several ways. It allows researchers to examine the time-varying effects of various sociodemographic exploratory factors on the purpose in life scores without imposing any pre-specified response-predictor relationships. This may lead to better understanding of purpose in life at various stages of life and therefore, leading to insightful contributions to research on healthy aging. We found a declining trend in sense of purpose during later life. This is a critical concern for older individuals because high purposefulness is associated with reduced risk for health related problems and mortality, possibly because purposeful individuals are more

likely to seek preventive healthcare services. Our results show that some life events that deviate from normative life schedules might compromise sense of purpose. For example, losing a spouse diminished purposeful thoughts throughout the second half of life as this can result in loss of support, affection, and family-based recreational activities. Moreover, having serious chronic illnesses slightly lowered the levels of purposefulness in early midlife when such major health events are unusual in the general population, and during some of old ages when health conditions are strongly related to social isolation and fear of dying. Yet, throughout most of our study window, having serious illnesses did not significantly lower the levels of purpose, which suggests that older adults might be resilient in the face of health crises. There was no significant impact of having ailing family members on purposeful thought which might be due to two reasons. First, in terms of measurement issues, the yes/no indicator of each family member's chronic illness or disability during the limited period might not fully capture the complexity of health conditions. Second, caring for a family member might provide unique opportunities to strengthen family bonding and foster a sense of fulfillment. Our study uniquely contributes to the literature by investigating how the developmental trajectory of purposefulness from midlife to older ages is associated with life histories of socioeconomic background. We found that purposeful thought varies across social mobility groups, particularly before age 60, with the most advantaged group exhibiting the highest level and the most disadvantaged showing the lowest level. This might be due to the opportunities and accumulating advantages beginning at youth of the persons in the stable high social class, which act as significant sources of purpose. Meanwhile, those who grow up in disadvantaged families are often facing higher financial challenges and more

exposure to negative life events than their higher-status counterparts, which may impede their ability to strive toward purposeful life pursuits. Although, early-life socioeconomic circumstances are influential, we found that participants in the upwardly mobile group exhibited significantly higher levels of purpose than the stable low and downward groups. This indicates that exposure to childhood financial disadvantages might not always be an obstacle to cultivating purposeful thoughts. However, a steeper decline in the sense of purpose during older ages for this particular group compared to the stable high group indicates that upward mobility can be an arduous process with potentially negative consequences for psychological and physical well-being, and individuals in this group might struggle to maintain high levels of purpose at the end of their lives. Our results may direct researchers to expand intervention programs that help to foster purposefulness among children and youths from disadvantaged families.

Chapter 5

Conclusions and Future Work

Motivated by the Women's Interagency HIV Study (WIHS) data, we proposed a time-varying joint model (TV-JM) to fully capture the dynamic patterns present in a longitudinal data. Our method can accurately estimate both response-predictor and response-response relationships as flexible functions of time in a joint modeling framework. We developed an estimation procedure via the Expectation-Maximization algorithm, where in the E-step, we approximated the underlying random effects, and in the M-step, we used local linear regression techniques to estimate the time-varying coefficients. We investigated the finite sample performance of our proposed TV-JM estimators through extensive simulation studies and proposed standard error estimates by using bootstrap techniques. Note that while our method was motivated by the WIHS data, it can be applied to a range of follow-up studies which involve longitudinal and event-time responses.

In Chapter 4, we presented a novel application of generalized varying coefficient models (GVCM) through the analysis of the Midlife in the United States (MIDUS) data.

Since the data contains approximately 46% missing observations, we did a thorough literature search and conducted extensive simulation studies to investigate the performance of GVCM in presence of missing data. Our simulation results agreed with the literature review that under missing at random mechanism, GVCM accurately captures the time-varying shapes of the model parameters.

In addition to the work presented in this dissertation, some future research is needed in time-varying joint modeling of longitudinal and time-to-event responses:

1. In the present work, the longitudinal outcome is assumed to follow a normal distribution. However, in many situations, the repeated measurements may be binary or count data. A future direction may be extension of this model to a generalized setting for longitudinal measurements belonging to the canonical exponential family.

2. In this dissertation, we have considered a single longitudinal and a single event-time outcome. It would be of interest to extend these methods for multiple repeated measure outcomes or competing risk survival outcomes.

3. In addition to a fully nonparameteric approach, an automated semiparametric method can be postulated to distinguish between the time-varying versus the time-invariant coefficients, which might significantly increase the computational efficiency.

Bibliography

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., et al. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, 330(10):657–662.

Andersen, P. K. (1982). Testing goodness of fit of cox's regression and life model. *Biometrics*, pages 67–77.

Andrinopoulou, E.-R., Eilers, P. H., Takkenberg, J. J., and Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using p-splines. *Biometrics*, 74(2):685–693.

Avalos, B., Klein, J., Kapoor, N., Tutschka, P., Klein, J., and Copelan, E. (1993). Preparation for marrow transplantation in hodgkin's and non-hodgkin's lymphoma using bu/cy. *Bone marrow transplantation*, 12(2):133–138.

Barrett, J. and Su, L. (2017). Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Statistics in medicine*, 36(9):1447–1460.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, pages 89–99.

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.

Carroll, R. J., Ruppert, D., and Welsh, A. H. (1998). Local estimating equations. *Journal of the American Statistical Association*, 93(441):214–227.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991). Local regression models. *Statistical Models in S (Chambers, J. M. and Hastie, T. J., eds)*, pages 309–376.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, pages 1445–1462.

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, pages 1003–1014.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

Elashoff, R. M., Li, G., and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762–771.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J., Lin, H., Zhou, Y., et al. (2006). Local partial-likelihood estimation for lifetime data. *The Annals of Statistics*, 34(1):290–325.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC press.

Greenwood, M. (1926). The natural duration of cancer (report on public health and medical subjects no 33). *London: Stationery Office*.

Harrell Jr., F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hong, Y., Su, L., Song, S., and Yan, F. (2021). Dynamic prediction of disease processes based on recurrent history and functional principal component analysis of longitudinal biomarkers: Application for ovarian epithelial cancer. *Statistics in Medicine*, 40(8):2006–2023.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037–1043.

Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian journal of statistics*, 31(4):515–534.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Hyde, J. (1980). Survival analysis with incomplete observations. *Biostatistics casebook*, pages 31–46.

Johansen, S. (1983). An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, pages 165–174.

Kalish, L. A., McIntosh, K., Read, J. S., Diaz, C., Landesman, S. H., Pitt, J., Rich, K. C., Shearer, W. T., Davenny, K., and Lew, J. F. (1999). Evaluation of human immunodeficiency virus (hiv) type 1 load, cd4 t cell level, and clinical class as time-fixed and time-varying markers of disease progression in hiv-1—infected children. *The Journal of infectious diseases*, 180(5):1514–1520.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American*

Statistical Association, 84(407):717–726.

Kauermann, G. (2000). Modeling longitudinal data with ordinal response by varying coefficients. *Biometrics*, 56(3):692–698.

Klein, J. P. (1991). Small sample moments of some estimators of the variance of the kaplan-meier and nelson-aalen estimators. *Scandinavian Journal of Statistics*, pages 333–340.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kürüm, E., Hughes, J., and Li, R. (2015). A semivarying joint model for longitudinal binary and continuous outcomes. *Canadian Journal of Statistics*, 44(1):44–57.

Kürüm, E., Hughes, J., Li, R., and Shiffman, S. (2018). Time-varying copula models for longitudinal data. *Statistics and its interface*, 11(2):203.

Kürüm, E., Li, R., Shiffman, S., and Yao, W. (2016). Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. *Statistica Sinica*, 26(3).

Kürüm, E., Li, R., Wang, Y., and Şentürk, D. (2014). Nonlinear varying-coefficient models with applications to a photosynthesis study. *Journal of agricultural, biological, and environmental statistics*, 19(1):57–81.

Lagakos, S. W., Barraj, L. M., and Gruttola, V. D. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75(3):515–523.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Li, K. and Luo, S. (2017). Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease. *Statistics in medicine*, 36(22):3560–3572.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American statistical Association*, 95(450):520–534.

Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Martins, R. (2021). A flexible link for joint modelling longitudinal and survival data ac-

counting for individual longitudinal heterogeneity. *Statistical Methods & Applications*, pages 1–21.

Matthews, D. E. and Farewell, V. T. (1982). On testing for a constant hazard against a change-point alternative. *Biometrics*, pages 463–468.

Piulachs, X., Andrinopoulou, E.-R., Guillén, M., and Rizopoulos, D. (2021). A bayesian joint model for zero-inflated integers and left-truncated event times with a time-varying association: Applications to senior health care. *Statistics in Medicine*, 40(1):147–166.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, 35(9):1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D. and Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.

Ryff, C. D. and Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of personality and social psychology*, 69(4):719.

Salazar, A., Ojeda, B., Dueñas, M., Fernández, F., and Failde, I. (2016). Simple generalized estimating equations (gees) and weighted generalized estimating equations (wgees) in longitudinal studies with dropouts: guidelines and implementation in r. *Statistics in medicine*, 35(19):3424–3448.

Song, X. and Wang, C. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics*, 64(2):557–566.

Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S., and Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6):1277–1300.

Sweeting, M. J. and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834.

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(429):27–37.

Turnbull, B. W. and Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, pages 367–375.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556.

Verbeke, G. and Molenberghs, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 62–153. Springer.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Ye, J., Li, Y., Guan, Y., et al. (2015). Joint modeling of longitudinal drug using pattern and time to first relapse in cocaine dependence treatment data. *Annals of Applied Statistics*, 9(3):1621–1642.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, pages 689–699.

Zhang, W. and Lee, S.-Y. (2000). Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis*, 74(1):116–134.

Appendix A

Details of the Gauss-Hermite Quadrature Approximation

Gauss-Hermite quadrature method is employed to approximate integrals of the form $\int_{-\infty}^{+\infty} e^{-z^2} g(z)dz$ by

$$\int_{-\infty}^{+\infty} e^{-z^2} g(z)dz \approx \sum_{r=1}^{m} w_r g(z_r), \tag{A.1}$$

where m is the number of quadrature points used and z_r are the roots of the Hermite polynomial

$$H_m(z) = (-1)^m e^{z^2} \frac{d^m}{dz^m} e^{-z^2},$$

with the associated weights $w_r = (2^{m-1} m! \sqrt{\pi}) / \left[m^2 \{H_{m-1}(z_r)\}^2 \right]$ and $r = 1, \ldots, m$.

In our estimation procedure, we use the transformation $z^2 = \xi_i^2 \Big/ 2\sigma_\xi^2$ to obtain the form (A.1) required for the Gauss-Hermite quadrature method in the E-step of the Expectation-Maximization (EM) algorithm. In order to estimate the posterior mean and variance of the subject-specific random effect ξ_i, we apply the quadrature method on both

the numerators and denominators in main text equation (3.5). For $m = 35$, the transformed integrations and their approximations at the current EM iteration are updated as

$$\int L_i(\xi_i, \boldsymbol{\theta}^*) d\xi_i = \int \exp(-z^2) g_1^*(z) dz \approx \sum_{r=1}^{35} w_r g_1^*(z_r), \tag{A.2}$$

$$\int \xi_i L_i(\xi_i, \boldsymbol{\theta}^*) d\xi_i = \int \exp(-z^2) g_2^*(z) dz \approx \sum_{r=1}^{35} w_r g_2^*(z_r), \tag{A.3}$$

$$\int (\xi_i - \xi_{i0})^2 L_i(\xi_i, \boldsymbol{\theta}^*) d\xi_i = \int \exp(-z^2) g_3^*(z) dz \approx \sum_{r=1}^{35} w_r g_3^*(z_r), \tag{A.4}$$

where

$$\begin{aligned} g_1^*(z) &= \frac{1}{\sqrt{\pi}}\ g_0^*(z) \\ g_2^*(z) &= z\sigma_\xi^* \sqrt{2/\pi}\ g_0^*(z) \\ g_3^*(z) &= \frac{1}{\sqrt{\pi}}\ (z\sigma_\xi^* \sqrt{2} - \xi_{i0})^2\ g_0^*(z), \end{aligned}$$

with

$$\begin{aligned} g_0^*(z) \quad = \quad & \left[\prod_{j=1}^{n_i} \{2\pi\sigma^{2*}(t_{ij})\}^{-1/2} \right] \exp\left[-\sum_{j=1}^{n_i} \frac{\left\{ Y_i(t_{ij}) - \mathbf{X}_i^{\mathrm{T}}(t_{ij})\boldsymbol{\beta}^*(t_{ij}) - z\sigma_\xi^*\sqrt{2} \right\}^2}{2\sigma^{2*}(t_{ij})} \right] \\ & \times \left(h_0^*(T_i) \exp\left[\left\{ \mathbf{X}_i^{\mathrm{T}}(T_i)\boldsymbol{\beta}^*(T_i) + z\sigma_\xi^*\sqrt{2} \right\} \gamma^*(T_i) + \mathbf{W}_i^{\mathrm{T}}(T_i)\boldsymbol{\eta}^*(T_i) \right] \right)^{\delta_i} \\ & \times \exp\left(-\int_0^{T_i} h_0^*(u) \exp\left[\left\{ \mathbf{X}_i^{\mathrm{T}}(u)\boldsymbol{\beta}^*(u) + z\sigma_\xi^*\sqrt{2} \right\} \gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u) \right] du \right), \end{aligned}$$

and c^* denotes the current estimate of any quantity c.

To estimate the posterior mean of ξ_i, that is, ξ_{i0}, approximations in (A.2) and (A.3) are computed and plugged into the first equation of (3.5) in the main text. The estimate of the posterior variance v_{i0} is obtained via first plugging in the estimated ξ_{i0} in (A.4) and then using the approximations in (A.2) and (A.4) in the second equation of (3.5) in the main text.

Appendix B

Details of the Maximization Step

In this section, we present further details on the Newton-Raphson algorithms employed to estimate the time-varying parameters in $\boldsymbol{\alpha}(t) = \{\boldsymbol{\beta}(t)^{\mathrm{T}}, \gamma(t), \boldsymbol{\eta}(t)^{\mathrm{T}}\}^{\mathrm{T}}$ and the parameters of the baseline hazard function. In addition, we present the likelihood-based standard error formula for the random effect variance.

B.1 Estimation of Time-Varying Parameters

In order to estimate the time-varying parameters $(\hat{\boldsymbol{\alpha}}_0^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}_1^{\mathrm{T}})^{\mathrm{T}}$, we employed local linear fitting techniques, in which, we maximized the expected local log-likelihood in main text equation (3.7) with respect to $\boldsymbol{\alpha}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \gamma_0, \boldsymbol{\eta}_0^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \gamma_1, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$ via a Newton-Raphson (NR) algorithm. Let $\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\}$ be the estimate of $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ at the current iteration of the NR algorithm and we updated $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$ according to

$$\left\{ \begin{array}{c} \boldsymbol{\alpha}_0^{(it+1)} \\ \boldsymbol{\alpha}_1^{(it+1)} \end{array} \right\} = \left\{ \begin{array}{c} \boldsymbol{\alpha}_0^{(it)} \\ \boldsymbol{\alpha}_1^{(it)} \end{array} \right\} - [\ell''\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\}]^{-1} \ell'\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\},$$

where $\ell'\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\}$ and $\ell''\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\}$ are the score function and Hessian of the approximated expected local log-likelihood (3.7) with respect to $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$, respectively, evaluated at the current estimates $\{\boldsymbol{\alpha}_0^{(it)}, \boldsymbol{\alpha}_1^{(it)}\}$. Let $\mathcal{A}_i^*(z) = m_{il}^*(z)\{\gamma_0^* + \gamma_1^*(z - t_0)\} + \mathbf{W}_i^{\mathrm{T}}(z)\{\eta_0^* + \eta_1^*(z - t_0)\}$ with $m_{il}^*(z) = \mathbf{X}_i^{\mathrm{T}}(z)\{\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_1^*(z - t_0)\} + \xi_{i0}^*$, where for any c, the estimate at the current iteration is denoted by c^*. The score functions and the hessian at the current NR iteration are presented below.

Score Functions

$$\begin{aligned}
\frac{\partial E\{\ell(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\beta}_0} &= \sum_{i=1}^{n} \Bigg(\sum_{j=1}^{n_i} \Big[\{\sigma^{2*}(t_{ij})\}^{-1} \mathbf{X}_i(t_{ij}) \big\{ Y_i(t_{ij}) - m_{il}^*(t_{ij}) \big\} \Big] K_{h_1}(t_{ij} - t_0) \\
&\quad + \Bigg[\delta_i \mathbf{X}_i(T_i) \left\{ \gamma_0^* + \gamma_1^*(T_i - t_0) \right\} \\
&\quad - \int_0^{T_i} h_0^*(u) \exp\left\{ \mathcal{A}_i^*(u) \right\} \mathbf{X}_i(u) \{\gamma_0^* + \gamma_1^*(u - t_0)\} \, du \Bigg] K_{h_2}(T_i - t_0) \Bigg),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial E\{\ell(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\beta}_1} &= \sum_{i=1}^{n} \Bigg(\sum_{j=1}^{n_i} \Big[\{\sigma^{2*}(t_{ij})\}^{-1} \mathbf{X}_i(t_{ij})(t_{ij} - t_0) \big\{ Y_i(t_{ij}) - m_{il}^*(t_{ij}) \big\} \Big] \\
&\quad \times K_{h_1}(t_{ij} - t_0) + \Bigg[\delta_i \mathbf{X}_i(T_i)(T_i - t_0) \big\{ \gamma_0^* + \gamma_1^*(T_i - t_0) \big\} \\
&\quad - \int_0^{T_i} h_0^*(u) \exp\left\{ \mathcal{A}_i^*(u) \right\} \mathbf{X}_i(u)(u - t_0) \big\{ \gamma_0^* + \gamma_1^*(u - t_0) \big\} \, du \Bigg] \\
&\quad \times K_{h_2}(T_i - t_0) \Bigg),
\end{aligned}$$

$$\frac{\partial E\{\ell(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\}}{\partial \gamma_0} = \sum_{i=1}^{n} \left(\delta_i m_{il}^*(T_i) - \int_0^{T_i} h_0^*(u) \exp\left\{ \mathcal{A}_i^*(u) \right\} m_{il}^*(u) \, du \right) K_{h_2}(T_i - t_0),$$

$$\frac{\partial E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial \gamma_1} = \sum_{i=1}^{n}\Bigg(\delta_i m^*_{il}(T_i)(T_i-t_0) - \int_0^{T_i} h^*_0(u)\exp\left\{\mathcal{A}^*_i(u)\right\}m^*_{il}(u)(u-t_0)\,du\Bigg)K_{h_2}(T_i-t_0),$$

$$\frac{\partial E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\eta}_0} = \sum_{i=1}^{n}\left(\delta_i \mathbf{W}_i(T_i) - \int_0^{T_i} h^*_0(u)\exp\left\{\mathcal{A}^*_i(u)\right\}\mathbf{W}_i(u)\,du\right)K_{h_2}(T_i-t_0),$$

$$\frac{\partial E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\eta}_1} = \sum_{i=1}^{n}\Bigg(\delta_i \mathbf{W}_i(T_i)(T_i-t_0) - \int_0^{T_i} h^*_0(u)\exp\left\{\mathcal{A}^*_i(u)\right\}(u-t_0)\mathbf{W}_i(u)\,du\Bigg)K_{h_2}(T_i-t_0).$$

Submatrices of the Upper-triangular Hessian Matrix

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0^{\mathrm{T}}} = -\sum_{i=1}^{n}\Bigg(\sum_{j=1}^{n_i}\{\sigma^{2*}(t_{ij})\}^{-1}\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})K_{h_1}(t_{ij}-t_0) + \Bigg[\int_0^{T_i} h^*_0(u)\exp\left\{\mathcal{A}^*_i(u)\right\}\mathbf{X}_i(u)\mathbf{X}_i^{\mathrm{T}}(u)\left\{\gamma^*_0+\gamma^*_1(u-t_0)\right\}^2 du\Bigg] \times K_{h_2}(T_i-t_0)\Bigg),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\Bigg(\sum_{j=1}^{n_i}\{\sigma^{2*}(t_{ij})\}^{-1}\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})(t_{ij}-t_0)^2K_{h_1}(t_{ij}-t_0) + \Bigg[\int_0^{T_i} h^*_0(u)\exp\left\{\mathcal{A}^*_i(u)\right\}\mathbf{X}_i(u)\mathbf{X}_i^{\mathrm{T}}(u)(u-t_0)^2 \times\left\{\gamma^*_0+\gamma^*_1(u-t_0)\right\}^2 du\Bigg]K_{h_2}(T_i-t_0)\Bigg),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\beta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\Bigg(\sum_{j=1}^{n_i}\{\sigma^{2*}(t_{ij})\}^{-1}\mathbf{X}_i(t_{ij})\mathbf{X}_i^{\mathrm{T}}(t_{ij})(t_{ij}-t_0)K_{h_1}(t_{ij}-t_0) + \Bigg[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}\mathbf{X}_i(u)\mathbf{X}_i^{\mathrm{T}}(u)(u-t_0) \times\left\{\gamma_0^*+\gamma_1^*(u-t_0)\right\}^2 du\Bigg]K_{h_2}(T_i-t_0)\Bigg),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0^2} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^{*2}(u)\,du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_1^2} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^{*2}(u)(u-t_0)^2\,du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0\partial\gamma_1} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^{*2}(u)(u-t_0)\,du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\eta}_0\partial\boldsymbol{\eta}_0^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}\mathbf{W}_i(u)\mathbf{W}_i^{\mathrm{T}}(u)\,du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\eta}_1\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}\mathbf{W}_i(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0)^2\,du\right] \times K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\eta}_0\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}\mathbf{W}_i(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0)\,du\right] \times K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\gamma_0} = \sum_{i=1}^{n}\Bigg(\delta_i\mathbf{X}_i(T_i) - \int_0^{T_i} h_0^*(u)\exp\big\{\mathcal{A}_i^*(u)\big\}\mathbf{X}_i(u) \times\Big[1+m_{il}^*(u)\big\{\gamma_0^*+\gamma_1^*(u-t_0)\big\}\Big]\,du\Bigg)K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\gamma_1} = \sum_{i=1}^{n}\Bigg(\delta_i\mathbf{X}_i(T_i)(T_i-t_0) - \int_0^{T_i} h_0^*(u)\exp\big\{\mathcal{A}_i^*(u)\big\}\mathbf{X}_i(u)(u-t_0) \times\Big[1+m_{il}^*(u)\big\{\gamma_0^*+\gamma_1^*(u-t_0)\big\}\Big]\,du\Bigg)K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0\partial\boldsymbol{\beta}_1^{\mathrm{T}}} = \left[\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\gamma_1}\right]^{\mathrm{T}},$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_1\partial\gamma_1} = \sum_{i=1}^{n}\Bigg(\delta_i\mathbf{X}_i(T_i)(T_i-t_0)^2 - \int_0^{T_i} h_0^*(u)\exp\big\{\mathcal{A}_i^*(u)\big\}\mathbf{X}_i(u)(u-t_0)^2 \times\Big[1+m_{il}^*(u)\big\{\gamma_0^*+\gamma_1^*(u-t_0)\big\}\Big]\,du\Bigg)K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\eta}_0^{\mathrm{T}}} = -\sum_{i=1}^{n}\Bigg[\int_0^{T_i} h_0^*(u)\exp\big\{\mathcal{A}_i^*(u)\big\}\mathbf{X}_i(u)\mathbf{W}_i^{\mathrm{T}}(u) \times\big\{\gamma_0^*+\gamma_1^*(u-t_0)\big\}\,du\Bigg]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\Bigg[\int_0^{T_i} h_0^*(u)\exp\big\{\mathcal{A}_i^*(u)\big\}\mathbf{X}_i(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0) \times\big\{\gamma_0^*+\gamma_1^*(u-t_0)\big\}\,du\Bigg]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\eta}_0\partial\boldsymbol{\beta}_1^{\mathrm{T}}} = \left[\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_0\partial\boldsymbol{\eta}_1^{\mathrm{T}}}\right]^{\mathrm{T}},$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}\mathbf{X}_i(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0)^2 \times\left\{\gamma_0^*+\gamma_1^*(u-t_0)\right\}du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0\partial\boldsymbol{\eta}_0^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^*(u)\mathbf{W}_i^{\mathrm{T}}(u)\,du\right]K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^*(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0)\,du\right] \times K_{h_2}(T_i-t_0),$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\boldsymbol{\eta}_0\partial\gamma_1} = \left[\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_0\partial\boldsymbol{\eta}_1^{\mathrm{T}}}\right]^{\mathrm{T}},$$

$$\frac{\partial^2 E\{\ell(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\}}{\partial\gamma_1\partial\boldsymbol{\eta}_1^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\exp\left\{\mathcal{A}_i^*(u)\right\}m_{il}^*(u)\mathbf{W}_i^{\mathrm{T}}(u)(u-t_0)^2\,du\right] \times K_{h_2}(T_i-t_0).$$

B.2 Estimation of the Baseline Hazard Parameter

For the estimation of the baseline hazard parameter $\boldsymbol{\theta}_{h_0}$, we maximized the approximated expected log-likelihood in main text equation (3.6) with respect to $\boldsymbol{\theta}_{h_0}$. Maximization is implemented via a Newton-Raphson algorithm and the updated estimator is obtained by $\boldsymbol{\theta}_{h_0}^{(r+1)} = \boldsymbol{\theta}_{h_0}^{(r)} - \{\mathcal{H}\ell_{h_0}^{(r)}\}^{-1}\mathbf{V}_{h_0}^{(r)}$, where r is the current iteration of the Newton-Raphson algorithm, $\mathbf{V}_{h_0}^{(r)}$ and $\mathcal{H}\ell_{h_0}^{(r)}$ are the score function and the Hessian of the approximated ex-

pected log-likelihood (3.6) with respect to $\boldsymbol{\theta}_{h_0}$. The respective score and the Hessian for the baseline hazard can be calculated as

$$\begin{aligned}\frac{\partial E\{\ell(\boldsymbol{\xi},\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}_{h_0}} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_{h_0}} \Bigg(\delta_i \log h_0^*(T_i) \\ &\qquad - \int_0^{T_i} h_0^*(u) \exp\{m_i^*(u)\gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u)\} du \Bigg), \text{ and}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 E\{\ell(\boldsymbol{\xi},\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}_{h_0} \partial \boldsymbol{\theta}_{h_0}^{\mathrm{T}}} &= \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}_{h_0} \partial \boldsymbol{\theta}_{h_0}^{\mathrm{T}}} \Bigg(\delta_i \log h_0^*(T_i) \\ &\qquad - \int_0^{T_i} h_0^*(u) \exp\{m_i^*(u)\gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u)\} du \Bigg).\end{aligned}$$

In our modeling framework, we employed a restricted cubic spline function to model the baseline hazard, which is given by

$$h_0(t) = \exp\left\{ \sum_{\kappa=1}^{K-2} \varphi_\kappa \omega_\kappa(t) + \varphi_{K-1} t + \varphi_K \right\}.$$

In vector notation, the baseline can be written as $h_0(t) = \exp\left\{\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\omega}(t)\right\}$, where $\boldsymbol{\varphi}^{\mathrm{T}} = \{\varphi_1, \ldots, \varphi_K\}$ constructs the vector of baseline hazard parameter $\boldsymbol{\theta}_{h_0}$, $\boldsymbol{\omega}(t) = \{\omega_1(t), \ldots, \omega_{K-2}(t),\ t,\ 1\}^{\mathrm{T}}$ with $\omega_\kappa(t) = (t-\vartheta_\kappa)_+^3 - \frac{(t-\vartheta_{K-1})_+^3(\vartheta_K-\vartheta_\kappa)}{(\vartheta_K-\vartheta_{K-1})} + \frac{(t-\vartheta_K)_+^3(\vartheta_{K-1}-\vartheta_\kappa)}{(\vartheta_K-\vartheta_{K-1})}$, for $\kappa = 1, \ldots, (K-2)$, and $(z)_+ = \max(0, z)$.

We investigated two more approaches to estimate the baseline hazard function: piecewise constant and linear spline. Note that although these approaches are not as flexible as the restricted cubic spline, they are more computationally straightforward to implement. The simplest way to model the baseline hazard function is using a piecewise constant. Under this approach, the baseline hazard is given by

$$h_0(t) = \exp\left\{ \sum_{\kappa=1}^{K} \varphi_\kappa . I(\vartheta_\kappa < t \leq \vartheta_{\kappa+1}) \right\},$$

where the time window is divided into K sub-intervals at time points $0 = \vartheta_1 < \vartheta_2 < \cdots < \vartheta_{K+1}$, such that, ϑ_{K+1} is larger than the largest observed time and φ_κ is the value of the baseline hazard in the interval $(\vartheta_\kappa, \vartheta_{\kappa+1}]$. The exponential in this model ensures that the baseline hazard is positive at all times. In vector notation, we can write the baseline as $h_0(t) = \exp\left\{\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\omega}(t)\right\}$, where $\boldsymbol{\omega}(t) = \{\omega_1(t), \ldots, \omega_K(t)\}^{\mathrm{T}}$ with $\omega_\kappa(t) = I(\varphi_\kappa < t < \varphi_{\kappa+1})$, for $\kappa = 1, \ldots, K$, and $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_K\}^{\mathrm{T}} = \boldsymbol{\theta}_{h_0}$.

Another method that can be used to model the log-baseline is given by a linear spline. The baseline risk here is given by the form

$$h_0(t) = \exp\left\{\varphi_1 + \varphi_2 t + \sum_{\kappa=3}^{K+2} \varphi_\kappa (t - \vartheta_{\kappa-2})_+\right\},$$

where $\boldsymbol{\varphi}^{\mathrm{T}} = \{\varphi_1, \ldots, \varphi_{K+2}\}$ is the vector of baseline hazard parameters $\boldsymbol{\theta}_{h_0}$, and the internal knots are given by $0 < \vartheta_1 < \cdots < \vartheta_K < \max(T_i)$. In vector notation, the baseline can be written as $h_0(t) = \exp\left\{\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\omega}(t)\right\}$, where $\boldsymbol{\omega}(t) = \{1, t, (t-\vartheta_1)_+, \ldots, (t-\vartheta_K)_+\}^{\mathrm{T}}$. Note that, since we can write all three methods mentioned above using the same vector notation, we can provide a general form of the score and the hessian for all three cases. With appropriate specification of $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}(\cdot)$ in the corresponding models, the score function and the Hessian can be presented as follows

$$\frac{\partial E\{\ell(\boldsymbol{\xi}, \boldsymbol{\theta})\}}{\partial \boldsymbol{\varphi}} = \sum_{i=1}^{n}\left[\delta_i \boldsymbol{\omega}(T_i) - \int_0^{T_i} h_0^*(u)\boldsymbol{\omega}(u)\exp\left\{m_i^*(u)\gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u)\right\}du\right],$$

$$\frac{\partial E\{\ell(\boldsymbol{\xi}, \boldsymbol{\theta})\}}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^{\mathrm{T}}} = -\sum_{i=1}^{n}\left[\int_0^{T_i} h_0^*(u)\boldsymbol{\omega}(u)\boldsymbol{\omega}(u)^{\mathrm{T}}\exp\left\{m_i^*(u)\gamma^*(u) + \mathbf{W}_i^{\mathrm{T}}(u)\boldsymbol{\eta}^*(u)\right\}du\right].$$

B.3 Likelihood-based Standard Error of the Random Effect Variance

To compute the likelihood-based standard error for the random effect variance σ_ξ^2, we calculate the second order derivative of the incomplete local log-likelihood in main text equation (3.4) and it has a closed form solution as follows

$$\begin{aligned}
\left.\frac{\partial^2 \ell(\boldsymbol{\theta})}{(\partial \sigma_\xi^2)^2}\right|_{\sigma_\xi^2=\widehat{\sigma}_\xi^2,\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} &= \left.\sum_{i=1}^{n} \frac{\partial^2}{(\partial \sigma_\xi^2)^2} \log\left\{\int L_i(\xi_i,\boldsymbol{\theta}) d\xi_i\right\}\right|_{\sigma_\xi^2=\widehat{\sigma}_\xi^2,\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\
&= \frac{n}{2\widehat{\sigma}_\xi^4} - \sum_{i=1}^{n} \frac{(\widehat{\xi}_{i0}^2 + \widehat{v}_{i0})}{\widehat{\sigma}_\xi^6} \\
&= \mathcal{H}_\xi(\widehat{\sigma}_\xi^2),
\end{aligned}$$

where $\widehat{c}$ denotes the estimate of c at the last EM iteration. The standard error is given by $\sqrt{-\left\{\mathcal{H}_\xi(\widehat{\sigma}_\xi^2)\right\}^{-1}}$.