The Pennsylvania State University
The Graduate School

JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA: NEW MODELS, COMPUTING TECHNIQUES AND APPLICATIONS

A Dissertation in
Statistics
by
Xiaoyu Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2014

The dissertation of Xiaoyu Liu was reviewed and approved* by the following:

Runze Li
Distinguished Professor of Statistics
Dissertation Advisor
Chair of Committee

David Hunter
Professor of Statistics

Zhibiao Zhao
Associate Professor of Statistics

Stephanie T. Lanza
Research Associate Professor of Health and Human Development

Aleksandra B. Slavkovic
Associate Professor of Statistics
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

Motivated from an empirical analysis of data collected by a smoking cessation study, this dissertation studies the methodology, computation and application of joint modeling of longitudinal and survival data, and extends the existing modeling framework to several new settings.

Firstly, we propose a joint model (JM) of survival data and multiple continuous longitudinal covariates, develop an estimation procedure using likelihood-based approach, and further establish the consistency and asymptotic normality of the resulting estimate. Computation for the proposed likelihood-based approach in the joint modeling framework is particularly challenging since the estimation procedure involves numerical integration over multi-dimensional space for the random effects. Existing numerical integration methods become ineffective or infeasible for JM. We introduce a numerical integration method based on computer experimental design for JM. We conduct Monte Carlo simulations to examine the finite sample performance of the proposed procedure and compare the new numerical integration method with the existing ones. We further illustrate the proposed procedure via an empirical study of smoking cessation data.

Secondly, we propose a general nonparametric JM to incorporate both the time-varying survival coefficients and the longitudinal process with an irregular trajectory. Such a model is more flexible than the existing parametric joint models, and requires more powerful computational capability. We employ B-splines to approximate the functional parameters and use a maximum joint likelihood approach for parameter estimation. The estimates are calculated by the newly introduced computing algorithm, the EM-DoIt algorithm, and simulation studies are conducted to demonstrate the feasibility of the proposed estimation and computing procedures. The proposed model is applied to a smoking cessation study to explore the dynamic structure of the longitudinal process and the possible time-varying relationships between the negative affect and time to lapse.

Finally, we propose a JM with discrete longitudinal covariates, which can also be fitted using the maximum joint likelihood approach, and implemented via the EM-DoIt algorithm. We conduct a few numerical studies to test the capability of the proposed approach in handling some specific types of longitudinal covariates, such as binary, count, and zero inflated discrete covariates.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my thesis advisor Dr. Runze Li, who has been a great mentor for me in both my research and my career development. Throughout my PhD study, Dr. Li has supported and encouraged me with his sharp ideas, helpful advices, and tremendous enthusiasm in research. This thesis would not have been possible without his guidance. I appreciate all his contribution, time and patience in guiding me and helping me to build up my confidence for research.

I would also like to thank my other committee members: Dr. David Hunter, Dr. Stephanie Lanza, and Dr. Zhibiao Zhao for their insightful comments and kind support. In particular, I have learnt a lot from Dr. Lanza about the application of my research to the field of public health. The applied projects I worked with her has stimulated my interest in the interdisciplinary study between statistics and social science. I am thankful for her valuable inputs.

In addition, I appreciate my home department, the Statistics Department at Penn State University, and the PhD program for not only providing me with the great opportunity for graduate study, but also offers a free and friendly atmosphere for me to learn, to communicate and to develop.

Last but not least, I would like to thank my parents: Mr. Mingfang Liu and Ms. Cuiping Yang for their endless support, encouragement and love during the past four years. I would like to give special thanks to my husband: Mr. Kai Yang, who has consistently taken care of me and helped me with his knowledge and love when I encountered difficulties.

This dissertation research is supported by NIDA, NIH grants P50 DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

Chapter 1

Introduction

1.1 A brief introduction of joint modeling

In clinical trials and medical studies, longitudinal and survival data are often collected together. For example, repeated measurements are usually recorded for participants until certain event of interest happens. In these studies the relationship between covariate processes and time-to-event is often the primary research interest. A famous example in pioneer work is the CD4 count data of AIDs patients studied by Tsiatis et al. (1995), in which the goal is to find out whether CD4 count, a reflection of immune status, can be taken as a surrogate biomarker for an active treatment of AIDs, or in other words, whether CD4 count has significant confounding effect with the active treatment on survival time. In this study, CD4 count was recorded periodically and repeatedly for the patients until their death, and thus has multiple measurements for each individual across time. On the other hand, the response, time-to-death, has only one record for each subject. Hence, the challenge occurs in modeling the relationship between two variables of different observational frequencies.

Intensive longitudinal data are a set of repeated measurements collected on a group of participants at subject-dependent time points over a study period. They can be conveniently collected via new technologies such as smart phones that record the instant measurements and thus facilitate the study of continuous covariate process with respect to both within-subject variation and between-subject heterogeneity. Unlike the traditional independent and identically distributed (iid) observations, longitudinal data have within-subject correlations. Hence they cannot be simply treated by linear regression as for iid data. Methodologies such as linear mixed-effects models and hierarchical linear modeling (HLM) (Raudenbush, 2002) have been developed for these data. These models and the extended methodologies for modeling longitudinal data are reviewed in detail in Chapter 2.

Survival data, distinct from longitudinal data, collect information regarding time-to-event and occurrence of censoring. Survival analysis has been well studied for decades to model the time-to-event data. Popular survival models such as the Cox model (Cox, 1972) and accelerated life model (Cox and Oakes, 1984) have been extensively applied to predict the risk of failure based on a series of subject-specific covariates, and meanwhile take into account the censoring due to death or dropouts. In this study we focus on the Cox model and its extension to incorporate longitudinal covariates and time-varying coefficients. The relevant literature is reviewed in Chapter 2.

Although the methodologies are well established individually for longitudinal data and survival data, the separate models are insufficient for the new data structure and research questions discussed at the beginning of this chapter. New modeling frameworks are required to address two issues. First, the longitudinal covariates collected for each individual are usually contaminated by measurement errors; second, the complete knowledge of a covariate history for all individuals are unavailable due to the intermittent measurements and the occurrence of events or censoring. Due to these limitations, modeling the longitudinal and survival data separately using mixed-effects models and varying-covariate survival models might lose information and result in biased parameter estimation and misleading statistical inferences (Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997). In order to meet such practical need, joint modeling techniques have been developed in the past two decades as a powerful statistical tool to fit the two submodels simultaneously via their shared information. Compared with the separate estimation procedure, joint modeling has greatly improved the model estimation, and provided more reliable inferences of the longitudinal-survival relationship.

The basic joint model setting consists of two submodels. The first is the mixed-effects model, which is used to address the incompleteness and measurement errors of the continuous longitudinal covariates, and is given by

$$\begin{aligned} W_{ij} &= X(t_{ij}) + e(t_{ij}), \\ X(t_{ij}) &= \boldsymbol{\rho}(t_{ij})^T \boldsymbol{b}_i, \end{aligned} \tag{1.1}$$

where t_{ij} is the jth observational time point of the ith subject. W_{ij} is the longitudinal covariates observed at time t_{ij} for the ith subject. It is composed of the true covariate process $X(t_{ij})$, and the measurement errors $e(t_{ij})$. By mixed-effects models, $X(t_{ij})$ is written as the product of a vector function of time, $\boldsymbol{\rho}(t_{ij})$, and a vector of subject-specific random effects, $\boldsymbol{b}_i$. Note that the major difference between the longitudinal data and the conventional repeated measurements is that the longitudinal data allow the observations to be collected at different frequencies and have different lengths for different individuals. Thus longitudinal data are more flexible and require more sophisticated methodologies than traditional regression to handle the between-subject heterogeneity contained in $\boldsymbol{b}_i$ and the within-subject variation represented in $e(t_{ij})$. In Section 2.5 we review the analytical approaches developed to model longitudinal data, which include estimation techniques for both parametric and nonparametric models for $X(t_{ij})$.

The second submodel is for survival data, which incorporate longitudinal predictors via the varying-covariate Cox model. It is of the form

$$\lambda(t; \boldsymbol{X}_i(t)) = \lambda_0(t) \exp\{\boldsymbol{X}_i(t)^T \boldsymbol{\beta}\}, \tag{1.2}$$

where $\lambda_0(t)$ is the baseline hazard, and $\boldsymbol{\beta}$ is the regression coefficients of interest. A more general model is to allow the coefficients to vary with time, i.e., $\boldsymbol{\beta}(t)$, to capture the time-varying covariate effects. In Section 2.5 we review the studies of such survival models and the techniques used to estimate the smooth function $\boldsymbol{\beta}(t)$.

The two submodels (1.1) and (1.2) are linked by the true covariate process $\boldsymbol{X}_i(t)$ which is unobservable, and troublesome for parameter estimation.

Basically, there are two broad types of model fitting procedures for joint modeling problems. The first type of methods link the two submodels by modifying partial likelihood equations of the Cox model to incorporate longitudinal predictors. Such strategies, including an early regression calibration approach (Tsiatis et al.,

1995) and conditional score approach (Tsiatis and Davidian, 2001; Song et al., 2002b), mainly focus on the regression coefficients in the Cox model and do not consider the parameter estimation in longitudinal models. Regression calibration is a “two-step” estimation approach which estimates the conditional expectations of true longitudinal predictors given the observed measurements in the first step, and substitute them into the partial likelihood as predictors to solve for the regression coefficients in the second step. No asymptotic property has been developed for this approach and it is considered highly dependent on the model assumption of the longitudinal processes (Wulfsohn and Tsiatis, 1997). The conditional score approach (Tsiatis and Davidian, 2001; Song et al., 2002b) treats the random effects in the mixed-effects models as nuisance parameters and estimates survival regression coefficients from a modified partial likelihood which conditions on a complete and sufficient estimator of the random effects. Although these partial likelihood-based methods fix the bias in the survival coefficients and are relatively fast in computation, they have a couple of drawbacks (Wulfsohn and Tsiatis, 1997; Yu et al., 2004). First, the survival information is not used in the estimation of the longitudinal process, which may lead to bias and loss of efficiency (Faucett and Thomas, 1996); second, the recovered biomarkers used in the survival submodel are treated as fixed, thus some degrees of uncertainty are lost (Yu et al., 2004).

The second type of methods link the two models through their joint likelihood, and the parameters are estimated by maximizing the likelihood function. Since the likelihood function involves integration over the random effects, it is challenging to find the maximizer. One solution is to use Bayesian techniques (Faucett and Thomas, 1996; Xu and Zeger, 2001b; Wang and Taylor, 2001; Brown and Ibrahim, 2003). Based on the joint likelihood, a Bayesian approach makes assumptions about prior distributions of all the parameters, and updates the parameter estimates using Gibbs sampling from the full conditional distributions of each parameter given the observed data and the current estimates of all the other parameters. The advantage of the Bayesian approach is that it is not constrained by the dimension random effects, whereas the drawback is that the procedure itself is quite computationally intensive. An alternative is the maximum joint likelihood approach (Wulfsohn and Tsiatis, 1997), which maximizes the joint likelihood via Expectation-Maximization (EM) algorithm, treating the unobservable random effects as missing values. This method has been widely used for simple joint models with a single longitudinal process. The maximum joint likelihood method is dimension-sensitive in computation, but has good theoretical properties. Zeng and Cai (2005) prove that in the single-covariate setting, the maximum likelihood estimates have consistency and asymptotic normality properties and are semiefficient.

1.2 Motivation of this dissertation research

This dissertation research was motivated by an empirical analysis of data from a smoking cessation study (Piper et al., 2009). At the beginning of the program, a group of N=1504 heavy smokers were recruited, randomly assigned to 5 treatment groups (1 placebo group and 4 active treatment groups), and provided information for a set of baseline covariates. During the study, the ecological momentary assessment (EMA) data were collected 4 times a day, 2 weeks before and after the actual quit date for each individual. As a collection method of intensive longitudinal data, EMA involves repeated sampling of the subjects' current behaviors and experiences in real time, and in subjects natural environments (Shiffman et al., 2008). In this smoking cessation study, EMA data were collected to record smokers' momentary feelings of cessation fatigue and withdrawal symptoms such as craving and negative affect before and after their attempt to quit.

One of the key research interests of this study is to examine the relationships between these longitudinal covariates and the survival outcomes such as time to relapse, which is defined as 7 consecutive days of smoking after quit. The longitudinal variables that have been considered as the potential predictors are of various types and have different trajectories. Some of them are continuous (e.g., craving and negative affect), while others are categorical (e.g., whether or not a stressful event occurred since the last prompt); some of them have simple trends over time, whereas others may have complex and irregular trajectories. These diversified features suggest various specific research questions such as

- What is the relationship between a continuous longitudinal process and the time-to-event?

- What is the relationship between multiple continuous longitudinal processes and the time-to-event?
- What is the relationship between a categorical longitudinal process and the time-to-event?
- How do we take into consideration the irregular longitudinal trajectory and time-varying relationship between a longitudinal process and the time-to-event?

These research questions are critical since they not only lead to the answer of whether a specific longitudinal factor is associated with the risk of cessation failure, but also motivate for a deeper understanding of the dynamics of the longitudinal processes themselves and their relationships with the survival outcome over time.

Despite their significance in science, most of these questions have not been addressed well by the existing joint modeling techniques. Although abundant studies have occurred in the related fields (Wulfsohn and Tsiatis, 1997; Song et al., 2002b; Hsieh et al., 2006; Faucett and Thomas, 1996; Faucett et al., 1998; Henderson et al., 2000; Tsiatis and Davidian, 2001), and many of these methods have been applied to the medical and public studies (Wang and Taylor, 2001; Yu et al., 2004; Liu, 2009; Yu and Ghosh, 2010), most of them only consider simple parametric joint models with a single continuous longitudinal process. Such models may be used to address the first question listed above, but are insufficient in answering the other questions, which require more flexible and complex joint models.

The development of joint modeling has been hindered by several obstacles. On one hand, the theoretical establishment is challenging since joint models include both constant coefficients and functional coefficients with infinite dimension, hence the conventional MLE arguments do not apply (Hsieh et al., 2006). One needs to consider empirical process-related theorems to establish the asymptotic property. On the other hand, the model implementation is also tricky due to both the high-dimensionality of the parameters and the complex form of the joint likelihood. The intensive computation involved in parameter estimation makes it difficult to extend the model to the more flexible settings. Most of the currently available JM packages in R and SAS can only handle simple joint models with restricted assumptions. Therefore, the goal of this dissertation is to study the existing joint modeling approaches in details, overcome the computational difficulties, and generalize the current modeling frameworks both theoretically and practically.

1.3 Contribution of this dissertation research

Based on the study of the existing joint modeling techniques, this dissertation advances the research of joint modeling in terms of model extension, methodology development, computation improvement and theory establishment. More specifically, its contributions are in the following four aspects:

First of all, we improve the computational efficiency of the popular joint model fitting method, the maximum joint likelihood approach. As mentioned before, it is computationally challenging to carry out the estimation procedure based on the maximum joint likelihood approach because the use of the EM algorithm to optimize the objective function involves numerical integration over a multi-dimensional space for the shared random effect in the E-step of the EM algorithm. The existing numerical integration techniques, though working well for the low-dimensional cases, tend to collapse when the dimension of random effects grows and the model becomes more complicated. In this dissertation, we propose an approach to carrying out the numerical integration based on the design of experiments-based interpolation technique (DoIt, Joseph 2012), and combine it with the EM algorithm to form a new computing method, the EM-DoIt algorithm, for the joint modeling framework. This new algorithm is introduced in Chapter 3 of this dissertation, where we conducted Monte Carlo simulations to compare the proposed numerical integration method with the existing ones under the setting of joint models with a single covariate process. Our numerical results indicate that the proposed method performs very well in terms of computing time and statistical estimation accuracy. The performance of the proposed numerical method was further examined with more complex joint models such as the joint model with multiple covariate processes in Chapter 3, the nonparametric joint model in Chapter 5, and the joint model with a discrete covariate process in Chapter 6. The numerical results imply that the proposed computing method works well in all these scenarios.

Facilitated by the increasing computational capability of the EM-DoIt algorithm, we are able to extend the parametric joint models from the basic single longitudinal covariate setting to a multiple longitudinal covariates setting. Motivated by an empirical analysis of smoking cessation data, we propose a joint model setting with multiple longitudinal covariate processes in Chapter 3. We develop an estimation procedure for the proposed joint model based on the joint likelihood approach. We systematically study the asymptotic property of the proposed estimation procedure. In Chapter 4, we establish the consistency and asymptotic normality of the resulting estimate by using the formulation of Zeng and Cai (2005), in which the authors established the sampling property of maximum likelihood estimate for joint model with a single longitudinal covariate process. The theoretical development for the joint model with multiple longitudinal processes is much more challenging than that for the one with a single longitudinal covariate process since we have to deal with the covariance among the multiple longitudinal covariates rather than variance for the single longitudinal covariate. Although the joint model with multiple covariates has been considered in previous studies using other parameter estimation methods (Xu and Zeger, 2001a; Song et al., 2002a; Huang et al., 2001; Ibrahim et al., 2004; Brown et al., 2005; Chi and Ibrahim, 2006; Albert and Shih, 2010; Hatfield et al., 2011), most of this literature focuses only on the application of the model. This is the first time that such model is fitted by the maximum joint likelihood approach and the asymptotic properties of the MLE are established.

Another challenge of joint modeling is to estimate standard errors of the resulting estimate. We propose an estimation method for the standard error by a bootstrap method. We conduct Monte Carlo simulations to examine finite sample performance of the proposed estimation procedures including estimation of the parameters and estimation of their standard errors. Our numerical results indicate that the proposed estimation procedure performs well with moderate sample size. We further apply the proposed estimation method to a smoking cessation study (Piper et al., 2009) to assess the relationship between time to lapse and multiple longitudinal measurements of withdrawal symptoms. We find that the results of joint models with multiple longitudinal covariates offer deeper insight into the applied study than the model with a single longitudinal covariate.

Based on the study of parametric joint modeling, we propose a unified nonparametric joint model setting that covers many of the existing parametric joint models as special cases in Chapter 5. Such nonparametric joint models are more flexible and useful in practice because they do not assume any parametric form for the longitudinal trajectories, and at the same time, they allow the associations between the longitudinal covariates and the survival outcome to vary with time. We propose to approximate the functional coefficients in the model using B-spline basis functions, and choose the number of basis functions by the model selection criteria such as AIC and BIC. Following the methodology used for the parametric joint modeling, we still adopt a maximum joint likelihood approach for parameter estimation. We conduct Monte Carlo simulations to demonstrate the performance of the proposed estimation procedure. The numerical results show that although the dimension of the random effects increases dramatically in the nonparametric model setting, the estimators are still obtainable using the EM-DoIt algorithm, and both the vector parameters and the functional coefficients are accurately estimated. Moreover, the numerical results also indicate that the misspecification of the complex longitudinal trajectories with simple parametric shapes would lead to severe bias in the estimate of survival coefficients. We further apply the proposed model fitting procedure to the real data of the smoking cessation study to explore the dynamic structure of the longitudinal process and the possible time-varying relationships between the negative affect and time to lapse. To the best of our knowledge, the nonparametric setting in both longitudinal and survival submodels has never been studied in the joint modeling literature.

Finally, we extend the continuous-covariate joint model to the model with a categorical longitudinal predictor. In Chapter 6, we propose to extend the current continuous joint model setting to the generalized joint models where the observed covariate process can take either the binary or the count values. We conduct simulation studies to show that the maximum joint likelihood approach, implemented via the EM-DoIt algorithm, is feasible in fitting such complicated models. Compared with most of the related existing studies that focus mainly on the ad-hoc models to the specific datasets, this work provides a general modeling framework and solution to such problems.

1.4 Organization of this dissertation

This dissertation consists of 6 chapters. Chapter 2 provides a comprehensive review of the related literature. Section 2.1 describes the basic joint modeling notations and parametric model settings that have been adopted by most of the relevant studies and this dissertation. Section 2.2 reviews 4 model fitting approaches that have been proposed and extensively applied. Since the computation is a critical issue for joint modeling, we review the three most popular existing computing methods used in the field in section 2.3, and introduce a relatively new numerical interpolation approach, the design of experiments-based interpolation technique (DoIt), in section 2.4. In section 2.5, we review the basic model fitting methods for varying-coefficient models, including the varying-coefficient longitudinal model, the varying-coefficient survival model, and joint models with varying-coefficients.

Chapter 3 proposes a parametric joint model with multiple longitudinal covariates. The model setting and the estimation method, the joint maximum likelihood approach, are presented in detail in section 3.2. We also propose a new computing method, EM-DoIt algorithm, in section 3.2. Numerical analysis are conducted in section 3.3, where we use two simulation examples to demonstrate the proposed estimation approach and the computing algorithm, and address the research questions of the smoking cessation study by analyzing the real data.

Chapter 4 establishes the asymptotic properties for the MLE obtained from the estimation approach proposed in Chapter 3. In section 4.1, we state the consistency, the asymptotic normality, and the efficiency properties of the resulting estimates based on a series of technical conditions. The detailed proofs are provided in section 4.2.

Chapter 5 proposes a nonparametric joint model where both the longitudinal trajectory and the time-varying survival coefficients are approximated nonparametrically using B-splines. Section 5.2 describes the model setting, the estimation approach and the computing algorithm. Section 5.3 conducts numerical studies including both the simulation examples and real data analysis.

Chapter 6 is composed of the extension of the current work and the future work. Based on the work of Chapter 3 through Chapter 5, in section 6.1, we extend the current joint model setting to include discrete longitudinal covariates such as binary or count variables. We present several simulation studies to present the feasibility of the proposed modeling framework and the computing scheme. In section 6.2, we discuss future work that can be considered within the joint modeling framework of this dissertation.

Chapter 2

Literature Review

In this chapter we introduce the modeling framework and review the estimating procedures developed to jointly model longitudinal and survival data. We focus on the joint likelihood method and the related computational issues, which are the cores of this dissertation.

2.1 Joint modeling framework

We first introduce the notation used throughout the rest of this dissertation. As mentioned in Chapter 1, in joint modeling of longitudinal and survival data, researchers are usually interested in exploring the relationships between survival time, baseline covariates and longitudinal covariates. The corresponding data are

$$(T_i, \boldsymbol{Z}_i, \boldsymbol{\mathcal{X}}_i),\ i = 1, 2, \ldots, n, \tag{2.1}$$

where T_i is the survival time for the ith individual, $\boldsymbol{Z}_i$ is a time-fixed vector of q baseline covariates observed as $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^T$. In longitudinal data, we denote by $\boldsymbol{X}(t) = (X_1(t), \ldots, X_p(t))$ the p-dimensional true covariate process. $\boldsymbol{\mathcal{X}}_i = (\boldsymbol{X}_i(t_{i1}), \ldots, \boldsymbol{X}_i(t_{i,N_i}))$ is the matrix of covariate process for individual i, with $\boldsymbol{X}_i(t_{ij})$ as the explanatory vector observed on covariate process $\boldsymbol{X}(t)$ at time t_{ij} on individual $i = 1, \ldots, n$.

In (2.1) both T_i and $\boldsymbol{\mathcal{X}}_i$ are difficult to observe in practice. In survival data the event time T_i is subject to censoring. Right censoring is assumed throughout this dissertation. Thus instead of T_i, $V_i = \min(T_i, C_i)$ is the observed event time, where C_i corresponds to the censoring time. Denote by Δ_i the failure indicator taking value 1 if the failure is observed (i.e., $T_i \leq C_i$) and 0 otherwise. For the ith individual the observed survival data is actually (V_i, Δ_i). Similarly, in longitudinal data, the true covariate process $\boldsymbol{X}(t)$ is usually subject to measurement errors and cannot be observed directly. To take into account the measurement errors, denote the observed covariate process by

$$\boldsymbol{\mathcal{W}}_i = (\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{iN_i}), \tag{2.2}$$

with

$$\boldsymbol{W}_{ij} = \boldsymbol{X}_i(t_{ij}) + \boldsymbol{e}_i(t_{ij}), \quad i = 1, \ldots, n; \quad j = 1, \ldots, N_i, \tag{2.3}$$

where $\boldsymbol{e}_i(t_{ij})$ is a zero-mean random error vector at time t_{ij} for the ith individual. Therefore, in practice the actual data observed on (2.1) become

$$D_o = (V_i, \Delta_i, \boldsymbol{Z}_i, \boldsymbol{\mathcal{W}}_i, \boldsymbol{t}_i), \tag{2.4}$$

where $\boldsymbol{t}_i = (t_{ij} : 1 \leq j \leq N_i \text{ and } t_{ij} \leq V_i)$ is the observed time points for the ith individual. Since the longitudinal process is observed until the event or censoring happens, the observed covariate process is also truncated at V_i, i.e., $\{\boldsymbol{W}_{ij} : t_{ij} \leq V_i\}$.

Most joint modeling research uses similar modeling framework as that proposed by Tsiatis et al. (1995), which was reviewed in detail in more recent literature (Tsiatis and Davidian, 2004; Li and Ren, 2011). This joint modeling framework is also adopted in this thesis and introduced in the following paragraphs.

Denote by $\boldsymbol{X}_i^H(t) = \{\boldsymbol{X}_i(s); 0 \leq s < t\}$ the history of the longitudinal process up to time t. The survival time is modeled by the Cox model:

$$\lambda(t; \boldsymbol{Z}_i, \boldsymbol{X}_i^H(t)) = \lambda_0(t) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}_Z + \boldsymbol{X}_i(t)^T \boldsymbol{\beta}_X), \tag{2.5}$$

where $\boldsymbol{\beta}_Z \in \mathbb{R}^q$ and $\boldsymbol{\beta}_X \in \mathbb{R}^p$ are the unknown regression parameters of baseline and longitudinal covariates, respectively, $\lambda_0(t)$ is an unspecified baseline hazard function, and $\lambda(t; \boldsymbol{Z}_i, \boldsymbol{X}_i^H(t))$ is the conditional hazard function of T given $\boldsymbol{Z} =$

$\boldsymbol{Z}_i, \boldsymbol{X}^H(t) = \boldsymbol{X}_i^H(t)$ in the following sense

$$\lambda(t; \boldsymbol{Z}_i, \boldsymbol{X}_i^H(t)) = \lim_{h\to 0} h^{-1} P\{t \le T_i < t + h | T_i \ge t, \boldsymbol{Z}_i, \boldsymbol{X}_i^H(t)\}. \tag{2.6}$$

In joint modeling literature it is usually assumed that the right censoring is noninformative, i.e.,

$$\begin{aligned}&\lim_{h\to 0} h^{-1} P\{t \le T_i < t + h, \Delta = 1 | T_i \ge t, \mathbf{Z}_i, \mathbf{X}_i^H(t)\}\\ =&\lim_{h\to 0} h^{-1} P\{t \le T_i < t + h | T_i \ge t, \mathbf{Z}_i, \mathbf{X}_i^H(t)\}.\end{aligned}$$

This assumption is critical because what we actually observe is the cause-specific hazard on the left-hand side of the above equation, not the right-hand side, which we shall model. Without this assumption it is impossible for us to link the observable hazard with (2.6), based on which all the fitting procedures are developed.

In longitudinal data, a standard approach used to model the true covariate process $X_i(t)$ is by the model:

$$X_i(t) = \boldsymbol{\rho}(t)^T \boldsymbol{b}_i, \tag{2.7}$$

where $\boldsymbol{\rho}(t)$ is a vector of functions of time t including basis functions of t, such as polynomial functions, as a special case. $\boldsymbol{b}_i$ is a vector of subject-specific random effects. Accordingly, the model for the observed longitudinal process $\mathbf{W}_i$ is given by the mixed-effects model

$$W_{ij} = X_i(t_{ij}) + e_i(t_{ij}) = \boldsymbol{\rho}(t)^T \boldsymbol{b}_i + e_i(t_{ij}). \tag{2.8}$$

In many previous studies (De Gruttola and Tu, 1994; Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997; Dafni and Tsiatis, 1998; Bycott and Taylor, 1998), a simple linear model $X_i(t) = b_{0i} + b_{1i}t$ has been used to specify the covariate process. In such case, $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$, and $\boldsymbol{\rho}(t) = (1, t)^T$.

Common assumptions on (2.7) and (2.8) require that the random effects $\boldsymbol{b}_i$ are independent of the baseline covariate $\boldsymbol{Z}_i$ and the errors $\boldsymbol{e}_i(t)$, and are normally distributed, representing the within-subject variation. Usually if $\boldsymbol{e}_i(t)$ is used to represent the deviation due to only the measurement errors and "local" variation, they can be assumed independent across time and participants. Otherwise, if $\boldsymbol{e}_i(t)$ also involves the variation caused by a longer-term within-subject autocorrelation process, the corresponding covariance matrix should be specified for $\boldsymbol{e}_i(t)$ to address such autocorrelation across time for each individual. For simplicity, $e_i(t_{ij})$ are often assumed to account only for the measurement error, and are independent and identically distributed with the normal distribution

$$e_i(t_{ij}) \sim N(0, \sigma^2),\ i = 1, \ldots, n; j = 1, \ldots, n_i. \tag{2.9}$$

The random effects $\boldsymbol{b}_i$ are usually assumed to be independent and identical from the multivariate normal distribution

$$\boldsymbol{b}_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \tag{2.10}$$

where d is the dimension of the vector function $\boldsymbol{\rho}(t)$.

An alternative approach considered by Taylor et al. (1994), Henderson et al. (2000), Wang and Taylor (2001), among others, is to characterize the true covariate process by

$$X_i(t) = \boldsymbol{\rho}(t)^T \boldsymbol{b}_i + U_i(t), \tag{2.11}$$

where $U_i(t)$ is a mean-zero stochastic process, usually taken to be independent of $\boldsymbol{b}_i$ and $\boldsymbol{Z}_i$. The involvement of $U_i(t)$ allows the covariate trend to vary with time and induces a within-subject autocorrelation structure that may be viewed as arising from evolving individual fluctuations in the process of a smooth trend. More philosophical considerations of $U_i(t)$ were reviewed by Tsiatis and Davidian (2004).

It is important to specify the measurement error in longitudinal process (2.8) because it addresses one of the two main concerns of joint models we present in Chapter 1. Without such specification, a naive solution is to use the raw measurements $\boldsymbol{W}(t)$ in survival analysis (2.5) to substitute $\boldsymbol{X}(t)$. However, this direct approach would lead to biased estimates. Prentice (1982) pointed out that measurement error can cause the estimated regression parameter in the time-dependent Cox model to be biased toward the null, and the magnitude of the bias is proportional to measurement error in the observed predictors.

The second concern deals with the incompleteness of the longitudinal process (Tsiatis and Davidian, 2004; Li and Ren, 2011). Note that (2.5) is different from the usual Cox model with time-varying covariates that can be estimated by partial likelihood (Cox, 1975)

$$\prod_{i=1}^{n}\left[\frac{\exp\{\boldsymbol{\beta}_X \boldsymbol{X}_i(V_i)+\boldsymbol{\beta}_Z^T\mathbf{Z}_i\}}{\sum_{j=1}^{n}\exp\{\boldsymbol{\beta}_X \boldsymbol{X}_j(V_i)+\boldsymbol{\beta}_Z^T\mathbf{Z}_j\}I(V_j\geq V_i)}\right]^{\Delta_i}. \tag{2.12}$$

It is clear that the equation (2.12) requires $\boldsymbol{X}_i(t)$ obtained for all $i = 1, \ldots, n$ at each observed failure time. If all individuals were measured at the same time, this would not be a problem. However, it is common that some individuals' covariates are not measured at other individuals' event times. For example, individuals may have different scheduled visiting time points. In that case, one individual's covariate value may be missing in the sense that there exists an event time which does not fall on this individual's schedule. An ad hoc approach, also known as the method of "Last Value Carried Forward (LVCF)", is to use the latest observation from the same individual, and treat it as if it were the current value of the covariate at the failure time. If the time-dependent covariate does not change sharply over time, this imputation will probably work well. Otherwise, this ad hoc approach is not guaranteed to lead to good estimates (Prentice, 1982).

In the following sections we review the methods developed to tackle these problems and estimate the parameters. All of them are based on the joint modeling framework (2.5) and (2.8), and the set of parameters of interest is

$$\Omega = (\lambda_0, \boldsymbol{\beta}_Z, \boldsymbol{\beta}_X, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2), \tag{2.13}$$

where the regression parameters $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ are of primary interest.

2.2 Estimation methods

Several estimation procedures have been developed to estimate the parameters in (2.13), especially the regression parameters $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ in the survival model (2.5). Below we summarize four commonly used methods: regression calibration method, Bayesian method, conditional score method and joint likelihood method.

2.2.1 Regression calibration method

Tsiatis et al. (1995) proposed the regression calibration method (also known as the "two-stage" method) to first approximate the covariate process $\boldsymbol{X}(t)$, and then plug the approximated values in (2.12) to solve for $\boldsymbol{\beta}$'s. The starting assumption for this inferential strategy is that neither the measurement error nor the timing of the visits prior to time t are prognostic. That is, given the true covariate history, the observed covariate history is independent of the hazard rate. This assumption implies that

$$\lambda(t|\boldsymbol{Z}, \boldsymbol{X}(t), \boldsymbol{W}(t)) = \lambda(t|\boldsymbol{Z}, \boldsymbol{X}(t)) = \lambda_0(t) \exp(\boldsymbol{Z}^T \boldsymbol{\beta}_Z + \boldsymbol{X}(t)^T \boldsymbol{\beta}_X). \tag{2.14}$$

Thus it follows by the law of conditional probability that

$$\begin{aligned} \lambda(t|\boldsymbol{Z}, \boldsymbol{W}(t)) &= \int \lambda(t|\boldsymbol{Z}, \boldsymbol{X}(t), \boldsymbol{W}(t)) dP(\boldsymbol{X}(t)|\boldsymbol{Z}, \boldsymbol{W}(t), V \geq t) \\ &= \lambda_0(t) \exp(\boldsymbol{Z}^T \boldsymbol{\beta}_Z) E\left[\exp(\boldsymbol{X}(t)^T \boldsymbol{\beta}_X)|\boldsymbol{W}(t), V \geq t\right], \end{aligned} \tag{2.15}$$

where $\mathbf{W}(t) = \{W(t_1), \ldots, W(t_J); t_J \leq t\}$. The goal is to estimate the conditional expectation $E\left[\exp(\boldsymbol{X}(t)^T \boldsymbol{\beta}_X)|\boldsymbol{W}(t), V \geq t\right]$ and substitute it in the partial likelihood equation (2.12) to solve for $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$.

In the first stage, the empirical Bayes estimator for $\boldsymbol{X}(t)$ is obtained from a standard fit of the mixed effects model defined by (2.8) for all subjects still in the risk set, i.e., satisfying $V_i \geq t$. With the model setting of (2.8) and the normal assumption of the random effects $\boldsymbol{b}_i$ it is convenient to obtain the joint normal distribution of $\{\boldsymbol{W}(t), \boldsymbol{X}(t)\}$, and the conditional distribution of $\boldsymbol{X}(t)$ given $\boldsymbol{W}(t)$. Denote the conditional mean by $\boldsymbol{\mu}_{\cdot|W}(t) = E(\boldsymbol{X}(t)|\boldsymbol{W}(t))$, and conditional variance by $\Sigma_{\cdot|W}(t) = \mathrm{Var}(\boldsymbol{X}(t)|\boldsymbol{W}(t))$. Then moment generating function of $\{\mathbf{X}(t)|\mathbf{W}(t)\}$ is obtained by

$$E\left\{\exp(\boldsymbol{X}(t)^T \boldsymbol{\beta}_X)|\boldsymbol{W}(t), V \geq t\right\} = \exp\{\boldsymbol{\mu}_{\cdot|W}(t)^T \boldsymbol{\beta}_X + \frac{1}{2} \boldsymbol{\beta}_X^T \Sigma_{\cdot|W}(t) \boldsymbol{\beta}_X\}. \tag{2.16}$$

In particular, if the longitudinal process is specified by the mixed-effects model as in (2.8) with the assumptions (2.9) and (2.10), for any time point t before the event time V, $\boldsymbol{X}(t)$ is a normal process with the mean $E(X(u)|V \leq t) =$

$\boldsymbol{\rho}(u)^T\boldsymbol{\mu}$ and covariance $\mathrm{cov}(X(u), X(v)) = \boldsymbol{\rho}(u)^T\boldsymbol{\Sigma}\boldsymbol{\rho}(v) \triangleq C_t(u, v)$ for any $u, v \leq t$. Thus for any time point t before the event time V, the joint distribution of $\{W(t), X(t)\} = \{W(t_1), \ldots, W(t_j), X(t)\}$ is normal with mean $\{\bar{\boldsymbol{\mu}}_w(t), \mu_x(t)\} \triangleq \{\boldsymbol{\rho}(t_1)^T\boldsymbol{\mu}, \ldots, \boldsymbol{\rho}(t_j)^T\boldsymbol{\mu}, \boldsymbol{\rho}(t)^T\boldsymbol{\mu}\}$, and variance

$$\boldsymbol{M}_t = \begin{pmatrix} C_t(t_1, t_1) & \ldots & C_t(t_1, t_j) & C_t(t_1, t) \\ \vdots & & & \\ C_t(t_j, t_1) & \ldots & C_t(t_j, t_j) & C_t(t_j, t) \\ C_t(t, t_1) & \ldots & C_t(t, t_j) & C_t(t, t) \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ \vdots & & & \\ 0 & \ldots & \sigma^2 & 0 \\ 0 & 0 & \ldots & 0 \end{pmatrix},$$

where t_j is the last observing time point before t. All the parameters can be specified by the estimates of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and σ^2 from the mixed-effects model. Accordingly, the conditional distribution of $\boldsymbol{X}(t)$ given $\boldsymbol{W}(t)$ at any t before V is also normal, and the conditional mean $\boldsymbol{\mu}_{\cdot|W}(t)$ and variance $\Sigma_{\cdot|W}(t)$ can be calculated from the joint mean and variance using the property of multivariate normal distribution.

In the second stage, the estimated value of the conditional expectation of (2.16) is plugged into the partial likelihood (2.12) to substitute $\exp\{\boldsymbol{\beta}_X\boldsymbol{X}(t)\}$ at each event time point. Then the regression coefficients $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ are estimated by maximizing the partial likelihood function.

As pointed out by Wulfsohn and Tsiatis (1997), this two-step modeling approach has several limitations. First, the assumption that the random effects are normally distributed in those at risk at each event time is probably unreasonable. If the covariate is predictive of survival time, patients with the steepest negative slope of the covariate trajectories may be at higher risk for mortality, and thus removed from the population early on. This may result in the random effects having a distribution shift toward a nonnormal distribution as time progresses. The violation of the normality assumption may cause biased estimation. Second, a first-order approximation is required in order to use polynomial growth curve models to simplify the partial likelihood to be maximized (Tsiatis et al., 1995). The validity of this approximation depends on the scaling of the covariate. Finally, the two-stage model does not use any survival information in modeling the covariate process, and thus information is not used as efficiently as it might be.

2.2.2 Bayesian method

Another well-developed estimation method for the joint modeling framework (2.5) and (2.8) is the Bayesian approach proposed by Faucett and Thomas (1996). Based on the density functions specified my model assumptions, the authors used a Markov chain Monte Carlo (MCMC) technique to estimate the posterior distribution of the unknown parameters in Ω given the observed data D_o.

In the Bayesian approach, Gibbs sampling is used to generate random samples from the joint posterior distribution of unknown parameters in a model one at a time, conditional on the observed data and other parameters. It is useful in joint modeling because the joint distribution of parameters is intractable, but the generation of samples from each full conditional distribution is feasible. Given a set of initial estimates for each of the unknown parameters, the authors generated samples in turn from the full conditional distributions of each unknown parameter conditional on the current assignment of all other parameters and data.

Based on Faucett and Thomas (1996), Xu and Zeger (2001b) considered generalizations of this MCMC approach, and allowed models of form (2.9). The author used the empirical characteristic statistics from the generated samples to estimate the parameters and draw statistical inferences based on them. Wang and Taylor (2001) fit a joint model to HIV data using MCMC and incorporating a longitudinal model of form (2.9). Brown and Ibrahim (2003) proposed a semiparametric Bayesian joint model of form (2.5) and (2.8) that furthermore makes no parametric assumption on the random effects. More discussion of the Bayesian approach can be found in the review of Tsiatis and Davidian (2004).

2.2.3 Conditional score method

Both the regression calibration and the Bayesian approach require specification of the distribution of random effect $\boldsymbol{b}_i$. To minimize reliance on parametric modeling assumptions, Tsiatis and Davidian (2001) developed a set of unbiased estimating equations that yields consistent and asymptotically normal estimators of the survival coefficients with no assumptions on $\boldsymbol{b}_i$. The idea of conditional score (Stefanski and Carroll, 1987) was employed to treat the random effects $\boldsymbol{b}_i$ as the "nuisance parameters" and estimate $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ conditional on an appropriate "sufficient statistic" for $\boldsymbol{b}_i$.

The conditional score approach is similar to the regression calibration approach in that it is also a "two-stage" inference technique requiring the estimation of the covariate process $\boldsymbol{X}_i(t)$ and using the partial likelihood equation to solve for the regression parameters. The differences between them are twofold. First, the regression calibration method we described before uses empirical Bayesian estimators for $\boldsymbol{X}_i(t)$ at each event time, while in the conditional score approach, ordinary least squares (OLS) estimators are used to approximate $\boldsymbol{X}_i(t)$. Second, the regression parameters in approximation method are estimated by partial likelihood for the risk $P\{t \leq V_i < t + dt|\boldsymbol{b}_i, \boldsymbol{Z}_i(t), \boldsymbol{X}_i^H(t), V_i \geq t\}$, where the distribution of $\boldsymbol{b}_i$ is required. For conditional score approach, however, the regression parameters are obtained by solving the partial likelihood equation for the risk $P\{t \leq V_i < t + dt|S_i(t, \beta, \sigma_e^2), \boldsymbol{Z}_i(t), \boldsymbol{X}_i^H(t), V_i \geq t\}$, where $S_i(t, \beta, \sigma_e^2)$ is a "sufficient statistic" for $\boldsymbol{b}_i$ based on $\hat{X}_i(t)$. Tsiatis and Davidian (2001) suggested that conditioning on $S_i(t, \beta, \sigma_e^2)$ would remove the dependence of the conditional distribution on the "nuisance parameter" $\boldsymbol{b}_i$.

More specifically, in the first stage, Tsiatis and Davidian (2001) obtained the OLS estimators $\hat{\boldsymbol{X}}_i(t)$ based on the observed data $\boldsymbol{W}_i(t)$, with the assumption that σ_e^2 is known. For instance, in one-dimensional simple linear case, $\hat{X}_i(t) = (1, t)^T \hat{\boldsymbol{b}}_i$, where $\hat{\boldsymbol{b}}_i = \{\boldsymbol{\rho}(t)^T \boldsymbol{\rho}(t)\}^{-1} \boldsymbol{\rho}(t)^T \boldsymbol{w}_i(t)$, and $\boldsymbol{\rho}(t)$ is the usual $\{n_i \times 2\}$ design matrix with first column all ones and second column t_{ij} for the ith subject. Note that $\hat{\boldsymbol{b}}$ and hence $\hat{X}_i(t)$ are defined only if there are at least two measurements prior to t. Define $Y_i(t) = I(V_i \geq t, t_{i2} \leq t)$. Assume that the distribution of "error" $e_i(t_{ij})$ at time t_{ij} is $N(0, \sigma_e^2)$, given that a measurement is taken at t_{ij}, i is at risk at t_{ij}, the measurement history prior to t_{ij}, $\boldsymbol{b}_i$ and $\boldsymbol{Z}_i$. The authors demonstrated that under these conditions and noninformative assumptions regarding the censoring and timing processes,

$$\{\hat{X}_i(t) \left| Y_i(t) = 1, \boldsymbol{b}_i, \boldsymbol{Z}_i\} \sim N(X_i(t), \sigma_e^2 \theta_i(t)) \right. , \tag{2.17}$$

where $\sigma_e^2 \theta_i(t)$ is the usual variance of the predicted value $\hat{X}_i(t)$, which depends on the time and the variance structure of $\boldsymbol{b}_i$. Define $dN_i(u) = I(u \leq V_i < u + du, \Delta_i = 1, t_{i2} \leq u)$, which puts point mass at time u for an observed event time after the second longitudinal measurement on subject i. The motivation for the conditional score estimating equations relies on identifying a "sufficient statistic" for $\boldsymbol{b}_i$.

At time u, given i is at risk, the conditional density for $\{dN_i(u) = r, \hat{X}_i(u) = x\}$ is

$$\begin{aligned}
&p\{dN_i(u) = r, \hat{X}_i(u) = x \mid Y_i(u) = 1, t_i(u), \boldsymbol{b}_i, \boldsymbol{Z}_i\} \\
&= p\{dN_i(u) = r \mid \hat{X}_i(u) = x, Y_i(u) = 1, t_i(u), \boldsymbol{b}_i, \boldsymbol{Z}_i\} \\
&\times p\{\hat{X}_i(u) = x \mid Y_i(u) = 1, t_i(u), \boldsymbol{\alpha}_i, \boldsymbol{Z}_i\}.
\end{aligned}$$

The first term on the right hand side of the above equation is a Bernoulli density with probability $\lambda_0(u)du \exp\{\beta X_i(u) + \boldsymbol{\eta}^T Z_i\}$, and the second term is a normal density of $N\{X_i(u), \sigma_e^2\theta_i(u)\}$. Substituting the associated densities into the above equation and simplifying yields

$$\begin{aligned}
&\exp\left[X_i(u)\left\{\frac{\beta\sigma_e^2\theta_i(u)dN_i(u) + \hat{X}_i(u)}{\sigma_e^2\theta_i(u)}\right\}\right] \\
&\times \frac{\lambda_0(u)\exp(\boldsymbol{\eta}^T\boldsymbol{Z}_i)^{dN_i(u)}}{\{2\pi\sigma_e^2\theta_i(u)\}^{1/2}} \exp\left\{-\frac{\hat{X}_i^2(u) + X_i^2(u)}{2\sigma_e^2\theta_i(u)}\right\}.
\end{aligned}$$

This contains a "sufficient statistic" for $\boldsymbol{b}_i$: $S_i(u, \beta, \sigma_e^2) = \beta\sigma_e^2\theta_i(u)dN_i(u) + \hat{X}_i(u)$. Conditional on this sufficient statistic and observed data, the survival process is

$$\begin{aligned}
&\lim_{du\to 0} du^{-1} p\{dN_i(u) = 1 \mid S_i(u, \beta, \sigma_e^2), \boldsymbol{Z}_i, t_i(u), Y_i(u)\} \\
&= \lambda_0(u)\exp\{\beta S_i(u, \beta, \sigma_e^2) - \beta^2\sigma_e^2\theta_i(u)/2 + \boldsymbol{\eta}^T\boldsymbol{Z}_i\}Y_i(u) \triangleq \lambda_0(u)E_{0i}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2).
\end{aligned}$$

Based on this equation, Tsiatis and Davidian (2001) proposed the estimating equations for β and $\boldsymbol{\eta}$ by equating "observed" and "expected" quantities in a spirit similar to such a derivation for the usual partial likelihood score equations:

$$\sum_{i=1}^n \int \{S_i(u, \beta, \sigma_e^2), Z_i^T\}^T\{dN_i(u) - E_{0i}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)\lambda_0(u)du\} = \mathbf{0},$$

$$\sum_{i=1}^n \{dN_i(u) - E_{0i}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)\lambda_0(u)du\} = \mathbf{0},$$

with $dN(u) = \sum_{j=1}^{n} dN_j(u)$ and $E_0^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2) = \sum_{j=1}^{n} E_{0j}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)$. The second equation yields

$$\hat{\lambda}_0(u)du = \frac{dN(u)}{E_0^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)}.$$

By substitution of $\hat{\lambda}_0(u)du$ in the first equation, the conditional score estimating equation for β and $\boldsymbol{\eta}$ are

$$\sum_{i=1}^{n} \int \left[\{S_i(u, \beta, \sigma_e^2), \boldsymbol{Z}_i^T\}^T - \frac{E_i^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)}{E_0^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)} \right] dN_i(u) = \mathbf{0}, \tag{2.18}$$

where $E_{1j}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2) = \{S_j(u, \beta, \sigma_e^2), \boldsymbol{Z}_i^T\}^T E_{0j}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)$, and $E_1^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2) = \sum_{j=1}^{n} E_{1j}^*(u, \beta, \boldsymbol{\eta}, \sigma_e^2)$.

This equation reduces to the partial likelihood score equations when $\sigma_e^2 = 0$ (i.e., $\hat{X}_i(u) = X_i(u)$). When σ_e^2 is unknown, Tsiatis and Davidian (2001) proposed an additional estimating equation for σ_e^2 based on residuals from individual least squares fits to the N_i measurements for each i and gave arguments indicating that the resulting estimators for β and $\boldsymbol{\eta}$ are consistent and asymptotically normal under assumptions specified above, with standard errors that may be derived based on the usual sandwich approach.

It is worth mentioning that the violation of the normality assumptions may not be a concern for some methods. For example, Song et al. (2002b) and Hsieh et al. (2006) investigated the robustness of joint likelihood procedures when the normal assumptions of the random effects are violated. They found the estimations of the parameters of interest are basically unbiased and as efficient as those estimated when normality is imposed, and the estimated cumulative hazards are almost identical in the two scenarios. For details on the rationale and the derivation of these estimating equations with their properties, see Tsiatis and Davidian (2004) for further discussions.

2.2.4 Joint likelihood method

The most popular method used to estimate joint models is the joint likelihood approach proposed by Wulfsohn and Tsiatis (1997), which simultaneously fits (2.5) and (2.8) linked by the random effects $\boldsymbol{b}_i$ via the joint likelihood. To specify the joint likelihood we need to assume the observed longitudinal process $\boldsymbol{W}_i(t)$ and the event processes $\{V_i, \Delta_i\}$ are independent given the random effects $\boldsymbol{b}_i$. For simplicity we first consider the one-dimensional covariate process $X(t)$, which can be extended to the multi-dimensional cases. With the above assumption, the joint likelihood of the observed data can be written as follows:

$$L(\Omega) = \prod_{i=1}^{n} L_i(\Omega) = \prod_{i=1}^{n} \int f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i} d\boldsymbol{b}_i, \tag{2.19}$$

where

$$f_{V_i,\Delta_i|\boldsymbol{b}_i} = \left\{\lambda_0(V_i) e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_i + \boldsymbol{Z}_i^T \boldsymbol{\beta}_Z}\right\}^{\Delta_i} \exp\left\{-\int_0^{V_i} \lambda_0(u) e^{\beta_X \boldsymbol{\rho}(u)^T \boldsymbol{b}_i + \boldsymbol{Z}_i^T \boldsymbol{\beta}_Z} du\right\}, \tag{2.20}$$

$$f_{W_i|\boldsymbol{b}_i} = (\frac{1}{\sqrt{2\pi\sigma^2}})^{N_i} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(w_{ij} - \boldsymbol{\rho}(t_{ij})^T \boldsymbol{b}_i)^2\right\}, \tag{2.21}$$

$$f_{\boldsymbol{b}_i} = \frac{1}{2\pi}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu})\right\}. \tag{2.22}$$

Using the joint likelihood function (2.19), there is no need to recover the true covariate process $X_i(t)$ as in the regression calibration and the conditional score approaches. The density for the survival data in equation (2.20) assumes that the current value of the covariate is the appropriate component of the covariate history to use in the model.

This approach also assumes that censoring is independent of the random effects. However, if this is not the case, the appropriate density for the censoring process should be incorporated in the model. If the censoring process that leads to dropout is not correctly modeled, the resulting estimates of the regression coefficients β_X and $\boldsymbol{\beta}_Z$ may be biased.

Usually the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_e^2, \beta_X$ and $\boldsymbol{\beta}_Z$ are estimated using parametric maximum likelihood and the baseline hazard $\lambda_0(u)$ using nonparametric maximum likelihood. The baseline hazard $\lambda_0(u)$ is assumed to take mass at each failure time, and its dimension is equal to the number of unique failure times.

The maximum likelihood approach is a popular modeling procedure for joint modeling problems, and has been extensively employed in many studies. However, the asymptotic properties for the resulting MLE has not been theoretically justified until Zeng and Cai (2005). In this study, the authors built up a joint modeling framework that is very similar to the one we specified in this section. The main difference is that in this work, the continuous covariate process $\boldsymbol{X}(t)$ is assumed to be fully observed instead of being recorded intermittently. Under this and other related assumptions, the authors proved that the MLE $\hat{\boldsymbol{\Omega}}$ for $\boldsymbol{\Omega}$ has the desired properties of strong consistency, asymptotic normality, and semiefficiency.

Solving for the MLE of the joint likelihood (2.19) is quite challenging because the integrals involved make it difficult to optimize. Wulfsohn and Tsiatis (1997) proposed to calculate the MLE of (2.19) using an expectation-maximization (EM) algorithm, where the unobserved random effects are treated as missing data, and the parameter estimates are updated iteratively until the algorithm converges. This method has been widely applied and proved feasible and robust in the later related studies (Henderson et al., 2000; Tsiatis and Davidian, 2001; Song et al., 2002b; Hsieh et al., 2006). This approach is also the main focus of this study, and we explain it in detail as follows.

Taking logarithm of the joint likelihood of (2.19),

$$\begin{aligned} l(\Omega) = \log(L(\Omega)) &= \Sigma_{i=1}^{n} \log \int f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i} d\boldsymbol{b}_i \\ &= \log(f(V_i, \Delta_i, W_i)). \end{aligned} \tag{2.23}$$

Let θ denote any of the parameters in Ω. Taking the derivative of $l(\Omega)$ with respect to θ and assuming the derivative and integral are interchangeable under certain conditions, it follows that

$$\begin{aligned} S(\theta) &= \frac{\partial l(\Omega)}{\partial \theta} = \frac{\partial}{\partial \theta} \log(f(V_i, \Delta_i, W_i)) \\ &= \frac{1}{f(V_i, \Delta_i, W_i)} \frac{\partial}{\partial \theta} \int f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i} d\boldsymbol{b}_i \\ &= \int \frac{\partial}{\partial \theta} \log(f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i}) \frac{f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i}}{f(V_i, \Delta_i, W_i)} d\boldsymbol{b}_i \\ &= \int \frac{\partial}{\partial \theta} \{\log(f_{V_i,\Delta_i|\boldsymbol{b}_i}) + \log(f_{W_i|\boldsymbol{b}_i}\} + \log(f_{\boldsymbol{b}_i})) \frac{f(V_i, \Delta_i, W_i, \boldsymbol{b}_i)}{f(V_i, \Delta_i, W_i)} d\boldsymbol{b}_i \\ &= \int \frac{\partial}{\partial \theta} \{\log(f_{V_i,\Delta_i|\boldsymbol{b}_i}) + \log(f_{W_i|\boldsymbol{b}_i}\} + \log(f_{\boldsymbol{b}_i})) f_{\boldsymbol{b}_i|D_o,\hat{\Omega}} d\boldsymbol{b}_i \\ &= \frac{\partial}{\partial \theta} \left\{ E(l_1(\boldsymbol{b}_i)) + E(l_2(\boldsymbol{b}_i)) + E(l_3(\boldsymbol{b}_i)) \right\}, \end{aligned} \tag{2.24}$$

where $E(\cdot)$ is conditional expectations of $\boldsymbol{b}_i$ given observed data D_o defined in (2.4) and the updated parameter estimates $\hat{\boldsymbol{\Omega}}$ in the EM algorithm, and

$$\begin{aligned}
l_1(\boldsymbol{b}_i) =& \log\{f_{V_i,\Delta_i|\boldsymbol{b}_i}\}\\
=&\Delta_i \log\{\lambda_0(V_i)\} + \Delta_i\left\{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_i + \boldsymbol{Z}_i^T\boldsymbol{\beta}_Z\right\} - \int_0^{V_i}\lambda_0(u)e^{\beta_X\boldsymbol{\rho}(u)^T\boldsymbol{b}_i+\boldsymbol{Z}_i^T\boldsymbol{\beta}_Z}du\\
=&\Delta_i \log\{\lambda_0(V_i)\} + \Delta_i\left\{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_i + \boldsymbol{Z}_i^T\boldsymbol{\beta}_Z\right\}\\
&-\sum_{j=1}^n \lambda_0(V_j)e^{\beta_X\boldsymbol{\rho}(V_j)^T\boldsymbol{b}_i+\boldsymbol{Z}_i^T\boldsymbol{\beta}_Z}I(V_i\geq V_j,\Delta_j=1),\\
l_2(\boldsymbol{b}_i) =& \log(f_{W_i|\boldsymbol{b}_i}) = -\frac{N_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{N_i}(w_{ij}-\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i)^2,\\
l_3(\boldsymbol{b}_i) =& \log(f_{\boldsymbol{b}_i}) = -\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{b}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}).
\end{aligned}$$

The corresponding conditional expectations in (2.24) are

$$\begin{aligned}
E\{l_1(\boldsymbol{b}_i)\} =& \Delta_i \log\{\lambda_0(V_i)\} + \Delta_i\beta_X E\left\{\boldsymbol{\rho}(V_i)^T\boldsymbol{b}_i + \boldsymbol{Z}_i^T\boldsymbol{\beta}_Z\right\}\\
&-\sum_{j=1}^n \lambda_0(V_j)E\left\{e^{\beta_X\boldsymbol{\rho}(V_j)^T\boldsymbol{b}_i+\boldsymbol{Z}_i^T\boldsymbol{\beta}_Z}I(V_i\geq V_j,\Delta_j=1)\right\}, \qquad (2.25)\\
E(l_2(\boldsymbol{b}_i)) =& -\frac{N_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{N_i}E(w_{ij}-\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i)^2, \qquad (2.26)\\
E(l_3(\boldsymbol{b}_i)) =& -\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}E(\boldsymbol{b}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}_i-\boldsymbol{\mu}). \qquad (2.27)
\end{aligned}$$

From (2.24) it is clear that to obtain the MLE of (2.19) the first task is to specify the conditional expectations of $l_k(\mathbf{b}_i)$, the functions of $\mathbf{b}_i$, in (2.25) through (2.27). This is what the E-step does in the EM algorithm. In the M-step, the log likelihood is maximized by taking partial derivatives of $E(l_1(\boldsymbol{b}_i)), E(l_2(\boldsymbol{b}_i)), E(l_3(\boldsymbol{b}_i))$ with respect to their corresponding parameters, i.e., calculating $S(\theta)$ in (2.24) (Flury and Zoppè, 2000). All the parameters in Ω except β_X and $\boldsymbol{\beta}_Z$ have the closed-form maximum likelihood estimates:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n E(\boldsymbol{b}_i), \qquad (2.28)$$

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n} E(\boldsymbol{b}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{b}_i - \hat{\boldsymbol{\mu}})^T, && (2.29)\\
\hat{\sigma}^2 &= \frac{1}{\sum_{i=1}^{n} N_i}\sum_{i=1}^{n}\sum_{j=1}^{N_i} E(w_{ij} - \boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i)^2, && (2.30)\\
\hat{\lambda}_0(u) &= \frac{\sum_{i=1}^{n}\Delta_i I(V_i = u)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(u)^T\boldsymbol{b}_j + \boldsymbol{\beta}_Z^T \boldsymbol{Z}_j} I(V_j \geq u)}, && (2.31)
\end{aligned}$$

where u only takes value at the event points. For other time points, $\hat{\lambda}_0(t) = 0$.

The MLE of β_X and $\boldsymbol{\beta}_Z$ are obtained by applying Newton-Raphson algorithm to the profile likelihood of $l_1(\boldsymbol{b}_i)$ after plugging in $\hat{\lambda}_0(t)$ in (2.31):

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + I^{-1}_{\hat{\beta}^{(k-1)}} S_{\hat{\beta}^{(k-1)}},$$

where $S_{\hat{\beta}^{(k-1)}}$ and $I_{\hat{\beta}^{(k-1)}}$ are the score and information equations of $\boldsymbol{\beta}$'s taking values at the $(k-1)$th updated estimates. The score and the information of β_X and the lth element of $\boldsymbol{\beta}_Z$ are calculated as below:

$$\begin{aligned}
S(\beta_X) =& \frac{\partial E(l_1(\boldsymbol{b}_i))}{\partial \beta_X}\\
=& \sum_{i=1}^{n}\Delta_i \left\{ E(\boldsymbol{\rho}(V_i)^T\boldsymbol{b}_i) - \frac{\sum_{j=1}^{n} E(\boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j) e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)} \right\},
\end{aligned} \tag{2.32}$$

$$\begin{aligned}
I(\beta_X) =& \frac{-\partial S(\beta_X)}{\partial \beta_X}\\
=& \sum_{i=1}^{n}\Delta_i \left\{ \frac{\sum_{j=1}^{n} E(\boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j)^2 e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)} \right.\\
& \left. - \left[\frac{\sum_{j=1}^{n} E(\boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j) e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)}\right]^2 \right\}, && (2.33)\\
S(\beta_{Z_l}) =& \frac{\partial E(l_1(\boldsymbol{b}_i))}{\partial \beta_{Z_l}}\\
=& \sum_{i=1}^{n}\Delta_i \left\{ Z_{il} - \frac{\sum_{j=1}^{n} Z_{il} E e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T\boldsymbol{b}_j + \boldsymbol{Z}_j^T\boldsymbol{\beta}_Z} I(V_j \geq V_i)} \right\}, && (2.34)\\
I(\beta_{Z_l}) =& \frac{-\partial S(\beta_{Z_l})}{\partial \beta_{Z_l}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{j=1}^{n} Z_{il}^2 E e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j + \boldsymbol{Z}_j^T \boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j + \boldsymbol{Z}_j^T \boldsymbol{\beta}_Z} I(V_j \geq V_i)} \right. \\
&\quad \left. - \left[\frac{\sum_{j=1}^{n} Z_{il} E e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j + \boldsymbol{Z}_j^T \boldsymbol{\beta}_Z} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j + \boldsymbol{Z}_j^T \boldsymbol{\beta}_Z} I(V_j \geq V_i)} \right]^2 \right\}. \qquad (2.35)
\end{aligned}
$$

In an EM algorithm, the above E-step and M-step are calculated iteratively until the algorithm converges. By (2.28) through (2.35), in each iteration, conditional expectations need to be evaluated for the following six functions of $\boldsymbol{b}$ for the ith subject, $i = 1, \ldots, n$:

$$
\begin{aligned}
g_1(\boldsymbol{b}_i) &= \boldsymbol{b}_i, \\
g_2(\boldsymbol{b}_i) &= \boldsymbol{b}_i \boldsymbol{b}_i^T, \\
g_3(\boldsymbol{b}_i) &= \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_i, \qquad (2.36) \\
g_4(\boldsymbol{b}_j) &= e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i, \\
g_5(\boldsymbol{b}_j) &= (\boldsymbol{\rho}(V_i)^T \boldsymbol{b}_i) e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i, \\
g_6(\boldsymbol{b}_j) &= (\boldsymbol{\rho}(V_i)^T \boldsymbol{b}_i)^2 e^{\beta_X \boldsymbol{\rho}(V_i)^T \boldsymbol{b}_j}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i.
\end{aligned}
$$

Since multi-dimensional integrals are involved in the conditional expectations $E\{g(\boldsymbol{b}_i)\}$ in the E-step, they need to be approximated by numerical integration techniques, which are quite time-consuming and cause instability in the EM results. This numerical challenge is confronted by all joint modeling approaches, especially when the random effects have high dimension. In the following section we discuss the numerical methods that are used to solve this problem.

2.3 Computing issue in joint likelihood

Despite the popularity of the EM algorithm in joint modeling, most of the applications are restricted to the simple parametric joint models, in which the longitudinal process is assumed to be captured well by low-dimensional random effects and event process by low-dimensional regression coefficients. Such constraint is caused by the difficulty in approximating the posterior expectations of the random effects in the E-step of the EM algorithm. The numerical methods applied to approximate these posterior densities and expectations can be divided into three categories. One is the deterministic interpolation techniques using the Gaussian quadrature points (Wulfsohn and Tsiatis, 1997; Song et al., 2002b); another is the random sampling methods based on Monte Carlo Markov Chain (Hsieh et al., 2006; Tseng et al., 2005; Ding and Wang, 2008); the other is Laplace approximation based on Taylor expansion of the integrand (Rizopoulos et al., 2009). In this and the following chapter, we will compare these methods both theoretically and numerically in terms of accuracy and computational time in low-dimensional parametric joint model settings. A newly proposed numerical method, design of interpolation techniques (Joseph, 2012), is introduced to the joint modeling framework as an alternative technique to approximate the posterior expectations.

2.3.1 Gaussian-Hermite quadrature

The most common numerical integration technique used for joint modeling is the Gaussian-Hermite quadrature proposed by Wulfsohn and Tsiatis (1997). The posterior density of the random effects of the ith subject can be written as

$$f(\boldsymbol{b}_i|D_o,\hat{\Omega}) = \frac{f(\boldsymbol{b}_i, V_i, \Delta_i|w_i, \boldsymbol{z}_i, \hat{\Omega})}{f(V_i, \Delta_i|w_i, \boldsymbol{z}_i, \hat{\Omega})} = \frac{f_{V_i,\Delta_i|\boldsymbol{b}_i} f_{\boldsymbol{b}_i|W_i}}{\int f_{V_i,\Delta_i|\boldsymbol{b}_i} f_{\boldsymbol{b}_i|W_i} d\boldsymbol{b}_i},$$

where $f_{\boldsymbol{b}_i|W_i}$ is the posterior density of $\boldsymbol{b}_i$ given w_i. Since both $\{W_i|\boldsymbol{b}_i\}$ and $\boldsymbol{b}_i$ are normally distributed by assumption, the posterior distribution of $\boldsymbol{b}_i|W_i$ is

$$\boldsymbol{b}_i|W_i \sim N(\boldsymbol{\mu} + D_{21}D_{11}^{-1}(W_i - \boldsymbol{\rho}(\boldsymbol{t}_i)^T\boldsymbol{\mu}),\ D_{22} - D_{21}D_{11}^{-1}D_{12}),$$

where $D_{11} = \boldsymbol{\rho}(\boldsymbol{t}_i)^T\boldsymbol{\Sigma}\boldsymbol{\rho}(\boldsymbol{t}_i) + \sigma^2 I_{n_i}$, $D_{21} = \boldsymbol{\Sigma}\boldsymbol{\rho}(\boldsymbol{t}_i)$, $D_{12} = D_{21}^T$, and $D_{22} = \boldsymbol{\Sigma}$. Thus the conditional expectation of $g(\boldsymbol{b}_i)$ can be written as

$$E(g(\boldsymbol{b}_i)) = \frac{\int g(\boldsymbol{b}_i) f_{V_i,\Delta_i|\boldsymbol{b}_i} f_{\boldsymbol{b}_i|\boldsymbol{W}_i} d\boldsymbol{b}_i}{\int f_{V_i,\Delta_i|\boldsymbol{b}_i} f_{\boldsymbol{b}_i|\boldsymbol{W}_i} d\boldsymbol{b}_i}. \tag{2.37}$$

In the Gaussian quadrature method, in order to get rid of the dependence structure of the random effects, $\boldsymbol{b}_i$ is transformed into $\boldsymbol{b}_i^* = q(\boldsymbol{b}_i)$ such that $\boldsymbol{b}_i^* \sim N(\boldsymbol{0}, \frac{1}{2}I)$. Now that $\boldsymbol{b}_i^*$ has independent elements, the conditional expectation can be approximated by evaluating the functions at the pre-specified Gaussian-Hermite quadrature points and calculating their weighted sum:

$$E(g(\boldsymbol{b}_i)) \approx \frac{\sum_{s=1}^{m} \sum_{t=1}^{m} g(q^{-1}(\boldsymbol{b}_i^*)) f_{V_i,\Delta_i|q^{-1}(\boldsymbol{b}_i^*)} w_s w_t}{\sum_{s=1}^{m} \sum_{t=1}^{m} f_{V_i,\Delta_i|q^{-1}(\boldsymbol{b}_i^*)} w_s w_t}, \tag{2.38}$$

where $\boldsymbol{b}_i^*$ takes m abscissa values on each of its two coordinates. w_s and w_t are the associated weights for the sth point of the first coordinate and the tth point of the second coordinate.

As a fundamental technique used for numerical integration in joint modeling, the Gaussian-Hermite quadrature method is implemented in the PROC NLMIXED procedure in SAS and the JM package in R (Rizopoulos, 2010, 2012), both of which can be used to fit low-dimensional parametric joint models. However, as the dimension of random effects increases, the number of quadrature points grows exponentially. This results in the soaring computing time and the potential failure to converge (Joseph, 2012).

Let $M = s \times m$ be the total number of interpolating points, with s being the number of integration dimension, and m being the number of points on each dimension. The convergence rate of the Gaussian quadrature method is of order $O(M^{-1})$, and thus higher accuracy can be achieved by adding more interpolating points. Although the Gaussian quadrature method with $m = 2$ worked well for the data in Wulfsohn and Tsiatis (1997), the JM package in R uses $m = 35$ as default. In my simulation study with two-dimensional $\boldsymbol{b}_i$, the performance of the standard Gaussian quadrature method of (2.38) with $m = 3$ is far from satisfactory. It performs worse than other numerical integration methods in terms of estimating accuracy and stableness. It is unstable because the numerator and the denominator in (2.38) are approximated separately and the denominator is subject to problems: the density function $f_{V_i,\Delta_i|q^{-1}(\boldsymbol{b}_i^*)}$ contains exponential terms, and can thus get to extremely large or small values during parameter updating for some outlier observations. This might hinder the calculation of $E(g(\boldsymbol{b}_i))$ and hence need special attention in programming.

2.3.2 Monte Carlo Markov Chain

Monte Carlo Markov Chain (MCMC) is another popular method used to evaluate the conditional expectations in the E-step in joint modeling (Henderson et al., 2000). The key is to sample $\boldsymbol{b}_i$ from appropriate posterior densities $f_{\boldsymbol{b}_i|D_o,\hat{\Omega}}$ that does not have a closed form. To make things easier, Tseng et al. (2005), Ding and Wang (2008) proposed to generate random samples $\{\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^M\}$ of size M for $\boldsymbol{b}_i$ from the partial posterior density $f_{\boldsymbol{b}_i|\boldsymbol{W}_i}$ and approximate $E(g(\boldsymbol{b}_i))$ by

$$E(g(\boldsymbol{b}_i)) \approx \frac{\sum_{l=1}^{M} g(\boldsymbol{b}_i^l) f_{V_i,\Delta_i|\boldsymbol{b}_i^l}}{\sum_{l=1}^{M} f_{V_i,\Delta_i|\boldsymbol{b}_i^l}}. \tag{2.39}$$

To ensure the accuracy of the approximation, M has to be large enough, usually taking the value of several hundreds or even thousands. Thus the estimating accuracy is abstained at the cost of computational time.

Instead of generating the MCMC samples from the partial posterior distribution $f_{\boldsymbol{b}_i|\boldsymbol{W}_i}$, we can also draw the sample $\{\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^M\}$ directly from the posterior distribution $f_{\boldsymbol{b}_i|\boldsymbol{W}_i,V_i,\Delta_i}$ by applying Metroplis-Hasting algorithm to the unnormalized posterior density

$$h(\boldsymbol{b}_i) = f_{V_i,\Delta_i|\boldsymbol{b}_i} \cdot f_{W_i|\boldsymbol{b}_i} \cdot f_{\boldsymbol{b}_i}.$$

Thus the posterior expectations of $g(\boldsymbol{b}_i)$ is approximated by

$$E(g(\boldsymbol{b}_i)) = \frac{\int g(\boldsymbol{b}_i) h(\boldsymbol{b}_i) d\boldsymbol{b}_i}{\int h(\boldsymbol{b}_i) d\boldsymbol{b}_i} \approx \frac{1}{M} \sum_{l=1}^{M} g(\boldsymbol{b}_i^l). \tag{2.40}$$

Since MCMC sampling is an updating procedure, a good choice of initial values contributes to a shorter "burn in" period, and reduces computational time. Since the complete log likelihood $l(\boldsymbol{b}_i) = \log(h(\boldsymbol{b}_i))$ is quadratic in $\boldsymbol{b}_i$ as N_i increases, the mode of $l(\boldsymbol{b}_i)$, which can be obtained by Newton-Raphson algorithm, serves as a good initial value for the Metroplis-Hasting sampling scheme.

The idea of using MCMC to approximate the posterior expectations of $g(\boldsymbol{b}_i)$ is straightforward and easy to implement in R with the package MHadaptive. The computing procedure is quite stable and can yield accurate estimates provided that the posterior samples are generated properly. However, even with good initial values, this approach still needs at least hundreds of sampling points, thus leading to much longer computational time compared with other methods where a few points are sufficient for the low-dimensional cases. Most of the time is spent in generating the posterior samples sequentially by MCMC schemes rather than calculating $g(\cdot)$ at each sample point.

2.3.3 Fully exponential Laplace

Recall that Laplace approximation is a mathematical technique to approximate integrals of the form

$$\int e^{af(x)}dx,$$

where a is a large number, and $f(x)$ is a twice-differentiable function. If $f(x)$ has a unique global maximum at x_0, then by Taylor expansion, for x close to x_0 we have

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + O((x-x_0)^3) \\ &\approx f(x_0) - \frac{1}{2}|f''(x_0)|(x-x_0)^2, \end{aligned} \tag{2.41}$$

because $f'(x_0) = 0$ and the second derivative is negative at the global maximum $f(x_0)$. This enables the approximation to the original integral:

$$\int e^{af(x)}dx \approx e^{af(x_0)} \int e^{-a|f''(x_0)|(x-x_0)^2/2}dx.$$

This approximated quantity has a Gaussian integral form with mean x_0 and variance $\frac{1}{a|f''(x_0)|}$, and can thus be approximated by

$$\int e^{af(x)}dx \approx \sqrt{\frac{2\pi}{a|f''(x_0)|}}e^{af(x_0)} \text{ as } a \to \infty. \tag{2.42}$$

This expansion is accurate to the order $O(1/a)$, and thus a large a is usually required to produce satisfactory results. In practice, a is related to the sample size of the data.

Rizopoulos et al. (2009) applied the idea of Laplace approximation to joint modeling. To see the link between Laplace approximation and joint modeling framework, consider again the posterior expectation in (2.40):

$$E(g(\boldsymbol{b}_i)) = \frac{\int g(\boldsymbol{b}_i)h(\boldsymbol{b}_i)d\boldsymbol{b}_i}{\int h(\boldsymbol{b}_i)d\boldsymbol{b}_i} = \frac{\int g(\boldsymbol{b}_i)e^{N_i l(\boldsymbol{b}_i)}d\boldsymbol{b}_i}{\int e^{N_i l(\boldsymbol{b}_i)}d\boldsymbol{b}_i}, \tag{2.43}$$

where $l(\boldsymbol{b}_i) = \frac{1}{N_i}\log(h(\boldsymbol{b}_i))$, and $\int e^{N_i l(\boldsymbol{b}_i)}d\boldsymbol{b}_i$ can be approximated using expression (2.42) at the mode $\hat{\boldsymbol{b}}_i$ of $l(\boldsymbol{b}_i)$ in both the numerator and denominator. This technique is referred to as "standard Laplace approximation" (Tierney et al., 1989) and the approximated posterior expectation is of the form

$$E(g(\boldsymbol{b}_i)) = g(\hat{\boldsymbol{b}}_i) + O(n_i^{-1}).$$

This approximation is not optimal in the context of longitudinal data since the observation number N_i can be quite small for many subjects and thus leads to the inaccurate estimates. To avoid this problem, instead of using "standard Laplace approximation", Rizopoulos et al. (2009) proposed to use "fully exponential Laplace approximation" (Tierney and Kadane, 1986) , which is to apply the standard Laplace methods in both the numerator and the denominator of (2.40), respectively. That is

$$E(g(\boldsymbol{b}_i)) = \frac{\int g(\boldsymbol{b}_i)h(\boldsymbol{b}_i)d\boldsymbol{b}_i}{\int h(\boldsymbol{b}_i)d\boldsymbol{b}_i} = \frac{\int e^{N_i l^*(\boldsymbol{b}_i)}d\boldsymbol{b}_i}{\int e^{N_i l(\boldsymbol{b}_i)}d\boldsymbol{b}_i}, \tag{2.44}$$

where $l(\boldsymbol{b}_i)$ is defined the same as in (2.43), and $l^*(\boldsymbol{b}_i) = \frac{1}{N_i}\{\log g(\boldsymbol{b}_i) + \log(h(\boldsymbol{b}_i))\}$. The "fully exponential Laplace approximation" differentiates from the "standard Laplace approximation" in that it evaluates the numerator and denominator separately at $\hat{\boldsymbol{b}}_i^*$ and $\hat{\boldsymbol{b}}_i$, which are the two modes of $l^*(\boldsymbol{b}_i)$ and $l(\boldsymbol{b}_i)$, respectively. In addition, it also requires $g(\cdot)$ to be positive. Tierney et al. (1989) proved that if the sequence of $g(\hat{\boldsymbol{b}}_i)$ is bounded away from 0, then $\hat{\boldsymbol{b}}_i^* - \hat{\boldsymbol{b}}_i = O(N_i^{-1})$. This results in the cancellation of $O(N_i^{-1})$ terms in the standard version of Laplace approximation, and the approximation accuracy is improved to order $O(N_i^{-2})$.

To extend $g(\cdot)$ to general forms, the approximation is applied to the cumulant-generating function $\log(Ee^{\boldsymbol{c}^T g(\boldsymbol{b}_i)})$. Then the required expectations can be calculated by $E(g(\boldsymbol{b}_i)) = \partial \log(Ee^{\boldsymbol{c}^T g(\boldsymbol{b}_i)})/\partial \boldsymbol{c}^T|_{c=0}$. See Theorem 2 in Tierney et al. (1989) for the specified terms of $O(N_i^{-2})$ in the approximations.

This idea was successfully adapted by Rizopoulos et al. (2009) to the joint modeling framework to estimate the posterior expectations in EM algorithms, and yield good estimating results for the subjects with even a few longitudinal observations. To obtain the estimates, they first calculate the global maximum, i.e., the posterior model $\hat{\boldsymbol{b}}_i$ of $h(\boldsymbol{b}_i)$ by Newton-Raphson algorithm. And the approximation at the mode is

$$E(g(\boldsymbol{b}_i)) = g(\hat{\boldsymbol{b}}_i) - \frac{1}{2}\mathrm{tr}(\Gamma) + O(1/N_i^2), \tag{2.45}$$

$$\begin{aligned} Var(g(\boldsymbol{b}_i)) &= g'(\hat{\boldsymbol{b}}_i)^T D_i^{-1} g'(\hat{\boldsymbol{b}}_i) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(-\Gamma\Gamma^T + D_i^{-1}\frac{\partial^2}{\partial \boldsymbol{c}^T \boldsymbol{c}} D_i^{(c)}|_{(c,b)=(0,\hat{b}_i)}\right) + O(1/N_i^3) \end{aligned} \tag{2.46}$$

where $\Gamma = D_i^{-1}\{\partial D_i^{(c)}/\partial c^T\}|_{(c,b)=(0,\hat{b}_i)}$, $D_i^{(c)} = -\partial^2[\log(h(b_i)) + c^T g(b_i)]/\partial b_i^T b_i$ is the negative second derivative, and $D_i = D_i^{(c)}|_{(c,b)=(0,\hat{b}_i)}$. For more details, please see Rizopoulos et al. (2009).

Compared with the Gaussian quadrature and the MCMC method, fully exponential

Laplace approximation is much faster in computing time because there is no need to generate either the interpolation or the sampling points. Since $h(\boldsymbol{b}_i)$ is quadratic and unimodal with large n_i in the joint modeling framework, the approximation using only one point, the posterior mode, performs well in estimating accuracy and numerical robustness. Thus, the fully exponential Laplace approximation has the potential to be applied to the large-dimensional cases of joint modeling. The main drawbacks of this technique, however, are multifold. First, unlike the numerical integration techniques and Monte Carlo related methods, the estimating accuracy of Laplace approximation does not rely on the evaluating points. Instead, it depends on the sample size of the data, which cannot be controlled. Hence the estimating errors are difficult to reduce and the approximating accuracy cannot be obtained at an arbitrary level as for the other methods introduced above. Second, the computation of Γ is very complicated because both $\{\partial D_i^{(c)}/\partial c^T\}|_{(c,b)=(0,\hat{b}_i)}$ and $\frac{\partial^2}{\partial \boldsymbol{c}^T \boldsymbol{c}} D_i^{(c)}|_{(c,b)=(0,\hat{b}_i)}$ involve the calculation of large-dimensional tensors. This may cause trouble when the dimension of $\boldsymbol{b}_i$ increases. In addition, Laplace approximation works well only for the unimodal symmetric functions. Although $h(\boldsymbol{b}_i)$ is quadratic with large N_i, this may not be the case when N_i is small, and the distribution is not necessarily symmetric. Therefore more general methods are needed to handle the expensive posterior densities.

2.4 Design of experiments-based interpolation technique

As stated in the previous section, computation of posterior expectations is a fundamental problem in the application of the EM algorithm in joint modeling, and this task becomes more challenging as the dimension of the random effects $\boldsymbol{b}_i$ increases. By comparing and summarizing the computing techniques used for joint modeling in Section 2.3, we find that most of the previously introduced approaches including the Gaussian Quadrature method, the MCMC approach and fully exponential Laplace approximation, are inadequate for large-dimensional and more complex joint modeling settings, for example, the nonparametric joint models. Hence a more efficient and stable computing method is needed if we want to extend joint modeling into a more flexible and sophisticated framework. In this section, we introduce a new computing approach into joint modeling. Originally proposed for Bayesian computation, this method is demonstrated via our simulation study to be not only capable of solving the computational problems of joint models, but also easy to implement in practice.

Recently, Joseph (2012) proposed a relatively new method to approximate the "expensive conditional densities" in Bayesian computation. It is named as design of experiments-based interpolation techniques, or DoIt. In his study, Joseph (2012) demonstrated the appealing features of DoIt by comparing its performance with other popular algorithms, and proved this technique works well in estimating posterior densities for complex hierarchical models where large-dimensional parameters are involved. Before introducing DoIt into joint modeling context, we first explain the approximation idea behind it.

The DoIt method borrows and extends the idea of Laplace approximation which approximates the posterior densities of interest via normal distributions. As mentioned previously, in Laplace approximation the posterior density $f_{\boldsymbol{b}_i|\boldsymbol{W}_i,T_i,\Delta_i}(\boldsymbol{b}_i)$ is approximated by the normal density $\phi(\boldsymbol{b}_i;\hat{\boldsymbol{b}}_i,\boldsymbol{D}_i^{-1})$, where $\hat{\boldsymbol{b}}_i$ is the mode of the density and $\boldsymbol{D}_i$ is the Fisher information matrix evaluated at the mode as defined in (2.46). Instead of using a single normal distribution for approximation, DoIt approximates the posterior densities of random effects $\boldsymbol{b}_i$ given the observed $\{\boldsymbol{W}_i(t),T_i,\Delta_i\}$ by the weighted sum of a sequence of normal densities with the means at a set of pre-specified evaluation points $(\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_M)$ of size M:

$$f_{\boldsymbol{b}_i|\boldsymbol{W}_i,T_i,\Delta_i}(\boldsymbol{b}_i)\approx\frac{1}{\sum_{l=1}^{M}c_l}\sum_{l=1}^{M}c_l\phi_l(\boldsymbol{b}_i), \tag{2.47}$$

where $\phi_l(\boldsymbol{b}_i)=\phi(\boldsymbol{b}_i;\boldsymbol{\nu}_l,\boldsymbol{D}_{\boldsymbol{i}}^{\boldsymbol{-1}})$ denote the normal density of $\boldsymbol{b}_i$ with mean $\boldsymbol{\nu}_l$ and variance $\boldsymbol{D}_i^{-1}$. c_l is the weight associated with $\phi_l(\cdot)$ and can be calculated by solving the linear equations

$$\boldsymbol{Q}\boldsymbol{c}=\boldsymbol{h},$$

where $h(\boldsymbol{b}_i)\propto f(\boldsymbol{W}_i,T_i,\Delta_i|\boldsymbol{b}_i)f(\boldsymbol{b}_i)$ is the unnormalized posterior densities of $\boldsymbol{b}_i$, and $\boldsymbol{h}=(h(\boldsymbol{\nu}_1),\ldots,h(\boldsymbol{\nu}_M))$ is a vector of $h(\cdot)$ taking values at the evaluation points $(\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_M)$. $\boldsymbol{Q}$ is an $M\times M$ matrix with ijth element being the unnormalized normal density

$$q(\boldsymbol{\nu}_i;\boldsymbol{\nu}_j,\boldsymbol{D}_i^{-1})=\exp\{-\frac{1}{2}(\boldsymbol{\nu}_i-\boldsymbol{\nu}_j)^T\boldsymbol{D}_i(\boldsymbol{\nu}_i-\boldsymbol{\nu}_j)\}.$$

Joseph (2012) pointed out that since $q(\boldsymbol{\nu};\boldsymbol{u},\boldsymbol{D}_i^{-1})$ is a positive definite function, $\boldsymbol{Q}^{-1}$ exists, provided $\boldsymbol{\nu}_i\neq\boldsymbol{\nu}_j$ for all i and j. Thus a unique solution of $\hat{\boldsymbol{c}}=\boldsymbol{Q}^{-1}\boldsymbol{h}$ is guaranteed.

The locations of $(\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_M)$ are determined by the pre-specified space-filling design points transformed to the parameter space of posterior distribution via $\hat{\boldsymbol{b}}_i$ and $\boldsymbol{D}_i^{-1}$. The desired level of accuracy can be obtained by increasing the size of M. Refer to Joseph (2012) for more details.

With the capability of approximating expensive posterior densities, DoIt can also be used to approximate the conditional expectations of real value functions of

$\boldsymbol{b}_i$, $g(\boldsymbol{b}_i)$

$$E\{g(\boldsymbol{b}_i)\} \approx \frac{1}{\sum_{l=1}^{M} c_l} \sum_{l=1}^{M} c_l E_l\{g(\boldsymbol{b}_i)\}, \tag{2.48}$$

where $E_l\{g(\boldsymbol{b}_i)\}$ is the expectation of $g(\boldsymbol{b}_i)$ with respect to the normal distribution $N(\boldsymbol{\nu}_l, \boldsymbol{D}_i^{-1})$.

Therefore, the DoIt approach well satisfies the need of the EM algorithm in joint modeling to calculate conditional expectations of $g(\boldsymbol{b}_i)$ specified in (2.36) in the E-step. By introducing the idea of DoIt into the computation of joint modeling, we find it produces good parameter estimates, and thus can be taken as an alternative to the existing computing techniques for joint models.

In addition, compared with the Gaussian Quadrature method, although the DoIt approximation of (2.48) also requires pointwise evaluation at M deterministic points, it does not suffer from "curse of dimensionality" like the Gaussian quadrature method because the number of M does not grow as fast as that for the Gaussian quadrature when the dimension increases. A common rule of thumb in the computer experiments' literature is to use $M = 10 \times dim(\boldsymbol{b}_i)$ (Loeppky et al., 2009). Joseph (2012) suggests $M = 50 \times dim(\boldsymbol{b}_i)$ for higher accuracy. However, we find with our simulation study that for the distributions with good shape, estimating accuracy can be achieved by small number of deterministic points. Compared with the MCMC approach, DoIt is much faster in computation in that the number of evaluation points is smaller, and the same set of basic design points is generated only once for all the subjects and then transformed according to each individual posterior distribution. In contrast, for MCMC, different sets of random samples have to be generated for different subjects and for each subject, the sample points are updated iteratively. This costs huge amount of computational time.

Considering all these facts, DoIt is a promising candidate for approximating the posterior expectations $E\{g(\boldsymbol{b}_i)\}$ in large dimensional cases. It is more flexible than the fully exponential Laplace method because it does not require the unimodal and symmetric assumption of the posterior distributions. Moreover, the computation is straightforward since no tensor calculation is involved as in the fully exponential Laplace.

Similar to other methods that depend on evaluation points, DoIt has the nice feature of being able to reduce the estimation error arbitrarily small by adding more points, i.e., increasing M (see Theorem 1 of Joseph (2012) for details). On the other hand, the implementation of DoIt requires specifying the deterministic points $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ by appropriate experimental designs. Different designs lead to different rates of convergence. Joseph (2012) proposed to use a space-filling design scheme such as minimax Latin Hypercube Design (MmLHD), followed by a sequential design to correctly locate the high-probability areas of the sampling space. Fang and Wang (1994) focused on number-theoretic-related design method (NTM) and provided theoretical and numerical comparisons among various space filling designs in terms of convergence rates. They pointed out that the convergence rate of LHD is $O(M^{-1/2})$, which is the same order as the application of simple random sample technique, and the only improvement is the variance of limiting distribution. Therefore from the viewpoint of numerical integration, LHD and its improvement versions still belong to the variance reduction technique of Monte Carlo method. By comparison, the convergence rate for the quadrature formula generated by some NTM, such as good lattice points set, is $O(M^{-1}\log^s M)$, where s is the number of dimension, if the integrand is a function with bounded total variation. Thus, NTM, when properly adjusted, might be more efficient than LHD in estimating the posterior distributions in joint modeling, especially when the integration dimension s is not so large. In this dissertation, we apply MmLHD as the design scheme for DoIt implementation since there is an existing R package *lhs*, which can be conveniently used to generate MmLHD samples.

2.5 Varying-coefficient models

In this section we review varying-coefficient models and the related model estimation procedures as a preparation for extending the parametric joint model to the nonparametric setting.

Varying-coefficient models were proposed by Hastie and Tibshirani (1993) as a form of nonparametric regression models and a generalization of additive models. They are defined as follows

$$Y = \beta_0(U) + \beta_1(U)X_1 + \cdots + \beta_p(U)X_p + \epsilon, \tag{2.49}$$

where ϵ is the random error satisfying $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. Compared with conventional linear regression models, the coefficients of the above models are replaced by smoothing nonparametric functions of other factors. This allows the predictors' effects to vary with other factors such as time. The varying-coefficient models cover a lot of special cases. For example, when $\beta(U)$ is a linear function in U, i.e., $\beta(U) = \beta_0 + \beta_1 U$, (2.49) reduces to linear regression with the interaction terms $\beta_1 UX$.

Various methods have been proposed to estimate the varying-coefficient models. Hastie and Tibshirani (1993) proposed to estimate $\beta_j(u)$ via smoothing splines and the penalized least squares approach, assuming that all the coefficient functions have the same degrees of smoothness. Fan and Zhang (1999) relaxed this assumption by allowing the coefficient functions to have different degrees of smoothness and proposed a two-step estimating approach to achieve the optimal convergence rate. Cai et al. (2000) extended the modeling framework of (2.49) to the generalized varying-coefficient models, accounting for categorical and count responses:

$$E\{y(t)|x(t)\} = \mu\{\beta_0(t) + \beta_1(t)x_1(t) + \cdots + \beta_d(t)x_d(t)\}.$$

The authors estimated the coefficient functions using local polynomial techniques. The asymptotic properties are established and a goodness-of-fit test is derived from a nonparametric maximum likelihood ratio type of test to detect whether the coefficient functions for certain covariates are constant over time and statistically significant in the model.

Varying-coefficient models have been extensively studied due to their flexible frameworks, easy interpretations, and most of all, the wide applications to various data such as longitudinal data (Hoover et al., 1998), ecological data (Cai et al., 2000) and nonlinear time series (Chen and Tsay, 1993; Cai et al., 2000). In this section we focus on their applications in longitudinal data analysis.

2.5.1 Time-varying effect models in longitudinal data

Longitudinal data have been an active research topic for decades. Before the introduction of varying-coefficient models (Hastie and Tibshirani, 1993; Faraway, 1997; Hoover et al., 1998; Wu et al., 1998), the parametric regression model, hierarchical linear models (HLM), have long been used, and is still in use, to analyze longitudinal data (Raudenbush, 2002; Diggle et al., 2002; Bollen and Curran, 2006; Hedeker and Gibbons, 2006). These models are designed for clustered data to account for both within- and between-subject correlations. Longitudinal data have the cluster structure since the observations are nested in the subjects. Usually the samples collected within a subject are correlated and samples between subjects are independent. Before diving into the time-varying effect models, we briefly review HLM, which are parametric and parsimonious in most cases. Usually HLM involves two levels of regression. The first level is of the typical regression form

$$Y_{ij} = \beta_{0i} + \beta_{1i} X_{ij} + e_{ij},$$

where the subscript ij refers to the jth observation of the ith individual, and this level of regression accounts for the within-subject variation. The second level of regression is on the coefficients β's and accounts for the between-subject variation:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} Z_i + u_{0i};\ \beta_{1i} = \gamma_{10} + u_{1i}.$$

In the above equations, γ_{00} refers to the overall intercept, representing the grand mean of the response variable across all the individuals when all the predictors are equal to 0. Z_i is the subject-specific predictor, and γ_{01} is the corresponding regression coefficient. u_{0i} refers to the random error component for the deviation of the intercept of the ith subject from the overall intercept. γ_{10} is the overall regression coefficient between the response and the first level predictor X_{ij}, and u_{1i} is the random error component for this slope representing the deviation of the ith subject. Usually $\boldsymbol{e}_i$ and $\boldsymbol{u}_i = (\boldsymbol{u}_{0i}, \boldsymbol{u}_{1i})$ are assumed to be normally distributed and independent of each other. Based on these assumptions, the conditional distribution $f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta}_i)$, $f(\boldsymbol{\beta}_i|Z_i, \boldsymbol{u}_i)$ and the prior distribution $f(\boldsymbol{u}_i)$ can be specified. The joint likelihood also involves the integral over the subject-level random components $\boldsymbol{u}_i$'s. Multiple software packages, such as PROC MIXED in SAS and *lme* in R, have been developed to solve the problem in practice, and the models have been extended to incorporate the generalized linear models with binary or count response.

The varying-coefficient models were introduced into longitudinal data as the time-varying effect model of the form (Hoover et al., 1998)

$$Y(t) = \beta_0(t) + \beta_1(t)X_1(t) + \cdots + \beta_p(t)X_p(t) + \epsilon(t), \tag{2.50}$$

where $Y(t)$ is the longitudinal response, and $\{X_k(t) : k = 1, \ldots, p\}$ are the longitudinal covariates. Terms $\beta_1(t), \ldots, \beta_p(t)$ are effects allowed to vary with time, and $\epsilon(t)$ is a zero mean stochastic process whose variance may change over time, while the correlation structure is invariant over time. Time-varying effect models are flexible in that they represent effects changing over time smoothly without assuming a specific functional relationship, such as linear or quadratic, between the effects and time. Note that the main difference between (2.49) and (2.50) is that the response, covariates and the error terms all change with time in (2.50), accounting for the dynamic process of the longitudinal data. Hoover et al. (1998) proposed to estimate $\boldsymbol{\beta}(t)$ using two nonparametric smoothing techniques: smoothing splines and locally weighted polynomials. They further derived the asymptotic properties for the established kernel estimators as a special case of local polynomials, and applied the model to the empirical HIV data. Wu et al. (1998) carefully studied the asymptotic distributions of the kernel estimates and constructed a class of approximate pointwise and simultaneous confidence sets. Huang et al. (2002, 2004) proposed to use regression splines to estimate the parameters of a varying-coefficient model with repeated measurements. They established the asymptotic consistency for these estimators and further derived the bootstrap confidence regions to draw inference. Eubank et al. (2004) developed "Bayesian" confidence intervals for the estimated coefficient curves based on smoothing spline techniques.

One extension of model (2.50) is the generalized varying-coefficient mixed-effects model considered by Zhang (2004). It is based on the nonparametric mixed-effects model (Rice and Wu, 2001; Wu and Zhang, 2002), and has the form

$$g\{E(Y(t)|X(t), \boldsymbol{Z}, \boldsymbol{b}_i)\} = \beta_0(t) + \beta_1(t)X(t) + \boldsymbol{Z}^T\boldsymbol{b}_i,$$

where $g(\cdot)$ is a known monotone and differentiable link function, $\beta_0(t)$ and $\beta_1(t)$ are the time-varying fixed effects and $\boldsymbol{b}_i$ are the random effects. Zhang (2004) leveraged the double penalized quasi-likelihood (DPQL) approach of Lin and Zhang (1999) to solve the intractable integration involved in evaluating the quasi-likelihood function to estimated the coefficient functions. He also developed a scaled chi-squared test for hypothesis testing whether an underlying varying coefficient function is a polynomial of a certain degree.

Fan et al. (2007) extended the varying-coefficient model (2.50) to a more generalized form of semi-varying effect partially linear model

$$Y(t) = \beta_1(t)X_1(t) + \dots \beta_p(t)X_p(t) + \alpha_1 Z_1(t) + \cdots + \alpha_q Z_q(t) + \epsilon(t),$$

where $Y(t)$, $X(t)$'s and $\beta(t)$'s are the same as in (2.50). $Z(t)$'s are another set of time covariates with unknown constant coefficients α's. This is a unified model that covers most of the existing regression models in the literature of longitudinal data including the nonparametric regression models (Lin and Carroll, 2001a,b; Wang, 2003; Rice and Wu, 2001; Wu and Zhang, 2002) and the partial linear model (Fan and Li, 2004). Fan et al. (2007) established kernel estimators for the nonparametric variance function, and proposed to estimate the parameters in correlation structure by either a quasi-likelihood approach or a minimum generalized variance method. They further developed a profile weighted least squares approach to estimate the regression coefficients and derived their corresponding asymptotic properties.

Parallel to the aforementioned literature of regression models with continuous response, another branch of studies focuses on regression models with discrete response, where generalized estimating equations (GEE) are usually used for parameter estimation. The GEE method was first proposed by Zeger and Liang (1986), and extended by researchers to various circumstances (Lin and Carroll, 2000; Qu et al., 2000; Wang, 2003). Starting with the parametric regression models, GEE only depends on the conditional mean $\mu_{ij} = E(y_{ij}|x_{ij})$ and variance $\sigma_{ij}^2 = \mathrm{Var}(y_{ij}|x_{ij}) = \phi V(\mu_{ij})$ which may contain a complex covariance structure with both between- and within-subject correlations. Under the framework of generalized linear models, assume that the mean μ_{ij} depends on x_{ij} through a known canonical link $\mu(\cdot)$,

$$\mu_{ij} = \mu\{\theta(x_{ij})\},$$

where $\theta(\cdot)$ is an unknown smooth function. In the conventional GEE (Zeger and

Liang, 1986), $\theta(x)$ is a linear function of x

$$\theta(x) = \beta_0 + \beta_1 x \triangleq \boldsymbol{g}(x)^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\boldsymbol{g}(x) = (1, x)^T$. $\boldsymbol{\beta}$ are estimated by solving the equations:

$$\sum_{i=1}^{n} \boldsymbol{G}_i^T \Delta_i \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\boldsymbol{G}_i = (\boldsymbol{g}(x_{i1}), \ldots, \boldsymbol{g}(x_{in_i}))^T$, $\Delta_i = diag\{\mu'\{\boldsymbol{g}^T(x_{ij})\boldsymbol{\beta}\}\}$. $\mu'(\cdot)$ is the first order derivative of $\mu(\cdot)$. $\boldsymbol{V}_i = \boldsymbol{A}_i^{-1/2} \boldsymbol{R} \boldsymbol{A}_i^{-1/2}$ is the covariance matrix with $\boldsymbol{A}_i$ being a $n_i \times n_i$ diagonal matrix with the jth diagonal element $V(\mu_{ij})$ and a given working correlation $\boldsymbol{R}$. GEE strategy has been extended to the nonparametric regression by Lin and Carroll (2000). See Fan and Li (2006) and Dziak et al. (2008) for the detailed review on GEE and the related techniques.

Qu et al. (2000) introduced the quadratic inference function (QIF) as an improvement of GEE. This approach uses data-driven weights that assign less weight to the estimating equations with larger variances. This study shows that if the working correlation is correctly specified then the QIF estimators have the asymptotic variances as low as GEE. If the working structure is incorrect, the QIF estimators are still optimal among the same linear class of estimating equations, whereas the GEE estimators with the same working correlation is not. This approach has also been applied to the nonparametric longitudinal data (Qu and Li, 2006; Dziak et al., 2008).

Based on the fundamental work of QIF, Qu and Li (2006) studied time-varying effect models under the generalized linear model framework and developed an efficient estimation procedure via penalized quadratic inference functions. The estimators inherit the advantage of QIF over conventional GEE estimators in terms of estimation efficiency. The authors further proposed a unified and efficient nonparametric hypothesis testing procedure and demonstrated the resulting test statistics have an asymptotic chi-squared distribution.

2.5.2 Cox model with time-varying effects

The Cox proportional hazard model (Cox, 1972) has been widely used to explore the relationship between the event time and the time-invariant covariates. In the conventional form of the Cox model, the coefficients are assumed to be constant, thus guaranteeing the hazard is proportional across time. However, this constant assumption may fail to capture the real cases when the covariates effects change over time. Zucker and Karr (1990) proposed the Cox model with time-varying covariate effects

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp\{\boldsymbol{\beta}(t)^T\boldsymbol{x}\}, \tag{2.51}$$

where the time-varying coefficients $\boldsymbol{\beta}(t)$ are assumed to be smooth functions of time and need to be estimated nonparametrically. Zucker and Karr (1990) proposed to estimate $\boldsymbol{\beta}(t)$ using penalized partial likelihood, and derived the weak uniform consistency and pointwise asymptotic normality of the estimators under certain regularity conditions. As an extension of this work, Hastie and Tibshirani (1993) developed a specialized algorithm based on natural spline basis to maximize the penalized partial likelihood. Murphy and Sen (1991) leveraged sieve estimation procedure for the model (2.51) and further established the sampling properties of the estimators. Gray (1992) considered using smoothing splines to estimate $\boldsymbol{\beta}(t)$ and proposed corresponding test statistics.

In recent studies, (2.51) is more frequently estimated by local partial likelihood. Cai and Sun (2003) carefully studied the pointwise asymptotic properties of the kernel estimators, and used the procedure as a diagnostic tool to uncover time-dependency or departure from the proportional hazard models. Tian et al. (2005) constructed simultaneous confidence bands for the kernel estimators using the "strong approximation techniques" combined with a novel perturbation method, and compared them with the pointwise confidence bands obtained by Cai and Sun (2003). Sun et al. (2009) developed empirical likelihood pointwise and simultaneous confidence bands for the time varying coefficients via local partial likelihood smoothing, and proved they perform better than the pointwise and simultaneous confidence bands in the previous studies (Cai and Sun, 2003; Tian et al., 2005). Fan et al. (2006) proposed a one-step local partial likelihood estimator which is used to facilitate the computation of the procedure and demonstrated to be as efficient as the fully iterated local partial likelihood estimator. They further introduced the penalized local likelihood estimator to select the important covariates into the model.

With the availability of longitudinal measurements, it is natural to extend (2.51) to incorporate the longitudinal covariates. A simplified version of this is

$$\lambda(t|\boldsymbol{x}^H(t)) = \lambda_0(t)\exp\{\boldsymbol{\beta}(t)^T\boldsymbol{x}(t)\}. \tag{2.52}$$

As we introduce in the first section of this chapter, problems occur when the longitudinal covariates are used to predict survival time because the survival models need the entire history of longitudinal processes, whereas the covariate processes themselves are often unobservable, and are intermittently measured at some time points and subject to measurement errors. Therefore, this problem goes back to the framework of joint modeling, and needs to be solved by the analytical approaches reviewed in section 2.2, combined with the smoothing techniques for the nonparametric coefficient functions $\boldsymbol{\beta}(t)$. So far there are few studies considering such a survival model with both time-varying covariates and time-varying effects. This is the main focus of the second half of this project.

2.5.3 Joint models with time-varying effects

Most of the existing approaches of joint modeling are developed for the models with parametric effects in both longitudinal and survival processes. However, the real cases may be more complex. Since the longitudinal covariates themselves vary with time, it is natural that the relationships between the different longitudinal processes and their effects on survival time would also change with time. Correspondingly, a more general form of the longitudinal process specified in (2.5) can be written in the form

$$W_i(t) = X_i(t) + e_i(t), \tag{2.53}$$

where $X_i(t)$ is assumed to be a smooth function of time without any parametric structure. Analogously, the general form of survival process with time-varying effects is

$$\lambda(t; \boldsymbol{Z}_i, \boldsymbol{X}_i^H(t)) = \lambda_0(t)\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta}_Z(t) + \boldsymbol{X}_i(t)^T\boldsymbol{\beta}_X(t)), \tag{2.54}$$

with $\boldsymbol{\beta}_X(t)$ and $\boldsymbol{\beta}_Z(t)$ being the time-varying effects of baseline covariates $\boldsymbol{Z}_i$ and time-varying coefficients $\boldsymbol{X}_i(t)$, respectively.

To the best of our knowledge, there is no literature considering the joint models with time-varying coefficients in both longitudinal and survival models as in equation (2.53) and (2.54). The most relevant study is Song and Wang (2008) which deals with the Cox model with longitudinal covariates and time-varying coefficients in (2.52). In that study, the authors proposed two estimation methods for the regression coefficient function $\boldsymbol{\beta}(t)$. The first is the corrected score approach based on Wang (2006), and the second is the conditional score approach developed by Tsiatis and Davidian (2001). Both of the methods were first developed to estimate the coefficients in parametric joint models. Song and Wang (2008) extended them to the nonparametric setting by constructing the local estimating equations similar to the local partial likelihood score functions in Cai and Sun (2003)

$$l'(\boldsymbol{\beta}) = \sum_{k=1}^{n} \int_0^{\tau} K_h(u-t) \left\{ \tilde{\boldsymbol{X}}_k(u, u-t) - \frac{G'(u, \boldsymbol{\beta})}{G(u, \boldsymbol{\beta})} \right\} dN_k(u),$$

where $\tilde{\boldsymbol{X}}_k(u, u-t) = \boldsymbol{X}_k(u) \otimes (1, u-t)$ with $\otimes$ being Kronecker product. $G_k(u, \boldsymbol{\beta}) = \exp\{\tilde{\boldsymbol{X}}_k(u, u-t)^T \boldsymbol{\beta}\}$ and $G(u, \boldsymbol{\beta}) = \sum_{k=1}^{n} I(V_k \geq u) G_k(u, \boldsymbol{\beta})$.

The corrected score approach substitutes the true covariate process $\boldsymbol{X}_k(t)$ in the function by the corrected least squares estimators; the conditional score approach replaces $\boldsymbol{X}_k(t)$ with a sufficient statistic of the random effects, and thus does not need the distribution specification of the random effects. Song and Wang (2008) proved that the two sets of estimators are asymptotically equivalent and further derived the large sample properties. They demonstrated via simulation that the corrected score method performs better than the conditional score method in practice.

Although Song and Wang (2008) extended the joint modeling framework to cover the varying-coefficient Cox models, they did not put much emphasis on the smooth function of $X_i(t)$ in the longitudinal part. On the other hand, there are joint modeling studies that only focus on the nonparametric form of $X_i(t)$ and leave β's to be constant in Cox models (Brown et al., 2005; Ding and Wang, 2008). Ding and Wang (2008) proposed to model $X_i(t)$ using a single random effect b_i

and a nonparametric underlying function $\mu(t)$ extended by the B-spline basis.

$$X_i(t) = b_i \mu(t),$$

with $E(b_i) = 1$, and

$$\mu(t) = E\{X_i(t)\} \approx \sum_{l=1}^{L} \gamma_l B_l(t).$$

The authors viewed longitudinal data as scattered realizations of functional data, and argued from the perspective of functional principal components that the single random effect b_i corresponds to the first eigenfunction, which explains more than 70% variation of the data, and thus the one random effect is sufficient to capture the subject effects on longitudinal covariates. This single random effect leads to only one dimension of integration and the computational difficulties are avoided.

Brown et al. (2005) proposed a more flexible nonparametric multivariate setting for joint modeling, with all the covariate processes expanded by B-spline basis associated with random effects plus a fixed baseline predictor

$$X_i(t_{ij}) = \sum_{k=1}^{q} \beta_{ik} B_k(t_{ij}) + x_i^T \alpha,$$

where $\beta_{ik} \sim N(b_{0k}, V_{0k})$, and α is a vector of parameters linking the vector of baseline covariates x_i to the longitudinal outcome. To handle the complex computation associated with such setting, the authors adopted the Bayesian approach and Gibbs sampler and adaptive rejection sampling to obtain samples from all the full conditional distributions to estimate the parameters. The number of knots is selected via Deviance Information Criterion (DIC) and Conditional Predictive Ordinate (CPO), which are the model selection criteria that can be easily applied to MCMC samples.

Chapter 3

Joint Likelihood Estimation for Joint Modeling Survival and Multiple Longitudinal Processes: Methodology and Application

3.1 Introduction

In the past two decades, driven by the need to explore the relationship between longitudinal covariate process and time-to-event in biomedical and public health research, statisticians have developed and modified a joint modeling approach to simultaneously analyze the two types of processes via their shared information (Wulfsohn and Tsiatis, 1997; Song et al., 2002b; Hsieh et al., 2006; Faucett and Thomas, 1996; Faucett et al., 1998; Henderson et al., 2000; Tsiatis and Davidian, 2001). Longitudinal and survival data are originally specified by mixed-effects models and hazard models, respectively. However, such naive separate estimation has been proved to yield great bias in regression coefficients due to the ignorance of measurement errors and missing observation at event time (Prentice, 1982; Tsiatis et al., 1995). To address such issue, the earliest joint modeling approach was developed to examining whether CD4 counts serve as a good biomarker for survival time of HIV patients (Tsiatis et al., 1995), and the later improvement in modeling techniques has been applied to more medical and public health studies (Wang and Taylor, 2001; Yu et al., 2004; Liu, 2009; Yu and Ghosh, 2010).

While most joint modeling literature focuses on the setting with a single longitudinal predictor in the survival model, fewer authors consider the model with multiple longitudinal predictors, which is a more flexible and useful model setting (Xu and Zeger, 2001a; Song et al., 2002a; Huang et al., 2001; Ibrahim et al., 2004; Brown et al., 2005; Chi and Ibrahim, 2006; Albert and Shih, 2010; Hatfield et al., 2011). The challenges for joint modeling with multiple longitudinal predictors are twofold. First, the number of random effects grows as the number of longitudinal predictors increases. This would lead to a higher dimension of the multiple integral in the joint likelihood function, which is already difficult to maximize even in the single-covariate case. Second, the correlations among the multiple longitudinal processes over time have to be considered in the model. This complicates the theoretical development.

Most of the related studies (Xu and Zeger, 2001a; Ibrahim et al., 2004; Brown et al., 2005; Chi and Ibrahim, 2006; Hatfield et al., 2011) resort to a Bayesian approach to handle the complex computation issue. Song et al. (2002a) and Albert and Shih (2010) avoided it by not considering the likelihood-based approaches, and instead applied conditional score and modified regression calibration techniques, respectively, to solve the problem. Huang et al. (2001) is the only literature we found that adopted the likelihood-based approach, and used the EM algorithm to maximize the complicated joint likelihood equation. However, the model setting considered in that study was specifically designed for the data of interest, and was quite different from the general joint modeling framework. For example, it employed discrete latent variables in the model and thus avoids the high-dimensional integral problems. Most of this literature focuses on the application of the methodology rather than theory establishment. Among them, Song et al. (2002a) is the only study that considered asymptotic properties by extending the related theories of the conditional score estimators from the single-covariate setting (Tsiatis and Davidian, 2001) to the multiple longitudinal-covariate setting.

Motivated by an empirical analysis of smoking cessation data, we proposed a joint model setting with multiple longitudinal covariate processes.We developed an estimation procedure for the proposed joint model based on the joint likelihood approach, and established the consistency and asymptotic normality of the resulting maximum likelihood estimates (MLE). This model setting is similar to Zeng and Cai (2005), in which the authors established the sampling property of maximum likelihood estimate for joint model with a single longitudinal covariate process. However, the theoretical establishment is more challenging due to the covariance structure among the multiple longitudinal covariates.

The estimation procedure was implemented using a new computing algorithm, the EM-DoIt algorithm, which combines the design of experiments-based interpolation technique (DoIt, Joseph 2012) with the EM algorithm to optimize the objective function with integrals. Since it is challenging to estimate standard errors of the resulting estimate in joint modeling, we propose an estimation method for the standard error by bootstrap method. We conduct Monte Carlo simulations to examine the proposed estimation approach and the computing algorithm. The numerical results show that the proposed method performs very well in terms of numerical and statistical estimation accuracy.

We further applied the proposed estimation method to a smoking cessation study (Piper et al., 2009) to assess the relationship between time to lapse and multiple longitudinal measurements of withdrawal symptoms. We find that the results of joint models with multiple longitudinal covariates offer deeper insights into the applied study than the model with a single longitudinal covariate.

The rest of this chapter is organized as follows. We describe the model setting, introduce the estimation approach, and explain the computing techniques in Section 3.2. Two simulation examples and a real data example are presented in Section 3.3. Conclusion and discussion are given in Section 3.4. Technical proofs of the asymptotic theories can be found in the next chapter.

3.2 Joint likelihood approach

In this section, we first present our model settings, and then propose the maximum likelihood approach to estimating the model parameters. A new numerical integration technique is introduced to deal with computational issues in maximizing the joint likelihood with integrals, and the asymptotic properties of the resulting estimates are established.

3.2.1 Model settings

Suppose that a random sample consists of observations of survival time to an event of interest, multiple longitudinal processes, and time-invariant covariates from n subjects. The multiple longitudinal processes of ith subject are observed at $t_{i1}, \ldots, t_{iN_i}$. The data that we are interested in modeling would be

$$(T_i, \mathbf{Z}_i, \mathcal{X}_i),\ i = 1, 2, \ldots, n, \tag{3.1}$$

where T_i is the time to the event, and $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})^T$ is a vector consisting of q-dimensional time-independent covariates. Denote by $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^T$ the p-dimensional longitudinal covariates recorded at time t, and $\mathcal{X}_i = (\mathbf{X}_i(t_{i1})^T, \ldots, \mathbf{X}_i(t_{i,N_i})^T)^T$ is a $N_i \times p$ matrix representing the longitudinal covariates for the ith subject at all the observational time points.

The major goal of joint modeling is to elucidate the relationship between the survival time T_i and the covariates $\{\mathbf{Z}_i, \mathcal{X}_i\}$. In practice, T_i and $\mathcal{X}_i$ may not be observable due to censoring and measurement error, respectively. In survival data, let C_i be the censoring time for the ith subject. Denote $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, the censoring indicator. Under the right-censoring scheme, the observed survival data for the ith subject is (V_i, Δ_i) instead of T_i. In longitudinal data, instead of the true longitudinal process $\mathcal{X}_i$, what can be observed are the realizations of the error-contaminated processes at times $t_{i1}, \ldots, t_{iN_i}$, which are denoted by $\mathcal{W}_i$ as follows

$$\mathcal{W}_i = (\mathbf{W}_i^T(t_{i1}), \ldots, \mathbf{W}_i^T(t_{iN_i}))^T, \tag{3.2}$$

where $\mathbf{W}_i(t) = (W_{i1}(t), \ldots, W_{ip}(t))^T$.

As a result, the observed data from the i-subject becomes

$$D_o = (V_i, \Delta_i, \mathbf{Z}_i, \mathcal{W}_i, \mathbf{t}_i), \tag{3.3}$$

where $\mathbf{t}_i = (t_{ij} : 1 \leq j \leq N_i \text{ and } t_{ij} \leq V_i)$ is the observed time points for the ith individual. Since the longitudinal process is observed until the event or censoring happens, the observed covariate process is also truncated at V_i, i.e., $\{\mathbf{W}_i(t_{ij}) : t_{ij} \leq V_i\}$.

To take into account the measurement error or biological variation, we specify the observed longitudinal covariate $W_{ik}(t)$ by the linear mixed-effects model

$$\begin{aligned} W_{ik}(t) &= X_{ik}(t) + e_{ik}(t), \\ X_{ik}(t) &= \tilde{\boldsymbol{\rho}}_k(t)^T \boldsymbol{\mu}_k + \boldsymbol{\rho}_k(t)^T \mathbf{b}_{ik}, \quad k = 1, \ldots, p, \end{aligned} \tag{3.4}$$

where $e_{ik}(t)$ is a random error with mean zero, and $X_{ik}(t)$ is a combination of fixed effect $\boldsymbol{\mu}_k$ and random effect $\mathbf{b}_{ik}$. $\boldsymbol{\rho}_k(t)$ and $\tilde{\boldsymbol{\rho}}_k(t)$ are the basis functions of time t, including polynomial functions as a special case. For example, $\boldsymbol{\rho}(t) = (1, t)^T$ yields the simplest linear function of time. The form of the function is flexible and can vary across the p longitudinal covariates, and their corresponding dimension (i.e., $\dim(\boldsymbol{\rho}_k(t)) = d_k$ and $\dim(\ \tilde{\boldsymbol{\rho}}_k(t)) = \tilde{d}_k$) would vary accordingly. $\mathbf{b}_{ik}$ is a $d_k \times 1$ vector of random effects accounting for the within-subject variation.

In practice, it is typical to assume that the observed longitudinal covariates $\mathcal{W}_i$ are independent for different individuals, but are correlated across time and across different covariates at the same time for the same person. In order to take into account these independence and correlation structures, we assume $\mathbf{b}_{ik}$ are independent across i and k, $\mathbf{e}_i(t) = (e_{i1}(t), \ldots, e_{ip}(t))^T$ are independent across i, and $\mathbf{b}_{ik}$ and e_{ik} are independent for all i and k. Thus the correlations of $W_{ik}(t)$ over time are modeled via the variance of $\mathbf{b}_{ik}$, and the within-subject correlations across the p covariates are modeled via the variance covariance structure of $\mathbf{e}_{i.}(t)$. We further assume that $\mathbf{b}_{ik} \sim N_{d_k}(\mathbf{0}, \boldsymbol{\Sigma}_{bk})$, and $\mathbf{e}_i(t) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_e)$.

If we write equation (3.4) in matrix form, it becomes

$$\mathbf{W}_i(t) = \tilde{\boldsymbol{\rho}}(t)^T \boldsymbol{\mu} + \boldsymbol{\rho}(t)^T \mathbf{b}_i + \mathbf{e}_i(t), \tag{3.5}$$

where $\boldsymbol{\rho}(t)$ and $\tilde{\boldsymbol{\rho}}(t)$ are $d \times p$- and $\tilde{d} \times p$-dimensional matrix containing p basis functions of time, respectively, i.e.,

$$\boldsymbol{\rho}(t) = \begin{pmatrix} \boldsymbol{\rho}_1(t)^T & & \\ & \ddots & \\ & & \boldsymbol{\rho}_p(t)^T \end{pmatrix}^T, \quad \tilde{\boldsymbol{\rho}}(t) = \begin{pmatrix} \tilde{\boldsymbol{\rho}}_1(t)^T & & \\ & \ddots & \\ & & \tilde{\boldsymbol{\rho}}_p(t)^T \end{pmatrix}^T.$$

Note that $d = \sum_{k=1}^p d_k$ and $\tilde{d} = \sum_{k=1}^p \tilde{d}_k$. Thus, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_p^T)^T$ is the $\tilde{d}$-

dimensional vector consisting of the p fixed effect vectors, and $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{ip}^T)^T$ is a d-dimensional vector consisting of the p random effect vectors. Assume that $\mathbf{b}_i \sim N_d(\mathbf{0}, \mathbf{\Sigma}_b)$, where $\mathbf{\Sigma}_b = \text{diag}(\mathbf{\Sigma}_{b1}, \ldots, \mathbf{\Sigma}_{b2})$.

For the survival data, conditional on $\mathbf{t}_i(t) = \{t_{ij} : t_{ij} \leq t\}$, $\mathbf{e}_{i\cdot\cdot}(t) = \{e_{ik}(t_{ij}) : t_{ij} \leq t, k = 1, \ldots, p\}$, $T_i \geq t$, and other covariates, the relationship between the time-to-event T_i and the covariates is assumed to follow a Cox model with the hazard function at t being

$$\begin{aligned} h_i(t) &= \lim_{h \to 0} h^{-1} P\{t \leq T_i < t + h | T_i \geq u, C_i, \mathbf{Z}_i, \mathbf{b}_i, \mathbf{t}_i(t), \mathbf{e}_{i\cdot\cdot}(t)\} \\ &= \lambda(t) \exp\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T \mathbf{b}_i) + \boldsymbol{\eta}^T \mathbf{Z}_i\}, \qquad (3.6) \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\eta} \in \mathbb{R}^q$ are the regression coefficients of the longitudinal processes and time-independent covariates, respectively, and $\lambda(t)$ is an unspecified baseline hazard function. Note that model (3.6) implies that censoring time C_i, observation time point t_{ij} and the error $\mathbf{e}_{i\cdot\cdot}$ are non-informative in predicting the time-to-event T_i. This is a basic assumption for our survival submodel.

Let $\mathbf{\Omega} = (\Lambda(t), \boldsymbol{\theta})$ be the parameters in joint models of (3.5) and (3.6), where $\boldsymbol{\theta}$ is the set of parameters in the parametric part. Specifically

$$\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_p, \text{Vec}(\mathbf{\Sigma}_e), \text{Vec}(\mathbf{\Sigma}_b), \boldsymbol{\beta}, \boldsymbol{\eta}),$$

where $\text{Vec}(\mathbf{\Sigma})$ is the vector consisting of all the elements in the upper triangular part of $\mathbf{\Sigma}$. $\Lambda(t)$ is the cumulative baseline hazard defined by $\Lambda(t) = \int_0^t \lambda(u) du$.

3.2.2 Joint Likelihood Method and EM algorithm

We use the maximum joint likelihood approach to estimate $\mathbf{\Omega}$ for the joint models with multiple longitudinal covariates. Based on models (3.5) and (3.6) and their associated assumptions, we can write out the density of $\mathbf{b}_i$, as well as the conditional densities of $\{V_i, \Delta_i|\mathbf{b}_i\}$ and $\{\mathcal{W}_i|\mathbf{b}_i\}$ in the form

$$\begin{aligned} f(V_i, \Delta_i|\mathbf{b}_i) =& \left\{\lambda_0(V_i)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{\cdot(V_i)^T}\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}\right\}^{\Delta_i} \\ &\times \exp\left\{-\int_0^{V_i}\lambda_0(u)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{\cdot(u)^T}\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}du\right\}, \\ f(\mathcal{W}_i|\mathbf{b}_i) =& \{(2\pi)^p|\boldsymbol{\Sigma}_e|\}^{-\frac{N_i}{2}} \\ &\times \exp\left\{-\frac{1}{2}\sum_{j=1}^{N_i}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i)^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i)\right\}, \\ f(\mathbf{b}_i) =& (2\pi)^{-d/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\{-\frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\}, \end{aligned} \tag{3.7}$$

where $\mathbf{W}_{ij}$, $\tilde{\boldsymbol{\rho}}_{ij}$ and $\boldsymbol{\rho}_{ij}$ are the vector of $\mathbf{W}_i(t) = (W_{i1}(t), \ldots, W_{ip}(t))^T$ and the matrices of $\tilde{\boldsymbol{\rho}}(t)$ and $\boldsymbol{\rho}(t)$ taking values at time t_{ij}, respectively.

Assume that the observed longitudinal processes $\mathcal{W}_i$ are independent of the observed survival process $\{V_i, \Delta_i\}$ given the random effects $\mathbf{b}_i$. By (3.7) the joint likelihood of the observed data $\mathbf{D}_o$ can be written as

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^n L_i(\boldsymbol{\Omega}) = \prod_{i=1}^n \int f(V_i, \Delta_i|\mathbf{b}_i)\cdot f(\mathcal{W}_i|\mathbf{b}_i)\cdot f(\mathbf{b}_i)d\mathbf{b}_i. \tag{3.8}$$

In order to obtain the maximum likelihood estimates (MLE) of Λ, we let $\lambda(t)$ take mass only at each event time T_i for which $\Delta_i = 1$. Thus the dimension of $\lambda(t)$ reduces from infinity to a finite value $\sum_{i=1}^n \Delta_i$. The MLE of $\boldsymbol{\theta}$ and $\lambda(t)$ at each event time point are obtained by maximizing a modified version of joint likelihood (3.8), where $f(\mathcal{W}_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ stay the same and $f(V_i, \Delta_i|\mathbf{b}_i)$ becomes

$$\begin{aligned} f(V_i, \Delta_i|\mathbf{b}_i) =& \left\{\lambda_0(V_i)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{\cdot(V_i)^T}\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}\right\}^{\Delta_i} \\ &\times \exp\left\{-\sum_{j=1}^n \lambda_0(V_j)e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{\cdot(V_j)^T}\mathbf{b}_i)+\boldsymbol{\eta}^T\mathbf{z}_i}I(V_i \geq V_j, \Delta_j = 1)\right\}. \end{aligned} \tag{3.9}$$

Maximizing the joint likelihood function (3.8) is challenging. Wulfsohn and Tsiatis (1997) proposed to use the expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the joint likelihood (3.8) with $p = 1$. In the EM algorithm, the unobserved random effects are treated as missing data, and the parameter estimates are updated iteratively between the expectation and maximization steps until the algorithm converges. This method has been extensively applied and its feasibility and robustness has been demonstrated in the later related studies (Henderson et al., 2000; Tsiatis and Davidian, 2001; Song et al., 2002b; Hsieh et al., 2006), but only for a single longitudinal process setting. Here we extend the EM method for joint likelihood with multiple longitudinal processes (i.e., $p > 1$).

Denote the logarithm of the joint likelihood contributed by the ith subject to be

$$l_i(\mathbf{\Omega}) = \log(L_i(\mathbf{\Omega})) = \log \int f(V_i, \Delta_i|\mathbf{b}_i) \cdot f(\mathcal{W}_i|\mathbf{b}_i) \cdot f(\mathbf{b}_i) d\mathbf{b}_i = \log\{f(V_i, \Delta_i, \mathcal{W}_i)\}.$$

Let θ denote a generic element in $\mathbf{\Omega}$. Take derivative of $l_i(\mathbf{\Omega})$ with respect to θ and assume the derivative and integral are interchangeable under certain conditions and after some algebra, we obtain

$$S_i(\theta) = \frac{\partial l_i(\mathbf{\Omega})}{\partial \theta} = \frac{\partial}{\partial \theta} \left\{ E(l_{1i}(\mathbf{b}_i)) + E(l_{2i}(\mathbf{b}_i)) + E(l_{3i}(\mathbf{b}_i)) \right\}, \tag{3.10}$$

where

$$l_{1i}(\mathbf{b}_{i.}) = \log\{f(V_i, \Delta_i|\mathbf{b}_{i.})\},$$

$$l_{2i}(\mathbf{b}_{i.}) = \log\{f(\mathbf{W}_{i..}|\mathbf{b}_{i.})\},$$

$$l_{3i}(\mathbf{b}_{i.}) = \log\{f(\mathbf{b}_{i.})\},$$

and $E(\cdot)$ is conditional expectations of $\boldsymbol{b}_i$ given observed data D_o defined in (3.3) and the updated parameter estimates $\hat{\mathbf{\Omega}}$ in the EM algorithm. By definitions of l_{ki}, $k = 1, 2$, and 3, their conditional expectations in (3.10) are

$$\begin{aligned} E\{l_{i1}(\mathbf{b}_i)\} =& \Delta_i \log\{\lambda_0(V_i)\} + \Delta_i \boldsymbol{\beta}^T E\left\{\boldsymbol{\rho}_{.}(V_i)^T \mathbf{b}_i\right\} + \boldsymbol{\eta}^T \mathbf{Z}_i \\ & - \sum_{j=1}^{n} \lambda_0(V_j) E\left\{ e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{.}(V_j)^T \mathbf{b}_i) + \boldsymbol{\eta}^T \mathbf{Z}_i} I(V_i \geq V_j, \Delta_j = 1) \right\}, \qquad (3.11) \\ E\{l_{i2}(\mathbf{b}_i)\} =& -\frac{pN_i}{2}\log(2\pi) - \frac{N_i}{2}\log(|\mathbf{\Sigma}_e|) \end{aligned}$$

$$-\frac{1}{2}\sum_{j=1}^{N_i} E(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i)^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i), \quad (3.12)$$

$$E\{l_{i3}(\mathbf{b}_i)\} = -\frac{p}{2}\log(2\pi)-\frac{1}{2}\log(|\boldsymbol{\Sigma}_b|)+E(\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i). \quad (3.13)$$

In order to obtain the MLE by setting $\sum_{i=1}^n S_i(\theta)$ equal to 0 and solving the equations, the EM algorithm first calculates the conditional expectations of the functions of $\mathbf{b}_i$ in (3.11), (3.12) and (3.13) in the E-step. Then in the M-step, the log likelihood is maximized by setting partial derivative of $E\{l_{i1}(\mathbf{b}_i)\}, E\{l_{i2}(\mathbf{b}_i)\}$, and $E\{l_{i3}(\mathbf{b}_i)\}$ with respect to their corresponding parameters in $\boldsymbol{\Omega}$ to be 0. It can be easily derived that all the parameters except $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ have the closed-form maximum likelihood estimates:

$$\hat{\boldsymbol{\mu}} = (\sum_{i=1}^n \tilde{\boldsymbol{\rho}}_{i.}^T\tilde{\boldsymbol{\rho}}_{i.})^{-1}\{\sum_{i=1}^n \tilde{\boldsymbol{\rho}}_{i.}^T E(\mathbf{W}_{i.}-\boldsymbol{\rho}_{i.}^T\mathbf{b}_i)\}, \quad (3.14)$$

$$\hat{\boldsymbol{\Sigma}}_b = \frac{1}{n}\sum_{i=1}^n E\mathbf{b}_i\mathbf{b}_i^T, \quad (3.15)$$

$$\hat{\boldsymbol{\Sigma}}_e = \frac{1}{\sum_{i=1}^n N_i}\sum_{i=1}^n\sum_{j=1}^{N_i} E(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\hat{\boldsymbol{\mu}}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i)(\mathbf{W}_{ij}-\tilde{\boldsymbol{\rho}}_{ij}^T\hat{\boldsymbol{\mu}}-\boldsymbol{\rho}_{ij}^T\mathbf{b}_i)^T, \quad (3.16)$$

$$\hat{\lambda}(u) = \frac{\sum_{i=1}^n \Delta_i I(V_i=u)}{\sum_{j=1}^n E e^{\hat{\boldsymbol{\beta}}^T(\boldsymbol{\rho}(u)^T\mathbf{b}_j)+\hat{\boldsymbol{\eta}}^T\mathbf{z}_j} I(V_j\geq u)}, \quad (3.17)$$

where u only takes value at the event time points. For other time points, $\hat{\lambda}(u)=0$.

The MLE of the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in the Cox model are obtained by applying the Newton-Raphson algorithm to the profile likelihood of $l_{i1}(\mathbf{b}_i)$ after plugging in $\hat{\lambda}(u)$ in (3.17):

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + I_\beta^{-1}(\hat{\boldsymbol{\beta}}^{(k-1)})S_\beta(\hat{\boldsymbol{\beta}}^{(k-1)}), \quad (3.18)$$

$$\hat{\boldsymbol{\eta}}^{(k)} = \hat{\boldsymbol{\eta}}^{(k-1)} + I_\eta^{-1}(\hat{\boldsymbol{\eta}}^{(k-1)})S_\eta(\hat{\boldsymbol{\eta}}^{(k-1)}), \quad (3.19)$$

where $S_\beta(\hat{\boldsymbol{\beta}}^{(k-1)})$, $S_\eta(\hat{\boldsymbol{\eta}}^{(k-1)})$ and $I_\beta(\hat{\boldsymbol{\beta}}^{(k-1)})$, $I_\eta(\hat{\boldsymbol{\eta}}^{(k-1)})$ are the score functions and information matrices of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, respectively, taking values at the $(k-1)$th updated estimates. The score and the information matrices of the lth element of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are given by the following equations

$$\begin{aligned}
S_{\beta_l}(\boldsymbol{\beta}) =& \sum_{i=1}^{n} \frac{\partial E\{l_{i1}(\mathbf{b}_{i.})\}}{\partial \beta_l} \\
=& \sum_{i=1}^{n} \Delta_i \left\{ E(\mathbf{b}_{il}^T \boldsymbol{\rho}_l(V_i)) - \frac{\sum_{j=1}^{n} E(\mathbf{b}_{jl}^T \boldsymbol{\rho}_l(V_i)) e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right\},
\end{aligned} \tag{3.20}$$

$$\begin{aligned}
I_{\beta_l}(\boldsymbol{\beta}) =& -\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_l} \\
=& \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{j=1}^{n} E(\mathbf{b}_{jl}^T \boldsymbol{\rho}_l(V_i))^2 e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right. \\
& \left. - \left[\frac{\sum_{j=1}^{n} E(\mathbf{b}_{jl}^T \boldsymbol{\rho}_l(V_i)) e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right]^2 \right\},
\end{aligned} \tag{3.21}$$

$$\begin{aligned}
S_{\eta_l}(\boldsymbol{\eta}) =& \sum_{i=1}^{n} \frac{\partial E(l_{i1}(\mathbf{b}_{i.}))}{\partial \eta_l} \\
=& \sum_{i=1}^{n} \Delta_i \left\{ Z_{il} - \frac{\sum_{j=1}^{n} Z_{il} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right\},
\end{aligned} \tag{3.22}$$

$$\begin{aligned}
I_{\eta_l}(\boldsymbol{\eta}) =& -\frac{\partial S(\boldsymbol{\eta})}{\partial \eta_l} \\
=& \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{j=1}^{n} Z_{il}^2 E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right. \\
& \left. - \left[\frac{\sum_{j=1}^{n} Z_{il} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)}{\sum_{j=1}^{n} E e^{\boldsymbol{\beta}^T(\mathbf{b}_{j.}^T \boldsymbol{\rho}_{.}(V_i)) + \boldsymbol{\eta}^T \mathbf{z}_j} I(V_j \geq V_i)} \right]^2 \right\}.
\end{aligned} \tag{3.23}$$

In an EM algorithm, the E-step and M-step are calculated iteratively until the algorithm converges. By (3.14) through (3.23), in each iteration, conditional expectations need to be evaluated for the following six functions of the random effects for the ith subject, $i = 1, \ldots, n$:

$$\begin{aligned}
g_1(\mathbf{b}_i) &= \mathbf{b}_i, \\
g_2(\mathbf{b}_i) &= \mathbf{b}_i \mathbf{b}_i^T, \\
g_3(\mathbf{b}_i) &= \mathbf{b}_i^T \boldsymbol{\rho}(V_i), \\
g_4(\mathbf{b}_j) &= e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i, \\
g_5(\mathbf{b}_j) &= \{\boldsymbol{\rho}(V_i)^T \mathbf{b}_j\} e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V_i)^T \mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i, \\
g_6(\mathbf{b}_j) &= \{\boldsymbol{\rho}(V_i)^T \mathbf{b}_j\}^2 e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_{.}(V_i)^T \mathbf{b}_j)}, \text{ for } j = 1, \ldots, n, \text{ and } V_j \geq V_i.
\end{aligned} \tag{3.24}$$

Since multi-dimensional integrals are involved in the conditional expectations $E\{g_k(\mathbf{b}_i)\}$ in the E-step, numerical integration techniques such as Gaussian-Hermite Quadrature (Wulfsohn and Tsiatis, 1997), Markov Chain Monte Carlo (Henderson et al., 2000; Tseng et al., 2005; Ding and Wang, 2008) and fully exponential Laplace approaches (Rizopoulos et al., 2009) have been applied to approximate the target expectations. Since most of these techniques are shown insufficient in estimating the joint models with large dimension of random effects, we propose to approximate the conditional expectations using design of experiments-based interpolation techniques (DoIt; Joseph 2012), which is further introduced in section 3.23.

Another challenge of joint modeling is to obtain standard errors (SE's) for the MLE. Louis (1982) suggested that the accurate variance estimation in the EM algorithm would require the calculation of the observed Fisher information matrix for the entire parameter set. However, this approach is impractical for our case considering the high dimensionality of $\boldsymbol{\Omega}$ mainly caused by $\lambda_0(t)$. On the other hand, if we use the second derivative of the profile likelihood $pl(\boldsymbol{\theta})$(i.e., substituting the estimate of $\lambda_0(t)$ into $l(\boldsymbol{\Omega})$) to calculate the SE's for $\boldsymbol{\theta}$, the resulting estimates have been shown to be biased and over-optimistic for statical inference. (Hsieh et al., 2006). Due to these limitations, in this study, we propose to estimate the SE's using bootstrap technique, the procedure of which will be explained in detail in simulation studies in Section 3.

3.2.3 Implementation

It has always been a challenge to approximate the large-dimensional integral numerically. In order to approximate the conditional expectations of $E\{g(\mathbf{b}_i)\}$ in (3.14) through (3.23), several techniques have been proposed. The Gaussian-Hermite quadrature method evaluates the function values at M different quadrature points and approximates the expectations using the weighted sums. This method is usually accurate for low dimensional integrals but the number of quadrature points increases exponentially with the dimension. The Markov Chain Monte Carlo method treats the posterior densities $f(\mathbf{b}_i|D_o, \hat{\mathbf{\Omega}})$ as transition probabilities of a Markov Chain from which the samples are drawn and updated. By the Law of Large Number, the sample means of $g(\mathbf{b}_i)$ converge to the expectations with a the rate of $O(M^{-1/2})$ in probability, where M is the number of random samples. Fully exponential Laplace (Tierney and Kadane, 1986; Tierney et al., 1989) does not rely on evaluation or sample points. Instead, it approximates the posterior distributions of $\mathbf{b}_{ik}$ by normal distributions with the posterior modes of $f(\mathbf{b}_i|D_o, \hat{\mathbf{\Omega}})$ as the means and the inverse Fisher information matrices as the variances. The approximation accuracy depends only on the sample size of each individual and was proved to be of the order $O(n_i^{-2})$. All of these methods, however, become either time-consuming or difficult to implement as the dimension of the random effects $\mathbf{b}_i$ increases. In order to extend joint models to more flexible settings, a more efficient and stable computing technique is needed to better implement the EM algorithm.

The design of experiments-based interpolation techniques (DoIt) was recently proposed by Joseph (2012) as a method to approximate the "expensive conditional densities" in Bayesian computation. The DoIt method borrows and extends the idea of Laplace approximation which approximates the posterior densities of interest via normal distributions. But instead of using a single normal distribution, DoIt also incorporates the idea of quadrature-based methods and approximates $f(\mathbf{b}_i|D_o, \hat{\mathbf{\Omega}})$ using the weighted sum of a sequence of normal densities with the means at M pre-specified evaluation points $(\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_M)$, i.e.,

$$f_{\mathbf{b}_i|D_o,\hat{\mathbf{\Omega}}}(\mathbf{b}_i) \approx \frac{1}{\sum_{l=1}^M c_l} \sum_{l=1}^M c_l \phi_l(\mathbf{b}_i), \tag{3.25}$$

where $\phi_l(\mathbf{b}_i) = \phi(\mathbf{b}_i; \boldsymbol{\nu}_l, \mathbf{D}_i^{-1})$ denote the normal density of $\mathbf{b}_i$ with mean $\boldsymbol{\nu}_l$ and variance $\mathbf{D}_i^{-1}$, the Fisher information matrix of $f(\mathbf{b}_i|D_o, \hat{\boldsymbol{\Omega}})$ evaluated at the mode. The c_l's are the weights associated with $\phi_l(\cdot)$ and can be calculated by solving the linear equations

$$\boldsymbol{Qc} = \mathbf{h},$$

where $h(\mathbf{b}_i) \propto f(\mathcal{W}_i, V_i, \Delta_i|\mathbf{b}_i)f(\mathbf{b}_i)$ is the unnormalized posterior densities of $\mathbf{b}_i$, and $\mathbf{h} = (h(\boldsymbol{\nu}_1), \ldots, h(\boldsymbol{\nu}_M))$ is a vector of $h(\cdot)$ taking values at the M evaluation points. $\mathbf{Q}$ is an $M \times M$ matrix with ijth element being the unnormalized normal density

$$q(\boldsymbol{\nu}_i; \boldsymbol{\nu}_j, \mathbf{D}_i^{-1}) = \exp\{-\frac{1}{2}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)^T \mathbf{D}_i(\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)\}.$$

Since $q(\boldsymbol{\nu}; \boldsymbol{u}, \mathbf{D}_i^{-1})$ is a positive definite function, $\mathbf{Q}^{-1}$ exists, provided $\boldsymbol{\nu}_i \neq \boldsymbol{\nu}_j$ for all i and j. Thus a unique solution of $\hat{\boldsymbol{c}} = \mathbf{Q}^{-1}\mathbf{h}$ is guaranteed.

With the capability to approximate the expensive posterior densities, DoIt can also be used to approximate the conditional expectations of real value functions of $\mathbf{b}_i$ in the E-step of the EM algorithm for joint modeling. Let $g(\mathbf{b}_i)$ be any real value function of $\mathbf{b}_i$. Then its conditional expectation given $\mathbf{D}_o$ and $\hat{\boldsymbol{\Omega}}$ is approximated by

$$E\{g(\mathbf{b}_i)\} \approx \frac{1}{\sum_{l=1}^M c_l} \sum_{l=1}^{M} c_l E_l\{g(\mathbf{b}_i)\}, \tag{3.26}$$

where $E_l\{g(\mathbf{b}_i)\}$ on the right-hand side is the expectation of $g(\mathbf{b}_i)$ with respect to the normal distribution $N(\boldsymbol{\nu}_l, \mathbf{D}_i^{-1})$. By introducing the idea of DoIt into the computation of joint modeling, we find it produces good parameter estimates, and thus can be taken as an alternative to the existing computing techniques for joint models.

The locations of $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ are determined by the pre-specified space-filling design points transformed to the parameter space of posterior distribution via $\hat{\mathbf{b}}_i$ and $\mathbf{D}_i^{-1}$. Similar to other methods that depend on evaluation points, DoIt has the nice feature of achieving desired level of accuracy by adding more points, i.e., increasing M (Joseph, 2012). However, the DoIt method does not suffer from "curse of dimensionality" like the Gaussian quadrature method because the number of M grows much slower to achieve the same level of accuracy when the dimension increases. A common rule of thumb in the literature of computer experiments is to use $M = 10 \times d_b$ (Loeppky et al., 2009), where d_b is the dimension of the integral. With good space-filling design points, the order of the convergence rate can reach $O(M^{-1} \log^{d_b} M)$ if the integrand is a function with bounded total variation (Fang and Wang, 1994). Although different space-filling designs can be considered to specify the locations of the deterministic points $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$, in this study we apply minimax Latin Hypercube Design (MmLHD; Joseph 2012) as the design scheme for DoIt implementation with the existing R package *LHS* which automatically generates MmLHD samples.

3.3 Numerical studies

In this section we conduct a simulation study and real data analysis to examine the estimation performance of the proposed approach.

3.3.1 Simulation studies

We consider joint models with both single and multiple longitudinal covariates. Under the framework of the joint likelihood approach with an EM algorithm, we compare different integral approximation techniques, using the existing R package *JM* (Rizopoulos, 2010) as benchmark. Below are the implementation details of each method.

(i) *JM* package (JM): we use the option "method = Cox-PH-GH" in the JM() function, in which an adaptive Gaussian-Hermite quadrature method with 35 quadrature points on each dimension of $\mathbf{b}_{ik}$ is used to calculate the conditional expectations in the E-step of the EM algorithm.

(ii) Gaussian-Hermite quadrature method (GHQ): we take 5 quadrature points on each dimension of $\mathbf{b}_{ik}$, and the total number of evaluation points are $M = 5^{d_b}$.

(iii) MCMC method: we use the R package *MHadaptive* to generate random samples from the unnormalized posterior density of $\mathbf{b}_{ik}$. For each individual 1000 MCMC samples are generated, with 100 of them as "burn-in" samples (i.e., M=900). Since the computing process is extremely slow with this method, a more relaxed EM stopping is used (i.e., $|\epsilon| < 10^{-2}$ for MCMC; $|\epsilon| < 10^{-4}$ for other methods, where $\epsilon = \|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k+1)}\|_\infty$ is the maximum dimension of the difference between the parameter estimates in the current and the previous iterations).

(iv) Fully exponential Laplace method (FEL): we adopt the technique described in Rizopoulos et al. (2009).

(v) DoIt method: the R package *lhs* is used to generate design points from MmLHD. The suggested number of design points $M = 10 \times d_b$ is used.

All the results are based on $N = 100$ replicates with $n = 100$ subjects in each data set.

Example 3.1. We consider a simple joint model with a single longitudinal covariate

$$W_i(t_{ij}) = X_i(t_{ij}) + e_i(t_{ij}) = b_{0i} + b_{1i}t_{ij} + e_i(t_{ij}),$$

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta X_i(t) + \eta Z_i\},$$

with the assumption

$$\mathbf{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ e_i(t_{ij}) \overset{i.i.d}{\sim} N(0, \sigma^2),$$

and $\boldsymbol{b}_i$ are independent of $e_i(t_{ij})$. This is a low-dimensional setting with $\dim(\mathbf{b}_i) = 2$, and $\boldsymbol{\rho}(t) = (1, t)^T$. The true values of the parameters are based on Hsieh et al. (2006). For the longitudinal process, we set $\boldsymbol{\mu} = (-4.9078, 0.500)^T$, $(\sigma_{11}, \sigma_{12}, \sigma_{22}) = (0.5, -0.001, 0.04)$, $\sigma^2 = 0.1$, where $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ are the upper diagonal elements of $\boldsymbol{\Sigma}$. The observational time $\mathbf{t}_i$ is generated first by $\mathbf{t}_i = seq(0, 12, 38)$ in R (i.e., 38 equally spaced time points between 0 and 12), and then truncated by the survival time of different individuals. For the event process, we assume the baseline hazard function is constant over time, i.e., $\lambda_0(t) = 1$. The true regression coefficients are $\beta = 1$ and $\eta = -1$. The censoring time for each subject is generated from an exponential distribution with mean 25, resulting in about 30% censoring. The survival time T_i is generated from the inverse survival function derived from the cumulative hazard $\Lambda_i(t)$ with the constant baseline hazard λ_0:

$$\begin{aligned}\Lambda_i(t) &= \int_0^t \lambda_0(u) \exp\{\beta X_i(u) + \eta Z_i\} \\ &= e^{\eta Z_i} \lambda_0 \int_0^t e^{\beta(b_{0i}+b_{1i}u)} du \\ &= \lambda_0 \frac{1}{\beta b_{1i}} e^{\beta b_{0i}+\eta Z_i} \left(e^{\beta b_{1i} t} - 1\right).\end{aligned}$$

Since the survival probability at time t is $S_i(t) = 1 - F_i(t) = \exp\{-\Lambda_i(t)\}$, $S_i(t)$ can be generated using the Monte Carlo samples $U_i \sim U(0,1)$, the uniform distribution, and $\exp\{-\Lambda_i(t)\} = U_i$. Thus,

$$T_i = \frac{1}{\beta_X b_{1i}} \log\{1 - \frac{\beta_X b_{1i} \log U_i}{\lambda_0 \exp(\beta_X b_{0i} + \beta_Z Z_i)}\}. \tag{3.27}$$

Note that since the logarithm is involved in (3.27), we need to monitor the sign of the expression inside log. Since U_i is between 0 and 1, $\log U_i$ is always negative. Thus we need $b_{1i} > 0$ to guarantee the positivity of $-\beta_X b_{1i} \log U_i$. In practice, most of the random samples of b_{1i} is positive with the setting of $\mu_2 = 0.500$ and $\Sigma_{22} = 0.04$. However, negative expression can occur with some specific random combinations of b_{1i} and U_i, for which the above equation to generate T_i is not well-defined. Such cases are treated as censoring in the simulation. However, these cases are very rare (usually less that 2%) with our parameter setting, and they hardly affect the distribution assumptions of the simulation.

After truncated by the observed event time V_i, the average number of longitudinal observations is $\bar{n}_. = 20$.

Table 3.1: Simulation results for Example 3.1 with $\bar{n}_{\cdot} = 20$

		JM	GHQ	MCMC	FEL	DoIt (M=10)
Time(s)	Median	14.60	4.67	584.10	9.97	9.41
	Bias	-0.0330	0.0022	-0.0012	-0.0195	-0.0043
$\beta = 1.0000$	SD	0.1270	0.1288	0.1276	0.1187	0.1278
	RMSE	0.1312	0.1288	0.1276	0.1203	0.1279
	Bias	0.0310	0.0106	0.0616	0.0345	0.0157
$\eta = -1.0000$	SD	0.3198	0.3187	0.2677	0.3106	0.3167
	RMSE	0.3213	0.3189	0.2747	0.3125	0.3171
	Bias	0.0063	-0.0034	-0.0034	-0.0006	-0.0004
$\mu_1 = -4.9078$	SD	0.1990	0.0749	0.0604	0.0729	0.0730
	RMSE	0.1991	0.0750	0.0605	0.0729	0.0730
	Bias	-0.0027	0.0028	0.0031	0.0000	-0.0001
$\mu_2 = 0.5000$	SD	0.0395	0.0231	0.0201	0.0224	0.0229
	RMSE	0.0400	0.0233	0.0203	0.0224	0.0229
	Bias	0.0967	0.0066	-0.0441	-0.0100	-0.0112
$\sigma_{11} = 0.5000$	SD	0.0965	0.0725	0.0732	0.0721	0.0719
	RMSE	0.1366	0.0728	0.0855	0.0728	0.0728
	Bias	-0.0134	-0.0069	0.0070	-0.0007	-0.0008
$\sigma_{12} = -0.0010$	SD	0.0208	0.0159	0.0145	0.0155	0.0156
	RMSE	0.0247	0.0173	0.0161	0.0170	0.0156
	Bias	0.0009	0.0048	-0.0033	-0.0008	-0.0009
$\sigma_{22} = 0.0400$	SD	0.0073	0.0067	0.0064	0.0061	0.0062
	RMSE	0.0074	0.0082	0.0072	0.0062	0.0063
	Bias	0.2731	0.0092	0.0166	-0.0049	-0.0006
$\sigma^2 = 0.1000$	SD	0.0096	0.0037	0.0099	0.0032	0.0033
	RMSE	0.2733	0.0099	0.0193	0.0059	0.0034

In Table 3.1, GHQ with 25 evaluation points takes the shortest computing time, showing its superiority over other methods when the dimension of random effects is small. DoIt with 10 evaluation points is the second fastest technique, followed by FEL, both faster than the standard *JM* package. MCMC is extremely slow, mostly because it requires a large number of random samples to be drawn for each individual in each iteration.

All the methods yield comparable estimates in terms of accuracy (i.e., bias) and efficiency (i.e., RMSE) in Table 3.1. The only exception is the JM estimates of σ^2, which has great bias compared with other methods. The reason for this is that JM package only focuses on estimating regression coefficients β and η in survival models using the joint information of the two processes. As for the parameters in the longitudinal model, it directly uses the linear mixed-effects models' estimates from the R function lme(), which does not consider the joint information from the survival part, and thus may lead to bias and inefficiency. For all the other methods, the estimates from lme() functions are only taken as initial values for the EM algorithm. The DoIt method with $M = 10$ evaluation points provides good estimating results in the table, and such good performance is consistent for all the parameters.

In our simulation studies, we use bootstrap technique to obtain the standard errors (SE's) of the maximum likelihood estimates. For each dataset replication, we randomly sample n_1 subjects with replacement from the uncensored group of individuals, and randomly sample n_2 subjects with replacement from the censored group. $n_2 = n - n_1$ is the number of censoring in the original replicate dataset. The sampling with replacement is conducted within each cluster so as to remain the same censoring rate as the original data. The two groups of random samples are then combined into a complete dataset and the parameter estimates are obtained using the maximum joint likelihood (MJL) method. Such estimating procedure is repeated $B = 100$ times for each replicate, and the standard deviations of these $B = 100$ parameter estimates are recorded as the bootstrap SE's of the replicate dataset.

Table 3.2 shows the performance of the bootstrap SE's of the $N = 100$ replicates for MJL with the DoIt algorithm. The second column, Mean Est, is the mean of the MLEs of the 100 replicates; the third column, SD_{Est}, is the standard deviations of the MLEs of the 100 replicates; the fourth column, SE_{Boot}, is the average of the bootstrap SE's across the 100 replicates; and the last column, $\mathrm{SD}_{SE.Boot}$, is the standard deviations of the bootstrap SE's across the 100 replicates. The results suggest that the average bootstrap SE's are very similar to the SD's for all the parameters. More specifically, all the SD's fall in the range of $\text{SE}_{Boot} \pm 2\text{SD}_{SE.Boot}$. This indicates that the bootstrap SE's are reliable.

Table 3.2: Performance of Bootstrap Standard Errors of Example 3.1

	True	Mean Est	SD_{Est}	SE_{Boot}	$\text{SD}_{SE.Boot}$
β	1.0000	0.9957	0.1278	0.1343	0.0238
η	-1.0000	-0.9843	0.3167	0.2894	0.0295
μ_1	-4.9078	-4.9082	0.0730	0.0697	0.0073
μ_2	0.5000	0.4999	0.0229	0.0202	0.0021
σ_{11}	0.5000	0.4888	0.0719	0.0692	0.0147
σ_{12}	-0.0010	-0.0018	0.0156	0.0149	0.0023
σ_{22}	0.0400	0.0391	0.0062	0.0058	0.0010
σ^2	0.1000	0.0994	0.0033	0.0031	0.0003

Example 3.2. In this example we consider a joint model with two longitudinal covariates

$$W_{i1}(t_{ij}) = X_{i1}(t_{ij}) + e_{i1}(t_{ij}) = b_{10i} + b_{11i}t_{ij} + e_{i1}(t_{ij}),$$

$$W_{i2}(t_{ij}) = X_{i2}(t_{ij}) + e_{i2}(t_{ij}) = b_{20i} + b_{21i}t_{ij} + e_{i2}(t_{ij}),$$

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

with the assumption

$$\mathbf{b}_{ik} = \begin{pmatrix} b_{k0i} \\ b_{k1i} \end{pmatrix} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{bk}),\ e_{ik}(t_{ij}) \overset{i.i.d}{\sim} N(0, \sigma_k^2),$$

where the random effects $\mathbf{b}_{ik}$ are independent of the measurement errors $e_{ik}(t_{ij})$, and the two longitudinal processes are independent of each other. The true values of the parameters are $\boldsymbol{\mu}_1 = (-5, 0.5)^T$, $\boldsymbol{\mu}_2 = (-2, 1)^T$, $Vec(\boldsymbol{\Sigma}_1) = (1, -0.001, 0.04)$, $Vec(\boldsymbol{\Sigma}_2) = (0.5, -0.001, 0.09)$, $\sigma_1^2 = \sigma_2^2 = 1$, $\beta_1 = 1$, $\beta_2 = 2$, $\eta_1 = -1$. They are specified in details in the output tables.

In this example, the dimension of the random effects becomes $\dim(\mathbf{b}_{i.}) = 4$. In the event process, we take $\lambda_0(t) = 1$. The average censoring rate is around 15%.

Using similar techniques as in example 1, we generate survival times as follows:

$$T = \frac{1}{(\beta_1 b_{11i} + \beta_2 b_{21i})} \log \left\{ 1 - \frac{(\beta_1 b_{11i} + \beta_2 b_{21i}) \log U_i}{\lambda_0 \exp(\beta_1 b_{10i} + \beta_2 b_{20i} + \eta Z_i)} \right\}.$$

Since *JM* package is not available with multiple longitudinal covariates and MCMC falls out of scope due to its extremely long computing time, we consider only GHQ, FEL and DoIt for this example.

In Table 3.3, GHQ becomes much slower than the other two methods because the number of evaluation points increases dramatically to $M = 5^4 = 625$. DoIt with $M = 10 \times 4 = 40$ uses the shortest computing time. The estimates of the three methods are of similar bias levels, indicating they are similar in terms of estimating accuracy. However, the estimates from the GHQ method, especially those of the longitudinal process, have much larger SD and RMSE than the corresponding estimates from FEL and DoIt. This suggests FEL and DoIt are more stable in estimation and provide more efficient estimates. Although FEL is comparable with DoIt in computing time and estimating performance, it involves the complex tensor computation (Rizopoulos et al., 2009). This makes it less appealing than DoIt for larger-dimensional problems.

Similar to Table 3.2, the results of Table 3.4 shows that the bootstrap SE's closely resemble the standard deviations in the second column, and thus can be regarded as valid standard error estimates of the MLE.

Table 3.3: Simulation results for Example 3.2 with $\bar{n}. = 20$

		GHQ	FEL	DoIt ($M = 10d$)
Time(s)	Median	440.8	122.6	118.5
	Bias (SD)	-0.0161(0.1769)	-0.0120 (0.1466)	0.0032 (0.1302)
$\beta_1 = 1.0000$	RMSE	0.1776	0.1471	0.1303
	Bias (SD)	0.0121(0.3119)	0.0304 (0.2461)	-0.0119 (0.1972)
$\beta_2 = 2.0000$	RMSE	0.3121	0.2480	0.1975
	Bias (SD)	0.0176 (0.2979)	0.0161 (0.3076)	-0.0032 (0.2715)
$\eta = -1.0000$	RMSE	0.2984	0.3080	0.2715
	Bias (SD)	0.0309 (0.5114)	-0.0174 (0.0951)	-0.0091 (0.1031)
$\mu_{11} = -5.0000$	MSE	0.5123	0.0967	0.1035
	Bias (SD)	-0.0025 (0.0537)	0.0002 (0.0185)	-0.0010 (0.0201)
$\mu_{12} = 0.5000$	RMSE	0.0538	0.0185	0.0201
	Bias (SD)	0.0173 (0.2136)	0.0008 (0.0699)	0.0132 (0.0665)
$\mu_{21} = -2.0000$	RMSE	0.2143	0.0699	0.0678
	Bias (SD)	-0.0109 (0.1050)	-0.0062 (0.0306)	-0.0006 (0.0276)
$\mu_{22} = 1.0000$	RMSE	0.1056	0.0312	0.0277
	Bias (SD)	-0.0057 (0.1809)	-0.0164 (0.1480)	-0.0323 (0.1297)
$\sigma_{111} = 1.0000$	RMSE	0.1810	0.1489	0.1337
	Bias (SD)	-0.0067 (0.0210)	0.0013 (0.0216)	-0.0012 (0.0189)
$\sigma_{112} = -0.0010$	RMSE	0.0220	0.0216	0.0189
	Bias (SD)	0.0055 (0.0082)	-0.0007 (0.0066)	-0.0004 (0.0064)
$\sigma_{122} = 0.0400$	RMSE	0.0099	0.0067	0.0064
	Bias (SD)	-0.0022 (0.0814)	-0.0153 (0.0616)	-0.0109 (0.0661)
$\sigma_{211} = 0.5000$	RMSE	0.0814	0.0635	0.0670
	Bias (SD)	-0.0088 (0.0208)	-0.0004 (0.0215)	-0.0020 (0.0220)
$\sigma_{212} = -0.0010$	RMSE	0.0226	0.0215	0.0220
	Bias (SD)	0.0070 (0.0177)	-0.0007 (0.0141)	-0.0010 (0.0143)
$\sigma_{222} = 0.0900$	RMSE	0.0190	0.0141	0.0144
	Bias (SD)	0.0088 (0.0115)	-0.0054 (0.0029)	0.0001 (0.0029)
$\sigma_1^2 = 0.1000$	RMSE	0.0145	0.0062	0.0030
	Bias (SD)	0.0087 (0.0115)	-0.0043 (0.0032)	0.0003 (0.0027)
$\sigma_2^2 = 0.1000$	RMSE	0.0144	0.0054	0.0027

Table 3.4: Performance of Bootstrap Standard Errors of Example 3.2

	True	Mean Est	SD_{Est}	SE_{Boot}	$SD_{SE.Boot}$
β_1	1.0000	1.0032	0.1302	0.1372	0.0309
β_2	2.0000	1.9881	0.1972	0.1980	0.0431
η	-1.0000	-1.0032	0.2715	0.2734	0.0501
μ_{11}	-5.0000	-5.0091	0.1031	0.0988	0.0171
μ_{12}	0.5000	0.4990	0.0201	0.0211	0.0039
μ_{21}	-2.0000	-1.9868	0.0665	0.0696	0.0114
μ_{22}	1.0000	0.9994	0.0276	0.0306	0.0053
σ_{111}	1.0000	0.9677	0.1297	0.1359	0.0366
σ_{112}	-0.0010	-0.0022	0.0189	0.0211	0.0039
σ_{122}	0.0400	0.0396	0.0064	0.0065	0.0016
σ_{211}	0.5000	0.4891	0.0661	0.0704	0.0159
σ_{212}	-0.0010	-0.0030	0.0220	0.0222	0.0044
σ_{222}	0.0900	0.0890	0.0143	0.0131	0.0034
σ_1^2	0.1000	0.1001	0.0029	0.0032	0.0005
σ_2^2	0.1000	0.1003	0.0027	0.0032	0.0006

3.3.2 A real data example

In this section, we illustrate the proposed model and estimation procedure by an empirical analysis of the data collected from a smoking cession study. Specifically, this data set was collected from a randomized, placebo-controlled clinical trial (N=1504) of five active smoking-cessation pharmacotherapies, in which daily smokers who were highly motivated to quit were recruited (Piper et al., 2009). The focus of this example is to examine the effects of multiple smoking withdrawal symptoms on the time to lapse (T_L, the first day of smoking following at least 1 day of abstinence) and time to relapse (T_RL, the first of 7 consecutive days of smoking following at least 1 day of abstinence) in a two-week post-quit study period (i.e. $t \in (0, 14)$). The withdrawal symptoms of interest include craving for smoking (Crav), negative affect (NA) and smoking cessation fatigue (feeling tired of quitting; Fatig). All of these symptoms are supposed to exhaust the self-control resources that prevent the participants from smoking when they are trying to quit. All the items are self-reported by the participants four times a day (i.e., morning, night and 2 random times). The five treatments are divided into three types: placebo, monotherapy and combined pharmacotherapy, and are treated as time-independent covariates in the joint model. We build two working data sets for analysis: one with T_L as the response of interest (i.e., Lapse Data), and the other with T_RL as the response of interest (i.e., Relapse Data). After data cleaning, N=794 subjects were used for the analysis with T_L as the response, and N=811 subjects were used for then analysis with T_RL as the response.

Participants without a record for the longitudinal covariates after the actual quit date or with the survival response equal to 0 were removed from the study. In addition, we only used the records between 0 and $\min(T\_L, 14)$ for Lapse Data and between 0 and $\min(T\_RL, 14)$ for Relapse Data. As a result, there are 794 subjects left in Lapse Data (83 placebo, 412 monotherapy and 299 combined pharmacotherapy) and the censor rate is 59.69%; 811 subjects are left in Relapse Data (86 placebo, 422 monotherapy and 303 combined pharmacotherapy) and the censor rate is 91.24%.

In the previous study (Liu et al., 2013) we modeled the three longitudinal processes nonparametrically and found all of them decrease linearly with time (see Figure 3.1). Thus in joint modeling framework we specify them using linear mixed-effects model of the following forms:

$$W_{\mathrm{Crav}i}(t_{ij}) = X_{\mathrm{Crav}i}(t_{ij}) + e_{ic}(t_{ij}) = b_{c0i} + b_{c1i}t_{ij} + e_{ic}(t_{ij}),$$

$$W_{\mathrm{NA}i}(t_{ij}) = X_{\mathrm{NA}i}(t_{ij}) + e_{in}(t_{ij}) = b_{n0i} + b_{n1i}t_{ij} + e_{in}(t_{ij}),$$

$$W_{\mathrm{Fatig}i}(t_{ij}) = X_{\mathrm{Fatig}i}(t_{ij}) + e_{if}(t_{ij}) = b_{f0i} + b_{f1i}t_{ij} + e_{if}(t_{ij}),$$

where

$$\mathbf{b}_{ci} = (b_{c0i}, b_{c1i})^T \sim N_2(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_{bc}),$$

$$\mathbf{b}_{ni} = (b_{n0i}, b_{n1i})^T \sim N_2(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_{bn}),$$

$$\mathbf{b}_{fi} = (b_{f0i}, b_{f1i})^T \sim N_2(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_{bf}),$$

and

$$e_{\mathrm{crav}i}(t_{ij}) \sim N(0, \sigma_c^2),$$

$$e_{\mathrm{cravi}}(t_{ij}) \sim N(0, \sigma_n^2),$$

$$e_{\mathrm{cravi}}(t_{ij}) \sim N(0, \sigma_f^2),$$

and all these random terms are assumed independent.

In this analysis we model the survival response using Cox's models with both single covariate process and multiple covariates processes. The single-covariate models are of the form:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_c X_{\mathrm{Crav}i}(t_{ij}) + \eta Z_i\},$$

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_n X_{\mathrm{NA}i}(t_{ij}) + \eta Z_i\},$$

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_f X_{\mathrm{Fatig}i}(t_{ij}) + \eta Z_i\},$$

and the multiple-covariates model is of the form

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta_c X_{\mathrm{Crav}i}(t_{ij}) + \beta_n X_{\mathrm{NA}i}(t_{ij}) + \beta_f X_{\mathrm{Fatig}i}(t_{ij}) + \eta Z_i\},$$

where Z_i is the treatment variable with three levels (i.e., 2 dummy variables in practice).

The goals of the analysis are: (a) to compare maximum joint likelihood method (with the DoIt algorithm) with the naive separate estimation, and (b) to compare the single-covariate-process model with the multiple-covariates-process model. The standard errors of maximum joint likelihood approach are calculated using the bootstrap techniques as discussed in simulation studies. The estimating results are presented in the following table.

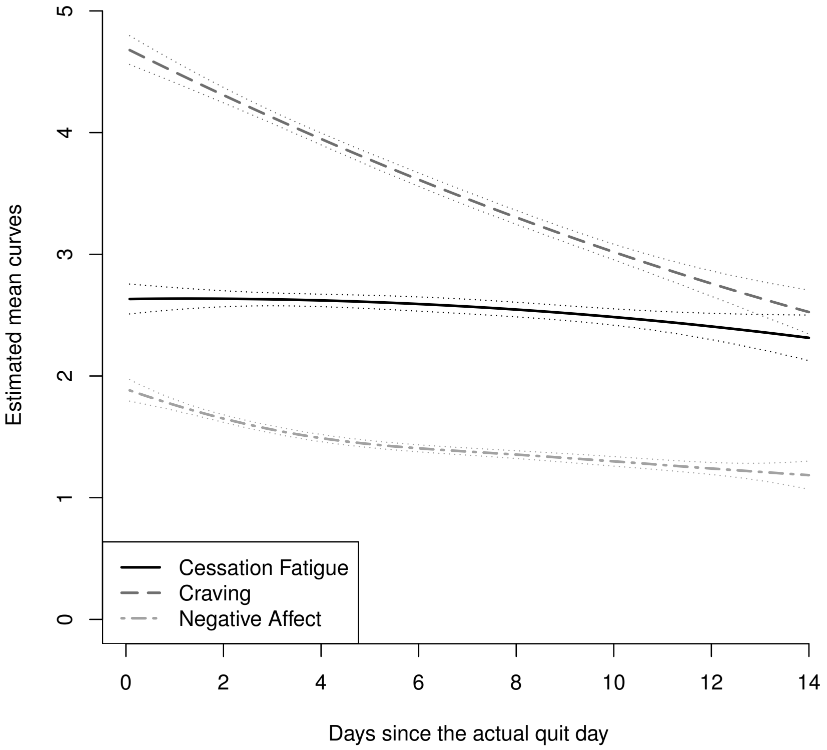

Figure 3.1. Nonparametric estimations of the three covariate processes

Table 3.5: Estimation for Lapse Data with Single Covariate Process

	Craving				Negative Affect				Fatigue			
	Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
η_1	-0.08	0.12	-0.22	0.16	-0.08	0.12	-0.22	0.16	-0.05	0.12	-0.20	0.17
η_2	-0.08	0.12	-0.45*	0.16	-0.10	0.12	-0.48*	0.16	-0.08	0.13	-0.48*	0.18
β	0.06*	0.01	0.13*	0.02	0.10*	0.02	0.15*	0.04	0.03*	0.01	0.03	0.02
μ_1	4.63*	0.11	4.64*	0.10	1.83*	0.05	1.84*	0.05	2.72*	0.11	2.72*	0.29
μ_2	-0.17*	0.01	-0.16*	0.01	-0.06*	0.005	-0.06*	0.005	-0.006	0.01	-0.005	0.01
Σ_{11}	7.86	-	7.89	0.33	2.00	-	2.00	0.14	8.96	-	8.95	1.05
Σ_{12}	-0.24	-	-0.23	0.03	-0.08	-	-0.08	0.01	-0.18	-	-0.18	0.07
Σ_{22}	0.04	-	0.04	0.005	0.01	-	0.01	0.001	0.10	-	0.10	0.02
σ^2	2.04	-	4.16	0.19	0.95	-	0.91	0.04	1.52	-	2.30	0.25

Table 3.5 represents the estimation results for the single-process models with days to lapse as the survival response and craving, negative affect, and fatigue as the single predictors in the three models, respectively. The estimation results indicate that the separation estimation and maximum joint likelihood method yield quit similar estimates for all the parameters in longitudinal models except σ^2, the variance of the random error. However, the parameter estimates for the survival models are quit different among the two models. Separate estimation suggests neither of the two active treatments are effective compared with placebo, whereas the MJL method suggests the combined pharmacotherapy has significant effect in reducing the risk of lapse. This is consistent with the knowledge that the estimates of separate estimation tends to bias towards the null (Prentice, 1982). The separate estimation also shows that all the three longitudinal covariates are significantly positively associated with the risk of lapse, whereas according to the MJL method, only Craving and Negative Affect are significant. Note that the absolute values of the coefficient estimates of the significant β's from MJL is much more larger than those from separate estimation.

---

[1] represents statistically significant at 0.05 level.

Table 3.6: Estimation for Lapse Data with Multiple Covariate Processes

		Separate Estimation		Joint Likelihood	
		Estimate	SE	Estimate	Bootstrap SE
	η_1	-0.07	0.12	-0.22	0.17
	η_2	-0.07	0.13	-0.45*	0.19
	β_{crav}	0.05*	0.01	0.13*	0.02
	β_{na}	0.05*	0.025	-0.004	0.05
	β_{fatig}	0.008	0.01	-0.001	0.02
Crav	μ_1	4.63*	0.10	4.64*	0.09
	μ_2	-0.17*	0.01	-0.16*	0.01
	Σ_{11}	7.86	-	7.89	0.36
	Σ_{12}	-0.24	-	-0.23	0.04
	Σ_{22}	0.04	-	0.04	0.01
	σ^2	2.04	-	4.16	0.16
NA	μ_1	1.83*	0.05	1.83*	0.05
	μ_2	-0.06*	0.004	-0.06*	0.005
	Σ_{11}	2.00	-	2.00	0.14
	Σ_{12}	-0.08	-	-0.08	0.01
	Σ_{22}	0.01	-	0.01	0.001
	σ^2	0.95	-	0.91	0.04
Fatig	μ_1	2.72*	0.11	2.72*	0.11
	μ_2	-0.005	0.01	-0.006	0.01
	Σ_{11}	8.96	-	8.98	0.55
	Σ_{12}	-0.18	-	-0.18	0.06
	Σ_{22}	0.10	-	0.10	0.02
	σ^2	1.52	-	2.19	0.12

Table 3.6 summarizes the results of joint models with three longitudinal covariates modeled simultaneously. Similar to the single-covariate models, the estimating results are close across the two methods for longitudinal parameters but show substantial difference for survival coefficients. The two active treatments are again estimated to be non-significant by the separate estimation, whereas the MJL method implies the combined therapy has significant effect in reducing the risk of lapse. Separate estimation shows both Craving and Negative Affect have significant positive effects on lapse, whereas MJL indicates that Craving is the only

significant longitudinal factor associated with lapse when the three longitudinal covariates are modeled together.

Compared with the results in Table 3.5, we found that the process of Negative Affect is identified to be a significant covariate for lapse in the single-covariate model, but becomes nonsignificant in the multiple-covariates model. The latter inference is consistent with the conjecture of the smoking cessation study that Negative Affect exerts influence on patients mainly through the feeling of Craving for smoking and research that showed that Craving, not Negative Affect, was the key mediator of treatment effects in this sample (Bolt et al., 2012). Thus, using multiple-covariates model instead of single-covariate models does provide more reasonable insight into the analysis in this case.

Table 3.7: Estimation for Relapse Data with Single Covariate Process

	Craving				Negative Affect				Fatigue			
	Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood		Separate Estimation		Joint Likelihood	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
η_1	-0.04	0.12	-0.57	0.32	-0.03	0.12	-0.56	0.30	-0.02	0.12	-0.57	0.32
η_2	-0.05	0.12	-0.66*	0.33	-0.04	0.11	-0.63	0.32	-0.04	0.12	-0.70*	0.33
β	0.04*	0.01	0.19*	0.04	0.08*	0.02	0.41*	0.07	0.002	0.01	0.06*	0.03
μ_1	4.63*	0.10	4.64*	0.11	1.82*	0.05	1.82*	0.05	2.74*	0.10	2.74*	0.12
μ_2	-0.15*	0.01	-0.15*	0.01	-0.05*	0.004	-0.05*	0.005	-0.003	0.01	-0.002	0.01
Σ_{11}	7.81	-	7.84	0.29	1.94	-	1.95	0.14	7.82	-	8.91	0.50
Σ_{12}	-0.22	-	-0.22	0.03	-0.08	-	-0.08	0.01	-0.22	-	-0.15	0.04
Σ_{22}	0.04	-	0.04	0.004	0.01	-	0.01	0.001	0.04	-	0.07	0.01
σ^2	2.11	-	4.45	0.15	0.98	-	0.97	0.04	2.10	-	2.40	0.13

Table 3.8: Estimation for Relapse Data with Multiple Covariate Processes

		Separate Estimation		Joint Likelihood	
		Estimate	SE	Estimate	Bootstrap SE
	η_1	-0.03	0.12	-0.55	0.33
	η_2	-0.03	0.12	-0.61	0.35
	β_{crav}	0.02*	0.01	0.11*	0.05
	β_{na}	0.05	0.03	0.31*	0.10
	β_{fatig}	0.01	0.01	-0.01	0.04
Crav	μ_1	4.64*	0.11	4.64*	0.10
	μ_2	-0.15*	0.01	-0.15*	0.01
	Σ_{11}	7.82	-	7.82	0.37
	Σ_{12}	-0.22	-	-0.22	0.03
	Σ_{22}	0.04	-	0.04	0.004
	σ^2	2.11	-	4.16	0.17
NA	μ_1	1.82*	0.05	1.83*	0.05
	μ_2	-0.05*	0.004	-0.06*	0.005
	Σ_{11}	1.94	-	1.94	0.14
	Σ_{12}	-0.08	-	-0.08	0.01
	Σ_{22}	0.01	-	0.01	0.001
	σ^2	0.98	-	0.97	0.04
Fatig	μ_1	2.74*	0.11	2.74*	0.12
	μ_2	-0.003	0.01	-0.003	0.01
	Σ_{11}	8.92	-	8.93	0.50
	Σ_{12}	-0.15	-	-0.16	0.04
	Σ_{22}	0.07	-	0.07	0.01
	σ^2	1.55	-	2.30	0.11

Table 3.7 and Table 3.8 present the estimation results for the models with time to relapse as the survival response. The results of the single-covariate models are summarized in Table 3.7, where all the three covariate processes are identified to be significantly positively associated with relapse by the MJL method. The combined pharmacotherapy is shown to have significant effect on reducing the risk of relapse by the MJL method when modeled with Craving or Fatigue. The two active treatments are still nonsignificant in separate estimation.

In Table 3.8 both the combined pharmacotherapy and the cessation fatigue are no longer significant by the MJL method, even though the absolute values of the coefficient estimates increase dramatically compared with those of the separate estimation. Note that in the Relapse Data we have an extremely high censoring rate (91.24%), which means the information might be insufficient to make accurate inference on the survival coefficients. However, the fact that the estimation results are different between the single-covariate models and the multiple-covariate model still indicates the necessity of applying them separately to different questions according to specific needs.

3.4 Discussion

We propose a likelihood-based method to estimate the joint models with multiple longitudinal covariates for the time-to-event response. Measurement errors and the covariance structure among the multiple longitudinal covariates are considered in the model. We established the asymptotic properties of the resulting MLE in Chapter 4. To circumvent the computational complexity involved in evaluating the large-dimensional integral in the likelihood function, we introduce the DoIt algorithm, which is implemented in R and combined with an EM algorithm to speed up the convergence. In real data analysis, we find that the models with multiple biomarkers (i.e., longitudinal covariates) generated different results from the models with single biomarker, and the former offered a more comprehensive interpretation of the data. Although in this data example we only demonstrate the feasibility of the proposed approach with three biomarkers, the algorithm can accommodate the data with a larger number of longitudinal predictors of interest. The computing feasibility makes it possible to consider more complex joint model settings in future work. For example, one may consider nonparametric mixed-effects model for the longitudinal processes, or generalized mixed-effects model for the non-continuous longitudinal processes. However, a limitation to our and the similar approach is that there was no explicit form of the standard errors for the estimators, the numerical solution was difficult, and we had to bootstrap the standard errors at the cost of additional computing time.

Chapter 4

Joint Likelihood Estimation for Joint Modeling Survival and Multiple Longitudinal Processes: Theory

In this chapter, we establish the asymptotic theorems from the MLE's obtained by the maximum joint likelihood approach proposed in Chapter 3. The technical conditions and the theorems are stated in section 4.1, and the technical proofs are given in section 4.2.

4.1 Sampling properties

Let $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ be the maximum likelihood estimators maximizing the likelihood equation given in (3.8). Zeng and Cai (2005) established the consistency and asymptotic normality of the joint maximum likelihood estimate for joint model with single covariate process. In this section, we adopt their statistical formulation. It is worth pointing out that since we have multivariate response in the longitudinal model, the covariance structure for the measurement errors in $\boldsymbol{\theta}$ is specified by the covariance matrix $\boldsymbol{\Sigma}_e$ instead of the scalar σ_e^2 for the single longitudinal response in Zeng and Cai (2005). This makes the theoretical proof much more challenging than that for single covariate process. For completeness, we state the technical conditions and theorems below and provide the proof details of Theorem 1 and

2 that follow and extend the work of Zeng and Cai (2005). The assumptions are made for any given subject in the study. The notations, if not respecified, follow the ones defined in section 2.

(A.1) Denote τ to be the end of the study time, $\mathcal{T}$ to be the longitudinal observation time, and $\mathcal{Z}$ to be the baseline covariates. In the interval $[0, \tau]$ and given the random effects $\mathbf{b}$, $\mathcal{T}$ and $\mathcal{Z}$ are conditionally independent of all the random variables in joint model (3.5) and (3.6).

(A.2) With probability one, every dimension of the functional vector $\boldsymbol{\rho}(t)$ and $\tilde{\boldsymbol{\rho}}(t)$ is continuously differentiable in $[0, \tau]$, and

$$\max_{t \in [0,\tau]} \|\boldsymbol{\rho}'(t)\| < \infty, \quad \max_{t \in [0,\tau]} \|\tilde{\boldsymbol{\rho}}'(t)\| < \infty.$$

In addition, the baseline covariates $\mathcal{Z}$ are bounded with probability one.

(A.3) Conditional on $\mathcal{T}$ and $\mathcal{Z}$, the censoring time C is non-informative for the joint model (i.e., given $\mathcal{T}$ and $\mathcal{Z}$, C is independent of T, $\mathcal{W}$ and $\mathbf{b}$).

(A.4) Let N be the number of longitudinal observations. There exists an integer n_0 such that $P(N \leq n_0) = 1$. In addition, with probability one,

$$P(N > 2d | \mathcal{T}, \mathcal{Z}, T) > 0, \quad P(N > \tilde{d} | \mathcal{T}, \mathcal{Z}, T) > 0,$$

where d is the total dimension of the random effects $\mathbf{b}$, and $\tilde{d}$ is the total dimension of the $(\tilde{\boldsymbol{\rho}}_1, \ldots, \tilde{\boldsymbol{\rho}}_p)$ for $\boldsymbol{\mu}$.

(A.5) The maximal right-censoring time is equal to τ.

(A.6) At time t_j, $j = 1, \ldots, N$, denote the value of $\boldsymbol{\rho}(t_j)$ and $\tilde{\boldsymbol{\rho}}(t_j)$ by $\boldsymbol{\rho}_j$ and $\tilde{\boldsymbol{\rho}}_j$, respectively. Remember that $\boldsymbol{\rho}_j$ is a $d \times p$ matrix, and $\tilde{\boldsymbol{\rho}}_j$ is a $\tilde{d} \times p$ matrix, where $d \geq p$ and $\tilde{d} \geq p$. Both $P(\boldsymbol{\rho}_j$ is full rank$) > 0$ and $P(\tilde{\boldsymbol{\rho}}_j$ is full rank$) > 0$ for all $j = 1, \ldots, N$. In addition, define the $d \times pN$ matrix $\mathbf{R} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_N)$, then $P(\mathbf{R}\mathbf{R}^T$ is full rank$) > 0$.

(A.7) There exist constant vectors $\boldsymbol{\beta}_c$ and $\boldsymbol{\eta}_c$ (of the same dimension as $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, respectively) such that, with positive probability, $\boldsymbol{\beta}_c^T(\boldsymbol{\rho}(t)^T\mathbf{1}) = \beta_{c1}(\boldsymbol{\rho}_1(t)^T\mathbf{1}) +$

$\cdots + \beta_{cp}(\boldsymbol{\rho}_p(t)^T\mathbf{1}) = g(t)$ and $\boldsymbol{\eta}_c^T\mathbf{Z} = 0$ for a deterministic function $g(t)$ for all $t \in [0, \tau]$, then $\boldsymbol{\beta}_c = \mathbf{0}$, $\boldsymbol{\eta}_c = \mathbf{0}$ and $g(t) = 0$.

(A.8) Let $\Theta \subseteq \mathbb{R}^{d_\theta}$ be the domain of $\boldsymbol{\theta}$, where d_θ is the dimension of $\boldsymbol{\theta}$. For any $\boldsymbol{\theta} \in \Theta$, assume $\|\boldsymbol{\theta}\| \leq M_0$, $\min_{\|\mathbf{e}\|=1} \mathbf{e}^T\boldsymbol{\Sigma}_e\mathbf{e} > M_0^{-1}, \min_{\|\mathbf{e}\|=1} \mathbf{e}^T\boldsymbol{\Sigma}_b\mathbf{e} > M_0^{-1}$ for a known positive constant M_0.

(A.9) The true baseline hazard function, denoted by $\lambda_0(t)$, is bounded and positive in $[0, \tau]$.

Remark Assumption (A.4) implies that for all the subjects, the number of longitudinal observations is bounded from above by n_0, and for at least some subjects the number of longitudinal observations is bounded from below by the $2d$. Assumption (A.8) indicates that Θ is a compact set. (A.8) and (A.9) together indicate the true hazard function is bounded and positive in $[0, \tau]$. Combined with assumption (A.2), this implies that $P(T > \tau|\mathcal{T}, \mathcal{Z}) > c_0$ for some positive constant c_0. Since (A.5) assumes that all the subjects surviving after τ censor at τ (i.e., $C = \tau$), this implies that $P(C \geq \tau|\mathcal{T}, \mathcal{Z}) = P(C = \tau|\mathcal{T}, \mathcal{Z}) > c_0$.

Recall that $\boldsymbol{\theta} = (\boldsymbol{\mu}^T, \mathrm{Vec}(\boldsymbol{\Sigma}_e), \mathrm{Vec}(\boldsymbol{\Sigma}_b), \boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ is the constant parameter set and $\Lambda(t) = \int_0^t \lambda(s)ds$ is the functional parameter of the likelihood. The observed likelihood function of $(\boldsymbol{\theta}, \Lambda)$ is

$$\begin{aligned} L(\boldsymbol{\theta}, \Lambda) = \prod_{i=1}^n \int_{\mathbf{b}} \Bigg\{ & (2\pi)^{-pN_i/2}|\boldsymbol{\Sigma}_e|^{-N_i/2} \\ & \times \exp\{-\frac{1}{2}\sum_{j=1}^{N_i}(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T\boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T\mathbf{b})\} \\ & \times \lambda(V_i)^{\Delta_i} \exp\left[\Delta_i\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}_i\} - \int_0^{V_i} e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(s)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}_i} d\Lambda(s)\right] \\ & \times (2\pi)^{-d/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\{-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}\}\Bigg\} d\mathbf{b}. \end{aligned} \tag{4.1}$$

In order to obtain the maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$, we let $\lambda(t)$ take mass only at the event times V_i for which $\Delta_i = 1$. Thus $\Lambda(t)$ becomes an increasing and right-continuous step function with jumps only at V_i, and the baseline hazard

$\lambda(t)$ in (4.1) is replaced by $\Lambda\{V_i\}$, the jump size of $\Lambda(\cdot)$ at V_i. The domain of $\Lambda(t)$ is denoted by $\mathcal{V}_n$, which consists of all the right-continuous step functions with jumps at V_i for which $\Delta_i = 1$. The domain depends on n because the number of the jumps of $\Lambda(t)$ is of order n. Denote the logarithm of the modified likelihood function by $l_n(\boldsymbol{\theta}, \Lambda)$.

Based on the above assumptions, the asymptotic properties of the maximum likelihood estimates are as follows:

Theorem 1 Under Assumptions (A.1) to (A.9), the maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ is strongly consistent under the product metric of the Euclidean norm and the supremum norm on $[0, \tau]$; that is,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0 \qquad a.s.$$

Theorem 2 Under Assumptions (A.1) to (A.9), $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\Lambda} - \Lambda_0)$ weakly converges to a Gaussian random element in $\mathbb{R}^{d_\theta} \times l^\infty[0, \tau]$, where d_θ is the dimension of $\boldsymbol{\theta}$ and $l^\infty[0, \tau]$ is the metric space of all bounded functions in $[0, \tau]$. Furthermore, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ weakly converges to a multivariate normal distribution with mean zero and its asymptotic variance attains the semiparametric efficiency bound for $\boldsymbol{\theta}_0$.

Theorem 3 Under Assumptions (A.1) to (A.9), $2\{pl_n(\hat{\boldsymbol{\theta}}) - pl_n(\boldsymbol{\theta}_0)\}$ weakly converges to a chi-square distribution with d_θ degrees of freedom and, moreover,

$$-\frac{pl_n(\hat{\boldsymbol{\theta}} + h_n\mathbf{e}) - 2pl_n(\hat{\boldsymbol{\theta}}) + pl_n(\hat{\boldsymbol{\theta}} - h_n\mathbf{e})}{nh_n^2} \xrightarrow{p} \mathbf{e}^T \boldsymbol{I} \mathbf{e},$$

where $pl_n(\boldsymbol{\theta}) = \max_{\Lambda\in\mathcal{V}_n} l_n(\boldsymbol{\theta}, \Lambda)$ is the profile likelihood function of $\boldsymbol{\theta}$. $h_n = O_p(n^{-1/2})$, $\mathbf{e}$ is any vector in $\mathbb{R}^{d_\theta}$ with unit norm, and $\boldsymbol{I}$ is the efficient information matrix for $\boldsymbol{\theta}_0$.

4.2 Technical Proofs

Below are the proofs of Theorem 1 and 2 based on the above conditions. We eliminate the proof of Theorem 3 as it follows the same arguments as in Zeng and Cai (2005).

Proof of Theorem 1

The proof of Theorem 1 is completed by verifying the following statements (i) to (iv).

(i) The maximizer of $l_n(\boldsymbol{\theta}, \Lambda)$, $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ exists for each n.

(ii) $\hat{\Lambda}(\tau)$ is bounded when n goes to infinity.

(iii) There exist a constant vector $\boldsymbol{\theta}^*$ and a right-continuous monotone function $\Lambda^*(t)$ such that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ and $\hat{\Lambda}(t)$ weakly converges to $\Lambda^*(t)$ for $t \in [0, \tau]$.

(iv) $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda^* = \Lambda_0$.

Note that all the above statements are made and hold for a fixed ω in the probability space, except for some null sets.

PROOF OF (i). Recall that by replacing $\lambda(V_i)$ with $\Lambda\{V_i\}$ in (4.1), the objective function becomes

$$
\begin{aligned}
& l_n(\boldsymbol{\theta}, \Lambda) \\
=& \sum_{i=1}^{n} \log \int_{\mathbf{b}} \Big\{ (2\pi)^{-pN_i/2} |\boldsymbol{\Sigma}_e|^{-N_i/2} \times \Lambda\{V_i\}^{\Delta_i} \\
& \times \exp\{-\frac{1}{2} \sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})\} \\
& \times \exp\left[\Delta_i \{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(V_i)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i\} - \sum_{j=1}^{n} I(V_j \le V_i) \Lambda\{V_j\} e^{\boldsymbol{\beta}^T (\boldsymbol{\rho}_i(V_j)^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i} \right] \\
& \times (2\pi)^{-d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\{-\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\} \Big\} d\mathbf{b},
\end{aligned}
$$

and $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ maximizes $l_n(\boldsymbol{\theta}, \Lambda)$ over the set $\{(\boldsymbol{\theta}, \Lambda) : \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}_n\}$. Since it is easy to verify that $l_n(\boldsymbol{\theta}, \Lambda)$ is concave, the existence of $\hat{\boldsymbol{\theta}}$ holds because its domain Θ is compact. Thus we only need to verify the existence of $\hat{\Lambda}$, which is satisfied when $\mathcal{V}_n$ is compact. Hence it suffices to prove that the jump size of Λ at V_i for which $\Delta_i = 1$ is finite.

Since for any $x > 0$, $e^x \geq 1 + x$, and $e^{-x} \leq (1+x)^{-1}$, for each i we have

$$\begin{aligned}
&\exp\left[-\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i}\right] \\
\leq&\left[1+\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i}\right]^{-1} \\
\leq&\left[\sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i}\right]^{-1} \\
\leq&\Lambda\{V_{j0}^{(i)}\}^{-1}e^{-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_{j0}^{(i)})^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}},
\end{aligned}$$

where $V_{j0}^{(i)}$ is any observed event time in the set $\{V_j : V_j \leq V_i, \Delta_j = 1, j = 1, \ldots, n\}$.

Hence, for any i such that $\Delta_i = 1$, take $V_{j0}^{(i)} = V_i$, the survival part of the likelihood (4.2) satisfies

$$\begin{aligned}
&\Lambda\{V_i\}^{\Delta_i}\exp\left[\Delta_i\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\} - \sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i}\right] \\
\leq&\Lambda\{V_i\}\exp\left[\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right] \times \Lambda\{V_i\}^{-1}\exp\left[-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right] = 1.
\end{aligned}$$

For those i with $\Delta_i = 0$, take any $V_{j0}^{(i)} \in \{V_j : V_j \leq V_i, \Delta_j = 1, j = 1, \ldots, n\}$, the second line of the likelihood (4.2) satisfies

$$\begin{aligned}
&\Lambda\{V_i\}^{\Delta_i}\exp\left[\Delta_i\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_i)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\} - \sum_{j=1}^{n} I(V_j \leq V_i)\Lambda\{V_j\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_j)^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i}\right] \\
\leq&\Lambda\{V_{j0}^{(i)}\}^{-1}\exp\left[-\{\boldsymbol{\beta}^T(\boldsymbol{\rho}_i(V_{j0}^{(i)})^T\mathbf{b})+\boldsymbol{\eta}^T\mathbf{Z}_i\}\right].
\end{aligned}$$

Accordingly, the likelihood (4.2) satisfies

$$
\begin{aligned}
l_n(\boldsymbol{\theta}, \Lambda) \le & \sum_{i=1}^{n} I(\Delta_i = 1) \log \int_{\mathbf{b}} \Big\{ (2\pi)^{-pN_i/2} |\boldsymbol{\Sigma}_e|^{-N_i/2} \\
& \times \exp\{-\sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})\} \\
& \times (2\pi)^{-d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\{-\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\} \Big\} \, d\mathbf{b}. \\
& + \sum_{i=1}^{n} I(\Delta_i = 0) \log \int_{\mathbf{b}} \Big\{ (2\pi)^{-pN_i/2} |\boldsymbol{\Sigma}_e|^{-N_i/2} \\
& \times \exp\{-\sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \boldsymbol{\mu} - \boldsymbol{\rho}_{ij}^T \mathbf{b})\} \\
& \times \Lambda\{V_{j0}^{(i)}\}^{-\Delta_i} \exp\Big\{ -(\boldsymbol{\beta}^T (\boldsymbol{\rho}_i (V_{j0}^{(i)})^T \mathbf{b}) + \boldsymbol{\eta}^T \mathbf{Z}_i) \Big\} \\
& \times (2\pi)^{-d/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\{-\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\} \Big\} \, d\mathbf{b}.
\end{aligned}
$$

Thus if $\Lambda\{V_{j0}^{(i)}\} \to \infty$ for some i and j_0, it follows that $l_n(\boldsymbol{\theta}, \Lambda) \to -\infty$. We conclude the jump size of Λ is finite. Therefore, the maximum likelihood estimate $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ exists. $\square$

PROOF OF (ii). Define $\hat{\xi} = \log \hat{\Lambda}(\tau)$ and $\tilde{\Lambda} = \hat{\Lambda}/e^{\hat{\xi}}$. Thus $\tilde{\Lambda}(\tau) = 1$. Note that after rescaling, $\tilde{\Lambda}(t) = \hat{\Lambda}(t)/\hat{\Lambda}(\tau)$ is the ratio of jump size at time t to that at time τ, and the range of $\tilde{\Lambda}$ is $[0, 1]$. To prove (ii), it is sufficient to show that $\hat{\xi}$ is bounded when n goes to infinity. By applying some algebra to (4.1), for any $\Lambda \in \mathcal{V}_n$, the average of the likelihood at $\hat{\boldsymbol{\theta}}$ over the n subjects is given by

$$
\begin{aligned}
& \frac{1}{n} l_n(\hat{\boldsymbol{\theta}}, \Lambda) \qquad (4.2) \\
= & - \frac{\sum_{i=1}^{n} N_i}{n} \log\{(2\pi)^{-p/2} |\hat{\boldsymbol{\Sigma}}_e|^{-1/2}\} - \log\{(2\pi)^{-d/2} |\hat{\boldsymbol{\Sigma}}_b|^{-1/2}\} \\
& + \frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i (\log \Lambda\{V_i\} + \hat{\boldsymbol{\eta}}^T \mathbf{Z}_i) - \sum_{j=1}^{N_i} \frac{1}{2} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \hat{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}_e^{-1} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \hat{\boldsymbol{\mu}}) \right] \\
& + \frac{1}{n} \sum_{i=1}^{n} \log \int_{\mathbf{b}} \exp\left\{ -\frac{1}{2} \mathbf{b}^T \boldsymbol{G}_i^{-1} \mathbf{b} + \boldsymbol{h}_i^T \mathbf{b} - \int_0^{V_i} e^{Q_{1i}(t, b, \hat{\theta})} d\Lambda(t) \right\} d\mathbf{b}, \qquad (4.3)
\end{aligned}
$$

where

$$\begin{aligned}
&\boldsymbol{G}_i^{-1} = \hat{\boldsymbol{\Sigma}}_b^{-1} + \sum_{j=1}^{N_i} \boldsymbol{\rho}_{ij} \hat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{\rho}_{ij}^T, \\
&\boldsymbol{h}_i^T = \Delta_i \left\{ \hat{\boldsymbol{\beta}} \boldsymbol{\rho}_i(V_i)^T \right\} + \sum_{j=1}^{N_i} (\mathbf{W}_{ij} - \tilde{\boldsymbol{\rho}}_{ij}^T \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}_e^{-1} \boldsymbol{\rho}_{ij}^T, \\
&Q_{1i}(t, b, \hat{\theta}) = \hat{\boldsymbol{\beta}} \left\{ \boldsymbol{\rho}_i(V_i)^T \mathbf{b} \right\} + \hat{\boldsymbol{\eta}}^T \mathbf{Z}_i.
\end{aligned}$$

Let $\tilde{\mathbf{b}} = \boldsymbol{G}_i^{-1/2}(\mathbf{b} - \boldsymbol{G}_i \boldsymbol{h}_i)$, thus the third part of the right-hand side of equation (4.2) becomes

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{2} \log |\boldsymbol{G}_i| + \frac{1}{2} \boldsymbol{h}_i^T \boldsymbol{G}_i \boldsymbol{h}_i + \log \int_{\tilde{\mathbf{b}}} \exp\{ -\frac{1}{2} \tilde{\mathbf{b}}^T \tilde{\mathbf{b}} - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \hat{\theta})} d\Lambda(t) \} d\tilde{\mathbf{b}} \right\},$$

where

$$\tilde{Q}_{1i}(t, \tilde{b}, \hat{\theta}) = \{ \boldsymbol{\beta} \boldsymbol{\rho}_i(t)^T \} \boldsymbol{G}_i^{1/2} \tilde{\mathbf{b}} + \{ \boldsymbol{\beta} \boldsymbol{\rho}_i(t)^T \} \boldsymbol{G}_i \boldsymbol{h}_i + \boldsymbol{\eta}^T \mathbf{Z}_i.$$

Since $\hat{\xi}$ maximizes the log-likelihood $l_n(\hat{\boldsymbol{\theta}}, e^{\xi}\tilde{\Lambda})$, we have $l_n(\hat{\boldsymbol{\theta}}, e^{\hat{\xi}}\tilde{\Lambda}) \geq l_n(\hat{\boldsymbol{\theta}}, e^{0}\tilde{\Lambda})$. It follows that

$$\begin{aligned}
0 \leq & n^{-1} l_n(\hat{\boldsymbol{\theta}}, e^{\hat{\xi}}\tilde{\Lambda}) - n^{-1} l_n(\hat{\boldsymbol{\theta}}, \tilde{\Lambda}) \\
= & \frac{1}{n} \sum_{i=1}^{n} \left[\Delta_i \hat{\xi} + \log \int_{\tilde{\mathbf{b}}} \exp \left\{ -\frac{\tilde{\mathbf{b}}^T \tilde{\mathbf{b}}}{2} - e^{\hat{\xi}} \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \hat{\theta})} d\tilde{\Lambda}(t) \right\} d\tilde{\mathbf{b}} \right. \\
& \left. - \log \int_{\tilde{\mathbf{b}}} \exp \left\{ -\frac{\tilde{\mathbf{b}}^T \tilde{\mathbf{b}}}{2} - \int_0^{V_i} e^{\tilde{Q}_{1i}(t, \tilde{b}, \hat{\theta})} d\tilde{\Lambda}(t) \right\} d\tilde{\mathbf{b}} \right].
\end{aligned} \tag{4.4}$$

Note that according to assumptions (A.2), (A.4) and the boundedness of $\boldsymbol{\theta}$, there exist some positive constants C_1, C_2, and C_3 such that

$$|\tilde{Q}_{i1}(t, \tilde{b}, \hat{\theta})| \leq C_1 \|\tilde{\mathbf{b}}\| + C_2 \sum_{j=1}^{N_i} \|\mathbf{W}_{ij}\| + C_3 \leq C_1 \|\tilde{\mathbf{b}}\| + C_2 n_0 \|\mathbf{W}_{ij}\| + C_3. \tag{4.5}$$

Since $\tilde{\mathbf{b}}$ is of a standard multivariate normal distribution, applying the above inequality (4.5), we obtain

$$
\begin{aligned}
&-\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-\int_0^{V_i}e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}\\
&\qquad=(2\pi)^{d/2}\log E_{\tilde{b}}\exp\left\{-\int_0^{V_i}e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\} \qquad (4.6)\\
&\qquad\geq(2\pi)^{d/2}\log E_{\tilde{b}}\exp\left\{-e^{C_1\|\tilde{\mathbf{b}}\|+C_2n_0\|\mathbf{W}_{ij}\|+C_3}\right\}.
\end{aligned}
$$

Using Jensen's inequality, the above expression satisfies

$$
\begin{aligned}
&(2\pi)^{d/2}\log E_{\tilde{b}}\exp\left\{-e^{C_1\|\tilde{\mathbf{b}}\|+C_2n_0\|\mathbf{W}_{ij}\|+C_3}\right\}\\
\geq&(2\pi)^{d/2}E_{\tilde{b}}\left\{-e^{C_1\|\tilde{\mathbf{b}}\|+C_2n_0\|\mathbf{W}_{ij}\|+C_3}\right\} \qquad (4.7)\\
=&-e^{C_2n_0\|\mathbf{W}_{ij}\|+C_4}
\end{aligned}
$$

for some constant C_4. Since $\mathbf{W}_{ij}$ is normally distributed, by the strong law of large numbers, there exist some positive constant C_5 such that

$$
\frac{1}{n}\sum_{i=1}^{n}e^{C_2n_0\|\mathbf{W}_{ij}\|+C_4}\to Ee^{C_2n_0\|\mathbf{W}_{1j}\|+C_4}=C_5 \qquad a.s.
$$

Thus

$$
-\frac{1}{n}\sum_{i=1}^{n}\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}-\int_0^{V_i}e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}
$$

is bounded by C_5 from above when n goes to infinity. Then (4.4) becomes

$$\begin{aligned}
0 &\le \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} + \frac{1}{n}\sum_{i=1}^{n}\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2} - e^{\hat{\xi}}\int_0^{V_i} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}} + C_5\\
&\le \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} + \frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2} - e^{\hat{\xi}}\int_0^{\tau} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}\\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}I(V_i\neq\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}d\tilde{\mathbf{b}} + C_5\\
&\le \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} + \frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2} - e^{\hat{\xi}}\int_0^{\tau} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}\\
&\qquad + C_6,
\end{aligned} \tag{4.8}$$

for some constant C_6. The last inequality follows by the integral of the unnormalized standard normal density of $\tilde{\mathbf{b}}$. Since for any $\Gamma \ge 0$, $x \ge 0$, we have $e^{x/\Gamma} \ge (1 + x/\Gamma)$, it follows that $e^{-x} \le (1 + x/\Gamma)^{-\Gamma}$. Using the similar arguments as in (4.6) and (4.7), the following inequality holds:

$$\begin{aligned}
&\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2} - e^{\hat{\xi}}\int_0^{\tau} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}d\tilde{\mathbf{b}}\\
&\le \int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{1 + \frac{e^{\hat{\xi}}}{\Gamma}\int_0^{\tau} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}^{-\Gamma}d\tilde{\mathbf{b}}\\
&\le \left(\frac{\Gamma}{e^{\hat{\xi}}}\right)^{\Gamma}\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{\int_0^{\tau} e^{\tilde{Q}_{1i}(t,\tilde{b},\hat{\theta})}d\tilde{\Lambda}(t)\right\}^{-\Gamma}d\tilde{\mathbf{b}}\\
&\le \left(\frac{\Gamma}{e^{\hat{\xi}}}\right)^{\Gamma}\int_{\tilde{\mathbf{b}}}\exp\left\{-\frac{\tilde{\mathbf{b}}^T\tilde{\mathbf{b}}}{2}\right\}\left\{e^{-C_1\|\tilde{\mathbf{b}}\| - C_2 n_0\|\mathbf{w}_{ij}\| - C_3}\right\}^{-\Gamma}d\tilde{\mathbf{b}}\\
&= \left(\frac{\Gamma}{e^{\hat{\xi}}}\right)^{\Gamma}(2\pi)^{d/2}E_{\tilde{b}}\left\{e^{\Gamma C_1\|\tilde{\mathbf{b}}\| + \Gamma C_2 n_0\|\mathbf{w}_{ij}\| + \Gamma C_3}\right\}\\
&= \left(\frac{\Gamma}{e^{\hat{\xi}}}\right)^{\Gamma}e^{\Gamma C_2 n_0\|\mathbf{w}_{ij}\| + C_7(\Gamma)},
\end{aligned}$$

where $C_7(\Gamma)$ is a deterministic function of Γ. Applying the above conclusion to

(4.8), we obtain

$$
\begin{aligned}
0 \leq & \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} + \frac{1}{n}\sum_{i=1}^{n} I(V_i = \tau)\Gamma\Big(\log\Gamma - \hat{\xi} + C_2 n_0\|\mathbf{W}_{ij}\| + C_7(\Gamma)\Big) + C_6 \\
\leq & \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} - \frac{\Gamma}{n}\sum_{i=1}^{n} I(V_i = \tau)\hat{\xi} + C_2 n_0 \frac{\Gamma}{n}\sum_{i=1}^{n}\|\mathbf{W}_{ij}\| + C_8(\Gamma)
\end{aligned}
\tag{4.9}
$$

where $C_8(\Gamma)$ is a deterministic function of Γ. By the strong law of large numbers,

$$
C_2 n_0 \frac{\Gamma}{n}\sum_{i=1}^{n}\|\mathbf{W}_{ij}\| + C_8(\Gamma) \to C_9(\Gamma), \qquad a.s.,
$$

where $C_9(\Gamma)$ is also a deterministic function of Γ. Thus for some sufficiently large n, it follows that

$$
0 \leq \frac{1}{n}\sum_{i=1}^{n}\Delta_i\hat{\xi} - \frac{\Gamma}{n}\sum_{i=1}^{n} I(V_i = \tau)\hat{\xi} + C_9(\Gamma).
$$

Take Γ large enough such that $\frac{1}{n}\sum_{i=1}^{n}\Delta_i \leq \frac{\Gamma}{2n}\sum_{i=1}^{n} I(V_i = \tau)$, then by assumption (A.5) and the strong law of large numbers,

$$
\begin{aligned}
& 0 \leq C_9(\Gamma) - \frac{\Gamma}{2n}\sum_{i=1}^{n} I(V_i = \tau)\hat{\xi} \\
& \hat{\xi} \leq 2C_9(\Gamma)/\frac{\Gamma}{n}\sum_{i=1}^{n} I(V_i = \tau) \to 2C_9(\Gamma)/\Gamma P(V = \tau) = B_0 \qquad a.s.
\end{aligned}
\tag{4.10}
$$

Thus $\hat{\xi}$ is bounded by some constant B_0. Since the above statement holds for every ω in the sample space except the null set, we conclude that with probability 1, $\hat{\Lambda}(\tau)$ is bounded for any n. $\square$

PROOF OF (iii). By the assumption (A.8) that $\Theta \in \mathbb{R}^{d_\theta}$ is a compact set, there exists a subsequence of $\hat{\boldsymbol{\theta}}_n$ and a constant vector $\boldsymbol{\theta}^* \in \Theta$ such that the subsequence converges to $\boldsymbol{\theta}^*$. If we can show in (iv) that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, the unique true parameter, let $\hat{\boldsymbol{\theta}}_{n_m}$ be a subsequence of $\hat{\boldsymbol{\theta}}_n$ that does not converge to $\boldsymbol{\theta}^*$. Thus, $\exists\ \delta_0 > 0$ and some large M s.t. $\|\hat{\boldsymbol{\theta}}_{n_m} - \boldsymbol{\theta}^*\| > \delta_0$ for all the $m > M$. However, since $\hat{\boldsymbol{\theta}}_{m_n} \in \Theta$

and the limit $\boldsymbol{\theta}^*$ is unique, there exists a sub-subsequence $\hat{\boldsymbol{\theta}}_{m_{n_k}}$ s.t. $\hat{\boldsymbol{\theta}}_{m_{n_k}} \to \boldsymbol{\theta}^*$ as $k \to \infty$. This contradict with the previous statement of the non-convergence of $\hat{\boldsymbol{\theta}}_{m_n}$. Thus all the subsequence of $\hat{\boldsymbol{\theta}}_n$ coverage to $\boldsymbol{\theta}^*$. We conclude that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ as $n \to \infty$.

Now consider $\hat{\Lambda}$. By Helly's Selection Theorem, there exists a subsequence of $\hat{\Lambda}_n(t)$ that weakly converges to some right-continuous monotone function $\Lambda^*(t)$ for each $t \in [0, \tau]$. Using the similar argument as for $\hat{\boldsymbol{\theta}}$, if we can show in (iv) that $\Lambda^* = \Lambda_0$, the unique true functional parameter, then $\hat{\Lambda}_n$ itself converges to Λ^* as n goes to infinity. Note that since the convergence holds for every ω in the sample space except the null sets, $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ and $\hat{\Lambda}_n \to \Lambda^*$ with probability 1. Thus the only thing left to prove is $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda^* = \Lambda_0$. $\square$

PROOF OF (iv). For a given subject, let $\mathbf{O} = \{\mathbf{W}_j, \boldsymbol{\rho}(s), \tilde{\boldsymbol{\rho}}_j, V, \Delta, \mathbf{Z}, j = 1, \ldots, N, 0 \le s \le t\}$ be the observed data. Denote

$$\begin{aligned} &G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda) \\ =&(2\pi)^{-pN/2}|\boldsymbol{\Sigma}_e|^{-N/2}(2\pi)^{d/2}|\boldsymbol{\Sigma}_b|^{-1/2} \\ &\times \exp\Bigg\{ -\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b}) - \frac{\mathbf{b}^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}}{2} \\ &+ \Delta\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}\} - \int_0^V e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(s)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}} d\Lambda(s)\Bigg\}. \end{aligned}$$

In addition, define

$$Q(V, \mathbf{O}; \boldsymbol{\theta}, \Lambda) = \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda)\exp\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}\}d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}, \Lambda)d\mathbf{b}}$$

to be the posterior expectation of $\exp\{\boldsymbol{\beta}^T(\boldsymbol{\rho}(V)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{Z}\}$ with respect to $\mathbf{b}$ given $(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$. For any measurable function $f(\mathbf{O})$, we use operator notation to define

$$\mathbf{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{O}_i), \quad \text{and}$$

$$\mathbf{P} f = \int f d\mathbf{P} = E[f(\mathbf{O})].$$

Thus $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$ is the associated empirical process based on $\mathbf{O}$. Define the class $\mathcal{F} = \{Q(V, \mathbf{O}; \boldsymbol{\theta}, \Lambda) : V \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}, \Lambda(0) = 0, \Lambda(\tau) \leq B_0\}$, where B_0 is the upper bound of $\Lambda(\tau)$ given in the proof of (ii), and $\mathcal{V}$ contains all the right-continuous monotone functions in $[0, \tau]$. Using the same techniques as in Appendix A.1 of Zeng and Cai (2005), it is easy to verify that $\mathcal{F}$ is P-Donsker.

Differentiating $l_n(\boldsymbol{\theta}, \Lambda)$ in (4.2) with respect to $\Lambda\{V_k\}$ and setting the equation to zero, we obtain

$$\hat{\Lambda}\{V_k\} = \frac{\Delta_k}{n\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\}|_{v=V_k}}. \tag{4.11}$$

Using the same structure, define

$$\bar{\Lambda}\{V_k\} = \frac{\Delta_k}{n\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}}. \tag{4.12}$$

Thus

$$\bar{\Lambda}(t) = \sum_{k=1}^{n} I(V_k \leq t)\bar{\Lambda}\{V_k\} = \frac{1}{n}\sum_{k=1}^{n} \frac{I(V_k \leq t)\Delta_k}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}}.$$

Since $(\boldsymbol{\theta}_0, \Lambda_0)$ maximizes $E[l_n(\boldsymbol{\theta}, \Lambda)]$, by taking the derivative of $E[l_n(\boldsymbol{\theta}, \Lambda)]$ with respect to Λ and setting the equation to 0, it can be verified that

$$\Lambda_0(t) = E\left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}}\right].$$

$$\begin{aligned}
&\sup_{t\in[0,\tau]} |\bar{\Lambda}(t) - \Lambda_0(t)| \\
&\leq \sup_{t\in[0,\tau]} \left| \frac{1}{n}\sum_{k=1}^{n} I(V_k \leq t)\Delta_k \left[\frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}\right]\Bigg|_{v=V_k} \right| \\
&\quad + \sup_{t\in[0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}}\right] \right| \\
&\leq \sup_{v\in[0,\tau]} \left| \frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
&\quad + \sup_{t\in[0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\frac{I(V_k \leq t)\Delta_k}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V_k}}\right] \right|
\end{aligned} \tag{4.13}$$

By Zeng and Cai (2005),

$$\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}$$

and

$$\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}$$

are bounded from below. Thus the first part of the above inequality (4.13) satisfies

$$\begin{aligned}
&\sup_{v\in[0,\tau]} \left| \frac{1}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} - \frac{1}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
= &\sup_{v\in[0,\tau]} \left| \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}} \right| \\
\leq &C_{10} \sup_{v\in[0,\tau]} \left| \mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} \right|.
\end{aligned} \tag{4.14}$$

for some constant C_{10}.

Using the same argument as in Appendix A.1 of Zeng and Cai (2005), it can be verified that $\{Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) : v \in [0, \tau]\}$ is a bounded Glivenko-Cantelli class. Since $\{I(V \geq v) : v \in [0, \tau]\}$ is also a Glivenko-Cantelli class, and the functional $(f, g) \mapsto fg$ for any two bounded functions f and g is Lipschitz continuous, $\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) : v \in [0, \tau]\}$ is a bounded Glivenko-Cantelli class. Thus when n goes to infinity

$$\sup_{v\in[0,\tau]} |\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}| \longrightarrow 0. \tag{4.15}$$

Hence the right-hand side of the above inequality (4.14) converge to 0 as n goes to infinity, and the first part of (4.13) disappears.

Similarly, since the class $\{I(V \leq t)/\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}|_{v=V} : t \in [0, \tau]\}$ is also a Glivenko-Cantelli class, the second part of the above inequality also converges to zero as n goes to infinity. Therefore by inequality (4.13) we conclude that

$$\sup_{t\in[0,\tau]} \|\bar{\Lambda}(t) - \Lambda_0(t)\| \to 0,$$

that is, $\bar{\Lambda}$ uniformly converges to Λ_0 in $[0, \tau]$.

Using the expressions of $\hat{\Lambda}$ and $\bar{\Lambda}$ in (4.11) and (4.12), we have

$$\frac{\hat{\Lambda}\{v\}}{\bar{\Lambda}\{v\}} = \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\}},$$

and accordingly,

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\}} d\bar{\Lambda}(v). \tag{4.16}$$

This implies that $\hat{\Lambda}(t)$ is absolutely continuous with respect to $\bar{\Lambda}(t)$.

Since $\{I(V \geq v) : v \in [0, \tau]\}$ and $\mathcal{F}$ are both Glivenko-Cantelli classes, $\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}, \Lambda) : v \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda \in \mathcal{V}, \Lambda_0(0) = 0, \Lambda(\tau) \leq B_0\}$ is also a Glivenko-Cantelli class. Thus,

$$\sup_{v \in [0,\tau]} |(\mathbf{P}_n - \mathbf{P})\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\}| \to 0 \qquad a.s. \tag{4.17}$$

Since $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ and $\hat{\Lambda}$ weakly converges to Λ^*, by bounded convergence theorem, for each $v \in [0, \tau]$, when n goes to infinity,

$$\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\} \to \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}. \tag{4.18}$$

Using assumption (A.2), it is easy to check that the derivative of $\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}$ with respect to v is uniformly bounded in $[0, \tau]$. Thus $\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}$ is equi-continuous. By the Arzela-Ascoli theorem which states that a bounded and equi-continuous functional sequence has uniformly convergent subsequence, we strengthen the conclusion in (4.18) to uniform convergence:

$$\sup_{v \in [0,\tau]} |\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}| \to 0, \quad n \to \infty.$$

This conclusion, together with (4.17), implies that

$$\sup_{v\in[0,\tau]} |\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\} - \mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}| \to 0. \quad (4.19)$$

By the conclusion of (4.19) and (4.15), it follows that

$$\frac{\hat{\Lambda}\{v\}}{\bar{\Lambda}\{v\}} = \frac{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}_n\{I(V \geq v)Q(v, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda})\}} \to \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}}, \quad (4.20)$$

uniformly in $[0, \tau]$.

Taking the limit with respect to n on both sides of (4.16), and applying the conclusion of (4.20), we obtain

$$\Lambda^*(t) = \int_0^t \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}} d\Lambda_0(v). \quad (4.21)$$

The above equation (4.21) indicates that both $\Lambda_0(t)$ and $\Lambda^*(t)$ are differentiable with respect to the Lebesgue measure. Denote $\lambda^*(t)$ to be the derivative of $\Lambda^*(t)$, and $\lambda_0(t)$ to be the derivative of $\Lambda_0(t)$. It follows from (4.21) that

$$\frac{\lambda^*(v)}{\lambda_0(v)} = \frac{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\}}{\mathbf{P}\{I(V \geq v)Q(v, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*)\}}.$$

Thus (4.20) implies that

$$\frac{\hat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \to \frac{\lambda^*(V)}{\lambda_0(V)} \quad (4.22)$$

uniformly in $[0, \tau]$. Integrating the denominator and numerator over V in $[0, v]$ on both sides in (4.22), we obtain that $\hat{\Lambda}(v)/\bar{\Lambda}(v)$ uniformly converges to $\Lambda^*(v)/\Lambda_0(v)$ in $[0, \tau]$. Since by (4.17) we already know that $\bar{\Lambda}(v)$ uniformly converges to $\Lambda_0(v)$ in $[0, \tau]$ for the denominators in (4.5), we conclude that $\hat{\Lambda}(v)$ uniformly converges to $\Lambda^*(v)$ in $[0, \tau]$.

Since $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ maximizes $l_n(\boldsymbol{\theta}, \Lambda)$, it follows that

$$0 \leq \frac{1}{n} l_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}) - \frac{1}{n} l_n(\boldsymbol{\theta}_0, \bar{\Lambda}) = \mathbf{P}_n \left[\Delta \log \frac{\hat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] + \mathbf{P}_n \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right]. \tag{4.23}$$

Similar as Zeng and Cai (2005), it is easy to verify that $\Delta \log[\hat{\Lambda}\{V\}/\bar{\Lambda}\{V\}]$ and $\log[\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) d\mathbf{b} / \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}]$ are both Glivenko-Cantelli classes. Thus

$$\begin{aligned} \sup_{V \in [0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\Delta \log \frac{\hat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] \right| &\to 0, \\ \sup_{V \in [0,\tau]} \left| (\mathbf{P}_n - \mathbf{P}) \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right] \right| &\to 0. \end{aligned} \tag{4.24}$$

Consider the first part on the right-hand side of equation (4.23). By (4.22) we know that $\hat{\Lambda}\{V\}/\bar{\Lambda}\{V\}$ uniformly converges to $\lambda^*(V)/\lambda_0(V)$, thus by applying the bounded convergence theorem, we obtain

$$\mathbf{P} \left[\Delta \log \frac{\hat{\Lambda}\{V\}}{\bar{\Lambda}\{V\}} \right] \to \mathbf{P} \left[\Delta \log \frac{\lambda^*(V)}{\lambda_0(V)} \right]. \tag{4.25}$$

Similarly, for the second part on the right-hand side of (4.23), since $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$, $\hat{\Lambda}$ uniformly converges to Λ^* and $\bar{\Lambda}$ uniformly converges to Λ_0, applying the bounded convergence theorem again, we obtain

$$\mathbf{P} \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \bar{\Lambda}) d\mathbf{b}} \right] \to \mathbf{P} \left[\log \frac{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*) d\mathbf{b}}{\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) d\mathbf{b}} \right]. \tag{4.26}$$

Combining the conclusions of (4.24), (4.25) and (4.26), and taking the limit with respect to n on both sides of (4.23), we obtain

$$\mathbf{P} \left[\log \left\{ \frac{\lambda^*(V)^{\Delta} \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*) d\mathbf{b}}{\lambda_0(V)^{\Delta} \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \theta_0, \Lambda_0) d\mathbf{b}} \right\} \right] \geq 0.$$

Since the measure $\mathbf{P}$ is with respect to the distribution with true parameter $(\boldsymbol{\theta}_0, \Lambda_0)$, the left-hand side of the above inequality is the negative Kullback-Leibler information. Then it follows that, with probability one,

$$\lambda^*(V)^\Delta \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}^*, \Lambda^*) d\mathbf{b} = \lambda_0(V)^\Delta \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) d\mathbf{b}. \tag{4.27}$$

According to assumption (A.8) and (A.9), $P(V \leq \tau, \Delta = 1 | \mathcal{T}, \mathcal{Z}) > 0$ with probability one. Thus equation (4.27) holds for the set $\{(V, \Delta) : V \in [0, \tau], \Delta = 1\}$. By assumption (A.5), $P(V = \tau, \Delta = 0 | \mathcal{T}, \mathcal{Z}) = P(C = \tau, \Delta = 0 | \mathcal{T}, \mathcal{Z}) > 0$ with probability one. Thus equation (4.27) also holds for the set $\{V = \tau, \Delta = 0\}$. However, since assumption (A.5), (A.8) and (A.9) imply that $P(C \geq \tau | \mathcal{T}, \mathcal{Z}) > c_0$ with probability one for some positive constant c_0, $P(V < \tau, \Delta = 0 | \mathcal{T}, \mathcal{Z}) = P(C < \tau, \Delta = 0 | \mathcal{T}, \mathcal{Z}) < P(C < \tau | \mathcal{T}, \mathcal{Z}) < 1 - c_0$ with probability one. Thus the set $\{(V, \Delta) : V \in [0, \tau), \Delta = 0\}$ might be a null set for which equation (4.27) may not hold. Zeng and Cai (2005) proved that equation (4.27) also holds for $\{(V, \Delta) : V \in [0, \tau), \Delta = 0\}$.

Let $\Delta = 0$ and $V = 0$ in equation (4.27). Since the expressions inside the integrals on both sides of (4.27) are quadratic functions of $\mathbf{b}$, after integrating over $\mathbf{b}$, we obtain that the following equation holds with probability one:

$$\begin{aligned}
&|\boldsymbol{\Sigma}_e^*|^{-N/2}|\boldsymbol{\Sigma}_b^*|^{-1/2}|\boldsymbol{\Sigma}_b^{*-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j|^{-1/2}\\
&\times\exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)\right.\\
&\left.+\frac{1}{2}\sum_{j=1}^{N}\left[(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right](\boldsymbol{\Sigma}_b^{*-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j)^{-1}\sum_{j=1}^{N}\left[\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)\right]\right\}\\
=&|\boldsymbol{\Sigma}_{0e}|^{-N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2}|\boldsymbol{\Sigma}_{0b}^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j|^{-1/2}\\
&\times\exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right.\\
&\left.+\frac{1}{2}\sum_{j=1}^{N}\left[(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu})^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right](\boldsymbol{\Sigma}_{0b}^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}\sum_{j=1}^{N}\left[\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j-\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]\right\}
\end{aligned}\tag{4.28}$$

Let $\mathbf{D}=(\boldsymbol{\Sigma}_b^{-1}+\sum_{j=1}^{N}\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_j)^{-1}$. The quadratic and linear terms of $\mathbf{W}_j$ in the exponential part yield

$$\begin{aligned}
&\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}^*\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j\right]\\
=&\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}\mathbf{W}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}_0\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j\right],
\end{aligned}\tag{4.29}$$

and

$$\begin{aligned}
&\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}^*)^T\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}^*\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_e^{*-1}\mathbf{W}_j\right]\\
=&\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j-\left[\sum_{j=1}^{N}(\tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\mathbf{D}_0\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j\right]
\end{aligned}\tag{4.30}$$

Letting $\mathbf{W}_1\neq 0$ and $\mathbf{W}_j=0$ for all $j=2,\ldots,N$, (4.29) and (4.30) become

$$\mathbf{W}_1^T(\boldsymbol{\Sigma}_e^{*-1}-\boldsymbol{\Sigma}_e^{*-1}\boldsymbol{\rho}_1^T\mathbf{D}^*\boldsymbol{\rho}_1\boldsymbol{\Sigma}_e^{*-1})\mathbf{W}_1=\mathbf{W}_1^T(\boldsymbol{\Sigma}_{0e}^{-1}-\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_1^T\mathbf{D}_0\boldsymbol{\rho}_1\boldsymbol{\Sigma}_{0e}^{-1})\mathbf{W}_1\tag{4.31}$$

$$(\tilde{\boldsymbol{\rho}}_1^T \boldsymbol{\mu}^*)^T (\boldsymbol{\Sigma}_e^{*-1} - \boldsymbol{\Sigma}_e^{*-1} \boldsymbol{\rho}_1^T \mathbf{D}^* \boldsymbol{\rho}_1 \boldsymbol{\Sigma}_e^{*-1}) \mathbf{W}_1 = (\tilde{\boldsymbol{\rho}}_1^T \boldsymbol{\mu}_0)^T (\boldsymbol{\Sigma}_{0e}^{-1} - \boldsymbol{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_1^T \mathbf{D}_0 \boldsymbol{\rho}_1 \boldsymbol{\Sigma}_{0e}^{-1}) \mathbf{W}_1 \tag{4.32}$$

Since $\mathbf{W}_1$ is arbitrary, (4.31) yields

$$\boldsymbol{\Sigma}_e^{*-1} - \boldsymbol{\Sigma}_e^{*-1} \boldsymbol{\rho}_1^T \mathbf{D}^* \boldsymbol{\rho}_1 \boldsymbol{\Sigma}_e^{*-1} = \boldsymbol{\Sigma}_{0e}^{-1} - \boldsymbol{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_1^T \mathbf{D}_0 \boldsymbol{\rho}_1 \boldsymbol{\Sigma}_{0e}^{-1}. \tag{4.33}$$

Combining this with (4.32), and using full rank assumption of $\tilde{\boldsymbol{\rho}}_j$ in (A.6), we obtain that $\boldsymbol{\mu}^* = \boldsymbol{\mu}_0$.

In order to prove $\boldsymbol{\Sigma}_e^* = \boldsymbol{\Sigma}_{0e}$ and $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{0b}$, we reexamine equality (4.29). Let $\mathbf{W} = (\mathbf{W}_1^T, \mathbf{W}_2^T, \ldots, \mathbf{W}_N^T)^T$, $\boldsymbol{\Sigma}_E^{*-1} = \mathbf{I}_N \otimes \boldsymbol{\Sigma}_e^{*-1}$ and $\boldsymbol{\Sigma}_{0E}^{-1} = \mathbf{I}_N \otimes \boldsymbol{\Sigma}_{0e}^{-1}$. Follow the assumption (A.4) and let $\mathbf{R} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_p)$. Thus equality (4.29) can be rewritten as

$$\mathbf{W}^T \boldsymbol{\Sigma}_E^{*-1} \mathbf{W} - \mathbf{W}^T \boldsymbol{\Sigma}_E^{*-1} \mathbf{R}^T \mathbf{D}^* \mathbf{R} \boldsymbol{\Sigma}_E^{*-1} \mathbf{W} = \mathbf{W}^T \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{W} - \mathbf{W}^T \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{R}^T \mathbf{D}_0 \mathbf{R} \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{W}.$$

Since both sides of the equality are quadratic forms of $\mathbf{W}$ and $\mathbf{W}$ is arbitrary in the space $\mathbb{R}^{pN}$, the equality can be reduced to the linear form

$$\boldsymbol{\Sigma}_E^{*-1} \mathbf{W} - \boldsymbol{\Sigma}_E^{*-1} \mathbf{R}^T \mathbf{D}^* \mathbf{R} \boldsymbol{\Sigma}_E^{*-1} \mathbf{W} = \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{W} - \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{R}^T \mathbf{D}_0 \mathbf{R} \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{W}. \tag{4.34}$$

Define two subspaces

$$\mathrm{Ker}^* = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{R} \boldsymbol{\Sigma}_E^{*-1} \mathbf{W} = \mathbf{0}\},$$

$$\mathrm{Ker}_0 = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{R} \boldsymbol{\Sigma}_{0E}^{-1} \mathbf{W} = \mathbf{0}\},$$

and the associated degrees of freedom are $\dim(\mathrm{Ker}^*) = \dim(\mathrm{Ker}_0) = pN - d$. According to assumption (A.4) that $N > 2d$, since $p \geq 1$, we have $pN > 2d$, thus by the inequality

$$\dim(\mathrm{Ker}^*) + \dim(\mathrm{Ker}_0) - \dim(\mathrm{Ker}^* \cap \mathrm{Ker}_0) = \dim(\mathrm{Ker}^* + \mathrm{Ker}_0) \leq pN,$$

we obtain that

$$\dim(\mathrm{Ker}^* \cap \mathrm{Ker}_0) \geq 2(pN - d) - pN = pN - 2d.$$

For any $\mathbf{W} \in \text{Ker}^* \cap \text{Ker}_0$, equation (4.34) reduces to

$$(\mathbf{\Sigma}_E^{*-1} - \mathbf{\Sigma}_{0E}^{-1})\mathbf{W} = \mathbf{0}.$$

Let $\mathbf{A} = \mathbf{\Sigma}_E^{*-1} - \mathbf{\Sigma}_{0E}^{-1}$ and $\mathbf{A}_0 = \mathbf{\Sigma}_e^{*-1} - \mathbf{\Sigma}_{0e}^{-1}$. Our goal is to prove that $\mathbf{A}_0 = \mathbf{0}$. Denote the kernel of operator $\mathbf{A}$ as

$$\text{Ker}(\mathbf{A}) = \{\mathbf{W} \in \mathbb{R}^{pN} : \mathbf{A}\mathbf{W} = \mathbf{0}\}.$$

Since $\text{Ker}^* \cap \text{Ker}_0 \subset \text{Ker}(\mathbf{A})$, we have

$$\dim(\text{Ker}(\mathbf{A})) \geq \dim(\text{Ker}^* \cap \text{Ker}_0) \geq pN - 2d. \tag{4.35}$$

Denote the ranges of the operator $\mathbf{A}$ and $\mathbf{A}_0$ as

$$\text{Ran}(\mathbf{A}) = \{\mathbf{A}\mathbf{W} : \mathbf{W} \in \mathbb{R}^{pN}\}, \quad \text{and}$$

$$\text{Ran}(\mathbf{A}_0) = \{\mathbf{A}_0\mathbf{W}_0 : \mathbf{W}_0 \in \mathbb{R}^{p}\},$$

respectively. Since $\mathbf{A} = \mathbf{I}_N \otimes \mathbf{A}_0$, it is easy to verify that

$$\dim(\text{Ran}(\mathbf{A})) = N\dim(\text{Ran}(\mathbf{A}_0)). \tag{4.36}$$

In addition, since for operator $\mathbf{A}$ we have $\dim(\text{Ker}(\mathbf{A})) + \dim(\text{Ran}(\mathbf{A})) = pN$, it follows from (4.35) and (4.36) that

$$N\dim(\text{Ran}(\mathbf{A}_0)) = \dim(\text{Ran}(\mathbf{A})) \leq pN - (pN - 2d) = 2d.$$

According to assumption (A.4) that $N > 2d$, we conclude that $\dim(\text{Ran}(\mathbf{A}_0)) = 0$. Therefore, $\mathbf{A}_0 = \mathbf{0}$, i.e., $\mathbf{\Sigma}_e^* = \mathbf{\Sigma}_{0e}$. Moreover, according to assumption (A.6) that $\mathbf{R}\mathbf{R}^T$ is full rank, it is easy to verify that $\mathbf{D}^* = \mathbf{D}_0$, which directly yields $\mathbf{\Sigma}_b^* = \mathbf{\Sigma}_{0b}$.

Next, let $\Delta = 0$ in equation (4.27), since all the density parts related to $\mathbf{W}_i$ cancel out by the above conclusions, we obtain

$$
\begin{aligned}
&E_{\mathbf{b}}\left[\exp\left\{-\int_0^V e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}}d\Lambda^*(t)\right\}\right]\\
=&E_{\mathbf{b}}\left[\exp\left\{-\int_0^V e^{\boldsymbol{\beta}_0^{T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^{T}\mathbf{z}}d\Lambda_0(t)\right\}\right],
\end{aligned}
$$

where $\mathbf{b} \sim N_d(\tilde{\mu}_b, \tilde{\boldsymbol{\Sigma}}_b)$ with

$$
\tilde{\mu}_b = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}\left[\sum_{j=1}^N \boldsymbol{\rho}_j \boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]
$$

and

$$
\tilde{\boldsymbol{\Sigma}}_b = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}.
$$

Treat $\sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j$ and $\sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j$ as fixed and $\sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j$ as parameter, $\mathbf{b}$ is complete sufficient statistics for $\sum_{j=1}^N \boldsymbol{\rho}_j^T \boldsymbol{\Sigma}_{0e}^{-1}\mathbf{W}_j$. Thus

$$
\exp\left\{-\int_0^V e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}}d\Lambda^*(t)\right\} = \exp\left\{-\int_0^V e^{\boldsymbol{\beta}_0^{T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^{T}\mathbf{z}}d\Lambda_0(t)\right\}.
$$

Equivalently,

$$
\lambda^*(t)e^{\boldsymbol{\beta}^{*T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}^{*T}\mathbf{z}} = \lambda_0(t)e^{\boldsymbol{\beta}_0^{T}(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^{T}\mathbf{z}}.
$$

Taking the logarithm on both sides of the equation and rearrange the terms, we obtain that there exits some function of time $\tilde{g}(t)$ such that

$$
\tilde{g}(t) = \log\lambda^*(t) - \log\lambda_0(t) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + (\boldsymbol{\eta}_0 - \boldsymbol{\eta}^*)^T\mathbf{Z},
$$

for any $\mathbf{b}$. According to assumption (A.7) and (A.11) we conclude that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\eta}^* = \boldsymbol{\eta}_0$ and $\Lambda^* = \Lambda_0$. $\square$

Proof of Theorem 2

Let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \Lambda) \in \Psi = \{(\boldsymbol{\theta}, \Lambda) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]}|\Lambda(t) - \Lambda_0(t)| \leq \delta\}$ for a fixed small δ. Note that Ψ is a convex set. Define a set

$$\mathcal{H} = \{(\mathbf{h}_1, h_2) : \|\mathbf{h}_1\| \leq 1, \|h_2\|_V \leq 1\},$$

where $\|h_2\|_V$ is the total variation of h_2 in $[0, \tau]$ defined as

$$\sup_{0=t_0 \leq t_1 \leq \cdots \leq t_N = \tau} \sum_{j=1}^{N} |h_2(t_j) - h_2(t_{j-1})|.$$

Recall that $l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$ is the log likelihood of a single subject. The associated Fr$\acute{e}$chet derivative is given by

$$f_{\psi,h} = l_\theta(\boldsymbol{\theta}, \Lambda)^T \mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta}, \Lambda)[h_2], \quad (\boldsymbol{\theta}, \Lambda) \in \Psi, (\mathbf{h}_1, h_2) \in \mathcal{H}, \tag{4.37}$$

where $l_\theta(\boldsymbol{\theta}, \Lambda) = \partial l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)/\partial \boldsymbol{\theta}$, and $l_\Lambda(\boldsymbol{\theta}, \Lambda)$ is the derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \Lambda_\epsilon)$ with respect to ϵ at $\epsilon = 0$, where $\Lambda_\epsilon(t) = \int_0^t (1 + \epsilon h_2(s)) d\Lambda_0(s)$. Define empirical processes

$$S_n(\boldsymbol{\psi})(\mathbf{h}_1, h_2) = \mathbf{P}_n f_{\psi,h},$$

$$S(\boldsymbol{\psi})(\mathbf{h}_1, h_2) = \mathbf{P} f_{\psi,h}.$$

By the definition of $f_{\psi,h}$, S_n and S are both maps from Ψ to $l^\infty(\mathcal{H})$ (i.e., the collection of all bounded functions from $\mathcal{H}$ to $\mathbb{R}$).

The asymptotic normality of $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ is established by checking the four conditions of the following theorem (Van Der Vaart and Wellner (1996), Theorem 3.3.1):

Theorem Let Ψ be a subset of a Banach space that contains the true parameter $\boldsymbol{\psi}_0$. Let S be a fixed map and S_n be a series of random maps, both of which map from Ψ to a Banach space such that

(a) $\sqrt{n}(S_n - S)(\hat{\boldsymbol{\psi}}_n) - \sqrt{n}(S_n - S)(\boldsymbol{\psi}_0) = o_p^*(1 + \sqrt{n}\|\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0\|)$;

(b) The sequence $\sqrt{n}(S_n - S)(\boldsymbol{\psi}_0)$ converges in distribution to a tight random element $\mathbf{Z}_0$;

(c) The function $\boldsymbol{\psi} \to S(\boldsymbol{\psi})$ is Fr$\acute{e}$chet differentiable at $\boldsymbol{\psi}_0$ with a continuously invertible derivative ∇S_{ψ_0} on its range;

(d) $S(\psi_0) = 0$. $\hat{\boldsymbol{\psi}}_n$ satisfies $S_n(\hat{\boldsymbol{\psi}}_n) = o_p^*(n^{-1/2})$ and $\hat{\boldsymbol{\psi}}_n$ converges in outer probability to $\boldsymbol{\psi}_0$.

Then, $\sqrt{n}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0) \Rightarrow -\nabla S_{\boldsymbol{\psi}_0}^{-1}\mathbf{Z}_0$.

We first check condition (a). According to Van Der Vaart and Wellner (1996), when the observations are independent and identical (iid), the theorem can be applied with $S_n(\boldsymbol{\psi})\mathbf{h} = \mathbf{P}_n f_{\psi,h}$ and $S(\boldsymbol{\psi})\mathbf{h} = \mathbf{P} f_{\psi,h}$, where $f_{\psi,h}$ is a measurable function indexed by Ψ and $\mathcal{H}$. In this case, for given $\boldsymbol{\psi} \in \Psi$,

$$\sqrt{n}(S_n - S)(\boldsymbol{\psi})\mathbf{h} = \sqrt{n}(\mathbf{P}_n - \mathbf{P})f_{\psi,h} \triangleq \{G_n f_{\psi,h} : \mathbf{h} = (\mathbf{h}_1, h_2) \in \mathcal{H}\}$$

is an empirical process indexed by the class $\{f_{\psi,h} : \mathbf{h} \in \mathcal{H}\}$. Thus, for the iid case, condition (a) in the above theorem becomes

$$\|G_n(f_{\hat{\psi}_n,h} - f_{\psi_0,h})\| = o_p^*(1 + \sqrt{n}\|\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_0\|). \tag{4.38}$$

Therefore, using the measurable function $f_{\psi,h}$ defined in (4.37), we only need to verify (4.38) instead of condition (a).

Lemma 3.3.5 in Van Der Vaart and Wellner (1996) provides sufficient conditions for (4.38). It claims that (4.38) holds if $\hat{\boldsymbol{\psi}}_n$ converges in outer probability to $\boldsymbol{\psi}_0$, and the following tow conditions are satisfied

(a.1) $\{f_{\psi,h} - f_{\psi_0,h} : \boldsymbol{\psi} \in \Psi, \mathbf{h} \in \mathcal{H}\}$ is P-Donsker,

(a.2) $\sup_{h\in\mathcal{H}} P(f_{\hat{\psi}_n,h} - f_{\psi_0,h})^2 \to 0$ as $\hat{\boldsymbol{\psi}} \to \psi_0$.

Since the convergence of $\hat{\boldsymbol{\psi}}_n$ to $\boldsymbol{\psi}_0$ is justified by Theorem 1, and both (a.1) and (a.2) are verified in the Appendix A.2 of Zeng and Cai (2005), equation (4.38) holds accordingly. Thus condition (a) is satisfied.

For condition (b), since it is easy to verify that the class $\{f_{\psi_0,h} : \mathbf{h} \in \mathcal{H}\}$ is P-Donsker (see (Zeng and Cai, 2005), Appendix A.2, for more details), there exists a tight Gaussian process $\mathbf{Z}_0 \in l^\infty(\mathcal{H})$ such that the empirical process $\sqrt{n}(S_n - S)(\boldsymbol{\psi}_0) = \sqrt{n}(\mathbf{P}_n - \mathbf{P})f_{\psi_0,h}$ converges to $\mathbf{Z}_0$ in distribution.

For condition (d), $S(\boldsymbol{\psi}_0) = 0$ and $S(\hat{\boldsymbol{\psi}}_n) = 0$ because $(\boldsymbol{\theta}_0, \Lambda_0)$ maximizes $\mathbf{P}l(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$, and $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ maximizes $\mathbf{P}_n(\mathbf{O}; \boldsymbol{\theta}, \Lambda)$. Thus with the consistency result of theorem 1, condition (d) is also satisfied.

It remains to verify condition (c). By the definition of Fr*é*chet differentiablility, $S(\boldsymbol{\psi})$ is Fr*é*chet differentiable at $\boldsymbol{\psi}_0$ if there exists a linear operator $A_{\boldsymbol{\psi}_0} : \Psi \mapsto l^\infty(\mathcal{H})$ such that

$$\begin{aligned} &S(\boldsymbol{\psi})(\mathbf{h}_1, h_2) - S(\boldsymbol{\psi}_0)(\mathbf{h}_1, h_2) \\ &= A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)(\mathbf{h}_1, h_2) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t \in [0,\tau]} |\Lambda(t) - \Lambda_0(t)|)(\|\mathbf{h}_1\| + \|h_2\|_V) \end{aligned} \tag{4.39}$$

for any $(\mathbf{h}_1, h_2) \in \mathcal{H}$. The existence of $A_{\boldsymbol{\psi}_0}$ is proved below.

Let $(\mathbf{h}_1^e, \mathbf{h}_1^b, \mathbf{h}_1^\mu, \mathbf{h}_1^\beta, \mathbf{h}_1^\eta)$ be the components of $\mathbf{h}_1$ corresponding to each of the parameters $(\boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_b, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta})$. Thus $f_{\psi,h} = l_\theta(\boldsymbol{\theta}, \Lambda)^T \mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta}, \Lambda)[h_2]$ can be written out explicitly in the expression

$$g_1(\mathbf{O}; \boldsymbol{\theta}, \Lambda)^T \mathbf{h}_1 - \int_0^V g_2(t, \mathbf{O}; \boldsymbol{\theta}, \Lambda)^T \mathbf{h}_1 d\Lambda(t) + \Delta h_2(V) - \int_0^V g_3(t, \mathbf{O}; \boldsymbol{\theta}, \Lambda) h_2(t) d\Lambda(t), \tag{4.40}$$

where

$$
\begin{aligned}
& g_1(\mathbf{O};\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 \\
&= \left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1} \\
&\quad \times \int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)\left[\frac{1}{2}\mathbf{b}^T\Sigma_b^{-1}\mathcal{D}_b\Sigma_b^{-1}\mathbf{b} - \frac{1}{2}tr(\boldsymbol{\Sigma}_b^{-1}\mathcal{D}_b)\right. \\
&\quad + \frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b})^T\boldsymbol{\Sigma}_e^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_e^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b}) \\
&\quad - \frac{N}{2}tr(\boldsymbol{\Sigma}_e^{-1}\mathcal{D}_e) \\
&\quad \left. + \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu} - \boldsymbol{\rho}_j^T\mathbf{b})^T\Sigma_e^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu} + \Delta\{\mathbf{h}_1^{\beta T}(\boldsymbol{\rho}(V)^T\mathbf{b}) + \mathbf{h}_1^{\eta T}\mathbf{Z}\}\right] d\mathbf{b},
\end{aligned}
$$

$$
\begin{aligned}
& g_2(t,\mathbf{O};\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 \\
&= \left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1} \\
&\quad \times \int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda) \times \left[\{\mathbf{h}_1^{\beta T}(\boldsymbol{\rho}(t)^T\mathbf{b}) + \mathbf{h}_1^{\eta T}\mathbf{Z}\}e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{z}}\right] d\mathbf{b},
\end{aligned}
$$

$$
\begin{aligned}
& g_3(t,\mathbf{O};\boldsymbol{\theta},\Lambda) \\
&= \left\{\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda)d\mathbf{b}\right\}^{-1}\int_{\mathbf{b}} G(\mathbf{b},\mathbf{O};\boldsymbol{\theta},\Lambda) \times e^{\boldsymbol{\beta}^T(\boldsymbol{\rho}(t)^T\mathbf{b}) + \boldsymbol{\eta}^T\mathbf{z}}d\mathbf{b}.
\end{aligned}
$$

Here, $\mathcal{D}_b$ and $\mathcal{D}_e$ are symmetric matrices such that $\text{Vec}(\mathcal{D}_b) = \mathbf{h}_1^b$ and $\text{Vec}(\mathcal{D}_e) = \mathbf{h}_1^e$, respectively.

For $j = 1, 2, 3$, we denote $\nabla_\theta g_j$ to be the derivative of g_j with respect to $\boldsymbol{\theta}$, and denote $\nabla_\Lambda g_j[\delta\Lambda]$ to be the derivative of g_j with respect to Λ along the path $\Lambda + \epsilon\delta\Lambda$. It is easy to check that for $j = 1, 2, 3$, the derivative of g_j with respect to Λ along the path $\Lambda + \epsilon\delta\Lambda$ can be expressed as $\nabla_\Lambda g_j[\delta\Lambda] = \int_0^t g_{j+3}(s,\mathbf{O};\boldsymbol{\theta},\Lambda)d\delta\Lambda(s)$ for some $g_k(s,\mathbf{O};\boldsymbol{\theta},\Lambda)$, $k = 4, 5, 6$. Thus, by the mean value theorem, for any $(\boldsymbol{\theta},\Lambda,\mathbf{h}_1,h_2)$ in $\Psi \times \mathcal{H}$,

$$
\begin{aligned}
& l_\theta(\boldsymbol{\theta},\Lambda)^T\mathbf{h}_1 + l_\Lambda(\boldsymbol{\theta},\Lambda)[h_2] - l_\theta(\boldsymbol{\theta}_0,\Lambda_0)^T\mathbf{h}_1 - l_\Lambda(\boldsymbol{\theta}_0,\Lambda_0)[h_2] \\
=& (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\left\{\nabla_\theta g_1(\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - \int_0^V \nabla_\theta g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})d\Lambda_0(t)\right\}\mathbf{h}_1 \\
& + \mathbf{h}_1^T\int_0^\tau I(t\le V)\Big\{g_4(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) \\
& - g_5(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V d\Lambda_0(s)\Big\}d(\Lambda-\Lambda_0)(t) \\
& - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)d\Lambda_0(t) \\
& - \int_0^\tau\Big\{I(t\le V)g_6(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V h_2(s)d\Lambda_0(s) \\
& + I(t\le V)g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)\Big\}d(\Lambda-\Lambda_0)(t),
\end{aligned}
$$

where $(\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) = \xi^*(\boldsymbol{\theta},\Lambda)+(1-\xi^*)(\boldsymbol{\theta}_0,\Lambda_0)$ for some $\xi^*\in[0,1]$. Thus by the definition of $S(\boldsymbol{\psi})(\mathbf{h}_1,h_2)$, it follows that

$$
\begin{aligned}
& S(\boldsymbol{\psi})(\mathbf{h}_1,h_2) - S(\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2) \\
=& (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{P}\left\{\nabla_\theta g_1(\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - \int_0^V \nabla_\theta g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})d\Lambda_0(t)\right\}\mathbf{h}_1 \\
& + \mathbf{h}_1^T\int_0^\tau \mathbf{P}\Big[I(t\le V)\Big\{g_4(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) - g_2(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda}) \\
& - g_5(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V d\Lambda_0(s)\Big\}\Big]d(\Lambda-\Lambda_0)(t) \\
& - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau \mathbf{P}I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)d\Lambda_0(t) \\
& - \int_0^\tau \mathbf{P}\Big\{I(t\le V)g_6(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})\int_t^V h_2(s)d\Lambda_0(s) \\
& + I(t\le V)g_3(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})h_2(t)\Big\}d(\Lambda-\Lambda_0)(t).
\end{aligned}
$$

Following the above equation, define $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)$ by the expression

$$\begin{aligned}
&A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)\\
=&(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{P}\left\{\nabla_\theta g_1(\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-\int_0^V\nabla_\theta g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)d\Lambda_0(t)\right\}\mathbf{h}_1\\
&+\mathbf{h}_1^T\int_0^\tau\mathbf{P}\left[I(t\le V)\left\{g_4(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)-g_2(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\right.\right.\\
&-\left.\left.g_5(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V d\Lambda_0(s)\right\}\right]d(\Lambda-\Lambda_0)(t)\\
&-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\int_0^\tau\mathbf{P}I(t\le V)\nabla_\theta g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)h_2(t)d\Lambda_0(t)\\
&-\int_0^\tau\mathbf{P}\left\{I(t\le V)g_6(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\int_t^V h_2(s)d\Lambda_0(s)\right.\\
&\left.+I(t\le V)g_3(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)h_2(t)\right\}d(\Lambda-\Lambda_0)(t).
\end{aligned}\tag{4.41}$$

According to the appendix A.3 of Zeng and Cai (2005), for $j=1,\ldots,6$, the following inequalities hold for some constant r_1 and r_2:

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\|g_j(t,\mathbf{O};\tilde{\boldsymbol{\theta}},\tilde{\Lambda})-g_j(t,\mathbf{O};\boldsymbol{\theta}_0,\Lambda_0)\|\\
&\le e^{r_1+r_2\sum_{j=1}^N\|W_j\|}\left\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\sup_{t\in[0,\tau]}|\Lambda(t)-\Lambda_0(t)|\right\}.
\end{aligned}$$

Using this inequality, it is convenient to check that equation (4.39) holds for $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$ defined in (4.41). Therefore, $S(\boldsymbol{\psi}_0)$ is Fr$\acute{e}$chet differentiable at $\boldsymbol{\psi}_0$, and we can denote

$$\nabla S_{\psi_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)=A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$$

to be the derivative of $S(\boldsymbol{\psi}_0)$ at $\boldsymbol{\psi}_0$. Note that similar to $S(\boldsymbol{\psi})$, ∇S_{ψ_0} is a function mapping from Ψ to $l^\infty(\mathcal{H})$. It remains to show that ∇S_{ψ_0} is continuously invertible on its range in $l^\infty(\mathcal{H})$.

From the definition of $A_{\boldsymbol{\psi}_0}(\boldsymbol{\psi}-\boldsymbol{\psi}_0)(\mathbf{h}_1,h_2)$ in (4.41), it is clear that ∇S_{ψ_0} can be rewritten into

$$\begin{aligned} &\nabla S_{\psi_0}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_1 - \Lambda_2)(\mathbf{h}_1, h_2) \\ &= (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d(\Lambda_1 - \Lambda_2)(t), \end{aligned} \tag{4.42}$$

where

$$\begin{aligned} \Omega_1[\mathbf{h}_1, h_2] =& \mathbf{h}_1^T \mathbf{P} \left\{ \nabla_\theta g_1(\mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) - \int_0^V \nabla_\theta g_2(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) d\Lambda_0(t) \right\} \\ & - \int_0^\tau \mathbf{P} I(t \le V) \nabla_\theta g_3(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) h_2(t) d\Lambda_0(t), \end{aligned}$$

$$\begin{aligned} \Omega_2[\mathbf{h}_1, h_2] =& \mathbf{h}_1^T \int_0^\tau \mathbf{P} \left[I(t \le V) \left\{ g_4(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) - g_2(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \right.\right. \\ & \left.\left. - \; g_5(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \int_t^V d\Lambda_0(s) \right\} \right] \\ & - \int_0^\tau \mathbf{P} \left\{ I(t \le V) g_6(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \int_t^V h_2(s) d\Lambda_0(s) \right\} \\ & - \mathbf{P}\{ I(t \le V) g_3(t, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \} h_2(t). \end{aligned}$$

From the above definitions, the operator $\Omega = (\Omega_1, \Omega_2)$ can be taken as a linear operator that maps from $\mathcal{H} \subset \mathbb{R}^d \times BV[0,\tau]$ to itself, where $BV[0,\tau]$ contains all the functions with finite total variation in $[0,\tau]$.

From equation (4.42), the operator $(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_1 - \Lambda_2)$ can be treated as a functional element in $l^\infty(\mathcal{H})$ via the definition

$$(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_1 - \Lambda_2)(\mathbf{h}_1, h_2) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \mathbf{h}_1 + \int_0^\tau h_2(t) d(\Lambda_1 - \Lambda_2)(t)$$

for any $(\mathbf{h}_1, h_2) \in \mathbb{R}^d \times BV[0,\tau]$. Thus by (4.42), the function ∇S_{ψ_0} can be regarded as a linear operator from $l^\infty(\mathcal{H})$ to itself, and for any $(\delta\boldsymbol{\theta}, \delta\Lambda) \in l^\infty(\mathcal{H})$ the norm of ∇S_{ψ_0} is given by

$$\begin{aligned}
\|\nabla S_{\psi_0}(\delta\boldsymbol{\theta},\delta\Lambda)\|_{l^\infty(\mathcal{H})} &= \sup_{(h_1,h_2)\in\mathcal{H}} |\nabla S_{\psi_0}(\delta\boldsymbol{\theta},\delta\Lambda)(\mathbf{h}_1,h_2)| \\
&= \sup_{(h_1,h_2)\in\mathcal{H}} \left|\delta\boldsymbol{\theta}^T\Omega_1[\mathbf{h}_1,h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1,h_2]d\delta\Lambda(t)\right| \\
&= \sup_{\Omega([h_1,h_2])\in\Omega(\mathcal{H})} |(\delta\boldsymbol{\theta},\delta\Lambda)\Omega[\mathbf{h}_1,h_2]| \\
&= \|(\delta\boldsymbol{\theta},\delta\Lambda)\|_{l^\infty(\Omega(\mathcal{H}))}.
\end{aligned}$$

Thus if we can find some positive constant ε such that $\varepsilon\mathcal{H}\subset\Omega(\mathcal{H})$, it follows that

$$\|(\delta\boldsymbol{\theta},\delta\Lambda)\|_{l^\infty(\Omega(\mathcal{H}))} \geq \varepsilon\|(\delta\boldsymbol{\theta},\delta\Lambda)\|_{l^\infty(\mathcal{H})},$$

and ∇S_{ψ_0} is hence continuously invertible.

According to Zeng and Cai (2005), in order to show that $\varepsilon\mathcal{H}\subset\Omega(\mathcal{H})$ for some ε (i.e., Ω is invertible), it is sufficient to verify that Ω is one-to-one. Since Ω is linear, it is left to prove that if $\Omega[\mathbf{h}_1,h_2]=0$, then $\mathbf{h}_1=0$ and $h_2=0$.

If $\Omega[\mathbf{h}_1,h_2]=0$, by choosing $\boldsymbol{\theta}_1-\boldsymbol{\theta}_2=\tilde{\varepsilon}\mathbf{h}_1$ and $\Lambda_1-\Lambda_2=\tilde{\varepsilon}\int h_2 d\Lambda_0$ in (4.42) for a small constant $\tilde{\varepsilon}$, we obtain that $\nabla S_{\psi_0}(\mathbf{h}_1,\int h_2 d\Lambda_0)[\mathbf{h}_1,h_2]=0$. Note that the left-hand side is the negative information matrix in the submodel with parameter $(\boldsymbol{\theta}_0+\tilde{\varepsilon}\mathbf{h}_1,\Lambda_0+\tilde{\varepsilon}\int h_2 d\Lambda_0)$. The corresponding score equation should also equal 0. That is, $l_\theta(\boldsymbol{\theta}_0,\Lambda_0)^T\mathbf{h}_1+l_\Lambda(\boldsymbol{\theta}_0,\Lambda_0)[h_2]=0$. Thus, using the notation $(\mathbf{h}_1^e,\mathbf{h}_1^b,\mathbf{h}_1^\mu,\mathbf{h}_1^\beta,\mathbf{h}_1^\eta),\mathcal{D}_b,\mathcal{D}_e$ defined above, together with the expression of (4.2), the following equation holds with probability one:

$$
\begin{aligned}
0 = &\int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \\
&\times \Bigg[\frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b \boldsymbol{\Sigma}_{0b}^{-1} \mathbf{b} - \frac{1}{2} tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b) - \frac{N}{2} tr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e) \\
&\quad + \frac{1}{2}\sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e \boldsymbol{\Sigma}_{0e}^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b}) \\
&\quad + \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \tilde{\boldsymbol{\rho}}_j^T \mathbf{h}_1^{\mu} + \Delta\{(\boldsymbol{\rho}(V)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta}\} \\
&\quad - \int_0^V \{(\boldsymbol{\rho}(t)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta}\} e^{\boldsymbol{\beta}_0^T (\boldsymbol{\rho}(t)^T \mathbf{b}) + \boldsymbol{\eta}_0^T \mathbf{z}} d\Lambda_0(t) \Bigg] d\mathbf{b} \\
&+ \int_{\mathbf{b}} G(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \left[\Delta h_2(V) - \int_0^V h_2(t) e^{\boldsymbol{\beta}_0^T (\boldsymbol{\rho}(t)^T \mathbf{b}) + \boldsymbol{\eta}_0^T \mathbf{z}} d\Lambda_0 \right] d\mathbf{b}.
\end{aligned}
\tag{4.43}
$$

Using the same argument as for equation (4.31) in the proof for consistency, we obtain that (4.43) holds for all $\{(V, \Delta) : \Delta = 0, V \in [0, \tau]\}$. Let $\Delta = 0$ and $V = 0$, then (4.43) becomes

$$
\begin{aligned}
0 = &\int_{\mathbf{b}} G_0(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0) \\
&\times \Bigg\{ \frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}_{0b}^{-1} \mathcal{D}_b \boldsymbol{\Sigma}_{0b}^{-1} \mathbf{b} - \frac{1}{2} tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b) - \frac{N}{2} tr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e) \\
&\quad + \frac{1}{2}\sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \mathcal{D}_e \boldsymbol{\Sigma}_{0e}^{-1} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b}) \\
&\quad + \sum_{j=1}^{N} (\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T \boldsymbol{\mu}_0 - \boldsymbol{\rho}_j^T \mathbf{b})^T \boldsymbol{\Sigma}_{0e}^{-1} \tilde{\boldsymbol{\rho}}_j^T \mathbf{h}_1^{\mu} + \Delta\{(\boldsymbol{\rho}(V)^T \mathbf{b})^T \mathbf{h}_1^{\beta} + \mathbf{Z}^T \mathbf{h}_1^{\eta}\} \Bigg\} d\mathbf{b},
\end{aligned}
\tag{4.44}
$$

where

$$\begin{aligned}
&G_0(\mathbf{b}, \mathbf{O}; \boldsymbol{\theta}_0, \Lambda_0)\\
=&(2\pi)^{(pN+d)/2}|\boldsymbol{\Sigma}_{0e}|^{-N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2}\\
&\times \exp\left\{-\frac{1}{2}\mathbf{b}^T(\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)\mathbf{b} + \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\mathbf{b}\right\} \qquad (4.45)\\
&\times \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right\}.
\end{aligned}$$

Thus, using the same technique as in the proof of consistency, $\mathbf{b}$ can be treated as a random vector from the normal distribution $N_d(\boldsymbol{\nu}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^{N}\boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1}$ and $\boldsymbol{\nu} = \boldsymbol{\Gamma}\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right]$. Hence equation (4.44) can be treated as the expectation of a quadratic function of $\mathbf{b}$, thus having the explicit form

$$\begin{aligned}
0 =&(2\pi)^{-pN/2}|\mathbf{D}|^{1/2}|\boldsymbol{\Sigma}_{0e}|^{N/2}|\boldsymbol{\Sigma}_{0b}|^{-1/2}\\
&\times \exp\left\{\frac{1}{2}\left[\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\Gamma}\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j\boldsymbol{\mu}_0)^T\right]\right.\\
&\left.-\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right\}\\
&\times\left\{\frac{1}{2}tr\left[(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T)\boldsymbol{\Gamma}\right] - \frac{1}{2}tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b) - \frac{N}{2}tr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e)\right.\\
&+\frac{1}{2}\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0) + \boldsymbol{\nu}^T\boldsymbol{\Sigma}_{0b}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}\boldsymbol{\nu}\\
&+\frac{1}{2}\boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\nu} - \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\nu}\\
&\left.+\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu} - \boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\right]\mathbf{h}_1^{\mu}\right\}
\end{aligned}$$

Rearranging the above equation and canceling the non-negative multipliers, we obtain

$$\begin{aligned}
0 =& tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}^{-1}\boldsymbol{\Gamma}) + tr(\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\Gamma}) - tr(\boldsymbol{\Sigma}_{0b}^{-1}\mathcal{D}_b) - Ntr(\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e) \\
&+ \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0) + \boldsymbol{\nu}^T\boldsymbol{\Sigma}_{0b}\mathcal{D}_b\boldsymbol{\Sigma}_{0b}\boldsymbol{\nu} \\
&+ \boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\right]\boldsymbol{\nu} - 2\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\mathcal{D}_e\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\boldsymbol{\nu} \\
&+ 2\sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu} - 2\boldsymbol{\nu}^T\left[\sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\right]\mathbf{h}_1^{\mu}
\end{aligned} \tag{4.46}$$

Since the right-hand side of (4.46) is a second-order polynomial of $(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)$ and $\mathbf{W}_j$ is arbitrary for all $j = 1, \ldots, N$, the first and second order terms are zero, respectively. Let $\mathbf{x}_j = \mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0$.

We first check the first-order terms $\mathbf{x}_j$. Let $\tilde{\mathbf{E}} = \sum_{j=1}^{N}\boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T$, we obtain

$$\begin{aligned}
0 &= \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\tilde{\boldsymbol{\rho}}_j^T\mathbf{h}_1^{\mu} - \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j^T\tilde{\mathbf{E}}\mathbf{h}_1^{\mu} \\
&= \sum_{j=1}^{N}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{0e}^{-1}(\tilde{\boldsymbol{\rho}}_j - \tilde{\mathbf{E}}\boldsymbol{\rho}_j)^T\mathbf{h}_1^{\mu}
\end{aligned} \tag{4.47}$$

Since $(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)$ is arbitrary in $\mathbb{R}^p$, it follows that

$$\boldsymbol{\Sigma}_{0e}^{-1}(\tilde{\boldsymbol{\rho}}_j^T - \boldsymbol{\rho}_j^T\tilde{\mathbf{E}})\mathbf{h}_1^{\mu} = \mathbf{0}, \quad \text{for } j = 1, \ldots, N.$$

Defining the $pN \times pN$ matrix $\mathbf{A} = \boldsymbol{\Sigma}_{0e}^{-1} \otimes I_N$, and the $pN \times \tilde{d}$ matrix $\mathbf{B} = (\tilde{\boldsymbol{\rho}}_1^T - \boldsymbol{\rho}_1^T\tilde{\mathbf{E}}, \ldots, \tilde{\boldsymbol{\rho}}_N^T - \boldsymbol{\rho}_N^T\tilde{\mathbf{E}})^T$, the above equation can be rewritten as

$$\mathbf{A}\mathbf{B}\mathbf{h}_1^{\mu} = \mathbf{0}.$$

Multipling both sides of the above equation by $\mathbf{A}^{-1}$, we obtain

$$\mathbf{B}\mathbf{h}_1^{\mu} = \mathbf{0}.$$

Then multipling both sides of the above equation by $\mathbf{B}^T$, we obtain

$$\mathbf{B}^T\mathbf{B}\mathbf{h}_1^{\mu} = \mathbf{0}.$$

By assumption (A.4), since $N > \tilde{d}$ and $p \geq 1$, it follows that $pN > \tilde{d}$. Then the $\tilde{d} \times \tilde{d}$ matrix $\mathbf{B}^T\mathbf{B}$ is of full rank. Thus we conclude that $\mathbf{h}_1^{\mu} = \mathbf{0}$.

Next, checking the second-order terms of $\mathbf{x}_j$ in (4.46), we obtain that

$$\begin{aligned}
0 &= \left(\sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma} \left(\sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{\Sigma}_{0e}^{-1} \mathbf{x}_j\right) \\
&+ \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{\Sigma}_{0e}^{-1} \mathcal{D}_e \mathbf{\Sigma}_{0e}^{-1} \mathbf{x}_j \\
&+ \left(\sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \left(\sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{\Sigma}_{0e}^{-1} \mathcal{D}_e \mathbf{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \left(\sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{\Sigma}_{0e}^{-1} \mathbf{x}_j\right) \\
&- 2 \left(\sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{\Sigma}_{0e}^{-1} \mathcal{D}_e \mathbf{\Sigma}_{0e}^{-1} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \left(\sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{\Sigma}_{0e}^{-1} \mathbf{x}_j\right).
\end{aligned} \tag{4.48}$$

Let $\mathbf{S} = \sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{\Sigma}_{0e}^{-1} \mathbf{x}_j$ and $\mathbf{E} = \mathbf{\Sigma}_{0e}^{-1} \mathcal{D}_e \mathbf{\Sigma}_{0e}^{-1}$. The above equation becomes

$$\begin{aligned}
0 &= \mathbf{S}^T \mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma} \mathbf{S} + \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{E} \mathbf{x}_j + \mathbf{S}^T \mathbf{\Gamma} \left(\sum_{j=1}^{N} \boldsymbol{\rho}_j \mathbf{E} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \mathbf{S} - 2 \left(\sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{E} \boldsymbol{\rho}_j^T\right) \mathbf{\Gamma} \mathbf{S} \\
&= \mathbf{S}^T \mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma} \mathbf{S} + \sum_{j=1}^{N} \left(\mathbf{x}_j^T \mathbf{E} \mathbf{x}_j + \mathbf{S}^T \mathbf{\Gamma} \boldsymbol{\rho}_j \mathbf{E} \boldsymbol{\rho}_j^T \mathbf{\Gamma} \mathbf{S} - 2\mathbf{x}_j^T \mathbf{E} \boldsymbol{\rho}_j^T \mathbf{\Gamma} \mathbf{S}\right) \\
&= \mathbf{S}^T \mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma} \mathbf{S} + \sum_{j=1}^{N} \left(\mathbf{x}_j - \boldsymbol{\rho}_j^T \mathbf{\Gamma} \mathbf{S}\right)^T \mathbf{E} \left(\mathbf{x}_j - \boldsymbol{\rho}_j^T \mathbf{\Gamma} \mathbf{S}\right).
\end{aligned} \tag{4.49}$$

By the definition of $\mathbf{S}$, in the above equation, $\mathbf{S}$ is an arbitrary vector in $\mathbb{R}^d$ because $\mathbf{x}_j$ is arbitrary in $\mathbb{R}^p$ and $\boldsymbol{\rho}_j$ is full rank by assumption (A.6). Thus, the right-hand side of (4.49) is the sum of $N + 1$ quadratic terms. Since both $\mathbf{E}$ and $\mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma}$ are symmetric and non-negative definite, it follows that $\mathbf{E} = \mathbf{\Sigma}_{0e}^{-1} \mathcal{D}_e \mathbf{\Sigma}_{0e}^{-1} = \mathbf{0}$ and $\mathbf{\Gamma} \mathbf{\Sigma}_{0b}^{-1} \mathcal{D}_b \mathbf{\Sigma}_{0b}^{-1} \mathbf{\Gamma} = \mathbf{0}$ Since $\mathbf{\Gamma}$ is invertible, we conclude that

$\mathcal{D}_e = \mathbf{0}$ and $\mathcal{D}_b = \mathbf{0}$.

Next, let $\Delta = 0$ in (4.43). Based on the above conclusions that $\mathbf{h}_1^\mu = \mathbf{0}$, $\mathcal{D}_e = \mathbf{0}$ and $\mathcal{D}_b = \mathbf{0}$, we obtain that

$$E_{\mathbf{b}}\left[\exp\left\{-\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}d\Lambda_0(t)\right\}\right.$$
$$\left.\times\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}\{(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t)\}d\Lambda_0(t)\right] = 0, \tag{4.50}$$

where $\mathbf{b}$ is from $N_d(\boldsymbol{\nu}, \boldsymbol{\Gamma})$ as in (4.44), with

$$\boldsymbol{\Gamma} = (\boldsymbol{\Sigma}_{0b}^{-1} + \sum_{j=1}^N \boldsymbol{\rho}_j^T\boldsymbol{\Sigma}_{0e}^{-1}\boldsymbol{\rho}_j)^{-1},$$

and

$$\boldsymbol{\nu} = \boldsymbol{\Gamma}\left[\sum_{j=1}^N \boldsymbol{\rho}_j\boldsymbol{\Sigma}_{0e}^{-1}(\mathbf{W}_j - \tilde{\boldsymbol{\rho}}_j^T\boldsymbol{\mu}_0)\right].$$

Since $\mathbf{b}$ is a complete and sufficient statistic for $\boldsymbol{\nu}$, it follows that

$$\int_0^V e^{\boldsymbol{\beta}_0^T(\boldsymbol{\rho}(t)^T\mathbf{b})+\boldsymbol{\eta}_0^T\mathbf{z}}\{(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t)\}d\Lambda_0(t) = 0,$$

which yields

$$(\boldsymbol{\rho}(t)^T\mathbf{b})^T\mathbf{h}_1^\beta + \mathbf{Z}^T\mathbf{h}_1^\eta + h_2(t) = 0,$$

with an arbitrary $\mathbf{b}$. Thus by assumption (A.7), we conclude that $\mathbf{h}_1^\beta = \mathbf{0}$, $\mathbf{h}_1^\eta = \mathbf{0}$ and $h_2(t) \equiv 0$.

Now that we have verified that conditions (a) to (d) of the theorem are satisfied, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\Lambda} - \Lambda_0)$ weakly converges to a tight random element in $l^\infty(\mathcal{H})$. Using the same argument as in Zeng and Cai (2005), we conclude that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \hat{\Lambda} - \Lambda_0)$ weakly converges to a Gaussian process in $l^\infty(\mathcal{H})$, and $\hat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$. $\square$

Chapter 5

Nonparametric Joint Modeling of Survival and Longitudinal Process: A Maximum Likelihood Approach

5.1 Introduction

In biomedical and public health study it is often of interest to explore the relationship between the time-to-event and the longitudinal covariates which are observed intermittently over a study period and usually contaminated by measurement errors. Several parametric joint modeling schemes have been developed in the literature to meet the practical need (Wulfsohn and Tsiatis, 1997; Song et al., 2002b; Hsieh et al., 2006; Faucett and Thomas, 1996; Faucett et al., 1998; Henderson et al., 2000; Tsiatis and Davidian, 2001). In these studies, the longitudinal process is often assumed to follow a simple linear mixed-effects model, and the survival process is specified by Cox model with varying covariates. The two submodels are usually linked by a small number of random effects shared by them. Various model fitting methods, including regression calibration (Tsiatis et al., 1995), maximum joint likelihood method (Wulfsohn and Tsiatis, 1997; Zeng and Cai, 2005; Hsieh et al., 2006), Bayesian approach (Faucett and Thomas, 1996; Xu and Zeger, 2001b; Wang and Taylor, 2001; Brown and Ibrahim, 2003) and conditional score method (Tsiatis and Davidian, 2001; Song et al., 2002a), have been proposed to estimate the model parameters.

Although these joint modeling frameworks and methodologies have been proved successful in many practical applications, they could be insufficient for more complex and versatile scenarios. For example, the mean trajectory of the longitudinal process can be irregular over time and thus difficult to be captured by linear models. Moreover, the shape of the longitudinal trajectory can vary greatly from person to person, and the variation may evolve over time. All these features would require more random effects in the model. In addition, the effects of the longitudinal and baseline covariates on the survival response might be time-varying instead of constant. In fact, a few studies have already shown that the misspecification of the longitudinal models would introduce extra bias into the survival regression coefficients (Brown et al., 2005; Ding and Wang, 2008), and the traditional proportional hazards model would be insufficient in delineating the dynamic association between the covariates and the survival response (Song and Wang, 2008).

Due to these limitations, the nonparametric model setting becomes an appealing alternative. Despite the abundant research in parametric joint modeling, the literature in nonparametric joint modeling is quite limited. Ding and Wang (2008) proposed a joint model with nonparametric multiplicative random effects submodel for the longitudinal process, where the mean longitudinal trajectory is specified by a fixed nonparametric function, and the variance across individuals is imposed by a single random variable multiplied to the mean trajectory. Brown et al. (2005) adopted a nonparametric mixed-effects model for the longitudinal process, which is a more flexible setting and allows more variation across individuals. Both of the two papers used polynomial splines to estimate the nonparametric curves for longitudinal processes, and they both assumed constant regression coefficients in the Cox models. Song and Wang (2008), on the other hand, focused on the nonparametric survival submodel, and allowed the regression coefficients in the Cox model to vary with time. They used both local smoothing techniques (Song and Wang, 2008) and polynomial splines (Song and Wang, unpublished manuscript) to fit the nonparametric curves, and argued that the latter is more computationally efficient than the former and nicely incorporate the cases with constant coefficients. However, the submodel for the longitudinal data in this literature is still considered to be parametric.

One of the main obstacles for the development of nonparametric joint models is the computational challenge. For the joint-likelihood-based methods, the numerical evaluation of multidimensional integrals is already challenging for the parametric models. And as the integral dimension grows dramatically in the nonparametric joint model settings, the computation becomes a great difficulty. To our best knowledge, there is no existing joint modeling study on nonparametric submodels for both the longitudinal and survival data.

In this chapter, we propose a nonparametric joint model, in which the longitudinal trajectory is delineated by a nonparametric mixed-effects model similar to Brown et al. (2005), and the survival process is specified by a varying-coefficient Cox model as in Song and Wang (2008). The nonparametric curves in both the submodels are estimated by the combinations of B-spline basis with either the fixed or the random coefficients. We propose to fit the model using the maximum joint likelihood approach, and the aforementioned computational challenge is tackled by a newly introduced algorithm, design of experiments-based interpolation technique (DoIt), which is efficient in approximating the multidimensional integrals with a number of deterministic accessing points that grows linearly with the dimension.

Section 5.2 describes the model setting, and introduces the estimation and model selection approach. Two simulation examples and a real data analysis are conducted in Section 5.3 to demonstrate the performance of the proposed modeling framework and estimation method. Conclusion and discussion are given in Section 5.4.

5.2 Model and Estimation

5.2.1 Model settings

For simplicity and without loss of generality we consider a single longitudinal process $\mathbf{W}$ and a vector of baseline covariates $\mathbf{Z}$ for a group of n individuals. We assume the survival process is subject to right censoring, and for the ith individual with event time T_i, the censoring time is C_i, and the observed event time is $V_i = \min\{T_i, C_i\}$. Denote by $\Delta_i = I(T_i \leq V_i)$ the censoring indicator. The longitudinal process is observed at N_i scattered time points $\mathbf{t}_i = \{t_{i1}, \ldots, t_{iN_i}, t_{iN_i} \leq V_i\}$, which gives $\mathbf{W}_i = \{W_{i1}, \ldots, W_{iN_i}\}$. Thus the observed data for the ith individual is denoted by

$$\mathbf{D}_{oi} = \{\mathbf{t}_i, \mathbf{W}_i, \mathbf{Z}_i, V_i, \Delta_i\}.$$

The observed longitudinal process is assumed to follow a nonparametric mixed-effects model given by

$$W_{ij} = \mu(t_{ij}) + f_i(t_{ij}) + e_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, N_i, \tag{5.1}$$

where $\mu(t_{ij}) = E\{W_{ij}\}$ is the population mean function of $\mathbf{W}$ observed at time t_{ij}. $f_i(t_{ij})$ is a random function with zero mean representing the ith individual's deviation from the mean at time t_{ij}. e_{ij} are the independent and identically distributed random errors with mean zero and variance σ_e^2, and are independent of the function $f_i(\cdot)$. Note that $\mu(t) + f_i(t)$ represents the unobservable true underling longitudinal process of the ith individual.

The Cox model with time-varying coefficients is adopted for the survival components of the joint model, with the conditional hazard function for the ith individual given by

$$\lambda_i(t|\mathbf{Z}_i, \{f_i(s), 0 \le s \le t\}) = \lambda_0(t) \exp\{\beta(t) f_i(t) + \boldsymbol{\eta}(t)^T \mathbf{Z}_i\}, \tag{5.2}$$

where $\lambda_0(t)$ is the baseline hazard, and $\beta(t)$ and $\boldsymbol{\eta}(t)$ are the time-varying regression coefficients for the longitudinal and baseline covariates, respectively. Note that if we take $\mu(t) + f_i(t)$ from (5.1) as the longitudinal covariates in the Cox's model, the first part $\beta(t)\mu(t)$ is non-identifiable across individuals, and thus can be absorbed into the baseline hazard $\lambda_0(t)$.

We propose to estimate the functional coefficients in the joint model by the linear combinations of B-spline basis functions as follows

$$\begin{aligned}
\mu(t) &\approx \sum_{l=1}^{d_\mu} \mu_l B_l^{(\mu)}(t) = \boldsymbol{\mu}^T \mathbf{B}^{(\mu)}(t), \\
f_i(t) &\approx \sum_{k=1}^{d_b} b_k B_k^{(b)}(t) = \mathbf{b}^T \mathbf{B}^{(b)}(t), \\
\beta(t) &\approx \sum_{p=1}^{d_\beta} \beta_p B_p^{(\beta)}(t) = \boldsymbol{\beta}^T \mathbf{B}^{(\beta)}(t), \\
\eta_k(t) &\approx \sum_{q=1}^{d_{\eta_k}} \eta_{kq} B_{kq}^{(\eta)}(t) = \boldsymbol{\eta}_k^T \mathbf{B}_k^{(\eta)}(t),
\end{aligned} \tag{5.3}$$

where $\mathbf{B}^{(\cdot)}(t) = (B_1^{(\cdot)}(t), \ldots, B_d^{(\cdot)}(t))$ are the B-spline basis functions that may vary across different parameter functions. The dimension d can also be different so as to specify different levels of smoothness for different parameter functions. The larger the d's are, the better the models fit, but at the cost of larger variance of the estimated curves, or known as over-fitting. In practice the optimal value of d's can be chosen by the model selection criteria such as AIC and BIC. The model selection procedure is explained in more details in Section 5.3. In the above equations, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{d_\mu})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_\beta})^T$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{d_\eta})^T$ are the coefficients for the fixed functional coefficients $\mu(t)$, $\beta(t)$ and $\eta(t)$, respectively. Once they are estimated, the parameter functions can be estimated by $\hat{\mu}(t) = \hat{\boldsymbol{\mu}}^T \mathbf{B}^{(\mu)}(t)$, $\hat{\beta}(t) = \hat{\boldsymbol{\beta}}^T \mathbf{B}^{(\beta)}(t)$, $\hat{\eta}_k(t) = \hat{\boldsymbol{\eta}}_k^T \mathbf{B}^{(\eta)}(t)$ accordingly. The random function $f_i(t)$ has the same form of spline approximation as the other functions, but with the random coefficient $\mathbf{b} = (b_1, \ldots, b_{d_b})$, which is assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_b$. Note that although we assume normal distribution for the random effects $\mathbf{b}_i$, previous studies have shown that maximum joint likelihood procedure is quite robust for the violation of normal assumption (Song et al., 2002b; Hsieh et al., 2006). With the given approximations, the parameter set of interest becomes

$$\boldsymbol{\Omega} = \{\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Sigma}_b, \sigma_e^2, \lambda_0(t)\}. \tag{5.4}$$

Note that except $\lambda_0(t)$, all the other parameters are parametric and can be estimated using the same method as that of parametric joint modeling.

5.2.2 The joint likelihood approach

Similar to the estimation for parametric joint modeling, we propose to use the maximum joint likelihood approach to estimate all the parameters. With the specified model setting and assumptions, we obtained the following joint likelihood function composed of three density functions

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \int f_{V_i,\Delta_i|\mathbf{b}_i} \cdot f_{W_i|\mathbf{b}_i} \cdot f_{\mathbf{b}_i} d\mathbf{b}_i, \tag{5.5}$$

where,

$$\begin{aligned} f_{\mathbf{b}_i} =& \frac{1}{2\pi}|\mathbf{\Sigma}_b|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{b}_i^T \mathbf{\Sigma}_b^{-1} \mathbf{b}_i\right\}, & (5.6) \\ f_{\mathbf{W}_i|\mathbf{b}_i} =& (\frac{1}{\sqrt{2\pi\sigma_e^2}})^{N_i} \exp\left[-\frac{1}{2\sigma_e^2}\sum_{j=1}^{N_i}\{W_{ij} - \mu(t_{ij}) - f_i(t_{ij})\}^2\right], & (5.7) \\ f_{V_i,\Delta_i|\mathbf{b}_i} =& \left[\lambda_0(V_i)\exp\left\{\beta(V_i)f_i(V_i) + \boldsymbol{\eta}^T(V_i)\mathbf{Z}_i\right\}\right]^{\Delta_i} \\ & \exp\left\{-\int_0^{V_i} \lambda_0(t)\exp\{\beta(t)f_i(t) + \boldsymbol{\eta}^T(t)\mathbf{Z}_i\}dt\right\}. & (5.8) \end{aligned}$$

We assume the baseline hazard $\lambda_0(t)$ takes mass only at the event time points at which $\Delta = 1$. Thus, the density function of (5.8) can be rewritten as

$$\begin{aligned} f_{V_i,\Delta_i|\mathbf{b}_i} =& \left[\lambda_0(V_i)\exp\left\{\beta(V_i)f_i(V_i) + \boldsymbol{\eta}(V_i)^T\mathbf{Z}_i\right\}\right]^{\Delta_i} \\ & \times \exp\left\{-\sum_{j=1}^{N} \lambda_0(V_j)\exp\{\beta(V_j)f_i(V_j) + \boldsymbol{\eta}^T(V_j)\mathbf{Z}_i\}I(V_i \geq V_j)\right\}. \end{aligned} \tag{5.9}$$

Since the functional parameters $\mu(t), f_i(t), \beta(t)$ and $\eta(t)$ in the joint likelihood equation are all substituted by their spline approximations in (5.3), the goal is to maximize the likelihood (5.4) with respect to the high-dimensional parametric parameter set

$$\tilde{\mathbf{\Omega}} = \{\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{\Sigma}_b, \sigma_e^2, \lambda_0(V_1), \ldots, \lambda_0(V_N)\}.$$

Optimizing (5.5) is challenging due to multidimensional integral of the unobservable random effects $\mathbf{b}_i$. As in the previous studies (Wulfsohn and Tsiatis, 1997; Ding and Wang, 2008), we adopt the EM algorithm, in which the estimation is conducted between the E-step and the M-step iteratively until the algorithm converges. Although the EM algorithm works well for the parametric joint models with a small dimension of $\mathbf{b}_i$, the convergence becomes extremely slow and quite unstable when the dimension of random effects increases (Ding and Wang, 2008). The greatest obstacle is the calculation of conditional expectations in the E-step that requires numerical approximation of multidimensional integral. In this paper, we use the DoIt method, which has been shown to be a efficient and robust numerical approach in the EM algorithm for joint models with a larger number of random effects. We explain the steps of the EM-DoIt algorithm in the following paragraphs.

Step 0 (initialization): Since an EM algorithm may be quite sensitive to the starting point, we adopt a revised "two-stage" method (Tsiatis et al., 1995) to set good initial values. In the first stage, the nonparametric mixed-effects model is fitted for the longitudinal process using B-spline approximation, which provides the initial estimates of $\hat{\boldsymbol{\mu}}^{(0)}$, $\hat{\boldsymbol{\Sigma}}_b^{(0)}$, $\hat{\sigma}_e^{2(0)}$, together with the best linear unbiased prediction (BLUP) of the random effects $\hat{\mathbf{b}}_i$. With these outcomes we could recover the nonparametric curves $\hat{\mu}(t)^{(0)}$ and $\hat{f}_i(t)^{(0)}$. In the second stage, we take $\hat{f}_i(t)^{(0)}$ into the varying-covariate Cox model, and estimate $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{\eta}}^{(0)}$ and $\hat{\lambda}_0(V_j)^{(0)}$. The initial values obtained from the "two-stage" method are quite close to the final EM outcomes, and thus makes the convergence faster.

Step 1 (E-step): The main task of this step is to evaluate a series of conditional expectations with the form of $E\{g(\mathbf{b}_i)|D_{oi}, \tilde{\boldsymbol{\Omega}}^{(k)}\}$ for all the individuals, where $g(\mathbf{b}_i)$ is a given function of the random effects $\mathbf{b}_i$, D_{oi} is the observed data for the ith individual, and $\tilde{\boldsymbol{\Omega}}^{(k)}$ is the updated estimation of $\tilde{\boldsymbol{\Omega}}$ from the kth (i.e., the previous) iteration. Since the conditional distribution of $\mathbf{b}_i$ given D_{oi} and $\tilde{\boldsymbol{\Omega}}^{(k)}$ does not have explicit expression, various numerical integration approaches have been investigated to estimate the integrals involved. They include the Gaussian-Hermite Quadrature method (Wulfsohn and Tsiatis, 1997; Hsieh et al., 2006; Song et al., 2002b), the Markov Chain Monte Carlo method (Henderson et al., 2000; Tseng et al., 2005; Ding and Wang, 2008), Fully exponential Laplace approximation (Rizopoulos et al., 2009), and so forth. Most of these methods, though efficient for low-dimensional cases, encounter difficulties when the integral dimension increases and may lead to the crash of the EM algorithm.

We propose to use a relatively new method, design of experiments-based interpolation techniques (DoIt, (Joseph, 2012)), to circumvent this difficulty. The main idea of DoIt is to estimate conditional distributions using the weighted means of a group of normal distributions with the modes at the pre-specified design points scattered in the evaluation subspace. Accordingly, the conditional distribution can be estimated as follows

$$E\{g(\boldsymbol{b}_i)|D_{oi}, \tilde{\boldsymbol{\Omega}}^{(k)}\} \approx \frac{1}{\sum_{l=1}^{M} c_l} \sum_{l=1}^{M} c_l E_l\{g(\boldsymbol{b}_i)\}, \tag{5.10}$$

where $E_l\{g(\boldsymbol{b}_i)\}$ is the expectation of $g(\boldsymbol{b}_i)$ with respect to the normal distribution $N(\boldsymbol{\nu}_l, \boldsymbol{D}_i^{-1})$, with $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ being the pre-specified evaluation points, and $\boldsymbol{D}_i$ being the Fisher information matrix evaluated at the mode of the unnormalized distribution of $\mathbf{b}_i$ given D_{oi} and $\tilde{\boldsymbol{\Omega}}^{(k)}$. This new numerical method has been shown to be much faster and more robust than the other approaches when the dimension of the random effects grows relatively large.

Step 2 (M-step): Take derivatives of the log likelihood with respect to each of the parameters in $\tilde{\boldsymbol{\Omega}}$, and set the equations to zero, with some algebra applied (see Chapter 2 for details), it is easy to derive the longitudinal parameter estimations have the explicit forms

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{B}_i^{(\mu)T} \mathbf{B}_i^{(\mu)}\right]^{-1} E\left\{\mathbf{B}_i^{(\mu)T}(\mathbf{W}_i - \mathbf{b}_i^T \mathbf{B}_i^{(b)})\right\}, \\
\hat{\sigma}_e^2 &= \frac{1}{\sum_{j=1}^{N_i}} \sum_{i=1}^{n} \sum_{j=1}^{N_i} E\left\{\left[W(t_{ij}) - \hat{\boldsymbol{\mu}}^T \mathbf{B}^{(\mu)}(t_{ij}) - \mathbf{b}_i^T \mathbf{B}^{(b)}(t_{ij})\right]^2\right\}, \\
\hat{\boldsymbol{\Sigma}}_b &= \frac{1}{n} \sum_{i=1}^{n} E\{\mathbf{b}_i \mathbf{b}_i^T\},
\end{aligned} \tag{5.11}$$

where the notation $E\{\cdot\}$ stands for the conditional expectation of a function of $\mathbf{b}_i$ given D_{oi} and $\tilde{\boldsymbol{\Omega}}^{(k)}$. In the first equation of $\hat{\boldsymbol{\mu}}$, $\mathbf{W}_i = (W_{i1}, \ldots, W_{iN_i})^T$ is a n_i dimensional vector, $\mathbf{B}_i^{(\mu)} = (\mathbf{B}^{(\mu)}(t_1)^T, \ldots, \mathbf{B}^{(\mu)}(t_{N_i})^T)^T$ is a $N_i \times d_\mu$ matrix, and $\mathbf{B}_i^{(b)} = (\mathbf{B}^{(b)}(t_1)^T, \ldots, \mathbf{B}^{(b)}(t_{N_i})^T)^T$ is a $N_i \times d_b$ matrix.

Similarly, given the estimation of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ from the last iteration the baseline hazard function at the jth event time point also has the explicit expression

$$\hat{\lambda}_0(V_j) = \frac{\sum_{i=1}^N \Delta_i I(V_i = V_j)}{\sum_{i=1}^N E_i \left\{ \exp[\hat{\beta}(V_j)\mathbf{b}_i^T \mathbf{B}^{(b)}(V_j) + \hat{\eta}(V_j) Z_i] I(V_i \geq V_j) \right\}}, \tag{5.12}$$

where $\hat{\beta}(V_j) = \hat{\boldsymbol{\beta}}^T \mathbf{B}^{(\beta)}(V_j)$ and $\hat{\eta}(V_j) = \hat{\boldsymbol{\eta}}^T \mathbf{B}^{(\eta)}(V_j)$. According to the model assumption, $\hat{\lambda}_0(t) = 0$ at $t \neq V_j, j = 1, \ldots, N$. Therefore, the parameter estimation of $(\boldsymbol{\mu}, \sigma_e^2, \boldsymbol{\Sigma}_b, \lambda_0(V_1), \ldots, \lambda_0(V_N))$ can be updated by directly plugging in the estimated $E\{g(\mathbf{b}_i)|D_{oi}, \tilde{\boldsymbol{\Omega}}\}$ from the E-step into the equations given above.

The estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, however, is less straightforward than in the conventional Cox model, and Newton-Raphson (NR) algorithm is needed to locate the maximizer using the iterative scheme: $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + I^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) S(\hat{\boldsymbol{\beta}}^{(k)})$, and $\hat{\boldsymbol{\eta}}^{(k+1)} = \hat{\boldsymbol{\eta}}^{(k)} + I^{-1}(\boldsymbol{\eta}^{(k)}) S(\boldsymbol{\eta}^{(k)})$. We provide the score functions and information matrices for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in the following equations

$$\begin{aligned}
S(\boldsymbol{\beta}) &= \sum_{i=1}^N \left[\mathbf{B}^{(\beta)}(V_i) \left\{ \Delta_i E\{\mathbf{b}_i^T \mathbf{B}^{(b)}(V_i)\} - \frac{\sum_{j=1}^N E_j g_{2i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} \right\} \right], \\
I(\boldsymbol{\beta}) &= -\sum_{i=1}^N \left[\mathbf{B}^{(\beta)}(V_i) \mathbf{B}^{(\beta)}(V_i)^T \right] \\
&\quad \times \left[\frac{\sum_{j=1}^N E_j g_{3i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} - \left\{ \frac{\sum_{j=1}^N E_j g_{2i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} \right\}^2 \right], \\
S(\boldsymbol{\eta}) &= \sum_{i=1}^N \left[\mathbf{B}^{(\eta)}(V_i) \left\{ \Delta_i Z_i - \frac{\sum_{j=1}^N Z_j E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} \right\} \right], \\
I(\boldsymbol{\eta}) &= -\sum_{i=1}^N \left[\mathbf{B}^{(\eta)}(V_i) \mathbf{B}^{(\eta)}(V_i)^T \right] \\
&\quad \times \left[\frac{\sum_{j=1}^N Z_j^2 E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} - \left\{ \frac{\sum_{j=1}^N Z_j E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)}{\sum_{j=1}^N E_j g_{1i}(\mathbf{b}_j) I(V_j \geq V_i)} \right\}^2 \right].
\end{aligned} \tag{5.13}$$

In the above equations, $E_j(\cdot)$ stands for conditional expectations with respect to $\mathbf{b}_j$, and for any $V_i \leq V_j$, $g_{1i}(\mathbf{b}_j)$, $g_{2i}(\mathbf{b}_j)$ and $g_{3i}(\mathbf{b}_j)$ are given by

$$\begin{aligned} g_{1i}(\mathbf{b}_j) &= \exp\{\boldsymbol{\beta}^T \mathbf{B}^{(\beta)}(V_i)\mathbf{b}_j^T \mathbf{B}^{(b)}(V_i) + \boldsymbol{\eta}^T \mathbf{B}^{(\eta)}(V_i) Z_j\}, \\ g_{2i}(\mathbf{b}_j) &= \mathbf{b}_j^T \mathbf{B}^{(b)}(V_i) g_{1i}(\mathbf{b}_j), \\ g_{3i}(\mathbf{b}_j) &= \left\{\mathbf{b}_j^T \mathbf{B}^{(b)}(V_i)\right\}^2 g_{1i}(\mathbf{b}_j). \end{aligned} \tag{5.14}$$

Since the Newton-Raphson algorithm in our case is imbedded in the EM algorithm, which is already an iterative procedure, we adopt one-step NR that does not require full convergence. In practice we found this technique faster and more stable than the fully iterative version of the NR algorithm.

Variance estimation is another great challenge for joint modeling problems due to the nature of the EM algorithm and the profile likelihood estimation involved for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. Hsieh et al. (2006) pointed out that due to the profile estimation of the survival parameters, the variance estimation based on the Fisher information matrix given in (5.13) is inaccurate and would lead to over-optimistic statistical inference. Louis (1982) suggested that the accurate variance estimation in the EM algorithm would require the calculation of the observed Fisher information matrix for the entire parameter set. However, this approach is infeasible for our case considering the high dimensionality of $\tilde{\boldsymbol{\Omega}}$. Therefore, we adopt the bootstrap technique as suggested by Hsieh et al. (2006) and Ding and Wang (2008) to obtain the valid variance estimation.

5.3 Numerical studies

5.3.1 Simulation studies

We conducted two simulation studies in this section. The major goal is to demonstrate the flexibility of the proposed nonparametric joint model settings and examine the performance of the proposed estimation approach and computing algorithm.

The first simulation example is relatively simple and considers only the nonparametric longitudinal submodel, where we show that misspecification of the irregular longitudinal process would cause bias for the survival regression coefficients. The second example is more complex, and takes into account both the nonparametric longitudinal submodel and the nonparametric survival submodel. In both examples we use the EM-DoIt approach to handle the moderate- to large-dimensional random effects introduced by spline approximation of random curves. In each example, we compare estimation performance among three models with different number of inner knots to represent widely varying degrees of smoothness. We also calculate standard errors of the estimated parameters using Bootstrap technique and compare them with the standard deviations obtained from data replicates.

Example 5.1. JM with nonparametric longitudinal part

In this example we consider joint modeling with only the longitudinal response modeled nonparametrically; the survival coefficients are constant. The models are of the form:

$$W_i(t) = \mu(t) + f_i(t) + \epsilon_i(t),$$

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta f_i(t) + \eta Z_i\},$$

where $\mu(t)$ is the average trajectory of $W_i(t)$ across all the individuals, $f_i(t)$ is the subject-specific random deviations from mean process $\mu(t)$, and $\epsilon_i(t)$ is the measurement error assumed to be independent and identically from $N(0, \sigma^2)$.

In data generation, we assume that the longitudinal response is measured from day 0 to day 12 (i.e., $t \in [0, 12]$), and each subject is assigned to either the treatment group (i.e., Z=1) or the placebo group (i.e., Z=0) with the same probability. For the constant parameters, we assume $\sigma^2 = 1$, $\lambda_0 = 0.2$, $\beta = 1$ and $\eta = -1$.

The fixed nonparametric function $\mu(t)$ in the longitudinal model is given by

$$\mu(t) = 4 + 5 \sin(\pi t/4),$$

and the random nonparametric function $f_i(t)$ is given by

$$f_i(t) = \mathbf{d}_i^T \mathbf{g}(t),$$

where

$$\mathbf{g}(t) = \begin{pmatrix} e^{0.03t}\sin(\pi t/20) \\ e^{0.03t}\cos(\pi t/20) \end{pmatrix}, \text{ and } \mathbf{d}_i \overset{i.i.d}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}_d), \text{ with } \mathbf{\Sigma}_d = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}.$$

The underlying function $\mathbf{g}(t)$ and the random factor $\mathbf{d}_i$ are designed in such a way that the variance of $f_i(t)$ increases with time, the correlation between $f_i(t)$ and $f_i(s)$ decreases as the two time points t and s fall apart, and the random deviation process $f_i(t)$ is independent across different individuals.

In joint modeling simulation with complex longitudinal process, it is always a challenge to generate the survival time T_i. Following the techniques introduced in Brown et al. (2005), we first randomly generate the survival probability $S(t)$ from the uniform distribution because $S(T) = 1 - F(T) \sim U(0, 1)$. The survival time T_i is then obtained numerically from $S(t) = \exp\{-\int_0^t \lambda(u)du\}$ via the combination of R functions *uniroot()* and *integrates().* The censoring time is generated from a uniform distribution $U(6, 24)$, and the average censoring rate is around 30%.

In model estimation, using the spline smoothing approach, the nonparametric functions $\mu(t)$ and $f_i(t)$ are approximated by the linear combination of B-spline basis as follows:

$$\mu(t) \approx \sum_{p=1}^{d_\mu} \mu_p B_p^{(\mu)}(t),$$

and

$$f_i(t) \approx \sum_{k=1}^{d} b_{ki} B_k^{(b)}(t),$$

where $\mathbf{B}^{(\mu)}(t) = (B_1^{(\mu)}(t), \ldots, B_{d_\mu}^{(\mu)}(t))^T$ and $\mathbf{B}^{(b)}(t) = (B_1^{(b)}(t), \ldots, B_d^{(b)}(t))^T$ are the different sets of B-spline basis for the two functions, respectively. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{d_\mu})^T$ are the fixed coefficients and b_{ki}'s are the random coefficients with the assumption $\mathbf{b}_i = (b_{1i}, \ldots, b_{di})^T \sim N_d(\mathbf{0}, \mathbf{\Sigma}_b)$. The numbers of the basis d_μ and d are adjusted in the simulation study to achieve different degrees of smoothness in fitting. With the cubic B-spline, the number of the inner knots is $(d_\mu - 4)$ for $\mu(t)$ and $(d - 4)$ for $f_i(t)$. In our study, these inner knots are located with equal space on the time span $[0, 12]$. We summarize the simulation results based on $n = 100$ and $n = 200$ subjects with $N = 100$ replicates in the following paragraphs.

Table 5.1 shows the means and standard deviations of the computing time and the number of iterations that the EM algorithm takes to converge under three combinations of d_μ and d (i.e., $(d_\mu = 6, d = 4)$, $(d_\mu = 7, d = 5)$, $(d_\mu = 8, d = 6)$). The results show that the mean computing time increases as the number of basis functions (d_μ, d) or the sample size n increases. However, the comparison between the two sample sizes indicates the dataset with more subjects (i.e., large n) takes fewer EM iterations to converge due to richer information. We also notice that among the three combinations, the case of $(d_\mu = 7, d = 5)$ takes the fewest EM iterations to converge in both $n = 100$ and $n = 200$ scenarios.

Table 5.1: Computing time and number of EM iterations

		$(6,4)$		$(7,5)$		$(8,6)$	
n		mean	SD	mean	SD	mean	SD
100	Computing time (s)	287.90	127.20	348.20	88.56	914.40	294.28
	EM iterations	39.03	16.05	30.04	7.80	32.79	8.92
200	Computing time (s)	628.90	182.37	841.40	164.50	1218.00	354.68
	EM iterations	35.17	7.36	28.83	4.69	31.67	5.33

For a coefficient function $f(\cdot)$, the performance of estimator $\hat{f}(\cdot)$ is assessed via the square root of average squared errors (RASE, Cai et al. 2000),

$$\text{RASE}^2 = n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \{\hat{f}(u_k) - f(u_k)\}^2, \tag{5.15}$$

where $\{u_k, k = 1, \ldots, n_{\text{grid}}\}$ are the grid points matching with the observation time points of all the individuals.

Table 5.2 presents the estimation results of $\mu(t)$ and $\sigma_f^2(t) = Var\{f_i(t)\}$ via median and median absolute deviations (MAD) of RASE of $N = 100$ replicates. The results are also presented for two sample sizes and three combinations of (d_μ, d). It can be seen from the table that the case of $(d_\mu, d) = (7, 5)$ have the smallest median and MAD RASE scores for both $n = 100$ and $n = 200$ scenarios. Hence it is the optimal choice for the number of basis functions used to approximate the nonparametric curves among the three combinations.

Table 5.2: RASE of estimated nonparametric functions

		$(6,4)$		$(7,5)$		$(8,6)$	
n	parameter	median	MAD	median	MAD	median	MAD
100	$\mu(t)$	0.7674	0.0267	0.0444	0.0188	0.0454	0.0193
	$\sigma_f^2(t)$	0.1405	0.1011	0.0538	0.1059	0.1093	0.0487
200	$\mu(t)$	0.7468	0.0193	0.0281	0.0102	0.0283	0.0102
	$\sigma_f^2(t)$	0.0841	0.0507	0.0444	0.0248	0.0464	0.0212

Figures 5.1 through 5.4 present the four estimated curves of $\mu(t)$, $Var\{f_i(t)\}$, $Cov\{f_i(t), f_i(6)\}$ and $Cov\{f_i(t), f_i(12)\}$, respectively. All of them are from the case of $(d_\mu = 7, d = 5)$ with $n = 100$. We choose to present the results of this case because it performs the best in Table 5.1 and Table 5.2 among all the three cases. The scenario of $n = 200$ yields similar figures and is not presented here. In Figure 5.1 through 5.4, the left panel presents the estimated curves from a typical sample data set (i.e., the data set with median RASE of the $N = 100$ replicates), and the right panel shows the mean estimated curves of the $N = 100$ replicates.

The estimated curves for $\mu(t)$ are shown in Figure 5.1, where they closely resemble the true curve in both panels by overlapping with the true curve in most parts and departing only at the peak and bottom points. In the right panel, the pointwise confidence intervals at time t is calculated using the expression

$$\hat{\boldsymbol{\mu}}^T \mathbf{B}^{(\mu)}(t) \pm 1.96 \times \sqrt{\mathbf{B}^{(\mu)}(t)^T \mathrm{Cov}(\hat{\boldsymbol{\mu}}) \mathbf{B}^{(\mu)}(t)},$$

where $\hat{\boldsymbol{\mu}}$ is the mean estimated value of $\boldsymbol{\mu}$ from the N data replicates, and $\mathrm{Cov}(\hat{\boldsymbol{\mu}})$ is estimated by the sample covariance of $\hat{\boldsymbol{\mu}}$ from the N data replicates. The true curve is completely covered by the confidence bands. All these patterns indicate that our estimation approach with B-spline smoothing performs well for $\mu(t)$ in this case.

For the random functions $f_i(t)$, since it is difficult to present the estimated curves for all the subjects in one figure, we instead display the fixed curves of $Var\{f_i(t)\}$, $Cov\{f_i(t), f_i(6)\}$ and $Cov\{f_i(t), f_i(12)\}$ in Figure 5.2 through 5.4 to demonstrate estimation performance. Although the estimation performance is not as good as that for $\mu(t)$ due to the randomness (i.e., less overlap between the estimated curves and the true curves), the estimated curves still recover the main trend of the true curves in all the figures. Compared with the estimated curves from the typical samples (i.e., the left panel), the mean estimated curves performs better as expected.

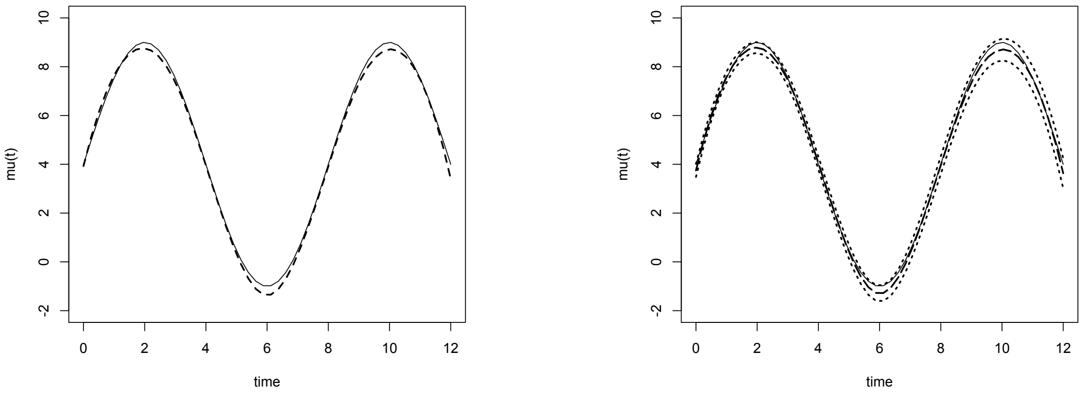

Figure 5.1. Estimated curves of $\mu(t)$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets, and the dotted curves are the pointwise confidence intervals.

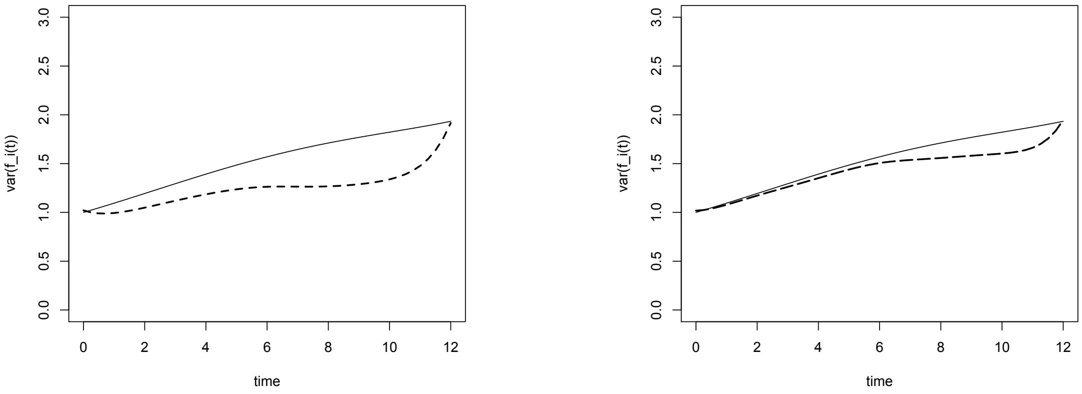

Figure 5.2. Estimated curves of $Var\{f_i(t)\}$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

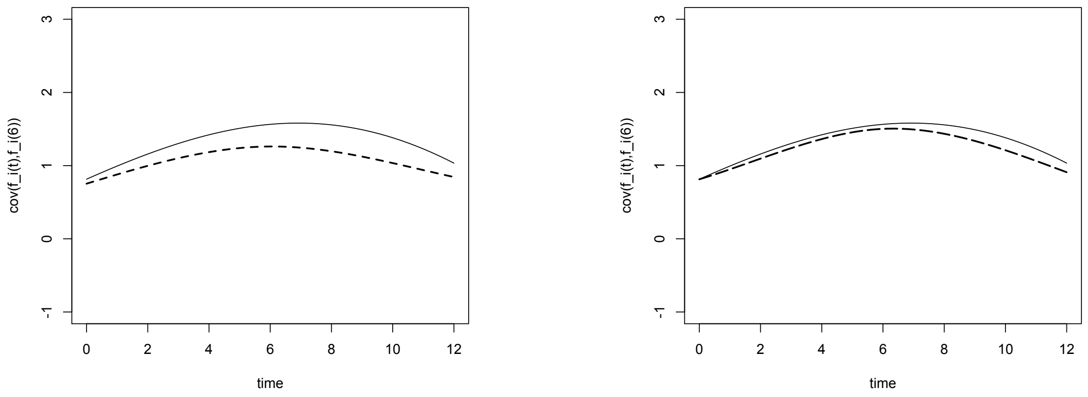

Figure 5.3. Estimated curves of $Cov\{f_i(t), f_i(6)\}$**.** The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

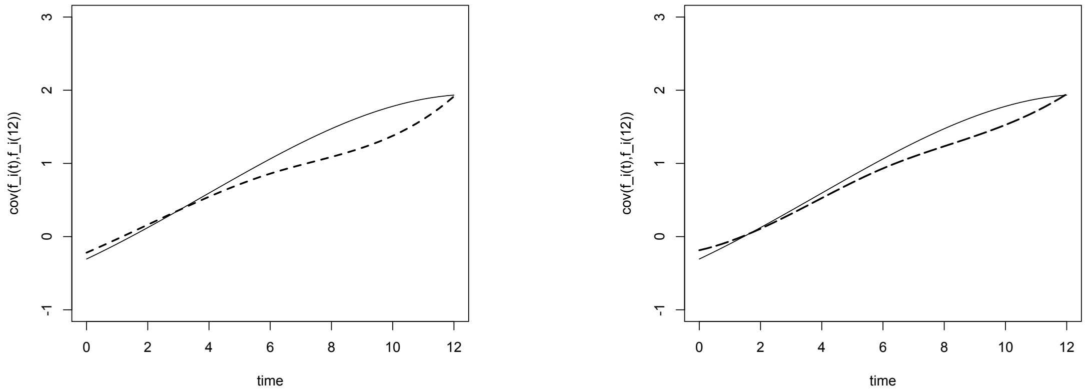

Figure 5.4. Estimated curves of $Cov\{f_i(t), f_i(12)\}$**.** The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

For the constant parameters, bias and RMSE are presented as a measure of estimation accuracy. Table 5.3 summarizes the simulation results for the constant parameters of the case ($d_\mu = 7, d = 5$) for the two sample sizes, and compared the results from the proposed nonparametric joint models with the results from a parametric joint model, where the longitudinal covariate is misspecified by a linear mixed-effects model. The results show that the estimation performs well for the nonparametric joint modeling in terms of both bias and RMSE for the constant variance σ^2 of the random error and the survival regression coefficients β and η. Compared with the nonparametric joint models, the simple parametric model not only fails to capture the actual trend of the longitudinal covariate (i.e., see the large bias of σ^2), but also introduces great bias for the survival coefficients β and η. This comparison example is consistent with the results of Brown et al. (2005) and Ding and Wang (2008), and explicitly shows that the misspecification of the longitudinal trajectories can be problematic in joint modeling problems and the nonparametric model setting is necessary in many real cases.

Table 5.3: Estimation results for constant parameters

	$n = 100$				$n = 200$			
	Nonparametric		Parametric		Nonparametric		Parametric	
parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β	-0.0118	0.1250	-0.7519	0.8091	-0.0136	0.1196	-0.7691	0.7978
η	0.0094	0.2491	0.2509	0.3633	-0.0016	0.1768	0.2905	0.3478
σ^2	0.0005	0.0309	8.9979	9.0134	0.0062	0.0200	8.9432	8.9507

Table 5.4: Standard errors of the estimators

parameter		$n = 100$			$n = 200$		
		SD	SE_{mean}	SE_{std}	SD	SE_{mean}	SE_{std}
$\mu(t)$	$t = 3$	0.1148	0.1135	0.0214	0.0849	0.0893	0.0078
	$t = 6$	0.1463	0.1478	0.0318	0.0983	0.1148	0.0126
	$t = 9$	0.1995	0.2091	0.0415	0.1342	0.1569	0.0207
β		0.1244	0.1328	0.0340	0.1188	0.0917	0.0250
η		0.2489	0.2499	0.0235	0.1768	0.1744	0.0172

Table 5.4 compares the parameter standard errors calculated using Bootstrap technique with the standard deviations of the $N = 100$ replicates with ($d_\mu = 7, d = 5$). We sample with replacement from each replicate for $Boot.N = 100$

times, and the Bootstrap standard errors of the given replicate are obtained by calculating the standard deviations of the parameter estimates of the 100 Bootstrap datasets. For the estimated nonparametric function $\hat{\mu}(t)$ the standard errors are evaluated at 3 time points equally spaced on $[0, 12]$. The standard deviations based on the $N = 100$ estimation of the parameters are displayed as SD in table 5.4. The average and the standard deviation of the 100 estimated standard errors are denoted as SE_{mean} and SE_{std} in the table, respectively.

It can be seen from the Table 5.4 that all the values of SD can be covered by $SE_{mean} \pm 1.96 SE_{std}$. This indicates that the standard errors calculated by Bootstrap technique performs well in this case.

Example 5.2. JM with nonparametric longitudinal and survival parts

In this example we consider joint modeling with both nonparametric longitudinal processes and time-varying survival coefficients. The model setting is the same to those specified at the beginning. Recall that the model is of the form

$$W_i(t) = \mu(t) + f_i(t) + \epsilon_i(t),$$

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta(t) f_i(t) + \eta(t) Z_i\},$$

where the nonparametric mixed-effects submodel for the longitudinal response is the same as that of Example 5.1, but the regression coefficients in the survival submodel are allowed to vary with time.

Similar to Example 5.1, the random process $f_i(t)$ is generated by

$$f_i(t) = \mathbf{d}_i^T \mathbf{g}(t),$$

and we set

$$\mathbf{g}(t) = \begin{pmatrix} e^{0.03t} \sin(\pi t/20) + 0.2 \\ e^{0.03t} \cos(\pi t/20) + 0.1 \end{pmatrix}, \text{ and } \mathbf{d}_i \overset{i.i.d}{\sim} N_2(\mathbf{0}, \boldsymbol{\Sigma}_d), \text{ with } \boldsymbol{\Sigma}_d = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix},$$

thus inheriting the main features of the nonparametric random curves in Example 5.1.

The true function of $\mu(t)$ is the same as in Example 5.1, i.e.,

$$\mu(t) = 4 + 5\sin(\pi t/4),$$

and the new nonparametric functions $\beta(t)$ and $\eta(t)$ are set as the following two continuous polynomials:

$$\beta(t) = \frac{1}{200}(t-7)^3 + 1.5 + \frac{1}{20}(t-8)^2 I(t > 8),$$

$$\eta(t) = \frac{1}{8}(t+2) - 3.5 + \frac{1}{400}(t-5)^3 I(t > 5).$$

Following the conclusions of Example 5.1 that $(d_\mu = 7, d = 5)$ performs the best of the three cases, in this example we directly adopt $(d_\mu = 7, d = 5)$ for the approximation of longitudinal process, and mainly focus on the estimation of $\beta(t)$ and $\eta(t)$. We hold the values of (d_μ, d) at $(7, 5)$, and compare the different degrees of smoothness of $\beta(t)$ and $\eta(t)$ via three combinations of basis functions: $(d_\beta = 4, d_\eta = 4), (d_\beta = 5, d_\eta = 4), (d_\beta = 5, d_\eta = 5)$. The estimation procedure is applied to $N = 100$ data replicates with $n = 200, 400$ sample size, respectively.

Table 5.5 presents the computing time and the number of iterations that the EM algorithm takes to converge. It can be seen that the first combination (i.e., $d_\beta = d_\eta = 4$) uses least computing time and number of EM iterations among the three.

Table 5.5: Computing time and number of EM iterations

		$(4,4)$		$(5,4)$		$(5,5)$	
n	parameter	mean	SD	mean	SD	mean	SD
200	Computing time (min)	13.29	5.94	14.45	6.52	15.97	7.03
	EM iterations	27.44	10.35	29.59	11.58	31.80	11.90
400	Computing time (min)	27.47	9.38	31.02	10.63	35.93	12.41
	EM iterations	21.15	7.25	27.63	8.21	29.88	9.42

Table 5.6 presents the median and median absolute deviations (MAD) of RASE of the estimated nonparametric functions. Focusing on $\beta(t)$ and $\eta(t)$, we notice that the first combination (i.e., $d_\beta = d_\eta = 4$) has better overall estimation than the other two cases in terms of median RASE. This suggests that the model with the fewest number of knots for $\beta(t)$ and $\eta(t)$ outperforms the other two models for

this case.

Table 5.6: RASE of estimated nonparametric functions

		(4, 4)		(5, 4)		(5, 5)	
n	parameter	Median	MAD	Median	MAD	Median	MAD
200	$\mu(t)$	0.0507	0.0214	0.0482	0.0199	0.0481	0.0204
	$\beta(t)$	0.0782	0.0445	0.0975	0.0561	0.1293	0.0689
	$\eta(t)$	0.4567	0.2751	0.4443	0.2548	0.6100	0.3123
	$\sigma_f^2(t)$	0.0964	0.0500	0.1038	0.0581	0.0959	0.0575
400	$\mu(t)$	0.0486	0.0175	0.0487	0.0168	0.0467	0.0168
	$\beta(t)$	0.0395	0.0244	0.0528	0.0231	0.0566	0.0248
	$\eta(t)$	0.2178	0.1395	0.2321	0.1413	0.3704	0.1991
	$\sigma_f^2(t)$	0.0514	0.0344	0.0549	0.0382	0.0519	0.0353

The estimated curves of $\mu(t), \beta(t), \eta(t), Var\{f_i(t)\}, Cov\{f_i(t), f_i(6)\}$ and $Cov\{f_i(t), f_i(12)\}$ were shown in Figures 5.5 through 5.10. The left panel of the figures present the estimated curves of a typical sample data set (i.e., the sample with the median RASE of the 100 replicates), and the right panel of the figures present the mean estimated curves of the 100 replicates. All the estimated curves are obtained from the combination of $(d_\mu, d, d_\beta, d_\eta) = (7, 5, 4, 4)$, which performs the best of the three models, with the sample size of $n = 400$. The other two cases and sample size of $n = 200$ have very similar estimated curves and are thus not presented here.

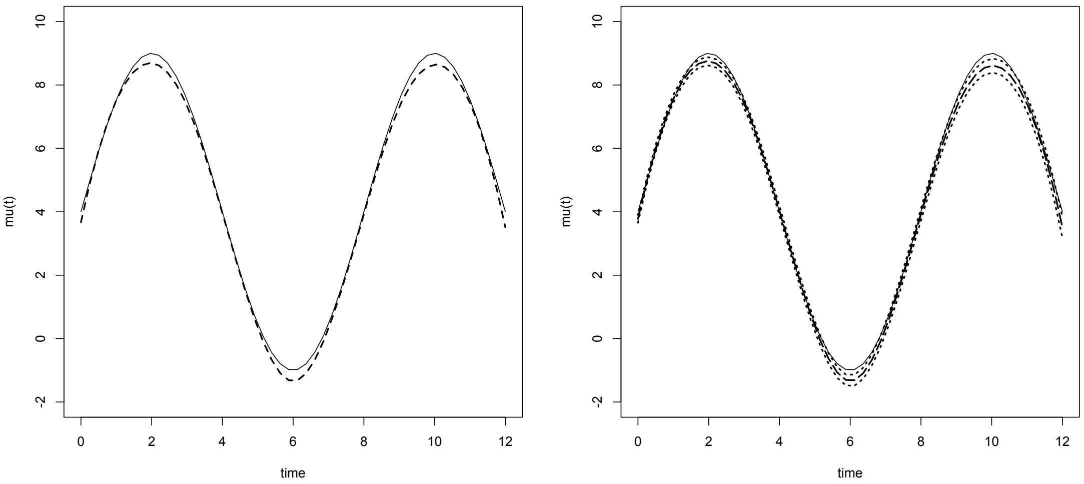

Figure 5.5. Estimated curves of $\mu(t)$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets, and the dotted curves are the pointwise confidence intervals.

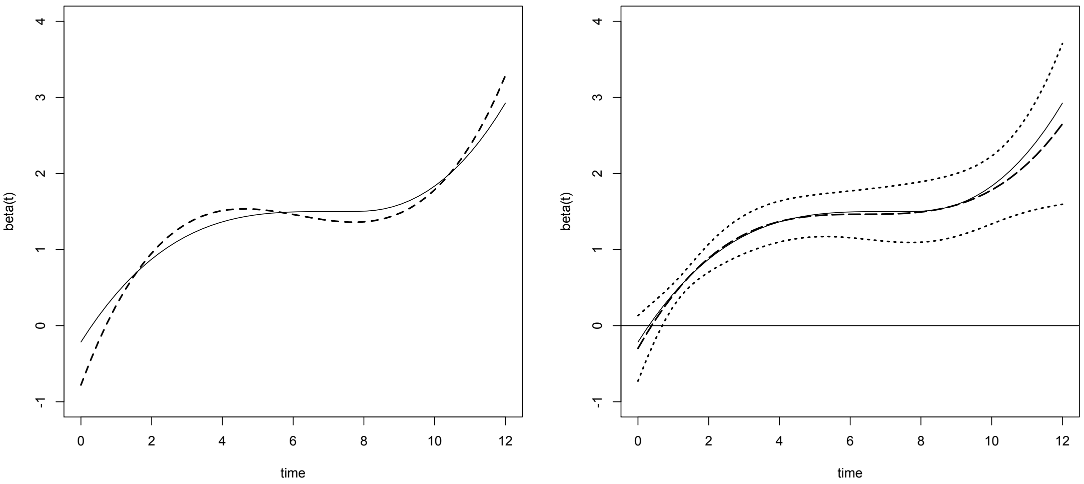

Figure 5.6. Estimated curves of $\beta(t)$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets, and the dotted curves are the pointwise confidence intervals.

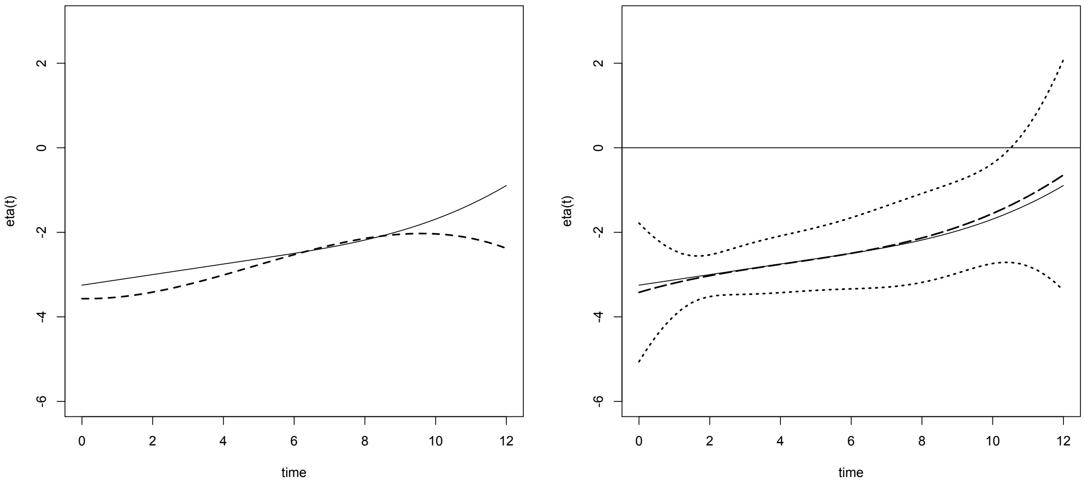

Figure 5.7. Estimated curves of $\eta(t)$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets, and the dotted curves are the pointwise confidence intervals.

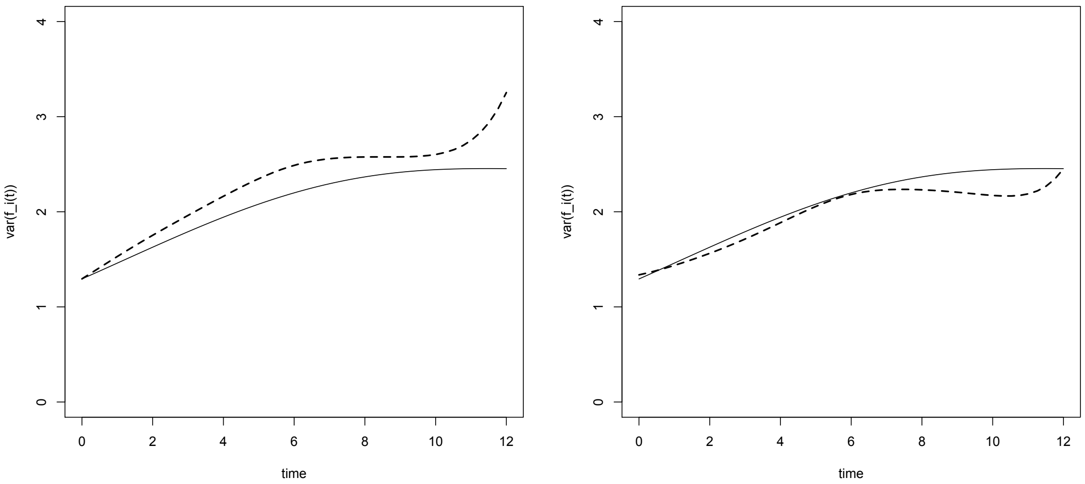

Figure 5.8. Estimated curves of $Var\{f_i(t)\}$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

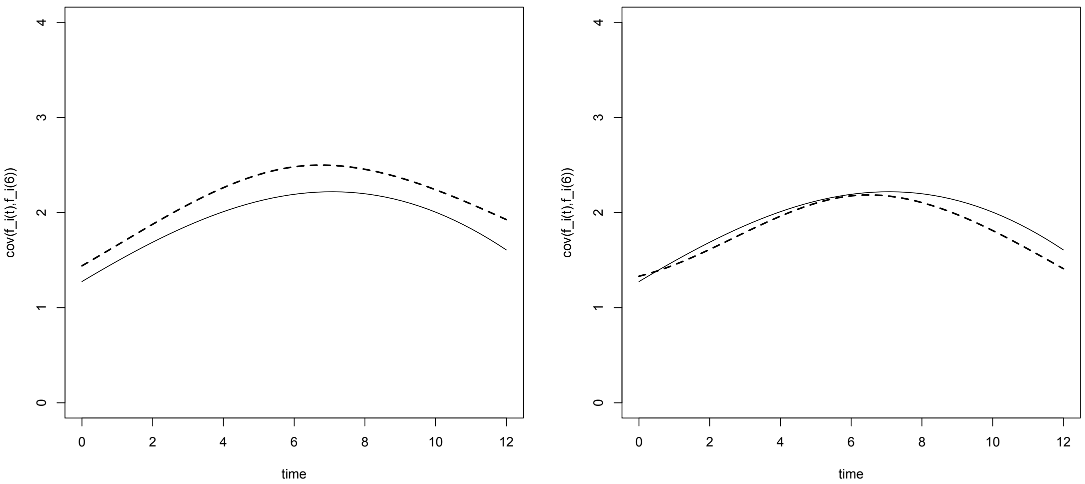

Figure 5.9. Estimated curves of $Cov\{f_i(t), f_i(6)\}$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

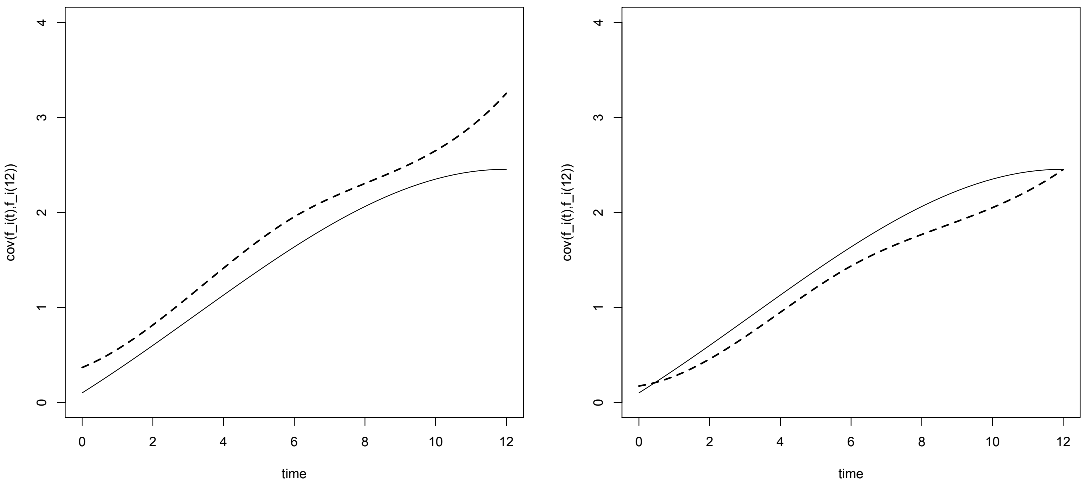

Figure 5.10. Estimated curves of $Cov\{f_i(t), f_i(12)\}$. The solid curves in both figures are the true curves. The dashed curve in the left figure is the estimated curves from the data with median RASE. The dashed curve in the right figure is the estimated mean curves from the 100 datasets.

Figure 5.5 is the estimated curves of $\mu(t)$. Similar to Example 5.1, the estimated curves in this example also closely resemble the true curves in both the left and right panels of the plot. Figure 5.6 and 5.7 present the estimated curves of $\beta(t)$ and $\eta(t)$, respectively. In the right panel, the estimated curves almost overlap with the true curves, which are all covered by the confidence bands across time. This indicates the good performance of our estimation procedure with B-spline smoothing. Note that the mean estimated curve (right panel) performs better than the median RASE estimated curves (left panel) as expected. The performance of the estimated random curves $f_i(t)$ are demonstrated in Figure 5.8 through 5.10 via its time-varying variance and covariance curves. It can be seen from the plots that although the estimated curves deviate from the true curves, the biases are small and the main trends of the curves are captured.

Table 5.7 presents the estimation performance of the constant parameter (only σ^2 in this case) in terms of bias and RMSE. In the table the performance of the three combinations of d_β and d_η is very similar.

Table 5.7: Estimation results for constant parameter

		$(4,4)$		$(5,4)$		$(5,5)$	
n	parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE
200	σ^2	-0.003	0.0228	-0.0015	0.023	-0.0014	0.0232
400	σ^2	0.0067	0.0162	0.0063	0.0161	0.0063	0.0161

Table 5.8 presents the standard errors of the estimated curves at three equally spaced time points in $[0, 12]$ for the model with $(d_\mu = 7, d = 5, d_\beta = 4, d_\eta = 4)$. The standard errors are calculated using Bootstrap technique described in Example 1. The mean and the standard deviations of the $N = 100$ standard errors are displayed in the columns of SD_{mean} and SD_{std}, respectively, and compared with the standard deviations calculated from the $N = 100$ replicates shown in the column of SD.

Table 5.8: Standard errors of the estimators

		$\mu(t)$			$\beta(t)$			$\eta(t)$		
n	time	SD	SE_{mean}	SE_{std}	SD	SE_{mean}	SE_{std}	SD	SE_{mean}	SE_{std}
200	$t = 3$	0.0975	0.0976	0.0324	0.1560	0.1548	0.0550	0.4203	0.3950	0.1402
	$t = 6$	0.1338	0.1222	0.0410	0.2286	0.2120	0.0796	0.6064	0.5341	0.2023
	$t = 9$	0.1610	0.1498	0.0518	0.3235	0.2775	0.1028	0.8693	0.7285	0.2924
400	$t = 3$	0.0715	0.0721	0.0068	0.1118	0.1210	0.0236	0.3024	0.3155	0.0598
	$t = 6$	0.0885	0.0875	0.0116	0.1707	0.1760	0.0459	0.4846	0.4699	0.1659
	$t = 9$	0.1121	0.1089	0.0165	0.2250	0.2283	0.0486	0.6978	0.6439	0.2312

5.3.2 A real data example

In this section, we illustrate the proposed model and estimation procedure by an empirical analysis of the data collected from a smoking cessation study. Specifically, this data set was collected from a randomized, placebo-controlled clinical trial (N=1504) of five active smoking-cessation pharmacotherapies, in which daily smokers who were highly motivated to quit were recruited (Piper et al., 2009). The focus of this example is to examine the effects of the longitudinal withdrawal symptom on the time to lapse (T_L, first smoking after quit) in a two-week post-quit study period (i.e. $t \in (0, 14)$). Accordingly, we treat time to lapse (T_L) as the survival response, and choose negative affect (NA) as the time-varying longitudinal covariate of interest. It is commonly believed that withdrawal symptom such as negative affect would exhaust the self-control resources that prevent the participants from smoking when they attempt to quit, and thus leads to the cessation failure. After data cleaning, N=794 subjects are used for the analysis in this study.

Figure 5.11 presents the mean curve of negative affect (NA) among the 794 subjects over time, which is estimated nonparametrically using penalized splines. The curvature shape in the figure indicates that a nonparametric submodel might be better option to capture the longitudinal trend than a parametric submodel. In addition, Figure 5.12 shows that the trajectories of NA vary greatly across individuals over time, indicating a high level of randomness. In order to take these factors into consideration, in this section, we use a nonparametric mixed effects model to fit NA, allowing for both nonparametric fixed and random curves.

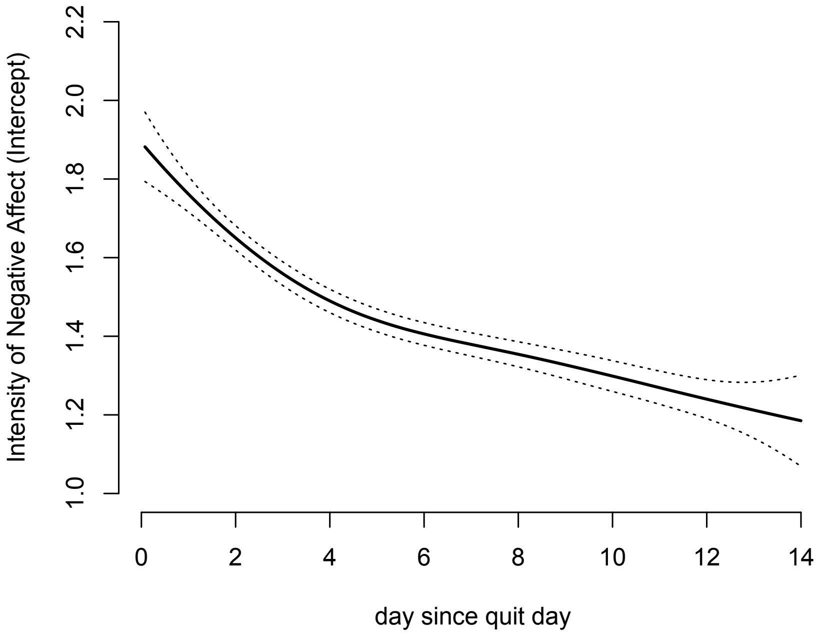

Figure 5.11. Nonparametric curve of Negative Affect. The solid curve is the mean trend, and the dashed curves are the pointwise confidence intervals

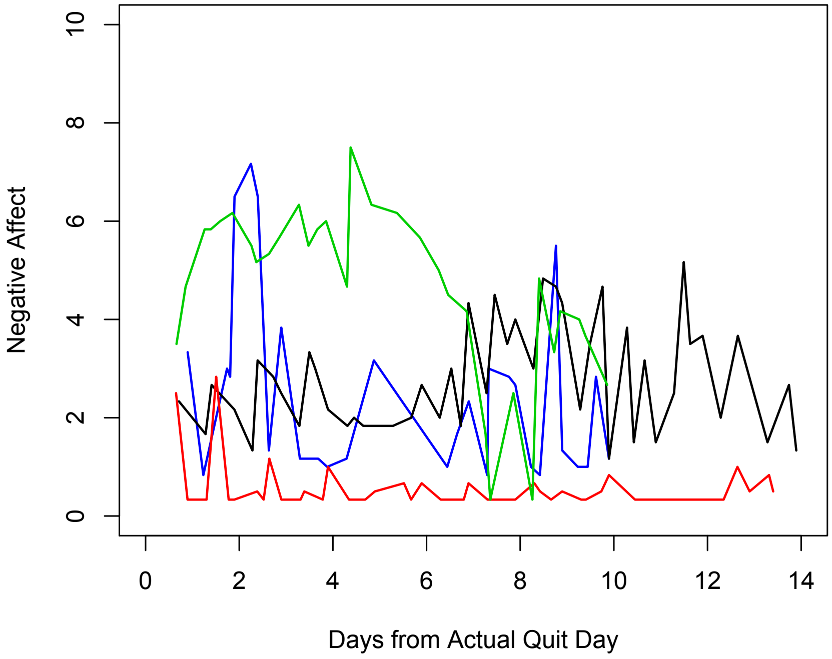

Figure 5.12. NA trajectories of 4 random individuals in the study

We consider the joint model

$$W_{\mathrm{NA}i}(t) = \mu(t) + f_i(t) + e_i(t),$$

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta(t)f_i(t) + \eta_1(t)Z_{1i} + \eta_2(t)Z_{2i}\},$$

where $\mu(t)$ is the smooth curve of average NA across population over time, and $f_i(t)$ is the ith individual's random deviation from the mean curve $\mu(t)$. $e_i(t)$ is the measurement error assumed to follow normal distribution $N(0, \sigma^2)$. It is assumed that $f_i(t)$ and $e_i(t)$ are independent. In the survival submodel, $f_i(t)$ presents the ith individual's level of Negative Affect, Z_1 and Z_2 are the two dummy variables corresponding to the monotherapy and the combined pharmacotherapy, respectively. $\beta(t), \eta_1(t)$ and $\eta_2(t)$ are the three time-varying coefficients for NA and two active treatments, respectively.

As mentioned in the previous section, we use the linear combination of cubic

B-spline basis functions to approximate all the nonparametric functions in the model, i.e.,

$$
\begin{aligned}
\mu(t) &\approx \sum_{p=1}^{d_\mu} \mu_p B_p^{(\mu)}(t),\\
f_i(t) &\approx \sum_{k=1}^{d} b_{ik} B_k^{(b)}(t),\\
\beta(t) &\approx \sum_{l=1}^{d_\beta} \beta_l B_l^{(\beta)}(t),\\
\eta_k(t) &\approx \sum_{j=1}^{d_{\eta_k}} \eta_{kj} B_j^{(\eta)}(t), \quad k = 1, 2.
\end{aligned}
$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{d_\mu})^T, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_\beta})^T, \boldsymbol{\eta}_k = (\eta_1, \ldots, \eta_{d_{\eta_k}}), k = 1, 2$ are the fixed coefficients of the corresponding spline basis, and $\mathbf{b}_i = (b_{i1}, \ldots, b_{ik})^T$ are the random coefficients following the normal distribution, $\mathbf{b}_i \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}_b)$. d_μ, d, d_β, d_{η_k} are the number of basis functions for $\mu(t), f(t), \beta(t)$ and $\eta_k(t)$, respectively. They are related to the number of inner knots (i.e., # basis function = # inner knots + 4), which are located with equal space in the study period. The number of basis functions can be selected via model selection criteria such as AIC and BIC.

In our analysis, we conduct both the naive separate estimation method and joint modeling method to estimate the parameters $(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\Sigma}_b, \boldsymbol{\beta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. Model selection criteria AIC, corrected AIC (AICc) and BIC are employed to identify the optimal number of basis functions for $\mu(t)$ and $f_i(t)$, and the estimating results are compared between the two methods for the optimal model.

Model selection

Model selection in this section is conducted using AIC, AICc and BIC criteria with the definitions:

$$
\begin{aligned}
\text{AIC} &= -2\log\mathcal{L} + 2K,\\
\text{AICc} &= AIC + \frac{2K(K+1)}{n-K-1},\\
\text{BIC} &= -2\log\mathcal{L} + K\log(n),
\end{aligned}
$$

where $\mathcal{L}$ is the joint likelihood, K is the number of unknown parameters and n is the sample size. We compare these criteria among the models with different numbers of knots for the approximation functions, and the model with the smallest criteria scores is preferred. The main challenge here is to evaluate the log likelihood $\log \mathcal{L}$, which involves the multiple integral of the random effects. We propose to estimate the log likelihood via the Monte Carlo method as follows.

With all the estimated parameters, the log likelihood of the model becomes

$$\sum_{i=1}^{n} \log \int_{\mathbf{b}} f_i(\mathbf{W}_i|\mathbf{b}; \hat{\mu}(t), \hat{\sigma}^2) f_i(V_i, \Delta_i|\mathbf{b}; \hat{\lambda}_0(t), \hat{\beta}, \hat{\eta}_1, \hat{\eta}_2) f(\mathbf{b}|\hat{\mathbf{\Sigma}}_b) d\mathbf{b},$$

where $f_i(\mathbf{W}_i|\mathbf{b}; \hat{\mu}(t), \hat{\sigma}^2)$ is the conditional density of the longitudinal response $\mathbf{W}_i$ given $\mathbf{b}$, $f_i(V_i, \Delta_i|\mathbf{b}; \hat{\lambda}_0(t), \hat{\beta}, \hat{\eta}_1, \hat{\eta}_2)$ is the conditional density of the survival response (V_i, Δ_1) given $\mathbf{b}$, and $f(\mathbf{b}|\hat{\mathbf{\Sigma}}_b)$ is the normal density of $\mathbf{b}$ with the estimated covariant matrix $\hat{\mathbf{\Sigma}}_b$. We treat first two density functions in the integral as a function of $\mathbf{b}$, denoted by $g_i(\mathbf{b})$, with the definition

$$g_i(\mathbf{b}) = f_i(\mathbf{W}_i|\mathbf{b}; \hat{\mu}(t), \hat{\sigma}^2) f_i(V_i, \Delta_i|\mathbf{b}; \hat{\lambda}_0(t), \hat{\beta}, \hat{\eta}_1, \hat{\eta}_2).$$

Hence the log likelihood can be estimated by generating M Monte Carlo samples of $\mathbf{b}$ from $N_k(\mathbf{0}, \hat{\mathbf{\Sigma}}_b)$, and calculating sample mean of $\sum_{i=1}^{n} \log g_i(\mathbf{b})$, i.e.,

$$\log \hat{L} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \log g_i(\mathbf{b}_m).$$

In this data example, we use $M = 100000$ random samples to estimate the log likelihood.

Since there are five nonparametric functions to approximate in the model, we have to select the optimal number of knots for these functions one by one. In order to save the computing time, before evaluating different models, it is better to get a rough idea of how these curves would look like and what would be a reasonable range for the number of knots, especially for the fixed curves of $\mu(t)$, $\beta(t)$ and $\eta_k(t)$.

As is shown in Figure 5.12, the trajectory of average NA, or $\mu(t)$, is quite smooth and does not vibrate greatly over time. This suggests that a small number of B-spline basis functions might be sufficient in approximating the curve $\mu(t)$. Thus we consider three reasonable scenarios for d_μ, i.e., $d_\mu = 4, 5, 6$, which correspond to 0, 1, and 2 inner knots, respectively.

For $\beta(t)$ and $\eta_k(t)$ in the survival model, we use a "quick and dirty" way to obtain the preliminary estimation of $\beta(t)$ and $\eta(t)$ via the varying-coefficient Cox model. This corresponds to the separate estimation method which can provide immediate but biased results. We use 4 B-spline basis functions for all the three curves tentatively ($d_\beta = d_{\eta 1} = d_{\eta 2} = 4$, corresponding to 0 inner knots). Despite the baiseness, we obtain the following figures (Figure 5.13), which, though not accurate, indicate that $\beta(t)$ might be constant over time, and the shapes of $\eta_1(t)$ and $\eta_2(t)$ are simple and can be estimated by simple models. Thus we consider constant and linear curves for these coefficient functions.

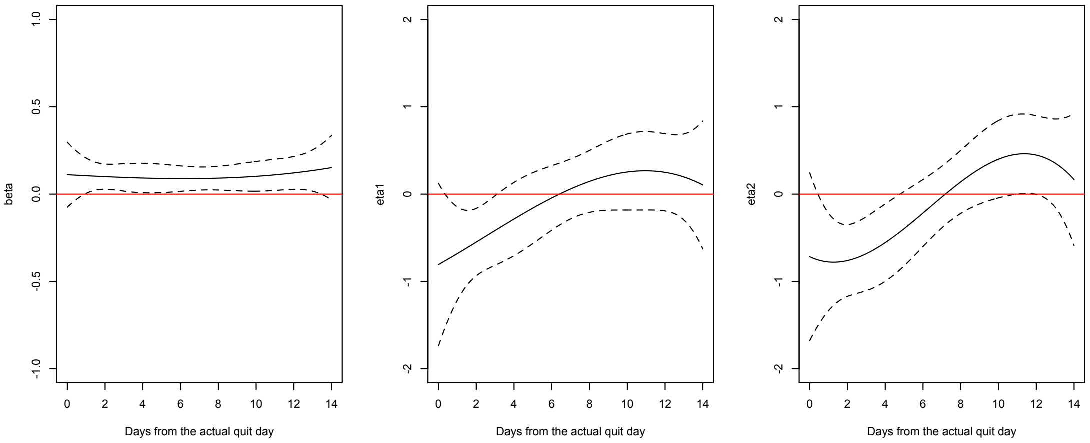

Figure 5.13. Estimated curves of $\beta(t)$, $\eta_1(t)$ and $\eta_2(t)$ from varying-coefficient Cox model

When it comes to the random curves $f_i(t)$, it is challenging in finding a reasonable range for the number of knots because we have little idea of how the curve would look like for each individual. Although Ding and Wang (2008) suggests that a small dimension of random effects (i.e., 1 or 2 random effects) is sufficient to capture the randomness of the longitudinal response, the snapshot of Figure 5.12 indicates that this may not be the case, and the level of randomness may vary over time. For example, in Figure 5.12, the variation of the four random curves is greater at the beginning than at the end of the study. Since our proposed algorithm, the EM-DoIt algorithm, is capable of estimating the joint models with large dimension of random effects, we consider the cases of $d = 5, 6, 7$ to incorporate a higher level of randomness in this case.

We conduct the model selection in two steps for the survival and the longitudinal part separately. In the first step, we hold the longitudinal submodel at $d_\mu = 6, d = 7$ – the largest number of basis functions considered, and choose the number of basis functions for $\beta(t)$ and $\eta(t)'s$ in the survival submodel. In the second step, we hold the survival submodel at the number of basis functions selected in step 1, and select the number of basis functions for $\mu(t)$ and $f_i(t)$ in the longitudinal submodel. The model selection results of the two steps are presented in Table 5.9 and Table 5.10, respectively.

Table 5.9: Model Selection Results for Survival Model (Step 1)

	$d_\beta = 1$				$d_\beta = 2$			
(d_{η_1}, d_{η_2})	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
$\log \hat{L}$	-28690	-28690	-28706	-28701	-28693	-28704	-28702	-28700
AIC	57456	57468	57489	57482	57464	57487	57485	57482
AICc	57460	57472	57493	57486	57469	57491	57489	57487
BIC	57631	57648	57669	57667	57645	57672	57669	57671

In Table 5.9, in the first two rows, $d = 1$ corresponds to a constant function and the $d = 2$ corresponds to a linear function. The results show that the log likelihood increases, and all the other criteria decrease as the number of d_β, d_{η_1} and d_{η_2} increase. This pattern suggests that the constant functions are sufficient, and the regression coefficients are constant over time. Therefore, we hold $\beta(t)$, $\eta_1(t)$ and $\eta_2(t)$ at constant, and obtain the following model selection results for $\mu(t)$ and $f_i(t)$ in Table 5.10.

Table 5.10: Model Selection Results for Longitudinal Model (Step 2)

	$d = 5$			$d = 6$				$d = 7$		
	$d_\mu = 4$	$d_\mu = 5$	$d_\mu = 6$	$d_\mu = 4$	$d_\mu = 5$	$d_\mu = 6$	$d_\mu = 7$	$d_\mu = 4$	$d_\mu = 5$	$d_\mu = 6$
$\log \hat{L}$	-26485	-26485	-26483	-26471	-26471	-26467	-26483	-26463	-26463	-26459
AIC	53016	53018	53015	53000	53002	52995	53029	52999	53000	52995
AICc	53018	53020	53017	53003	53005	52998	53032	53002	53004	52999
BIC	53123	53129	53131	53134	53141	53139	53177	53165	53171	53170

In Table 5.10, the model with $d = 6, d_\mu = 6$ was selected by AIC and AICc as the optimal model, and the simplest model with $d = 5, d_\mu = 4$ was selected by BIC. We decide to follow the criteria of AIC and AICc, and take $d = 6, d_\mu = 6$ as the number of bases functions for $f_i(t)$ and $\mu(t)$, respectively, in the final model.

Estimation Results

Based on the model selection results, $d_\mu = 6, d = 6, d_\beta = 1, d_{\eta_1} = 1, d_{\eta_2} = 1$ is selected for the final model. In this subsection, we compare the estimation results between the separate estimation method and the maximum joint likelihood approach for the chosen model. The standard errors of the maximum joint likelihood (MJL) method are calculated via bootstrap by evaluating standard deviations of the estimates from the $B = 100$ iid datasets sampled with replacement form the original Lapse data.

The estimating results in Table 5.11 show that there is little difference of the longitudinal parameter estimates (i.e., $\boldsymbol{\mu}, \boldsymbol{\Sigma}_b, \sigma^2$) between separate estimation and MJL. However, the estimated survival parameters (i.e., β and η's) are quite different between the two methods. Similar to what we have discovered for the parametric joint model settings, separate estimation yields coefficient estimates that are biased towards the null, and thus tend to be nonsignificiant, whereas the MJL method produces more reliable estimates. In this case, neither of the two active treatments is identified as effective by separate estimation, whereas the MJL result indicates that the combined pharmacotherapy has significant effect in reducing the risk of Lapse. Both the two methods find the longitudinal covariate NA significantly positively associated with Lapse. The estimated curves of $\mu(t)$ and $var\{f_i(t)\}$ from the selected joint model are shown in Figure 5.14 and 5.15, respectively. The pointwise confidence intervals are obtained from the standard deviations of bootstrap estimates at each time point. The estimated curves from separate estimation method are very similar and thus not presented here.

Table 5.11: Parameter Estimation for Lapse Data

	Separate Estimation		Maximum Joint Likelihood	
	Estimate	SE	Estimate	Bootstrap SE
η_1	-0.0760	0.1204	-0.2251	0.1739
η_2	-0.0966	0.1243	-0.4837*	0.2013
β	0.1039*	0.0230	0.1455*	0.0432
μ_1	1.8768*	0.0790	1.8681*	0.0777
μ_2	-0.0194	0.1173	0.0012	0.1275
μ_3	-0.5699*	0.0742	-0.5669*	0.0810
μ_4	-0.3264*	0.1181	-0.2992*	0.1265
μ_5	-0.8214*	0.0992	-0.8437*	0.1055
μ_6	-0.6205*	0.1101	-0.6327*	0.1054
Σ_{11}	2.4248	-	3.0036*	0.3183
Σ_{22}	3.3232	-	4.5763*	0.7801
Σ_{33}	1.8867	-	2.2230*	0.2789
Σ_{44}	3.2772	-	4.3666*	0.8018
Σ_{55}	1.9466	-	2.7018*	0.4460
Σ_{66}	1.6896	-	2.2069*	0.3702
σ^2	0.9144	-	0.8313*	0.0423

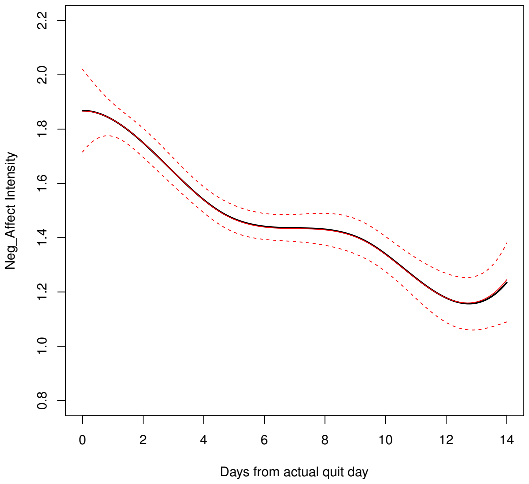

Figure 5.14. Estimated curves of $\mu(t)$ from MJL. The solid curve is the estimated mean trajectory of NA. The dashed curves are the pointwise confidence intervals estimated using bootstrap.

---

[1] represents statistically significant at 0.05 level.

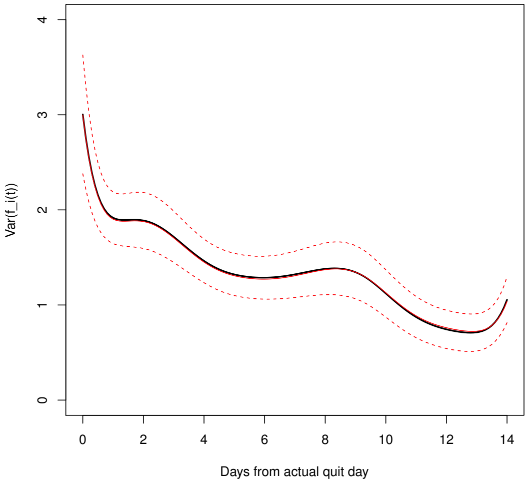

Figure 5.15. Estimated curves of $var\{f_i(t)\}$ from MJL. The solid curve is the estimated mean trajectory of NA. The dashed curves are the pointwise confidence intervals estimated using bootstrap.

5.4 Discussions

In this chapter, we have presented a nonparametric joint modeling framework which is flexible and robust in fitting complex longitudinal processes and time-to-event simultaneously. Such model is capable of capturing the irregular shape of longitudinal trajectories, and at the same time allows survival coefficients in the Cox's model to vary with time. Cubic B-splines are employed to approximate the parameter functions in the model. An advantage of using spline smoothing technique is that it flexibly incorporates the constant functions as a special case, and the nonparametric functions can be easily reduced to constant parameters as shown in the real data analysis.

We propose to estimate the parameters using the classical maximum joint likelihood approach implemented via an EM algorithm. Thanks to the new numerical integration technique of DoIt, we are able to handle the multidimensional integrals in the likelihood successfully and estimate the high-dimensional parameters which were unobtainable by the conventional EM algorithm in the previous studies. Other estimation methods such as the Bayesian approach can also be investigated for the current model setting. Although we considers only a single longitudinal process as an illustration in this paper, more longitudinal covariates or biomarkers can be added to the nonparametric model as the computational capability further improves.

Chapter 6

Extension and Future Work

6.1 Extension: joint modeling with discrete longitudinal covariates

The joint models studied in Chapter 3 and Chapter 5 consider only the continuous longitudinal processes as the predictors for the survival outcome. A useful extension of such model setting is to include the discrete longitudinal covariates such as binary or count measurements. In the smoking cessation study discussed in the previous chapters, the candidate discrete longitudinal covariates include the binary variable of whether a stress event has happened since the last prompt, and the count variable such as the number of cigarettes smoked since the last prompt. All these covariates are potential predictors for the risk of relapse or the abstinence failure. Although a naive approach is to use observed measurements $W_i(t)$ directly as the predictor in the Cox model, as discussed in Chapter 1, these observations are not available at all the event time points. Thus joint modeling approaches can be considered for such cases in order to obtain accurate estimations and reliable inferences.

The literature of joint modeling discrete longitudinal processes and survival time is quite limited. Most of the studies design ad-hoc models for the specific data of interest, and few of them consider a generic model and methodology as in the case of the joint models with continuous longitudinal processes. Faucett et al. (1998) proposed a Markov model for binary longitudinal covariates, which is used as a predictor in the proportional hazards model for the survival outcome. Xu and Zeger (2001b) considered a latent model, where the latent process is assumed to follow a Gaussian stochastic process, and links with the longitudinal measurement and the survival outcome via generalized linear model (GLM) and proportional hazards model, respectively. Huang et al. (2002) developed a survival model for bivariate event times, which is jointly modeled with the binary longitudinal measurements through some discrete latent variables and logistic regression models. Cowling et al. (2006) explored the relationship between the survival time and the recurrent events such as epileptic seizures which is measured by count variable in the cohort study. The author assumes the count measurement follows a Poisson distribution and the survival time follows the Pareto distribution, and the event rates of two processes are linked by the same set of random effects and baseline covariates. Rizopoulos et al. (2008) used a mixed-effects logistic regression model for the binary longitudinal covariates with excess zeros and a mixed-effects accelerated failure time model for the survival time, and the two models share the same random effects.

6.1.1 Model Setting

In this chapter we propose a generalized joint model that accommodates both binary and count longitudinal covariates, which is estimated by the maximum joint likelihood approach with the EM-DoIt algorithm. All the notations are the same as in Chapter 2, with the only exception of W_{ij}, which is now the discrete longitudinal covariate of the ith individual observed at time t_{ij}. We assume that for each individual, the observed longitudinal observations are linked with a linear predictor $X_i(t)$ via the generalized linear models given by

$$g[E\{W_i(t)|X_i(t)\}] = X_i(t), \tag{6.1}$$

where $g(\cdot)$ is the link function that varies for different types of observations. $X_i(t)$ can be thought as an underlying latent process that captures the heterogeneity across subjects, and can be modeled by time and other covariates. To be consistent with the previous model settings, here we consider a nonparametric model with time as the only predictor,

$$X_i(t) = \mu(t) + f_i(t), \tag{6.2}$$

where $\mu(t)$ is the fixed mean trajectory of all the individuals, and $f_i(t)$ is the random curve representing the ith individual's deviation from the mean process. Note that in the simplest scenario, (6.2) reduces to the linear mixed-effects model of the form

$$X_i(t) = \mathbf{b}_i^T \boldsymbol{\rho}(t), \tag{6.3}$$

where $\boldsymbol{\rho}(t)$ is a vector function of t, and $\mathbf{b}_i$ is the random effects of the same dimension of $\boldsymbol{\rho}(t)$ and assumed to follow the multivariate normal distribution $N(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$.

Given the underlying process $X_i(t)$ and the baseline covariates Z_i, the conditional hazard is assumed to follow a varying-coefficient Cox model given by

$$\lambda_i(t|X_i(t), Z_i, t \le T_i) = \lambda_0(t) \exp\{\beta(t) X_i(t) + \eta(t) Z_i\}. \tag{6.4}$$

Note that the fixed part $\mu(t)$ in $X_i(t)$ is non-identifiable across individuals and can thus be combined with the baseline hazard $\lambda_0(t)$. Hence the model becomes

$$\lambda_i(t|f_i(t), Z_i, t \le T_i) = \lambda_0(t) \exp\{\beta(t) f_i(t) + \eta(t) Z_i\}, \tag{6.5}$$

which is exactly the same as the survival submodel used in Chapter 5. Note that for the simple cases, the time-varying coefficients $\beta(t)$ and $\eta(t)$ can be reduced to constant form of β and η.

Similar to the nonparametric joint modeling approach proposed in Chapter 5, the nonparametric functions in the current model setting can be approximated by linear combinations of B-splines. For example, the random function $f_i(t)$ is approximated by the B-spline basis functions with random coefficient vectors $\mathbf{b}_i$. The associated joint likelihood function is of the form

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \int f_{V_i, \Delta_i | \mathbf{b}_i} \cdot f_{W_i | \mathbf{b}_i} \cdot f_{\mathbf{b}_i} d\mathbf{b}_i, \tag{6.6}$$

where the density function of the random effects and the conditions density function of the survival outcome given the random effects are the same as we specified in Chapter 5, given by

$$
\begin{aligned}
f_{\mathbf{b}_i} =& \frac{1}{2\pi}|\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{b}_i^T \boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\right\}, \\
f_{V_i,\Delta_i|\mathbf{b}_i} =& [\lambda_0(V_i)\exp\{\beta(V_i)f_i(V_i) + \eta(V_i)Z_i\}]^{\Delta_i} \\
& \times \exp\left\{-\int_0^{V_i} \lambda_0(t)\exp\{\beta(t)f_i(t) + \eta(t)Z_i\}dt\right\},
\end{aligned}
\tag{6.7}
$$

whereas the conditional density function of the discrete longitudinal covariates $f_{W_i|\mathbf{b}_i}$ is no longer the normal density. It is determined by different types of W_i and its associated link functions. The specific forms of $f_{W_i|\mathbf{b}_i}$ for various scenarios are provided in the simulation examples.

The logarithm of (6.6) can be maximized using the same EM-DoIt algorithm as for the previous continuous cases. Since the discrete longitudinal covariates contain less information than the continuous covariates, the EM algorithm requires longer time to converge, and the computation becomes more intensive.

In this chapter, we apply the EM-DoIt algorithm to the parametric generalized model settings, where the longitudinal processes are assumed to follow submodel (6.1) and (6.3), and the survival outcome is modeled by (6.5) with constant regression coefficients. The future work would include the simulation and real data analysis for the nonparametric joint model setting. In addition, to the best of our knowledge, none of the existing literature has studied the asymptotic properties of the estimators of the generalized joint models. Thus, the theoretical establishment of the MLE proposed in this chapter would be another challenging task for the future work.

6.1.2 Simulation results

In this section we consider the joint models with discrete longitudinal processes, where the longitudinal covariates are modeled by the generalized mixed-effects models. We discuss three simulation examples, including three types of discrete longitudinal covariates: binary, count, and zero-inflated count observations. All the simulations are conducted using the EM-DoIt algorithm proposed in Chapter 3.

Example 6.1. In this example, we consider a 2-dimensional joint model with binary longitudinal process. In this setting, $\boldsymbol{W}_i(t)$ is a longitudinal process taking binary values with mean $E(W_i(t_{ij})) = \pi_{ij}$. The longitudinal and event processes satisfy the model

$$\text{logit}(\pi_{ij}) = X_i(t_{ij}) = b_{0i} + b_{1i}t_{ij},$$

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_X X_i(t) + \beta_Z Z_i\}$$

with the assumption

$$\boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

This is a low-dimensional setting with $\dim(\boldsymbol{b}_i) = 2$, and $\boldsymbol{\rho}(t) = (1, t)^T$. Since the longitudinal process $W_i(t)$ is binary, it is characterized by a generalized mixed-effects model with the logit link function. $X_i(t)$ is the underlying log-odds of $W_i(t)$ and is used in the Cox model as the time-varying covariate. The survival time is generated via the equation

$$T_i = \frac{1}{\beta_X b_{1i}} \log\{1 - \frac{\beta_X b_{1i} \log U_i}{\lambda_0 \exp(\beta_X b_{0i} + \beta_Z Z_i)}\}. \tag{6.8}$$

with $\lambda_0 = 1$. The average censoring rate is around 8%. All the parameters are specified in the table. In this scenario the estimating results are based on $N = 100$ data sets with $n = 500$ subjects in each data set.

Since the longitudinal covariates $W_i(t)$ is no longer a Gaussian process, the joint likelihood changes. Although the density functions of (6.7) remain the same, the logarithm of $f_{W_i|\mathbf{b}_i}$ in (6.6) changes to the form

$$l_2(\boldsymbol{b}_i) = \log(f_{W_i|\boldsymbol{b}_i}) = \sum_{j=1}^{N_i}(w_{ij}\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i) - \log\{1 + \exp(\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i)\},$$

and this makes the joint likelihood more complicated, and thus the integration techniques such as the Gaussian-Hermite Quadrature method and fully exponential Laplace method fail. However, DoIt is much easier in computation for such cases, and the algorithm for the continuous joint modeling can be extended directly into the nonlinear cases. In the following table, we show the simulation results obtained from the EM-DoIt algorithm which is also used in both Chapter 3 and Chapter 5.

Table 6.1: Simulation results for Example 6.1 by DoIt with M=10

	$\bar{n}. = 10$			$\bar{n}. = 20$		
	Bias	SD	RMSE	Bias	SD	RMSE
$\beta_X = 1.0000$	-0.1330	0.0693	0.1420	-0.0290	0.0615	0.0678
$\beta_Z = -1.0000$	0.0643	0.0923	0.1125	0.0490	0.0953	0.1072
$\theta_1 = -2.0000$	0.0465	0.0940	0.1049	0.0274	0.0574	0.0636
$\theta_2 = 1.0000$	-0.0133	0.0641	0.0655	-0.0095	0.0393	0.0404
$\sigma_{11} = 1.0000$	0.2261	0.2247	0.3188	0.0799	0.0983	0.1266
$\sigma_{12} = -0.0100$	-0.1059	0.0933	0.1411	-0.0471	0.0417	0.0629
$\sigma_{22} = 0.5000$	0.0501	0.0514	0.0718	0.0069	0.0217	0.0227
Median time(s)	2860			1373		

In Table 6.1 we notice that the DoIt method takes much longer time than it does in Example 3.1 for the continuous case even though this is also a low-dimensional setting. The increase of time is caused by two facts. First, as we mentioned above, the likelihood becomes more complicated than the previous one for the continuous Gaussian process. Thus it takes longer time to numerically locate the mode $\hat{\boldsymbol{b}}_i$ and evaluate the Fisher information $\boldsymbol{D}_i$ at the mode. Second, the binary covariate provides much less information for calculation than the continuous covariate. Therefore, although we increase the subject number from $n = 100$ to $n = 500$, the EM algorithm still takes much more iterations to coverage than in the previous scenarios.

As for the estimating performance, the EM-DoIt algorithm yields good estimates for all the parameters when there are more personal observations (i.e., $\bar{n}. = 20$) in terms of bias, SD, and RMSE. However, when there are fewer personal observations (i.e., $\bar{n}. = 10$) the estimating performance deteriorates due to the lack of information.

Example 6.2. In this example, we consider a 2-dimensional joint model with count longitudinal process. In this scenario, $\boldsymbol{W}_i(t)$ is a longitudinal process taking count values with the mean $E(W_i(t_{ij})) = \tau_{ij}$. The longitudinal processes satisfy the Poisson mixed-effects model

$$\log(\tau_{ij}) = X_i(t_{ij}) = b_{0i} + b_{1i}t_{ij}.$$

The event process and the assumptions are the same as that in Example 6.1. This setting is almost the same as that of Example 6.1 except that we assume the longitudinal covariate is from a Poisson process so that the generalized mixed-effects model has a Poisson link function. $X_i(t)$ is the underlying log mean of $W_i(t)$ and represent the strength of the observed longitudinal covariate. The corresponding logrithm of $f_{W_i|\mathbf{b}_i}$ in (6.6) becomes

$$l_2(\boldsymbol{b}_i) = \log(f_{W_i|\boldsymbol{b}_i}) = \sum_{j=1}^{n_i} w_{ij}\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i - \exp\{\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i\} - \log w_{ij}.$$

Similar to Example 6.1, since this complex likelihood causes great difficulty for approximation using GHQ and FEL, in this scenario we still only consider the DoIt method. The survival time is generated from (6.8) with $\lambda_0 = 0.5$. The average censoring rate is around 18%. All the parameters are specified in Table 6.2. The estimating results are based on $N = 100$ data sets with $n = 100$ subjects in each of them.

Table 6.2: Simulation results for Example 6.2 by DoIt with M=10

	$\bar{n}. = 5$			$\bar{n}. = 20$		
	Bias	SD	RMSE	Bias	SD	RMSE
$\beta_X = 1.0000$	-0.0014	0.2400	0.2400	0.0289	0.1650	0.1675
$\beta_Z = -1.0000$	-0.0615	0.2811	0.2877	-0.0181	0.2630	0.2636
$\theta_1 = -1.0000$	0.0469	0.1179	0.1268	0.0067	0.0923	0.0925
$\theta_2 = 1.0000$	-0.0122	0.0711	0.0721	0.0011	0.0448	0.0448
$\sigma_{11} = 0.5000$	0.0607	0.1814	0.1913	0.0114	0.1224	0.1229
$\sigma_{12} = -0.0010$	-0.0462	0.0702	0.0840	-0.0077	0.0468	0.0474
$\sigma_{22} = 0.5000$	0.0233	0.0408	0.0469	0.0031	0.0228	0.0230
Median time(s)	74.19			36.23		

Note that the computation in this example is much faster in this scenario than in Example 6.1. This is mainly because the count covariate process resembles continuous process much more than the binary covariates. Thus a moderate number of subjects $n = 100$ provide sufficient information for calculation and the EM algorithm takes fewer iterations to converge. Again, the algorithm consumes less time for the data set with more individual observations (i.e., $\bar{n}.$) compared with the data set with fewer individual observations (i.e., $\bar{n}. = 5$) due to the richness of the information. In both cases the EM-DoIt algorithm yields good parameter estimates in terms of bias, SD and RMSE.

Example 6.3. For the last simulation scenario, we consider a 2-dimensional joint models with zero-inflated count longitudinal process. In this scenario, $\boldsymbol{W}_i(t)$ is a longitudinal process taking value 0 (i.e. defect) with probability π, where π is a known constant. The other values are the count values from the Poisson distribution with mean $E(W_i(t_{ij})) = \tau_{ij}$. Thus the longitudinal process $\boldsymbol{W}_i(t)$ satisfies the distribution:

$$
\begin{aligned}
P(W_i(t_{ij}) = 0) &= \pi + (1-\pi)e^{-\tau_{ij}}, \\
P(W_i(t_{ij}) = k) &= (1-\pi)\frac{\tau_{ij}^k e^{-\tau_{ij}}}{k!}, \ \ k = 1, 2, \ldots .
\end{aligned}
$$

The parameter τ_{ij} is specified by the Poisson mixed-effects model as in Example 6.2, and the event process and the model assumptions also follow those of Example 6.1 and 6.2. Since the distribution of $W_i(t)$ is a combination of binary and Poisson distributions, the corresponding logarithm of $f_{W_i|\mathbf{b}_i}$ in (6.6) becomes even more complex

$$
\begin{aligned}
l_2(\boldsymbol{b}_i) =& \log(f_{W_i|\boldsymbol{b}_i}) \\
=& -\sum_{j=1}^{n_i} \log(1 + e^{\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i}) + \sum_{w_{ij}=0} \log\{e^{\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i} + \exp(-e^{\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i})\} \\
&+ \sum_{w_{ij}>0} \boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i - \exp\{\boldsymbol{\rho}(t_{ij})^T\boldsymbol{b}_i\}.
\end{aligned}
$$

We consider only DoIt in this scenario. The survival time is generated via (6.8) with $\lambda_0 = 0.5$. The average censoring rate is around 18%. All the parameters are specified in Table 6.3 and the results are based on $N = 100$ data sets with $n = 300$ subjects in each of them.

Table 6.3: Simulation results for Example 6.3 by DoIt with M=10, $\bar{n}. = 20$

	$\pi = 0.1$			$\pi = 0.3$		
	Bias	SD	RMSE	Bias	SD	RMSE
$\beta_X = 1.0000$	0.0029	0.1164	0.1164	-0.0357	0.1222	0.1273
$\beta_Z = -1.0000$	-0.0073	0.1532	0.1534	0.0347	0.1375	0.1418
$\theta_1 = -1.0000$	0.0152	0.0585	0.0604	0.0059	0.0577	0.0580
$\theta_2 = 1.0000$	-0.1020	0.0319	0.1069	-0.3004	0.0319	0.3021
$\sigma_{11} = 0.5000$	0.0287	0.0617	0.0680	0.0615	0.0802	0.1011
$\sigma_{12} = -0.0010$	-0.0218	0.0220	0.0310	-0.0324	0.0293	0.0437
$\sigma_{22} = 0.5000$	0.0099	0.0107	0.0148	0.0072	0.0124	0.0143
Median time(s)	178.6			212.5		

As expected, in the zero-inflated Poisson setting, the EM-DoIt algorithm method takes more computation time than the Poisson scenario but not as much as the binary setting. And as we increase π from 0.1 to 0.3, more observations of 0 occurs for $W_i(t)$, thus less information can be used in estimation and longer time is taken. In both cases, the proposed method yields good estimates in terms of bias, SD and RMSE. It tends to produce more accurate and efficient estimates when the defect rate π is smaller.

In all the simulation examples we discussed above, the proposed maximum joint likelihood method produces reasonably good estimating results. In particular, the EM-DoIt algorithm is capable of calculating the MLE when the joint likelihood functions become more complex in the cases of discrete longitudinal processes, and the computing time is reasonable. Such computational capability makes it possible for us to extend to current simulation setting to the more complicated nonparametric generalized joint model setting in the future work.

6.2 Future work

6.2.1 Asymptotic theories for nonparametric joint modeling

We proposed nonparametric joint model settings in Chapter 5, and conducted numerical studies and real data analysis. However, the asymptotic theories for the associated MLE have not been established yet. Similar to the theories developed for the parametric joint modeling in Chapter 4, the theories of nonparametric joint modeling can be built using the tool of empirical process. In this section, we state the desired conclusion of the consistency property and provide proof outline. The complete proof and the asymptotic normality will be established in future work.

To prepare for the theory development, we restate the model setting here. For the longitudinal part, let $W_i(t)$ denote the observed covariate process which is modeled by the nonparametric mixed effects model

$$W_i(t) = X_i(t) + \epsilon_i(t), i = 1, \ldots, n, \tag{6.9}$$

where $X_i(t)$ is the true covariate process and $\epsilon_i(t)$ is the zero-mean measurement error with variance σ^2. $X_i(t)$ is composed of two parts

$$X_i(t) = \mu(t) + f_i(t), \tag{6.10}$$

where $\mu(t) = E\{X_i(t)\}$ is the fixed process representing the average coefficient level across all the subject over time, and $f_i(t)$ is the random part representing the subject-specific trajectory deviations from the mean. Both $\mu(t)$ and $f_i(t)$ can be approximated by a linear combination of spline basis

$$\begin{aligned} \mu(t) &\approx \sum_{p=1}^{p_n} \mu_p B_p^{(\mu)}(t) = \boldsymbol{\mu}^T \mathbf{B}^{(\mu)}(t), p_n \to \infty, n \to \infty, & (6.11) \\ f_i(t) &= \sum_{k=1}^{d} b_{ik} B_k^{(b)}(t) = \mathbf{b}_i^T \mathbf{B}^{(b)}(t), & (6.12) \end{aligned}$$

where $\mathbf{B}^{(\mu)}(t)$ and $\mathbf{B}^{(b)}(t)$ are two sets of spline basis. $\boldsymbol{\mu}$ is a p_n dimensional constant coefficient vector and $\mathbf{b}_i$ is a d dimensional random coefficient vector assumed to be normally distributed with mean $\boldsymbol{\mu}_b$ and covariance $\boldsymbol{\Sigma}_b$. Note that we allow the dimension p_n to diverge with n while d be finite.

The survival data (V_i, Δ_i) is modeled by the varying-coefficient Cox model of the form

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta(t)X_i(t) + \boldsymbol{\eta}(t)^T\mathbf{Z}_i\}, i = 1, \dots, n, \tag{6.13}$$

where $\beta(t)$ and $\boldsymbol{\eta}(t)$ are the coefficient functions for the longitudinal process $X_i(t)$ and the K dimensional baseline covariates $\mathbf{Z}_i$. The coefficient functions can also be approximated by linear combination of spline basis

$$\beta(t) \approx \sum_{q=1}^{q_n} \beta_q B_q^{(\beta)}(t) = \boldsymbol{\beta}^T\mathbf{B}_q^{(\beta)}(t), \tag{6.14}$$

$$\eta_k(t) \approx \sum_{r=1}^{r_k} \eta_{kr} B_r^{(\eta_k)}(t) = \boldsymbol{\eta}_k^T\mathbf{B}_r^{(\eta_k)}(t), k = 1, \dots, K, \tag{6.15}$$

where $\boldsymbol{\beta}$ is a q_n dimensional constant coefficient vector and $\boldsymbol{\eta}_k$ is a r_k dimensional constant coefficient vector. We assume both q_n and $k_n = \max_{1\le k\le K} r_k$ diverge with n.

With the above model setting, the parameter set of interest is

$$\boldsymbol{\Omega} = (\boldsymbol{\theta}, \Lambda(t), \mu(t), \beta(t), \boldsymbol{\eta}(t)^T),$$

where the constant parameter $\boldsymbol{\theta} = (\sigma^2, vec(\boldsymbol{\Sigma}_b)) \in \boldsymbol{\Theta} \subset \mathbb{R}^{\dim(\theta)}$, with $vec(\boldsymbol{\Sigma}_b)$ being the distinct elements of $\boldsymbol{\Sigma}_b$. $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard function belongs to the class $\mathcal{V}$ which contains all the non-decreasing functions on $[0, \tau]$, where τ is the end of study time. $\mu(t)$, $\beta(t)$ and each column of $\boldsymbol{\eta}(t) = (\eta_1(t)^T, \dots, \eta_K(t)^T)^T$ are the coefficient functions belonging to the class $\mathcal{C}^r[0, \tau]$ that consists of all the rth order $(r \ge 1)$ continuous differentiable functions on $[0, \tau]$. The parameter set $\boldsymbol{\Omega}$ is defined on the product space of $\boldsymbol{\Theta} \times \mathcal{V} \times \mathcal{C}^r[0, \tau]^{(K+2)}$.

Based on the specified models, the observed joint likelihood can be written out in the form

$$
\begin{aligned}
l_n(\mathbf{\Omega}) = & \sum_{i=1}^{n} \log \int_{\mathbf{b}} \Big[(2\pi\sigma_e^2)^{-N_i/2} \\
& \times \exp\left\{ -\frac{1}{2\sigma_e^2} (\mathbf{W}_i - \boldsymbol{\mu} - \mathbf{B}^{(b)T}\mathbf{b})^T (\mathbf{W}_i - \boldsymbol{\mu} - \mathbf{B}^{(b)T}\mathbf{b}) \right\} \\
& \times \Lambda\{V_i\}^{\Delta_i} \exp\left\{ \Delta_i \left(\beta(V_i)X_i(V_i) + \boldsymbol{\eta}(V_i)^T \mathbf{Z}_i \right) - \int_0^{V_i} e^{\beta(t)X_i(t) + \boldsymbol{\eta}(t)^T \mathbf{Z}_i} d\Lambda(t) \right\} \\
& \times (\sqrt{2\pi})^{-d} |\mathbf{\Sigma}_b|^{-1/2} \exp\{ -\frac{1}{2} \mathbf{b}^T \mathbf{\Sigma}_b^{-1} \mathbf{b} \} \Big] \, d\mathbf{b},
\end{aligned}
\tag{6.16}
$$

where $\mathbf{W}_i = (W_i(t_1), \ldots, W_i(t_{N_i}))^T$ is the $N_i \times 1$ vector of observed longitudinal responses for subject i. $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_{N_i}))^T$ is the functional parameter $\mu(t)$ taking value at the N_i observational time points. $\mathbf{B}^{(b)} = (\mathbf{B}^{(b)}(t_1), \ldots, \mathbf{B}^{(b)}(t_{N_i}))$ is the $d \times N_i$ matrix of the B-spline basis used to express $f_i(t)$ at all the N_i time points.

The maximum likelihood estimator, denoted by $\hat{\mathbf{\Omega}} = (\hat{\boldsymbol{\theta}}, \hat{\Lambda}(t), \hat{\mu}(t), \hat{\beta}(t), \hat{\boldsymbol{\eta}}(t)^T)$, maximizes the observed joint likelihood over the space $\mathbf{\Theta} \times \mathcal{V}_n \times \mathcal{C}_n[0,\tau]^{(k+1)}$, where $\mathcal{V}_n$ consists of all the right-continuous step functions with positive jumps only at V_i for which $\Delta_i = 1$, and $\mathcal{C}_n[0,\tau]$ is the spline space that contains all the continuous functions on $[0,\tau]$ that can be expressed as a linear combination of B-spline basis, and the dimensional of the space is related to n.

Note that the likelihood function (6.16) is almost the same as the one derived in Chapter 4 for the parametric joint modeling except the functional parameters $(\mu(t), \beta(t), \boldsymbol{\eta}(t))$. Thus we will establish a similar set of asymptotic properties, but including the functional parameters. It is important to establish the consistency property. Ideally, we would expect that under some regularity conditions, the maximum likelihood estimator $\hat{\Omega}$ converges to the true parameter Ω_0, where the convergence of $\Lambda(t)$ will be achieved under the superior norm, and other parameter functions under L_2 norm. That is,

$$
\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| + \|\hat{\mu}(t) - \mu(t)\| + \|\hat{\beta}(t) - \beta(t)\| + \sum_{k=1}^{K} \|\hat{\eta}_k(t) - \eta_k(t)\| \to 0 \ a.s.
$$

And we would expect the rate of convergence for each terms with L_2 norm as follows

$$
\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\sqrt{1/n}),
$$

$$\|\hat{\mu}(t) - \mu(t)\| = O_p(\sqrt{p_n/n}),$$
$$\|\hat{\beta}(t) - \beta(t)\| = O_p(\sqrt{q_n/n}),$$
$$\|\hat{\eta}_k(t) - \eta_k(t)\| = O_p(\sqrt{k_n/n}).$$

The preliminary thoughts on the proof outline of the above consistency property are presented in the following paragraphs. Following the proof of Chapter 4, the above consistency theory can be proved by verifying the following steps:

(i) For each n, the maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{\mu}, \hat{\beta}, \hat{\boldsymbol{\eta}})$ exists.

(ii) When n goes to infinity, $\hat{\Lambda}(\tau)$ is bounded with probability one.

(iii) When n goes to infinity, $\hat{\mu}(t), \hat{\beta}(t), \hat{\eta}_k(t)$ are uniformly bounded and equicontinuous on $[0, \tau]$ with probability one.

(iv) With the assumption that $\boldsymbol{\Theta}$ is compact, there is a convergent subsequence of $\hat{\boldsymbol{\theta}}$, denoted to be $\boldsymbol{\theta}^*$. Similarly, since the functional space of $\hat{\Lambda}(t)$ is compact by (ii), by Helly's selection theorem, there is a weakly convergent subsequence, and we denote the limit function as $\Lambda^*(t) \in \mathcal{V}$. Since the functional space of $\hat{\mu}(t), \hat{\beta}(t), \hat{\eta}_k(t)$ is also compact by (iii), according to Arzela-Ascoli theorem, there exist uniformly convergent subsequences, and we denote the limit functions as $\mu^*(t), \beta^*(t), \eta_k^*(t)$, all belong to $\mathcal{C}^r[0, \tau]$.
We claim that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, $\Lambda^*(t) = \Lambda_0(t)$, $\mu^*(t) = \mu_0(t)$, $\beta^*(t) = \beta_0(t)$ and $\eta_k^*(t) = \eta_0(t), k = 1, \ldots, K$.

Due to the involvement of the functional parameters, the proof of this consistency theory is more difficult than that of the parametric joint models in Chapter 4. The challenges are twofold. First, we need step (iii) to establish the uniform convergence of the three functional parameters. However, it is not easy to verify the uniform boundedness and equicontinuity of $\hat{\mu}(t), \hat{\beta}(t), \hat{\eta}_k(t)$ as n goes to infinity. Second, similar to the proof in Chapter 4, we need to verify that the class

$$\mathcal{F} = \{Q(s, \mathbf{O}; \boldsymbol{\theta}, \Lambda, \mu, \beta, \boldsymbol{\eta}) : s \in [0, \tau], \Lambda(t) \in \mathcal{A}, \mu(t), \beta(t), \eta_k(t) \in C^r[0, 2]\}$$

is P-Donsker. To achieve this, the main task is to prove the likelihood function is Lipchitz continuous with respect to $\mu(t), \beta(t), \eta_k(t)$. This can also be challenging due to the complexity form of the likelihood. All of these challenges remain to be solved in the future work.

6.2.2 Completion and extension of joint modeling with categorical longitudinal covariates

In section 6.1, we proposed the joint modeling with categorical longitudinal covariates. Although the model settings were specified and a few simple numerical studies were conducted to test the feasibility of the proposed EM-DoIt algorithm, this project is far from completion. In the future work, we plan to refine the current study both theoretically and numerically. We will conduct more complicated simulation examples and carry out real data analysis. In the smoking cessation study we discussed in the previous chapters, the researchers are also interested in whether some specific discrete longitudinal covariates are associated with the risk of relapse. For example, the number of cigarets smoked since the last prompt, or whether there is any stressful event happened since the last prompt. We can apply our proposed methodology to address such research questions. In addition, it would be ideal to establish the asymptotic theories for the maximum likelihood estimators obtained from the estimation.

The model framework in section 6.1 can be further extended in two directions. First, one may consider building the a joint model with multiple discrete longitudinal covariates. This is similar to the case of continuous longitudinal covariates in Chapter 4. Moreover, the joint model with both the continuous and the discrete longitudinal covariates can also be considered, but the correlations among the different types of the longitudinal covariates need to be carefully specified. Second, following the idea of Chapter 5, one may think of extending the parametric joint model setting in section 6.1 to a nonparameteric joint model with discrete longitudinal covariates, where the mean processes of the longitudinal covariates are allowed to have irregular shapes, and the survival coefficients are allowed to vary with time. Such models, thought sophisticated, would be more computational challenging.

Bibliography

Albert, P. S. and J. H. Shih (2010). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics 4*(3), 1517.

Bollen, K. A. and P. J. Curran (2006). *Latent curve models: A structural equation perspective*, Volume 467. Wiley. com.

Bolt, D. M., M. E. Piper, W. E. Theobald, and T. B. Baker (2012). Why two smoking cessation agents work better than one: role of craving suppression. *Journal of consulting and clinical psychology 80*(1), 54.

Brown, E. and J. Ibrahim (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics 59*(2), 221–228.

Brown, E. R., J. G. Ibrahim, and V. DeGruttola (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics 61*(1), 64–73.

Bycott, P. and J. Taylor (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in Medicine 17*(18), 2061–2077.

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

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*(451), 941–956.

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in cox regression models. *Scandinavian Journal of Statistics 30*(1), 93–111.

Chen, R. and R. S. Tsay (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association 88*(421), 298–308.

Chi, Y.-Y. and J. G. Ibrahim (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics 62*(2), 432–445.

Cowling, B., J. Hutton, and J. Shaw (2006). Joint modeling of event counts and survival times. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(1), 31–39.

Cox, D. D. R. and D. Oakes (1984). *Analysis of Survival Data*, Volume 21. CRC Press.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Dafni, U. G. and A. A. Tsiatis (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics 54*(4), 1445–1462.

De Gruttola, V. and X. M. Tu (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics 50*(4), 1003–1014.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of longitudinal data*. Oxford University Press.

Ding, J. and J.-L. Wang (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics 64*(2), 546–556.

Dziak, J. J., R. Li, and A. Qu (2008). An overview on quadratic inference function approaches for longitudinal data. In X. L. J. Fan and J. Liu (Eds.), *New Developments in Biostatistics and Bioinformatics*, pp. 49–72. World Scientific Publishing Co. Singapore and Higher Education Press, Beijing China.

Eubank, R., C. Huang, Y. M. Maldonado, N. Wang, S. Wang, and R. Buchanan (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), 653–667.

Fan, J., T. Huang, and R. Li (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association 102*(478), 632–641.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*(467), 710–723.

Fan, J. and R. Li (2006). An overview on nonparametric and semi parametric techniques for longitudinal data. In J. Fan and H. Koul (Eds.), *Frontiers of Statistics*, pp. 277–304. London: Imperial College Press.

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

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Fang, K.-T. and Y. Wang (1994). *Number theoretic methods in statistics*, Volume 51. CRC Press.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics 39*(3), 254–261.

Faucett, C. L., N. Schenker, and R. M. Elashoff (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association 93*(442), 427–437.

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

Flury, B. and A. Zoppè (2000). Exercises in em. *The American Statistician 54*(3), 207–209.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association 87*(420), 942–951.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(4), 757–796.

Hatfield, L. A., M. E. Boye, and B. P. Carlin (2011). Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of Biopharmaceutical Statistics 21*(5), 971–991.

Hedeker, D. and R. D. Gibbons (2006). *Longitudinal Data Analysis*, Volume 451. John Wiley & Sons.

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

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

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

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*(3), 763–788.

Huang, W., S. L. Zeger, J. C. Anthony, and E. Garrett (2001). Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association 96*(455), 906–914.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica 14*(3), 863–884.

Joseph, V. R. (2012). Bayesian computation using design of experiments-based interpolation technique. *Technometrics 54*(3), 209–225.

Li, R. and J. Ren (2011). An overview on joint modeling of censored survival time and longitudinal data. In T. Cai and X. Shen (Eds.), *Analysis of High-Dimensional Data*, Volume 2, pp. 195–220. Beijing: Higher Education Press.

Lin, X. and R. J. Carroll (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.

Lin, X. and R. J. Carroll (2001a). Semiparametric regression for clustered data. *Biometrika 88*(4), 1179–1185.

Lin, X. and R. J. Carroll (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association 96*(455), 1045–1056.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(2), 381–400.

Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine 28*(6), 972–986.

Loeppky, J. L., J. Sacks, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*(4), 366–376.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 226–233.

Murphy, S. A. and P. K. Sen (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications 39*(1), 153–180.

Piper, M. E., S. S. Smith, T. R. Schlam, M. C. Fiore, D. E. Jorenby, D. Fraser, and T. B. Baker (2009). A randomized placebo-controlled clinical trial of 5 smoking cessation pharmacotherapies. *Archives of General Psychiatry 66*(11), 1253–1262.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika 69*(2), 331–342.

Qu, A. and R. Li (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics 62*(2), 379–391.

Qu, A., B. G. Lindsay, and B. Li (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika 87*(4), 823–836.

Raudenbush, S. W. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Volume 1. Sage.

Rice, J. A. and C. O. Wu (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics 57*(1), 253–259.

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

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-event Data: With Applications in R*, Volume 6. CRC Press.

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

Rizopoulos, D., G. Verbeke, E. Lesaffre, and Y. Vanrenterghem (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics 64*(2), 611–619.

Shiffman, S., A. A. Stone, and M. R. Hufford (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology 4*, 1–32.

Song, X., M. Davidian, and A. A. Tsiatis (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics 3*(4), 511–528.

Song, X., M. Davidian, and A. A. Tsiatis (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics 58*(4), 742–753.

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

Stefanski, L. A. and R. J. Carroll (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika 74*(4), 703–716.

Sun, Y., R. Sundaram, and Y. Zhao (2009). Empirical likelihood inference for the cox model with time-dependent coefficients via local partial likelihood. *Scandinavian Journal of Statistics 36*(3), 444–462.

Taylor, J. M., W. Cumberland, and J. Sy (1994). A stochastic model for analysis of longitudinal aids data. *Journal of the American Statistical Association 89*(427), 727–736.

Tian, L., D. Zucker, and L. Wei (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association 100*(469), 172–183.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Tierney, L., R. E. Kass, and J. B. Kadane (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association 84*(407), 710–716.

Tseng, Y.-K., F. Hsieh, and J.-L. Wang (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika 92*(3), 587–603.

Tsiatis, A., V. Degruttola, and M. Wulfsohn (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.

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

Tsiatis, A. A. and M. Davidian (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica 14*(3), 809–834.

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence*. Springer.

Wang, C. (2006). Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica 16*(1), 235.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika 90*(1), 43–52.

Wang, Y. and J. M. G. Taylor (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association 96*(455), 895–905.

Wu, C. O., C.-T. Chiang, and D. R. Hoover (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association 93*(444), 1388–1402.

Wu, H. and J.-T. Zhang (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association 97*(459), 883–897.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*(1), 330–339.

Xu, J. and S. L. Zeger (2001a). The evaluation of multiple surrogate endpoints. *Biometrics 57*(1), 81–87.

Xu, J. and S. L. Zeger (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 375–387.

Yu, B. and P. Ghosh (2010). Joint modeling for cognitive trajectory and risk of dementia in the presence of death. *Biometrics 66*(1), 294–300.

Yu, M., N. J. Law, J. M. Taylor, and H. M. Sandler (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica 14*(3), 835–862.

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 42*(1), 121–130.

Zeng, D. and J. Cai (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics 33*(5), 2132–2163.

Zhang, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics 60*(1), 8–15.

Zucker, D. M. and A. F. Karr (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics 18*(1), 329–353.

XIAOYU LIU

Vita

610 Toftrees Ave, Apt 354
State College, PA 16803
(814) 954 - 2129
vera.xy.liu@gmail.com

EDUCATION

Pennsylvania State University, University Park, PA, **Ph.D., Statistics**, 2014
Dissertation: Joint Modeling of Longitudinal and Survival Data
Advisor: Dr. Runze Li

Fudan University, Shanghai, P.R.China, **B.S., Statistics**, 2010

EMPLOYMENT

Research Assistant, the Methodology Center, Pennsylvania State University 2013-2014
Supervisers: Dr. Runze Li, Dr. Donna L. Coffman

Teaching Assistant, Department of Statisitcs, Pennsylvania State University 2011-2013

Statistician Summer Intern, Capital One Financial Corporation, Richmond, VA 2013 summer

PUBLICATIONS

Liu, Xiaoyu, Runze Li, Stephanie T. Lanza, Sara A. Vasilenko, and Megan Piper. "Understanding the role of cessation fatigue in the smoking cessation process." *Drug and alcohol dependence* 133, no. 2 (2013): 548-555.

Lanza, Stephanie T., Sara A. Vasilenko, **Xiaoyu Liu**, Runze Li, and Megan E. Piper. "Advancing the understanding of craving during smoking cessation attempts: A demonstration of the time-varying effect model." *Nicotine & Tobacco Research* (2013): ntt128.

Vasilenko, Sara A., Megan E. Piper, Stephanie T. Lanza, **Xiaoyu Liu**, Jingyun Yang, and Runze Li. "Time-Varying Processes Involved in Smoking Lapse in a Randomized Trial of Smoking Cessation Therapies." *Nicotine & Tobacco Research* 16, no. Suppl 2 (2014): S135-S143.

Coffman, Donna L., Bethany C. Bray, John J. Dziak, **Xiaoyu Liu**, Stephanie T. Lanza. "Estimating Average Causal Effects of Latent Class Treatments on Binary, Count, and Continuous Outcomes". Submitted.

HONORS AND AWARDS

Traveling Awards for SRNT Pre-conference Workshop for Early Career Scientists, NIDA, 2012.

Distinguished University Graduate Fellowship, Penn State University, 2010-2011.

Outstanding Undergraduate Scholarship, Fudan University, 2009-2010.

Peoples Scholarship, Fudan University, 2006-2009.