JOINT MODELING OF LONGITUDINAL DATA AND TIME-TO-EVENT DATA

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2022

Jijun Zhu
School of Mathematics

Contents

Abstract **7**

Declaration **8**

Copyright Statement **9**

Acknowledgements **10**

1 Introduction **12**

1.1 Background 12

1.2 Methods for Longitudinal and Time-to-Event Data 13

1.2.1 Analysis for time-to-event data 14

1.2.2 Analysis for longitudinal data 16

1.2.3 Analysis for longitudinal and time-to-event data 17

1.3 FMCox PH Models 18

1.3.1 Model definition 18

1.3.2 Cox-assisted clustering algorithm 20

1.4 Structure of the Thesis 22

2 Conditional Score Method for FMCox PH Model with Multiple Longitudinal Covariates Measured with Error **24**

2.1 Notations and models 25

2.2 Inference procedure and large sample property 27

2.2.1 EM-Algorithm with conditional score method 27

2.2.2 large-sample property 33

2.3 Simulation 36

2.3.1 Simulation study scenario 1 36

2.3.2 Simulation study scenario 2 41
2.4 Real analysis 45
2.5 Discussion 48

3 FMCox PH Model with Time-varying Coefficients 50
3.1 The local partial likelihood technique 51
3.2 Notations and models 52
3.3 Estimation procedure and asymptotic results 54
3.3.1 Proposed algorithm 54
3.3.2 Estimation of the baseline hazard function 56
3.3.3 Concavity of the local linear partial likelihood 57
3.3.4 Asymptotic theory 58
3.4 Simulation 66
3.4.1 Simulation study scenario 1 66
3.4.2 Simulation study scenario 2 69
3.5 Real data analysis 71
3.6 Discussion 73

4 Conclusion and Discussion 75

Word count

List of Tables

2.1 estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 1, with the mean of the time-independent covariate $\mu = 0$. 39

2.2 estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 1, with the mean of the time-independent covariate $\mu = 5$. 39

2.3 Comparison among the estimation results for FMCox PH models, proposed joint FMCox PH models and the oracle estimators, with the mean of the time-independent covariate $\mu = 0$ 40

2.4 Comparison among the estimation results for FMCox PH models, proposed joint FMCox PH models and the oracle estimators, with the mean of the time-independent covariate $\mu = 5$ 40

2.5 estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 2. 43

2.6 Comparison between the estimation results for the proposed joint FM-Cox PH models and the oracle estimators in the scenario 2. 44

2.7 Estimation results for the AIDS study. 47

3.1 Simulation results for MAD based on 100 replicates for scenario 1 . . . 67

3.2 Simulation results for MAD based on 100 replicates for scenario 2 . . . 71

List of Figures

2.1 20 samples of the longitudinal processes in simulation study scenario 1 37

2.2 the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 1, with the mean of the time-independent covariate $\mu = 0$. 38

2.3 the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 1, with the mean of the time-independent covariate $\mu = 5$. 38

2.4 50 samples of the longitudinal processes in simulation study scenario 2 41

2.5 the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 2. 42

2.6 20 samples of the log CD4 counts processes in AIDS study. 46

2.7 the Kaplan-Meier survival curves for the two latent classes in the AIDS study. 48

3.1 the Kaplan-Meier survival curves for the given latent classes in the simplest simulation study with the mean of the time-independent covariate $\mu = 0$. 67

3.2 estimation of $\eta_1(t)$ for the simulated models 68

3.3 estimation of $\eta_2(t)$ for the simulated models 68

3.4 the Kaplan-Meier survival curves for the given latent classes in the given scenario with the mean of the time-independent covariate $\mu = -0.1$. . . 69

3.5 estimation of $\eta_1(t)$ for the simulated models 70

3.6 estimation of $\eta_2(t)$ for the simulated models 70

3.7 estimation of $\eta_1(t)$ for AIDS data 72

3.8 estimation of $\eta_2(t)$ for AIDS data 72

3.9 the Kaplan-Meier survival curves for the two latent classes in the AIDS study. 73

The University of Manchester

Jijun Zhu
Doctor of Philosophy
Joint Modeling of Longitudinal Data and Time-to-Event Data
January 12, 2022

In medical studies, most common diseases including cancer are heterogeneous that they vary in etiology, pathogensis, and prognosis, which we have limited knowledge. Some longitudinal biomarkers include important information from the past history and provide feedback to the future events. It is frequent to collect both repeated measures of longitudinal processes and the time to an event of interest simultaneously. The existing literature considers heterogeneity of survival data analysis under the FMCox PH models, joint analysis of longitudinal and time-to-event data under the standard joint models. In Chapter 2, the joint FMCox PH models and the corresponding estimation procedures have been proposed to deal with the longitudinal and time-to-event data analysis, considering the heterogeneity. The consistency of the proposed estimators has also been proved in Chapter 2. In Chapter 3, we further develop FMCox PH models with time-varying coefficients, which could explore the time-efficient associations of the covariates of interest. The local partial likelihood technique has been reviewed. This approach and one-step method have been used in the estimation procedures for the proposed models in Chapter 3. And the asymptotic results for the estimators have also been provided in Chapter 3. For these two projects, two simulation scenarios have been provided, the first simulation scenario focuses on comparing the performance of FMCox PH models and the proposed models in this thesis, the second simulation scenario explains the performance of the proposed models in more general cases. And the proposed joint FMCox PH models and the FMCox PH models with time-varying coefficients have also been used for the AIDS study analysis.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Jianxin Pan, for his excellent guidance, caring, patience, continuous support and encouragement throughout my PhD study and research. I met Prof. Pan during my master study period, he taught me several important statistical courses and supervised my master's thesis, which made me interested in statistics and decided to apply for the further study. With his academic expertise, Prof. Pan has offered me valuable guidance on my research work and academic career development. He taught me the principle of good statistical analysis - to solve the practical problems with the simplest yet effective statistical methods, which exerts a strong influence on my PhD research and on my view towards statistics. I would also like to express my appreciation to Dr. Christiana Charalambous and Prof. Korbinian Strimmer, who were my examiners of PhD annual progress report and have provided constructive comments and advice on the content of my thesis. And I really thank my external examiner Prof. Dankmar Bohning and internal examiner Dr. Yang Han for their specific, constructive and instructive comments which will make my thesis more rigorous.

I have an unforgettable study experience in the University of Manchester. I would like to thank the Faculty of Science and Engineering for the Department of Mathematics Scholarship which covers the tuition fee during my PhD. And I also want to thank the China Scholarship Council which provides me the maintenance for the last three years.

Thanks my dear office mates: Jiaming Shen, Chunyu Wang, Peng Zheng, Ruoxuan Zhang and other PhD students for their academic support and valuable discussion, they are not only my colleagues but also my dear friends.

I would like to express my sincere thanks to all my friends, thanks for encouraging me all the time, as well as the company and communication when I was suffering

anxiety. Firstly, thanks Xiaoxiao Yu, Yueying Chang, Xiaochen Zhang, Jiyuan Liu and Da Guo who have study experience in the University of Manchester for taking good care of me when we were staying together and we truly had a lot of wonderful moments which left me an indelible memory. I also want to thank the friends who live and study in other countries: Mingming Zhao, Shuxian Xu, Xiaoyan Wang and Xiangya Cao, they are my best friends and we met each other during my undergraduate study period, thanks for their support in academic and spirit. As for other friends who live in China and from all walks of life: Wenwen Li, Zimin Ke and Jingyuan Su thanks for your patience when I was fallen into depression and many support in other fields.

I would like to express my special thanks to Ke Chen and Danni Zheng, who are my idols but also my special friends. I met them in the last year of my PhD, I was suffering from depression. They bring me a lot of happiness and encouragement.

Last but not the least, I would like to thank my parents Yuzhong Zhu and Hong Zhou for their forever and unconditional love and support. Without their encouragement, deepest care and love, I could not complete my PhD. Everyday I love them more, today more than yesterday.

Chapter 1

Introduction

1.1 Background

It is common to collect both longitudinal data and life time data from the same medical study. These two types of data are often associated with each other in some ways. Practical interests include the trajectory of longitudinal biomarkers and the failure time process, together with their interrelationships. For example, trajectory of CD4 counts and time to AIDS are always collected simultaneously, and in studies of prostate cancer trajectory of Prostate Specific Antigen (PSA) and time to disease recurrence are always obtained at the same time. Wu et al. (2012) proposed several situations where the joint analysis of longitudinal and time-to-event data are needed:

1. survival data analysis including time-dependent covariates with measurement error (Wulfsohn and Tsiatis, 1997);

2. longitudinal data analysis including informative dropout (Dupuy and Mesbah, 2002);

3. the longitudinal process and event time process are linked through latent components (Henderson et al., 2000).

For these different issues, the main objective and corresponding approaches are different. Particular emphasis is placed on the first situation in this thesis.

Most common diseases including cancer are heterogeneous that they vary in etiology, pathogenesis, and prognosis, which we have limited knowledge. For example, the

patients may have different prognoses to the same treatment, but we usually do not know why some patients are more responsive to a certain treatment than the others. In the medical research, a highly plasusible and widely accepted reason is the existence of various subtypes of a perceived same disease (Curtis et al., 2012; Koboldt et al., 2012; Schlicker et al., 2012).

Therefore, the risk factors and treatment responses may differ among different disease subtypes, identifying disease subtypes and applying the certain treatments are important objectives in precision medicine. For example, Roustaei et al. (2018) proposed an approach to study the latent heterogeneous problem in CPCRA study, compared with joint latent class model (JLCM) proposed by Liu et al. (2015) and separate approach. In the clinical trial from the Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) study (Abrams et al., 1994), 467 patients infected with the human immunodeficiency virus (HIV) were randomized into treatment daddanoisene (ddI) and zalcitabines (ddC). In this study, repeated measures of CD4 cell counts and the time of infection or death were recorded at the same time. The main objective is estimating the association between CD4 count and death rate, and whether the given treatment have different effects on the patients in different latent classes. And this real data will be analyzed in this thesis as well.

In the following sections, methods for longitudinal and time-to-event data and FMCox PH models are introduced. Evaluating the parameters which link the time-independent covariates, biomarker measures and event time is of primary interest.

1.2 Methods for Longitudinal and Time-to-Event Data

For the first situation mentioned before, the longitudinal data are observed as time-dependent covariates with measurement errors in the event time process. Various approaches have been proposed e.g. likelihood approach (Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997), conditional score method (Tsiatis and Davidian, 2001) and corrected score method (Wang, 2006). As for the second situation, the dependency of dropout and censoring on the longitudinal response need to be properly addressed in the longitudinal process. And the shared parameter model (Vonesh et al., 2006) is

the most useful tool for the last situation in which the potential association between the longitudinal and time-to-event process can be fully characterized. This section covers the notations, definition, important concepts and some properties for the first situation, laying the groundwork for the next sections.

1.2.1 Analysis for time-to-event data

Assuming that the event time $\tilde{T}$ is continuous, the survival function which is primarily used to describe the distribution of $\tilde{T}$, is defined as

$$\mathcal{S}(t) = \Pr(\tilde{T} > t) = \int_t^\infty f(s)ds,$$

where $f(\cdot)$ denotes the corresponding probability density function. Another important function in survival analysis is the hazard function, which describes the instantaneous risk for an event in the time interval $[t, t + dt)$ and is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq \tilde{T} < t + dt \mid \tilde{T} \geq t)}{dt}, \quad t > 0,$$

which can be written as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq \tilde{T} < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

The survival function can be expressed in terms of the risk function as

$$\mathcal{S}(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(s)ds\right\}, \tag{1.1}$$

where $\Lambda(\cdot)$ is the cumulative hazard function which describes the accumulated risk up until time t.

Ideally, the true event time is known, in this case all the information is contained in the event time $\tilde{T}$. But in practice, censoring must be taken into account. For each subject i, let $T_i = min(\tilde{T}_i, C_i)$ be the observed event time, with $\tilde{T}_i$ being the true event time and C_i being the right-censored event time. $\delta_i = I(\tilde{T}_i \leq C_i)$ indicates whether or not the observed time T_i is censored.

In general, estimation of the distribution of $\tilde{T}$ using the available information $\{\tilde{T}_i, C_i\}$ is the main objective in survival analysis. The most well-known nonparametric estimator of survival function was proposed by Kaplan and Meier (1958), which does

not need any assumptions for the underlying distribution of the event time. Another nonparametric estimator named the Nelson-Aalen estimator ((Aalen, 1976; Altshuler, 1970; Fleming and Harrington, 2011; Nelson, 1972)) is used for the cumulative hazard function.

Kaplan-Meier estimator

Let $t_1 < t_2 < ... < t_{k-1} < t_k$ denote the observed event times in one sample. The probability of survival time could be written as the product of the conditional probabilities, using the law of total probability:

$$\Pr\left(\tilde{T} > t\right) = \Pr\left(\tilde{T} > t \mid \tilde{T} > t_k\right) \times \Pr\left(\tilde{T} > t_k \mid \tilde{T} > t_{k-1}\right) \times \dots$$

Using this expansion and accounting for the censoring, the survival probability at the unique event time could be obtained:

$$\hat{\mathcal{S}}_{KM}(t) = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i}, \tag{1.2}$$

where r_i is the number of subjects at risk at t_i and d_i is the number of events at t_i. The variance of $\hat{\mathcal{S}}_{KM}(t)$ could be calculated by Greenwood's formula ((Greenwood et al., 1926; Kalbfleisch and Prentice, 2011)). And a better approach is to calculate a confidence interval for $\log \Lambda(t)$ so that the confidence interval for $\mathcal{S}(t)$ could be proposed.

Nelson-Aalen estimator

Using the same notation, the similar nonparametric estimator for the cumulative hazard function was proposed:

$$\hat{\Lambda}_{NA}(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i}. \tag{1.3}$$

Breslow (1972) suggested the following estimator for the survival function based on the relation (1.1) and Nelson-Aalen estimator:

$$\hat{\mathcal{S}}_B(t) = \exp\left\{-\hat{\Lambda}_{NA}(t)\right\} = \prod_{i:t_i \leq t} \exp\left(-d_i/r_i\right). \tag{1.4}$$

These two nonparametric estimators of the survival function are asymptotically equivalent.

Cox PH models

Although the nonparametric estimators introduced before have provided robust statistical inference for the time-to-event data, the event time usually depends on some covariates of interest in practice. Cox (1972) proposed the proportional hazards models (Cox PH models) in the modern survival analysis, which assumes that covariates have effect on the hazard for an unique event:

$$\lambda_i(t) = h_0(t) \exp\left(\boldsymbol{\eta}^T \boldsymbol{X}_i\right), \tag{1.5}$$

where $\boldsymbol{X}_i = (x_{i1}, ..., x_{ip})^T$ denotes the $p-$vector of covariates which are assumed to be associated with the hazard of each subject, $\boldsymbol{\eta}$ is the corresponding regression coefficients and $h_0(t)$ is called baseline hazard function.

From the PH Cox model (1.5), the assumptions of the distribution of the event time $\tilde{T}$ are included in the specification of the baseline hazard function, e.g. the baseline hazard function is $h_0(t) = \phi\sigma_t t^{\sigma_t - 1}$ if $\tilde{T}$ follows the Weibull distribution. If the baseline hazard function has a specific assumption, the estimation of all the parameters θ in the given model (1.5) could be obtained by maximizing the corresponding log-likelihood function:

$$\ell(\theta) = \sum_{i=1}^{n} \left[\delta_i \log f\left(T_i; \theta\right) + (1 - \delta_i) \log \mathcal{S}_i\left(T_i; \theta\right)\right], \tag{1.6}$$

which could be rewritten in terms of the hazard function, using the relation (1.1):

$$\ell(\theta) = \sum_{i=1}^{n} \left[\delta_i \log \lambda_i\left(T_i; \theta\right) - \int_0^{T_i} \lambda_i(s; \theta) ds\right]. \tag{1.7}$$

However, Cox (1972) proposed another estimation method for the parameters of interest η, which did not require the specific assumption of the baseline hazard function $h_0\cdot$, that is, without specifying the distribution of event time $\tilde{T}$. In this case, the semiparametric model (1.5) has been widely used, and the parameters η could be estimated by maximizing the partial log-likelihood function:

$$p\ell(\eta) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\eta}^\top \boldsymbol{x}_i - \log\left\{\sum_{T_j \geq T_i} \exp\left(\boldsymbol{\eta}^\top \boldsymbol{x}_j\right)\right\}\right]. \tag{1.8}$$

1.2.2 Analysis for longitudinal data

As described above, estimating the parameters which characterize the association between the event occurrences and covariates of interest is usually the main objective in

time-to-event data analysis. The methods introduced in section 1.2.1 are used in the case just considering time-independent covariates, e.g. gender or medical status measured at baseline observation time. As for the case which includes the time-dependent covariates, the observations of the covariates are always collected intermittently.

If the repeated measurement of the time-dependent covariates are available, the observations are appropriate to be used in the analysis of time-to-event data (Fisher and Lin, 1999). There are two common situations used for this problem. In one situation, the event time is associated with the current values of the time-dependent covariates, which could be assessed by the longitudinal biomarker (Prentice, 1989). In another situation, the event time is associated with the time-dependent covariates just based on the random effects which are used for characterizing the longitudinal data (Tsiatis and Davidian, 2001; Wang, 2006). In order to characterize the trajectory of the longitudinal covariate, linear mixed effects models are most useful.

For each subject i, let $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$ denote the response vector, $\boldsymbol{X}_i$ denote the design matrices for fixed effects and $\boldsymbol{Z}_i$ denote the design matrices for random effects. Assuming $\boldsymbol{\beta}$ as the coefficients for the fixed effects and $\boldsymbol{\alpha}_i$ as the random effects for each subject i, the linear mixed effects model was proposed by Laird and Ware (1982):

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, m,$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^{\mathrm{T}}$ are usually assumed to follow multivariate normal distribution $N(0, \Sigma_i)$, and is independent of the random effects $\boldsymbol{\alpha}_i$.

1.2.3 Analysis for longitudinal and time-to-event data

Based on the linear mixed-effects models and Cox PH models presented in the previous two sections, the joint modeling framework for longitudinal and time-to-event data (Faucett and Thomas, 1996; Henderson et al., 2000; Tsiatis and Davidian, 2004; Wulfsohn and Tsiatis, 1997) was proposed to characterize the association between the biomarkers and event time.

Let $Z_i(t)$ be the true and unobserved measurement of longitudinal biomarker at time t. The Cox PH model with time-independent and time-dependent covariates could be written as:

$$\lambda_i(t) = h_0(t) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{x}_i + \gamma Z_i(t)\right\}, \tag{1.9}$$

where γ quantifies the strength of the association between the biomarker and the risk of the event and $\boldsymbol{x}_i$ is the vector of baseline covariates.

Therefore, to measure the effect of the longitudinal covariate to the event time, the unobserved covariate $Z_i(t)$ needs to be estimated through the observed longitudinal data Y_i, using the linear mixed effects model:

$$\begin{aligned} Y_i(t) &= Z_i(t) + \epsilon_i(t) \\ &= \boldsymbol{x}_i^T(t)\boldsymbol{\beta} + \boldsymbol{z}_i^T(t)\boldsymbol{\alpha}_i + \epsilon_i(t), \quad \epsilon_i(t) \sim N\left(0, \sigma^2\right), \end{aligned} \tag{1.10}$$

where $\boldsymbol{x}_i(t)$ and $\boldsymbol{\beta}$ denote the fixed-effects part, $\boldsymbol{z}_i(t)$ and $\boldsymbol{\alpha}_i$ denote the random-effects part. Assuming that the event time and the longitudinal observations for each subject are independent if the distribution of the random effects $\boldsymbol{\alpha}_i$ is known, these two models are associated through the joint distribution:

$$f\left(y_i, T_i, \delta_i\right) = \int f\left(y_i \mid \boldsymbol{\alpha}_i\right) \left\{\lambda\left(T_i \mid \boldsymbol{\alpha}_i\right)^{\delta_i} S\left(T_i \mid \boldsymbol{\alpha}_i\right)\right\} f\left(\boldsymbol{\alpha}_i\right) d\boldsymbol{\alpha}_i, \tag{1.11}$$

where α_i is the vector of random effects, $T_i = min(\tilde{T}_i, C_i)$ is the observed event time, $\delta_i = I(\tilde{T}_i \leq C_i)$ indicates whether or not the observed time T_i is censored, $f(\cdot)$ is the density function and $S(\cdot)$ is the survival function.

1.3 FMCox PH Models

Finite mixture models are typically used to deal with heterogeneity in many fields by assuming a separate distribution for each sub-class (Fraley and Raftery, 2002; McLachlan and Basford, 1988; Muthén and Masyn, 2005; Qin and Self, 2006; Wedel and DeSarbo, 1995). A finite mixture model is a statistical model that assumes the presence of latent classes, within an overall population. And every latent class could be fit with its own model. It is natural to develop Finite mixture Cox PH model for time-to-event data. Eng and Hanlon (2014) used the EM algorithm and proposed a Cox-assisted clustering algorithm for FMCox PH model. And this mixture relaxes the PH assumption that hazards are proportional within their given clusters.

1.3.1 Model definition

Let (T_i, δ_i, x_i), $i = 1, ..., n$ be an independent right-censored sample, $\boldsymbol{T} = (T_1, ..., T_n)$, $\boldsymbol{\Delta} = (\delta_1, ..., \delta_n)$ and $\boldsymbol{x}_i$ denote the vectors of observed event time, indicators and

regression covariates. To account for heterogeneity, each subject is assumed to arise from the K latent classes with probability π_k, $k = 1, ..., K$, $\sum_k \pi_k = 1$. Therefore, the FMCox PH models are proposed by assuming Cox PH model within each sub-class k, so that the covariates x effects the event time in different latent class k log-linearly via a class-specific hazard:

$$\log \lambda_k(t \mid x) = \log h_{0k}(t) + \boldsymbol{\eta}_k^T \boldsymbol{x}. \tag{1.12}$$

Recall that the right-censored observation following Cox PH model has the density function as

$$f_k(\boldsymbol{T}, \boldsymbol{\delta} \mid \boldsymbol{x}) = \left[h_{0k}(T) \exp\left(\boldsymbol{\eta}_k^T \boldsymbol{x}\right)\right]^{\delta} \exp\left[-H_{0k}(T) \exp\left(\boldsymbol{\eta}_k^T \boldsymbol{x}\right)\right], \tag{1.13}$$

where $h_{0k}(\cdot)$ and $H_{0k}(\cdot)$ are the baseline hazard function and baseline cumulative hazard function for the latent k-th class. Therefore, the joint density function of (T_i, δ_i) for each subject i could be written as

$$f\left(T_i, \delta_i \mid x_i\right) = \sum_{k=1}^{K} \pi_k f_k\left(T_i, \delta_i \mid x_i\right). \tag{1.14}$$

The mixture likelihood may be written as

$$f(T, \Delta \mid x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k\left(T_i, \delta_i \mid x_i\right). \tag{1.15}$$

The log-form of the mixture likelihood based on the observed data is,

$$logf(T, \Delta \mid x) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k\left(T_i, \delta_i \mid x_i\right). \tag{1.16}$$

In order to obtain the MLE estimators of the fixed effects η_k for $k = 1, 2, ..., K$, this observed log-form mixture density function need to be maximized intuitively. But the score function from it is difficult to calculate, which is similar to the case of Guassian mixture distribution. Therefore, the EM-algorithm (Dempster et al., 1977) can be used to find the MLEs of η_k, if the unobserved latent indicator variable c_{ik} according to whether or not the i-th subject comes from the k-th latent class, is known. Assume that the latent class $U = (U_1, U_2, ..., U_n)$, where $U_i \sim$ Multinomial(π), $U_i \in \{1, 2, \ldots, K\}$ and $c_{ik} = 1_{\{U_i = k\}}$, are observed, the complete data likelihood of the complete data $(T_i, \delta_i, x_i, U_i)$ could be written as:

$$f(T, \Delta \mid \boldsymbol{x}, U) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k f_k \left(T_i, \delta_i \mid x_i\right)\right]^{c_{ik}}, \tag{1.17}$$

where $\boldsymbol{U} = (c_{ik})_{n \times K}$ is a $n \times K$ matrix with all elements either 0 or 1. Eng and Hanlon (2014) proposed maximizing this mixture density function via the EM algorithm framework (Dempster et al., 1977) to estimate the regression coefficients η_k and baseline hazard functions $h_{0k}(\cdot)$.

Therefore, the interpretation of the FMCox PH model is organizing observations into clusters which don't have a known prior. This type of clustering should not be confused with the case where the observations are obtained from several known sources e.g. different hospitals. Instead, observations are gathered according to their best-fitting subclass model.

1.3.2 Cox-assisted clustering algorithm

Based on the observed data (T_i, δ_i, x_i) for each subject i, let

i. the parameters of interest as the mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, the baseline hazard functions $\boldsymbol{h} = \{h_{01}(t), \ldots, h_{0K}(t)\}$ and the coefficient vectors $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$;

ii. observed event time $\boldsymbol{T} = (T_1, ..., T_n)$, indicator variables $\boldsymbol{\Delta} = (\delta_1, ..., \delta_n)$ and time-independent covariates vectors $\boldsymbol{x} = (x_1, ..., x_n)$.

If we know the classification parameters $c_{ik} = 1_{\{U_i = k\}}$ i.e. whether the i-th subject comes from the k-th latent class, the mixture likelihood (1.17) from complete data with mixing parameters $\boldsymbol{\pi}$ and class-specific parameters $\boldsymbol{h}$ and $\boldsymbol{\eta}$ could be separated into a mixing distribution part and a component distribution part as

$$\begin{aligned}
\log f(\boldsymbol{\pi}, \boldsymbol{h}, \boldsymbol{\eta}; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{U} \mid \boldsymbol{x}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log f_k \left(T_i, \delta_i \mid x_i\right) \\
&= \log L_1(\boldsymbol{\pi}; \boldsymbol{U}) + \log L_2(\boldsymbol{h}, \boldsymbol{\eta}; \boldsymbol{T}, \boldsymbol{\Delta} \mid \boldsymbol{U}, \boldsymbol{x}),
\end{aligned}$$

where the first part is simply

$$\log L_1(\boldsymbol{\pi}; \boldsymbol{U}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} c_{ik} \right) \log \pi_k, \tag{1.18}$$

and the second part is

$$logL_2(\boldsymbol{h}, \boldsymbol{\eta}; \boldsymbol{T}, \boldsymbol{\Delta} \mid \boldsymbol{U}, \boldsymbol{x}) = \sum_{k=1}^{K} \sum_{i=1}^{n} [\delta_i c_{ik} \log h_{0ki} + \delta_i c_{ik} x_i^T \eta_k - c_{ik} H_{0ki} \exp\left(x_i^T \eta_k\right)]. \tag{1.19}$$

In order to obtain the MLE of the parameters of interest, the EM approach is needed here, which has the main idea that plugging $\hat{c}_{ik} = \mathbb{E}\left(c_{ik} \mid T_i, \delta_i, \boldsymbol{x}\right)$ into the complete data mixture density function (1.17). Supposing that the current values of the parameters of interest are $\pi_k^{(m)}, h_{0ki}^{(m)}, H_{0ki}^{(m)}$ and $\eta_k^{(m)}$ at the m-th iteration. The proposed algorithm proceeds as follow:

E-step the conditional expectation of $c_{ik}^{(m+1)}$ is:

$$\hat{c}_{ik} = \mathbb{E}\left(c_{ik} \mid T_i, \delta_i, \boldsymbol{x}\right) = \frac{\pi_k^{(m)} \left[h_{0ki}^{(m)} \exp\left(x_i^T \eta_k^{(m)}\right)\right]^{\delta_i} \exp\left[-H_{0ki}^{(m)} \exp\left(x_i^T \eta_k^{(m)}\right)\right]}{\sum_{k'} \pi_{k'}^{(m)} \left[h_{0k'i}^{(m)} \exp\left(x_i^T \eta_{k'}^{(m)}\right)\right]^{\delta_i} \exp\left[-H_{0k'i}^{(m)} \exp\left(x_i^T \eta_{k'}^{(m)}\right)\right]}, \tag{1.20}$$

which is calculated according to the application of Bayes rule. Note that the update of the conditional expectation of c_{ik} only depends on the current estimates of π and η, if the baseline hazard functions are assumed to be the same across clusters.

M-step update of the mixing proportions π is straightforward:

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \hat{c}_{ik}}{n}. \tag{1.21}$$

To update the unspecific baseline hazard function h, the profile estimates which are similar to Breslow (1974) have been proposed:

$$h_{0k}^{(m+1)}\left(T_i\right) = \frac{\hat{c}_{ik}}{\sum_{j:T_j \geq T_i} \hat{c}_{jk} \exp\left(x_j^T \eta_k^{(m+1)}\right)}, \tag{1.22}$$

$$H_{0k}^{(m+1)}\left(T_i\right) = \sum_{l:T_l \leq T_i} \frac{\hat{c}_{lk}}{\sum_{j:T_j \geq T_l} \hat{c}_{jk} \exp\left(x_j^T \eta_k^{(m+1)}\right)}. \tag{1.23}$$

Therefore, the profile objective is a partial likelihood weighted by the conditional expectation $\hat{c}_{ik}$:

$$\log L_2(h(\eta), \eta; T, \delta, \hat{U} \mid x) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_i \left\{ \hat{c}_{ik} x_i^T \eta_k - \log \sum_{j:T_j \geq T_i} \exp\left[\hat{c}_{jk} x_j^T \eta_k\right] \right\}. \tag{1.24}$$

Finally, iterates between the E-step and M-step until the increment in log mixture density function is small.

The reasonable initial values for the parameters of interest are important in the EM algorithm, and have effects on the convergence speed. Eng and Hanlon (2014) proposed setting initial value for every classification indicator variable c_{ik} randomly, using multiple starts and picking the best fitting results.

Classification rule

Given K and the corresponding parameters of interest $\boldsymbol{\pi}$, $\boldsymbol{\eta}$ and $\boldsymbol{h}$ obtained from the Cox-assisted clustering algorithm, the sub-population that a sample (T, δ, x) belongs to could be determined by

$$\underset{k}{\operatorname{argmax}} \left\{ f\left(\boldsymbol{\pi}, \boldsymbol{h}, \boldsymbol{\eta}; c_{ik} = 1 \mid (\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{x})\right) \right\},$$

where

$$f\left(\boldsymbol{\pi}, \boldsymbol{h}, \boldsymbol{\eta}; c_{ik} = 1 \mid (\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{x})\right) = \frac{\pi_k f_k\left(T, \delta \mid x, h_{0k}(\cdot), \eta_k\right)}{\sum_{k'=1}^{K} \pi_k f_{k'}\left(T, \delta \mid x, h_{0k'}(\cdot), \eta_{k'}\right)}$$

is the posterior probability that the sample belongs to the k-th sub-class, given the parameters of interest.

1.4 Structure of the Thesis

In Chapter 1, the background of longitudinal and time-to-event data has been introduced, and literature for longitudinal data analysis, survival data analysis and joint modeling of these two outcomes has been reviewed at the same time. However, the latent heterogeneity arise in the survival data analysis and many approaches have been proposed to focus on the association between the event time and some covariates of interest and deal with the effects of the heterogeneity. FMCox PH models are one of the most common models to be used in this case. FMCox PH models and the corresponding estimation procedures also have been introduced in Chapter 1.

Based on the background of analysis for longitudinal and time-to-event data and FMCox PH models, it is natural to notice that the latent heterogeneity arise in the longitudinal and time-to-event data as well. So the joint FMCox PH models have been proposed in Chapter 2, which could estimate the association between both the event

time and the time-independent covariates of interest and the longitudinal processes. Considering the complexity of the standard likelihood technique used for the proposed models, conditional score method has been proposed in the estimation procedures in Chapter 2, which did not need any assumptions on the random effects. The consistency of the proposed estimators has also been proved in Chapter 2. And two simulation scenarios have been provided, the first simulation scenario focuses on comparing the performance of FMCox PH models and the proposed models in this thesis, the second simulation scenario explains the performance of the proposed models in more general cases. At the end of Chapter 2, the proposed joint FMCox PH models have been used for the AIDS study.

In Chapter 3, we further develop FMCox PH models with time-varying coefficients, which could explore the time-efficient associations of the covariates of interest. The local partial likelihood technique has been reviewed. This approach and one-step method have been used in the estimation procedures for the proposed models in Chapter 3. And the asymptotic results for the estimators have also been provided in Chapter 3. Two simulation scenarios have been provided, the first one focuses on the performance compared with FMCox PH models, the second one explains the performance of the proposed models in more general cases. At the end of Chapter 3, the proposed FMCox PH models with time-varying coefficients have also been used for the AIDS data analysis.

In Chapter 4, we give a full discussion of the new models proposed in this thesis, including the practical problems and limitations. Further extension and future work is also addressed.

Chapter 2

Conditional Score Method for FMCox PH Model with Multiple Longitudinal Covariates Measured with Error

In many medical studies, the repeated measures of a biomarker and the event time of interest always are recorded simultaneously, and unobserved heterogeneity will also make effects on the time-to-event data analysis. As been reviewed in chapter 1, the existing literature considers heterogeneity of time-to-event data under the FMCox PH model (Eng and Hanlon, 2014), joint analysis of longitudinal and time-to-event data under the standard joint models (Faucett and Thomas, 1996; Henderson et al., 2000; Tsiatis and Davidian, 2004; Wulfsohn and Tsiatis, 1997). It is natural to extend the FMCox PH model to the case with multiple time-dependent covariates measured with error, where the time-dependent covariates could be regarded as the longitudinal processes or the biomarkers and be modeled based on the longitudinal data analysis. So in this chapter, the FMCOx PH models with both time-independent covariates and time-dependent covariates measured with error (called the joint FMCox PH models in this thesis) and the estimation procedure for the proposed models will be presented.

The rest of this chapter is organized as follows, the definitions and notations of the joint FMCox PH models will be introduced at first in Section 2.1. The proposed joint modeling framework for longitudinal and time-to-event data with latent heterogeneity,

consists of longitudinal sub-model, Cox PH sub-model and latent classification indicator variable. Section 2.2 describes the estimation procedure and the large sample property of the proposed estimators. Simulation study and real data analysis are presented in Section 2.4 and 2.5. Finally, the discussion of this work will be given in Section 2.6.

2.1 Notations and models

Longitudinal sub-model

For each subject i, $i = 1, ..., n$, assume that the L longitudinal covariate process $Z_{il}(u)$, $l = 1, ..., L$, follow the following models,

$$Z_{il}(u) = \boldsymbol{\alpha}_{il}^T \boldsymbol{f}_l(u), \tag{2.1}$$

where $\boldsymbol{f}_l(u)$ is a $(q_l \times 1)$ vector of functions of time u, $\boldsymbol{\alpha}_{il}$ is a $(q_l \times 1)$ vector of random effects, and $\boldsymbol{f}_l(u)$ and $\boldsymbol{\alpha}_{il}$ may be different for each l.

The longitudinal covariate process $Z_{il}(u)$ which will be regarded as the time-dependent covariates in the Cox PH sub-models, are not observed directly. Instead, we have the longitudinal measurements $W_{il}(t_{ilj})$ for the l-th covariate at time t_{ilj}, $j = 1, 2, ..., m_l$, and use the the linear mixed effects models (Laird and Ware, 1982)

$$W_{il}(t_{ilj}) = Z_{il}(t_{ilj}) + e_{ilj}, \tag{2.2}$$

where measurement error e_{ilj} are assumed to be normally distributed with 0 mean, variance σ_{ll}, for $l = 1, ..., L$, $j = 1, ..., m_{il}$. For the different longitudinal measurements $W_{il}(t_{ilj})$ and $W_{il'}(t_{il'j'})$, $l, l' = 1, ..., L$, $j = 1, ..., m_{il}$, $j' = 1, ..., m_{il'}$, let $cov(e_{ilj}, e_{il'j'}) = \sigma_{ll'} I(t_{ilj} = t_{il'j'})$, where $\sigma_{ll'}$ is the covariance between the measurement errors e_{ilj} and $e_{il'j'}$ at the same time point. Otherwise, we assume that all covariance between the measurement errors at different time point equals 0.

Survival sub-model

For each subject i, let $T_i = min(\tilde{T}_i, C_i)$ be the observed event time, with $\tilde{T}_i$ being the unobserved true event time and C_i being the right censored event time. $\delta_i = I(\tilde{T}_i \leq C_i)$

indicates whether or not the observed time T_i is censored. The Cox PH model with time-independent and time-dependent covariates for subject i in the k-th sub-class is:

$$\begin{aligned}\lambda_{ik}(u) &= \lim_{du\to 0} du^{-1} Pr_k\{u \le T_i < u + du | T_i \ge u, \boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{U}_i, \boldsymbol{e}_i(u), \boldsymbol{t}_i(u)\} \\ &= h_{0k}(u) \exp\{\boldsymbol{\gamma}_k^T \boldsymbol{G}(u, \boldsymbol{\alpha}_i) + \boldsymbol{\eta}_k^T \boldsymbol{x}_i\}.\end{aligned} \tag{2.3}$$

Here, $h_{0k}(u)$ is an unspecified baseline hazard function of the k-th sub-class; $\boldsymbol{G}(u, \alpha_i) = \boldsymbol{G}(u)\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_{i1}^T \boldsymbol{f}_1(u), ..., \boldsymbol{\alpha}_{iL}^T \boldsymbol{f}_L(u))^T$ is a $(s \times 1)$ vector whose elements are functions of u and $\boldsymbol{\alpha}_i$; for k-th sub-class $\boldsymbol{\gamma}_k$ and $\boldsymbol{\eta}_k$ are the corresponding $(s \times 1)$ and $(p \times 1)$ effect parameters of time-dependent covariates $\boldsymbol{G}(u, \alpha_i)$ and time-independent covariates $\boldsymbol{Z}_i$, respectively; $\boldsymbol{t}_i(u) = (t_{ilj} \leqslant u; l = 1, \ldots, L)$ denotes the observation times up to and including time u; and $\boldsymbol{e}_i(u) = (e_{ilj} : t_{ilj} \leqslant u; l = 1, \ldots, L)$. Our main interest focuses on estimation of every $\boldsymbol{\gamma}_k$ and $\boldsymbol{\eta}_k$ for the different sub-classes.

Latent variable

Let $\boldsymbol{U} = (c_{ik})_{n\times K}$ be a $n \times K$ matrix with all elements either 0 or 1:

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ belongs to sub-class } k \\ 0 & \text{otherwise} \end{cases},$$

with the mixing probability:

$$\mathbb{P}\{c_{ik} = 1\} = \pi_k \quad \text{and} \quad \sum_k^K \pi_k = 1.$$

Joint FMCox PH models

Based on the observed data $(T_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{W}_i)$, the subject i belongs to sub-class k follows a Cox PH model, and has the following density function, under the conditional independent assumption that the longitudinal observations $\boldsymbol{W}_i$ and observed event time T_i are independent if the distribution if the random effects $\boldsymbol{\alpha}_i$ is known (Tsiatis and Davidian, 2004):

$$f_k(T_i, \delta_i, \boldsymbol{W}_i | \boldsymbol{x}_i) = \int \{\lambda_{ik}(T_i | \boldsymbol{\alpha}_i, \boldsymbol{x}_i)^{\delta_i} \exp[-\Lambda_{ik}(T_i | \boldsymbol{\alpha}_i, \boldsymbol{x}_i)]\} f(\boldsymbol{W}_i | \boldsymbol{\alpha}_i) p(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i. \tag{2.4}$$

Therefore the log-likelihood based for the observed data, which is similar to the FMCox PH models, could be obtained as,

$$l_{obs}(\boldsymbol{\Pi}, \boldsymbol{H}, \boldsymbol{\Theta}; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{W}|\boldsymbol{X}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k \left(T_i, \delta_i, \boldsymbol{W}_i|\boldsymbol{x}_i\right). \tag{2.5}$$

It is difficult to optimize this observed log-likelihood function w.r.t. the parameters $\boldsymbol{\gamma}_k$ and $\boldsymbol{\eta}_k$ for k-th sub-class. The EM algorithm is helpful if we know the unobserved latent indicator random variables c_{ik}. Then the complete log-likelihood is,

$$l_c(\boldsymbol{\Pi}, \boldsymbol{H}, \boldsymbol{\Theta}; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{W}, \boldsymbol{U}|\boldsymbol{X}) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log f_k \left(T_i, \delta_i, \boldsymbol{W}_i|\boldsymbol{x}_i\right). \tag{2.6}$$

2.2 Inference procedure and large sample property

However, since the joint likelihood function may be complicated, a main challenge for the standard likelihood method is computation. Note that the density function (2.4) for k-th sub-class will be a integral w.r.t. the random effects, which makes maximizing the complete log-likelihood (2.6) is also too complex. The estimation procedures of the FMCox PH model (1.12) can not be directly used for the joint FMCox PH model proposed in this chapter. To remove the complication which comes from the latent random effects, the conditional score method (Tsiatis and Davidian, 2001) which does not need any distributional assumption on the random effects,is proposed to replace the partial likelihood weighted estimations in the M-step of the cox-assisted clustering algorithm described in section 1.3.2.

This approach exploits the conditional score idea of standard joint models for longitudinal and time-to-event data Tsiatis and Davidian (2001). And this idea proposed the unbiased estimating equations for the fixed effects in every sub-class are based on treating the random effects $\boldsymbol{\alpha}_i$ as 'nuisance parameters' and conditioned on an appropriate 'sufficient statistic'.

2.2.1 EM-Algorithm with conditional score method

Before introducing the iteration steps in the algorithm, the 'sufficient statistic' which replace the random effects in the calculations need to be proposed at first.

'sufficient statistic'

Assume that $\omega = \{\sigma_{ll'} : l \geq l'\}$ is known. Define the at risk process:

$$Y_i(u) = I\left(T_i \geqslant u, m_{il}(u) \geqslant q_l, l = 1, \ldots, L\right).$$

Conditional on $Y_i(u) = 1, \alpha_i, Z_i, t_i(u)$, define the counting process increment be

$$\mathrm{d}N_i(u) = I\left(u \leqslant T_i < u + \mathrm{d}u, \delta_i = 1, m_{il}(u) \geqslant q_l, l = 1, \ldots, L\right).$$

And for the subject i, which is assumed to be only in one sub-class, the counting process increment $dN_i(u)$ is distributed as Bernouli with probability (for k-th sub-class), based on (2.3):

$$Pr_k(dN_i(u) = 1 | Y_i(u) = 1, c_{ik} = 1, \boldsymbol{\alpha}_i, \boldsymbol{Z}_i, \boldsymbol{t}_i(u)) = h_{0k}(u)\mathrm{d}u \exp\left\{\boldsymbol{\gamma}_k^T \boldsymbol{G}(u)\boldsymbol{\alpha}_i + \boldsymbol{\eta}_k^T \boldsymbol{x}_i\right\}.$$

Therefore we can rewrite density function for subject i in sub-class k (i.e. $(T_i, \delta_i, \boldsymbol{x}_i)$ in sub-class k), if the unobserved random effects $\boldsymbol{\alpha}_i$ are known, the new density function for $(T_i, \delta_i, \boldsymbol{x}_i)$ i.e. $f_k\left(T_i, \delta_i | \boldsymbol{x}_i, \boldsymbol{\alpha}_i\right)$ is:

$$\begin{aligned} &f_k\left(T_i, \delta_i | \boldsymbol{x}_i, \boldsymbol{G}(u, \boldsymbol{\alpha}_i)\right) = \\ &\prod_{\text{all grid points } u} Pr_k\{dN_i(u) = 1\}^{dN_i(u)}[1 - Pr_k\{dN_i(u) = 1\}]^{1 - dN_i(u)}. \end{aligned} \tag{2.7}$$

Let $\hat{\alpha}_{il}(u)$ be the ordinary least-squares estimator of α_{il} based on all the longitudinal data measured before time u for the l-th covariate for subject i.

In order to get $\hat{\boldsymbol{\alpha}}_{il}$, we need at least q_l observations (which is similar to the ordinary least square problem, the rank of the design matrix is q_l,i.e. $m_{il}(u) \geq q_l$ for each l. Note that $m_{il}(u)$ denote the number of time points in $t_{il}(u)$.

Define

$$\boldsymbol{F}_{il} = \begin{pmatrix} \boldsymbol{f}_l^T(t_{il1}) \\ \boldsymbol{f}_l^T(t_{il2}) \\ \ldots \\ \boldsymbol{f}_l^T(t_{ilm_{il}(u)}) \end{pmatrix} \implies \boldsymbol{Z}_{il} = \boldsymbol{F}_{il}\boldsymbol{\alpha}_{il}, \tag{2.8}$$

where $\boldsymbol{f}_l(u)$ is a $(q_l \times 1)$ vector of functions of time u and $\boldsymbol{F}_{il}$ is a $(m_{il}(u) \times q_l)$ design matrix, $\boldsymbol{Z}_{il}$ is a $m_{il}(u)$-vector and $\boldsymbol{\alpha}_{il}$ is a q_l-vector. and we have:

$$\boldsymbol{\alpha}_i = \begin{pmatrix} \boldsymbol{\alpha}_{i1} \\ \boldsymbol{\alpha}_{i2} \\ \ldots \\ \boldsymbol{\alpha}_{iL} \end{pmatrix} \tag{2.9}$$

is a q-vector, where $q = \sum_k q_l$.

For $\forall l$, i.e. any random effect $\boldsymbol{\alpha}_{il}$ of time-dependent covariate $\boldsymbol{Z}_{il}(u)$, we can calculate the least-square estimator of $\boldsymbol{\alpha}_{il}$, and obtain the following properties:

$$\mathbb{E}(\hat{\boldsymbol{\alpha}}_{il}) = \boldsymbol{\alpha}_{il}$$

$$Var(\hat{\boldsymbol{\alpha}}_{il}) = \sigma_{ll}\{\boldsymbol{F}_{il}^T\boldsymbol{F}_{il}\}^{-1}\boldsymbol{F}_{il}^T I_{ill}\boldsymbol{F}_{il}\{\boldsymbol{F}_{il}^T\boldsymbol{F}_{il}\}^{-1},$$

where $\boldsymbol{I}_{ill}$ is a $m_{il}(u) \times m_{il}(u)$ matrix whose (j, j') entry is $I(t_{ilj} = t_{ilj'})$, for $j, j' = 1, ..., m_{il}(u)$. As for any $\forall l$ and l' we have:

$$\Gamma_{ill'}(\omega)_{q_l \times q_{l'}} = cov(\hat{\alpha}_{il}, \hat{\alpha}_{il'}) = \sigma_{ll}\{\boldsymbol{F}_{il}^T\boldsymbol{F}_{il}\}^{-1}\boldsymbol{F}_{il}^T I_{ill'}\boldsymbol{F}_{il'}\{\boldsymbol{F}_{il'}^T\boldsymbol{F}_{il'}\}^{-1},$$

where $\omega = \{\sigma_{ll'} : l \geqslant l'\}$ and $\boldsymbol{I}_{ill'}$ is a $m_{il}(u) \times m_{il'}(u)$ matrix whose (j, j') entry is $I(t_{ilj} = t_{il'j'})$, for $j = 1, ..., m_{il}(u), j' = 1, ..., m_{il'(u)}$.

For $\boldsymbol{G}(u, \boldsymbol{\alpha}_i) = \boldsymbol{G}(u)\boldsymbol{\alpha}_i$, we can obtain that $\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i$ have the normal distribution $N\{\boldsymbol{G(u)}\boldsymbol{\alpha}_i, \boldsymbol{\Sigma}_i(u, \omega)\}$, where the covariance matrix of $\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i$ is $\boldsymbol{G}(u)\boldsymbol{\Gamma}_i(\omega)\boldsymbol{G}^T(u)$, and $\boldsymbol{\Gamma}_i(\omega)$ is the covariance matrix of $\hat{\boldsymbol{\alpha}}_i$:

$$\boldsymbol{\Gamma}_i(\omega)_{q\times q} = \begin{bmatrix} \Gamma_{i11}(\omega) & \Gamma_{i12}(\omega) & \cdots & \Gamma_{i1L}(\omega) \\ \Gamma_{i21}(\omega) & \Gamma_{i22}(\omega) & \cdots & \Gamma_{i1L}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{iL1}(u, \omega) & \Gamma_{iL2}(\omega) & \cdots & \Gamma_{iLL}(\omega) \end{bmatrix}.$$

The joint conditional probability mass function of $\{\mathrm{d}N_i(u), \boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i(u)\}$ in k-th sub-class, given $\{Y_i(u) = 1, c_{ik} = 1, \boldsymbol{\alpha}_i, \boldsymbol{Z}_i, \boldsymbol{t}_i(u)\}$ up to order du:

$$\begin{aligned}
&Pr_k\{dN_i(u) = r | Y_i(u) = 1, c_{ik} = 1, \boldsymbol{G}(u)\boldsymbol{\alpha}_i(u), \boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{t}_i(u)\} \\
&\times Pr\{\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i(u) = \boldsymbol{G}(u)\boldsymbol{\alpha}_i(u) | Y_i(u) = 1, c_{ik} = 1, \boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{t}_i(u)\} \\
&= \exp\left\{\boldsymbol{S}_{ik}^T(u, \gamma_k, \omega)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{G}(u)\boldsymbol{\alpha}_i\right\} \frac{\left\{h_{0k}(u)\mathrm{d}u \exp\left(\boldsymbol{\eta}_k^T\boldsymbol{x}_i\right)\right\}^{\mathrm{d}N_i(u)}}{(2\pi)^{s/2}\left|\boldsymbol{\Sigma}_i(u, \omega)\right|^{1/2}} \\
&\times \exp\left\{-\frac{\hat{\boldsymbol{\alpha}}_i^T G^T(u)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i + \boldsymbol{\alpha}_i^T\boldsymbol{G}^T(u)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{G}(u)\boldsymbol{\alpha}_i}{2}\right\},
\end{aligned}$$

where $\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) = \boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i + \mathrm{d}N_i(u)\boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k$, is a complete sufficient statistic for $\boldsymbol{\alpha}_i$ in k-th subtype. It suggests that conditioning on $\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)$ would remove the dependence of the conditional hazard for k-th subtype on the random effects and proof will be presented in the rest of this section.

Theorem 2.2.1. *The density function conditional on the sufficient statistic* $\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)$ *is independent of the random effects* $\boldsymbol{\alpha}$.

Proof. Rewriting $\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i$ as the following form:

$$\boldsymbol{G}(u)\hat{\boldsymbol{\alpha}}_i = \boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) - \mathrm{d}N_i(u)\boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k$$

Therefore, the joint conditional p.m.f could be written as:

$$\{h_{0k}(u)du\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}^{dN_i(u)} \times \mathcal{K}_k\{u, \boldsymbol{\alpha}_i, \omega, \boldsymbol{\gamma}_k\}\times$$
$$\exp\{-\frac{[\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) - \mathrm{d}N_i(u)\boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k]^T\boldsymbol{\Sigma}_i^{-1}(u, \omega)[\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) - \mathrm{d}N_i(u)\boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k]}{2}\}^{,}$$

where

$$\mathcal{K}_k\{u, \boldsymbol{\alpha}_i, \omega, \boldsymbol{\gamma}_k\} = \exp\{\frac{2\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{G}(u)\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^T\boldsymbol{G}^T(u)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{G}(u)\boldsymbol{\alpha}_i}{2}\}$$
$$\times (2\pi)^{-s/2}|\boldsymbol{\Sigma}_i^{-1}(u, \omega)|^{-1/2}.$$

Hence, for each sub-class k, the information in $\boldsymbol{\alpha}_i$ have been represented in $\mathcal{K}_k\{u, \boldsymbol{\alpha}_i, \omega, \boldsymbol{\gamma}_k\}$. Therefore, we have the probability for the counting process increment in sub-class k:

$$Pr_k\{dN_i(u) = 1|\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega), \boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u), c_{ik} = 1\}$$
$$= \frac{\int Pr_k\{dN_i(u) = 1, \boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u), c_{ik} = 1\}p(\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u))d\boldsymbol{\alpha}_i}{num + \int Pr_k\{dN_i(u) = 0, \boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u), c_{ik} = 1\}p(\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u))d\boldsymbol{\alpha}_i}$$

The numerator of the probability is

$$\{h_{0k}(u)du\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\} \times \int \mathcal{K}_k\{u, \boldsymbol{\alpha}_i, \omega, \boldsymbol{\gamma}_k\}p\{\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u)\}d\boldsymbol{\alpha}_i \times Y_i(u)\times$$
$$\exp\{-\frac{[\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) - \boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k]^T\boldsymbol{\Sigma}_i^{-1}(u, \omega)[\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega) - \boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k]}{2}\}.$$

The denominator of the probability is

$$num + \int \mathcal{K}_k\{u, \alpha_i, \omega, \boldsymbol{\gamma}_k\}p\{\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u)\}d\boldsymbol{\alpha}_i$$
$$\times Y_i(u) \times \exp\{-\frac{\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)}{2}\}.$$

Let

$$A_k = \exp\{-\frac{\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega)\boldsymbol{\Sigma}_i^{-1}(u, \omega)\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega)}{2}\}Y_i(u)$$
$$\times \int \mathcal{K}_k\{u, \boldsymbol{\alpha}_i, \omega, \boldsymbol{\gamma}_k\}p\{\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u)\}d\boldsymbol{\alpha}_i,$$

$$B_k = \{h_{0k}(u)\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}\exp\{\frac{2\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega)\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u, \omega)\boldsymbol{\gamma}_k}{2}\}$$

up to order du, we have:

$$\begin{aligned} Pr_k\{dN_i(u)=1|\boldsymbol{S}_{ik}(u,\gamma_k,\omega),\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u),c_{ik}=1\} &= \frac{B_k\times du}{1+B_k\times du}\times Y_i(u) \\ &= [1-\frac{1}{1-(-B_k\times du)}]\times Y_i(u) \\ &= B_k\times du\times Y_i(u)+o(du) \end{aligned}$$

Therefore, the hazard function for subject i in the k-th subtype, conditioning on $\boldsymbol{S}_{ik}(u,\gamma_k,\omega)$ and up to order du, is:

$$\begin{aligned} \lambda_{ik}(u) &= \lim_{du\to 0} du^{-1}Pr_k\{dN_i(u)=1|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u),c_{ik}=1\} \\ &= \{h_{0k}(u)\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}\exp\{\frac{2\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega)\boldsymbol{\gamma}_k-\boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u,\omega)\boldsymbol{\gamma}_k}{2}\}Y_i(u). \end{aligned} \tag{2.10}$$

The probability for the counting process increment in sub-class k is:

$$\begin{aligned} &Pr_k\{dN_i(u)=1|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u),c_{ik}=1\} \\ &\quad=\{h_{0k}(u)du\cdot\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}\exp\{\frac{2\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega)\boldsymbol{\gamma}_k-\boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u,\omega)\boldsymbol{\gamma}_k}{2}\}Y_i(u). \end{aligned} \tag{2.11}$$

The density function $f_k\left(T_i,\delta_i|\boldsymbol{x}_i,\boldsymbol{\alpha}_i\right)$ could be rewritten as the following form conditioning on the 'complete sufficient statistic' $\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega)$:

$$\begin{aligned} &f_k(dN_i(u)|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i) \\ &\quad=\prod_{\text{all grid points } u} Pr_k\{dN_i(u)=1|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u),c_{ik}=1\}^{dN_i(u)} \\ &\quad\times[1-Pr_k\{dN_i(u)=1|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u),c_{ik}=1\}]^{1-dN_i(u)}, \end{aligned} \tag{2.12}$$

which is independent of the random effects.

□

Estimation procedures in iterations

From the results above, we can obtain the complete log-likelihood conditional on the 'complete sufficient statistic' $\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega)$:

$$\ell(\boldsymbol{\Pi},\boldsymbol{H},\boldsymbol{\gamma},\boldsymbol{\eta};\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{U}|\boldsymbol{X},\boldsymbol{F})=\sum_{i=1}^{n}\sum_{k=1}^{K}c_{ik}[\log\pi_k+\log f_k(dN_i(u)|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),x_i)].$$

It is natural to consider the EM algorithm to estimate the parameters. Given the parameter estimates in the m-th iteration:

E-Step compute $\boldsymbol{U}^{(m+1)}$ according to the posterior probability:

$$c_{ik}^{(m+1)} = \frac{\pi_k^{(m)} f_k^{(m)}\left(dN_i(u)|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i\right)}{\sum_{k'=1}^{K}\pi_{k'}^{(m)} f_{k'}^{(m)}\left(dN_i(u)|\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i\right)}.$$

M-Step Update $\boldsymbol{\Pi}^{(m+1)}$ according to

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} c_{ik}^{(m+1)}}{n}.$$

As for the updating procedures for $\boldsymbol{\gamma}_k$, $\boldsymbol{\eta}_k$ and h_{ok}, we need to maximize the last term in the complete log-likelihood:

$$\ell(\boldsymbol{\gamma},\boldsymbol{\eta},\boldsymbol{H}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_u c_{ik}[dN_i(u)\log Pr_k + (1-dN_i(u))\log(1-Pr_k)].$$

Note that: $\frac{\partial\ell}{\partial\gamma_k} = \frac{\partial\ell}{\partial Pr_k}\times\frac{\partial Pr_k}{\partial\gamma_k}$ and $\frac{\partial\ell}{\partial\eta_k} = \frac{\partial\ell}{\partial Pr_k}\times\frac{\partial Pr_k}{\partial\eta_k}$, where

$$\frac{\partial Pr_k}{\partial\gamma_k} = \{h_{0k}(u)du\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}\exp\{\frac{2\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega)\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u,\omega)\boldsymbol{\gamma}_k}{2}\}Y_i(u)\times\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega),$$

$$\frac{\partial Pr_k}{\partial\eta_k} = \{h_{0k}(u)du\exp(\boldsymbol{\eta}_k^T\boldsymbol{x}_i)\}\exp\{\frac{2\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega)\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u,\omega)\boldsymbol{\gamma}_k}{2}\}Y_i(u)\times\boldsymbol{x}_i,$$

$$\frac{\partial\ell}{\partial Pr_k} = \sum_{i=1}^{n}\int c_{ik}[\frac{dN_i(u)}{Pr_k} - \frac{1-dN_i(u)}{1-Pr_k}].$$

Let $E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega) = \exp\{\boldsymbol{\eta}_k^T\boldsymbol{x}_i + \boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega)\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_i(u,\omega)\boldsymbol{\gamma}_k/2\}Y_i(u)$,

$$E_{0k}(u,\gamma_k,\eta_k,\omega) = \sum_{i=1}^{n}E_{0ik}(u,\gamma_k,\eta_k,\omega)c_{ik}.$$

Let $E_{1ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega) = (\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_k,\omega),\boldsymbol{x}_i^T)^T E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)$,

$$E_{1k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega) = \sum_{i=1}^{n}E_{1ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)c_{ik}.$$

Therefore the conditional score estimating equations for γ_k and η_k are:

$$\begin{aligned}\sum_{i=1}^{n}\int c_{ik}[dN_i(u) - E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)h_{ok}(u)du]\times\boldsymbol{S}_{ik}(u,\boldsymbol{\gamma}_k,\omega) = 0\\ \sum_{i=1}^{n}\int c_{ik}[dN_i(u) - E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)h_{ok}(u)du]\times\boldsymbol{x}_i = 0.\end{aligned} \tag{2.13}$$

And the proposed estimator for $h_{ok}(u)du$ in the $(m+1)$-th iteration could be obtained:

$$\hat{h}_{ok}(u)du = \frac{\sum_{i=1}^{n} c_{ik}^{(m+1)} dN_i(u)}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)} \tag{2.14}$$

replace $h_{ok}(u)du$ in equations (2.13) by (2.14), we can obtain the conditional score estimating equations for $\boldsymbol{\gamma}_k^{(m+1)}$ and $\boldsymbol{\eta}_k^{(m+1)}$:

$$\sum_{i=1}^{n} \int [(\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega), \boldsymbol{x}_i^T)^T - \frac{E_{1k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}] c_{ik}^{(m+1)} dN_i(u) = 0. \tag{2.15}$$

Generally, the parameters ω are unknown, under the assumptions, it may be estimated based on least-squares fits to all the data on each covariate for each subject when possible (i.e. $m_{ik} > q_k$). It is shown that an unbiased estimator for ω is $\hat{\omega}$, with element $\sigma_{kk'}$ estimated by

$$\hat{\sigma}_{kk'} = \frac{\sum_{i=1}^{n} \boldsymbol{I}\left(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0\right) \boldsymbol{R}_{ik}^T \boldsymbol{A}_{ikk'}^* \boldsymbol{R}_{ik'}}{\sum_{i=1}^{n} \boldsymbol{I}\left(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0\right) \operatorname{tr}\left\{\boldsymbol{P}_{ik} \boldsymbol{A}_{ikk'}^* \boldsymbol{P}_{ik'} \boldsymbol{A}_{ikk'}^{*T}\right\}},$$

where

- $\boldsymbol{P}_{ik} = \boldsymbol{I}_{m_{ik}} - \boldsymbol{F}_{ik}\left(\boldsymbol{F}_{ik}^T \boldsymbol{F}_{ik}\right)^{-1} \boldsymbol{F}_{ik}^T$;
- $\boldsymbol{R}_{ik} = \boldsymbol{P}_{ik} \boldsymbol{W}_{ik} = \boldsymbol{P}_{ik} \boldsymbol{e}_{ik}$;
- Suppose covariates k and k' are observed in common at $m_{ikk'} > 0$ time points;
- $\boldsymbol{A}_{ik}$ is the $(m_{ikk'} \times m_{ik})$ matrix of zeros and ones that identifies the residuals for covariate k at the common time points, define $\boldsymbol{A}_{ik'}$ similarly, and $\boldsymbol{A}_{ikk'}^* = \boldsymbol{A}_{ik}^T \boldsymbol{A}_{ik'}$ $(m_{ik} \times m_{ik'})$.

Therefore, the final conditional score estimating equations for the parameters of interest in the $(m+1)$-th iteration, with the replacement of ω by its unbiased estimator $\hat{\omega}$ introduced above, are:

$$\sum_{i=1}^{n} \int [(\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \hat{\omega}), \boldsymbol{x}_i^T)^T - \frac{E_{1k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \hat{\omega})}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \hat{\omega})}] c_{ik}^{(m+1)} dN_i(u) = 0. \tag{2.16}$$

2.2.2 large-sample property

We give a brief proof showing that solving the conditional score estimating equations (2.15) with the parameters $\boldsymbol{\omega}$ known should yield consistent estimators for $(\boldsymbol{\gamma}_k, \boldsymbol{\eta}_k)$. We demonstrate this property via simulations in Section 2.3.

Theorem 2.2.2. *Under regularity conditions, as $n \to \infty$, a solution to the conditional score estimating equations (2.15), say $(\hat{\boldsymbol{\gamma}}_k, \hat{\boldsymbol{\eta}}_k)$, exists uniquely in a neighborhood of true parameters $(\boldsymbol{\gamma}_{0k}, \boldsymbol{\eta}_{0k})$ with probability 1.*

Lemma 2.2.3. ***Glivenko-Cantelli lemma**: Assume that $X_1, X_2, ..., X_n$ are i.i.d random variables in $\mathbb{R}$ with common cumulative distribution function $F(x)$. The empirical distribution for $X_1, X_2, ..., X_n$ is*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}(x),$$

and

$$\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 \quad \text{almost surely.}$$

Proof. Defining $\bar{S}_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega) = \frac{E_{1k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}{E_{0k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}$ to be the weighted average of vectors $\{\boldsymbol{S}_{ik}(u, \boldsymbol{\gamma}_k, \omega), \boldsymbol{x}_i^T\}^T$ among individuals i in sub-class k at risk at time u, and letting $\mu_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)$ denote the probabilistic limit of $\bar{S}_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)$, by adding and subtracting common terms, we can rewrite the conditional score estimating equations (2.16) as:

$$\sum_{i=1}^{n}\int [(\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega), \boldsymbol{x}_i^T)^T - \bar{\boldsymbol{S}}_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)]\{c_{ik}dN_i(u) - \frac{c_{ik}E_{0ik}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}\sum_{i=1}^{n} c_{ik}dN_i(u)\}, \tag{2.17}$$

which could be rewritten as:

$$\sum_{i=1}^{n}\int [(\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_k, \omega), \boldsymbol{x}_i^T)^T - \mu_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)]\{c_{ik}dN_i(u) - \frac{c_{ik}E_{0ik}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}\sum_{i=1}^{n} c_{ik}dN_i(u)\}, \tag{2.18}$$

$$+\sum_{i=1}^{n}\int [\mu_k(u, \boldsymbol{\gamma}_k, \eta_k, \omega) - \bar{\boldsymbol{S}}_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)]\{c_{ik}dN_i(u) - \frac{c_{ik}E_{0ik}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}{E_{0k}(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)}\sum_{i=1}^{n} c_{ik}dN_i(u)\}. \tag{2.19}$$

If set equal to zero with the function $\mu_k(u, \boldsymbol{\gamma}_k, \boldsymbol{\eta}_k, \omega)$ known, (2.18) is an unbiased estimating equation for $(\boldsymbol{\gamma}_k, \boldsymbol{\eta}_k)$, which follows as, at the true values $(\boldsymbol{\gamma}_{0k}, \boldsymbol{\eta}_{0k})$, (2.18) is a sum of independent and identically distributed zero-mean random vectors. Taking the expectation inside the integral for the i-th summand and conditioning on $[(\boldsymbol{S}_{ik}^T(u, \boldsymbol{\gamma}_{0k}, \omega), \boldsymbol{x}_i^T)^T, \boldsymbol{x}_i, \boldsymbol{t}_i(u), Y_i(u)]$, which yields

$$\int E\{[(\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_{0k},\omega),\boldsymbol{x}_i^T)^T-\mu_k(u,\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k},\omega)]$$
$$\times(E[c_{ik}dN_i(u)|\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_{0k},\omega),\boldsymbol{x}_i^T)^T,\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u)]-\frac{c_{ik}E_{0ik}(u,\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k},\omega)}{E_{0k}(u,\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k},\omega)}\sum_{i=1}^n c_{ik}dN_i(u))\}, \tag{2.20}$$

as

$$E[c_{ik}dN_i(u)|\boldsymbol{S}_{ik}^T(u,\boldsymbol{\gamma}_{0k},\omega),\boldsymbol{x}_i^T)^T,\boldsymbol{x}_i,\boldsymbol{t}_i(u),Y_i(u)]=\frac{c_{ik}E_{0ik}(u,\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k},\omega)}{E_{0k}(u,\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k},\omega)}\sum_{i=1}^n c_{ik}dN_i(u),$$

is the conditional intensity in section 2.2, the inner expectation is equal to zero, so that (2.20) is zero, demonstrating the unbiasedness. That n^{-1} times (2.19) converges in probability to zero uniformly in a neighbourhood $\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})$ of $(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})$ follows from the inequality

$$\sup_{\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})}|\int[\mu_k(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)-\bar{S}_k(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)]$$
$$\times n^{-1}\sum_{i=1}^n\{c_{ik}dN_i(u)-\frac{c_{ik}E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}{E_{0k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}\sum_{i=1}^n c_{ik}dN_i(u)\}| \tag{2.21}$$

$$\leq\sup_{\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})}[\sup_u\{|\mu_k(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)-\bar{S}_k(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)|\}] \tag{2.22}$$

$$\times[n^{-1}\sum_{i=1}^n\int c_{ik}dN_i(u)+n^{-1}\sum_{i=1}^n\sup_{\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})}\{\int\frac{c_{ik}E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}{E_{0k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}\sum_{i=1}^n c_{ik}dN_i(u)\}]. \tag{2.23}$$

The first term in (2.23) is bounded by 1, and the second converges to

$$E\{\sup_{\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})}\int\frac{c_{ik}E_{0ik}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}{E_{0k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)}\sum_{i=1}^n c_{ik}dN_i(u)\}$$

in probability. Uniform covergence of $n^{-1}E_{0k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)$ and $n^{-1}E_{1k}(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)$, and hence of $\bar{\boldsymbol{S}}_k(u,\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k,\omega)$, both in u and $(\boldsymbol{\gamma}_k,\boldsymbol{\eta}_k)$ in $\mathcal{N}(\boldsymbol{\gamma}_{0k},\boldsymbol{\eta}_{0k})$, could be established by a modification of the Glivenko-Cantelli lemma, thus showing convergence in probability to zero of (2.19).

Therefore, the behaviour of the estimators solving conditional score equations (2.15) could be dictated by (2.18). Since (2.18) is an unbiased estimating equation, under regularity conditions, a consistent sequence of solutions to it exists Andersen and Gill (1982), indicating the existence of consistent solutions to (2.18).

□

2.3 Simulation

In this section, two simulation studies have been proposed to assess the performance of the joint FMCox PH models and the corresponding estimation method. Scenario 1 mimicked the simulation studies given by Eng and Hanlon (2014), which focuses on the special type of the joint FMCox PH models proposed in this thesis. And from the scenario 1, the performance of the joint FMCox PH models and the FMCox PH model (Eng and Hanlon, 2014) could be compared directly. Scenario 2 was aimed to illustrate the performance of the proposed estimator for the more general parameters settings.

For the following simulation studies, recall the proposed joint FMCox PH models:

Longitudinal sub-model:

$$\begin{aligned} Z_i(u) &= \boldsymbol{f}^T(u)\boldsymbol{\alpha}_i \quad \text{e.g. } Z_i(u) = \alpha_{0i} + \alpha_{1i}u \\ W_i(u) &= Z_i(u) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{2.24}$$

where $\boldsymbol{\alpha}_i$ are the random effects, α_{0i} is the random intercept and α_{1i} is the random slope.

Survival sub-model:

$$T_i = min(\tilde{T}_i, C_i) \text{ and } \delta_i = I(T_i \leq C_i)$$

hazard function (i.e. the Cox PH model) for subject i in sub-class k is:

$$\lambda_{ik}(u) = h_{0k}(u)\exp\{\gamma_k Z_i(u) + \eta_k x_i\} \tag{2.25}$$

Latent variable:

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ belongs to sub-class } k \\ 0 & \text{otherwise} \end{cases}$$

with the probability

$$\mathbb{P}\{c_{ik} = 1\} = \pi_k \quad \text{and} \quad \sum_k \pi_k = 1$$

2.3.1 Simulation study scenario 1

From the simulation study proposed by Eng and Hanlon (2014), the main objective is estimating the fixed effect of the time-independent covariate x. To compare the performance of the proposed joint FMCox PH models and the FMCox PH models in Eng

and Hanlon (2014), scenario 1 focuses on the situation of a single time-independent covariate x with the corresponding parameters η_k of interest and a given time-dependent covariate $Z(u)$ with the corresponding coefficients as $\gamma_k = 0$.

Assume that the number of latent classes $K = 2$, all the $2n$ observations have a single covariates $(x_1, ..., x_{2n}) \sim N(\mu \cdot 1_{2n}, I_{2n})$. The relationship between the event time of interest and covariate X is controlled by η, where the first class has $\eta_1 = \eta$ and the second class has $\eta_2 = -\eta$. Let the baseline hazard function for each class be $h_{ok}(u) = 1$. Hence, the survival time for the ith subject is generated by $\tilde{T}_i = \frac{U_i}{\exp(x_i \eta_k)}$, where $U_i \sim$ Exponential(1). The censoring time is generated from $C_i \sim$ Uniform$(0, \lambda)$, where λ depends on the choice of μ and β and a target censoring rate. The corresponding observed survival time is $T_i = min(\tilde{T}_i, C_i)$.

Set $n = 500$ subjects in each class and set $\eta = 3$ so that $\eta_1 = 3$ and $\eta_2 = -3$. To target 40% censoring rate, λ is set $\lambda = exp(0.99)$ for $\mu = 0$ and $\lambda = exp(12.83)$ for $\mu = 5$. As for the time-dependent covariate, which could be observed as the longitudinal process $W(u)$, could be generated from the linear mixed model:

$$Z_i(u) = \alpha_{0i} + \alpha_{1i} u \quad u = 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4.$$
$$W_i(u) = Z_i(u) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2 = 0.69^2)$$

where the random effects $(\alpha_{0i}, \alpha_{1i})^T$ are generated from the multivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (8.03, -0.16)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 0.87 & -0.001 \\ -0.001 & 0.02 \end{pmatrix}$. And the coefficients $\gamma_k, k = 1, 2$ for the time-dependent covariate $Z(u)$ are set equals 0 in this simulation.

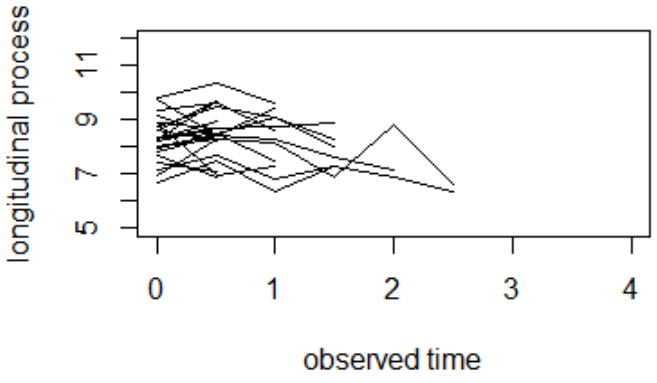

Figure 2.1: 20 samples of the longitudinal processes in simulation study scenario 1

As for the simulated event time, the Kaplan-Meier survival curve could be plotted in Figure 2.2 and Figure 2.3.

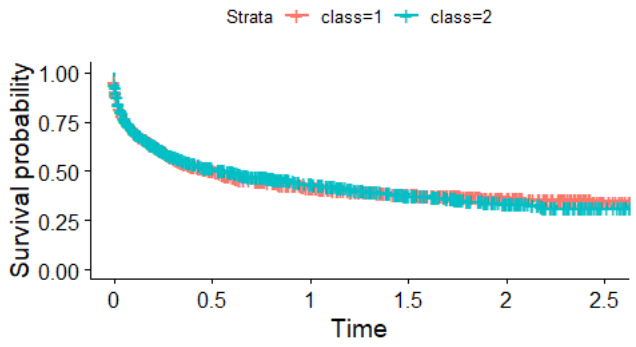

Figure 2.2: the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 1, with the mean of the time-independent covariate $\mu = 0$.

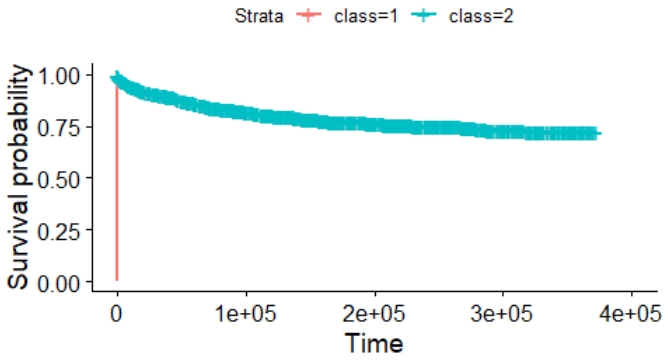

Figure 2.3: the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 1, with the mean of the time-independent covariate $\mu = 5$.

From the Figure 2.2 and 2.3, it is obvious that the event times in the two specific latent classes have two extreme presented form, the first one is that the event times have no significant difference between the two classes, and the other one is that the event times have the difference in ten thousand times the order of magnitude. In this case, the standard joint modeling of longitudinal and time-to-event data will be used at first, which did not consider the latent class, the estimation results are presented in Table 2.1 and Table 2.2. Therefore, this method couldn't provide the estimation results for the parameters γ_k. And the estimation results for the parameters η_k are the same in different latent class, which is contrary to the facts.

Table 2.1: estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 1, with the mean of the time-independent covariate $\mu = 0$.

	coef	se(coef)	z	$Pr(>\|z\|)$
X	0.01556	0.03131	0.497	0.619
	exp(coef)	exp(-coef)	lower .95	upper .95
X	1.016	0.9846	0.9552	1.08

Table 2.2: estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 1, with the mean of the time-independent covariate $\mu = 5$.

	coef	se(coef)	z	$Pr(>\|z\|)$
X	-0.25971	0.02903	-8.947	$< 2e-16$
	exp(coef)	exp(-coef)	lower .95	upper .95
X	0.7713	1.297	0.7286	0.8164

From these estimation results, it is necessary to take the latent class into consideration. FMCox PH models (Eng and Hanlon, 2014) just include the time-independent covariate x and estimate the corresponding coefficients η_k. The joint FMCox PH models proposed in this thesis will include both time-independent covariate x and time-dependent covariate $Z(u)$ and estimate the corresponding coefficients η_k and γ_k.

We study the same scenarios over 500 simulations. In the following tables, we report the estimated mixing probability π_k, the estimated η_k from the FMCox PH models (Eng and Hanlon, 2014), estimated η_k and γ_k from the proposed joint FMCox PH models (3.36) and the oracle estimator, which is named under the situation if the classification parameters c_{ik} are known. Classification accuracy is also represented in the Table 2.3 and Table 2.4, which means the proportion of observations assigned to their correct class.

The results imply that the proposed joint FMCox PH model and its corresponding EM-algorithm with conditional score method work well when the dataset include time-independent X and time-dependent $Z(u)$ with the corresponding coefficients equal to 0, and has heavy censoring rate. The bias between the proposed estimators ($\eta_{joint,1}$, $\eta_{joint,2}$, $\gamma_{joint,1}$, $\gamma_{joint,2}$) and the oracle estimators is believed from the algorithm greedily reinforcing. And the bias between the proposed estimators and the results from the FMCox PH model, which just include the time-independent covariate, are quite small. Based on these conclusion, the joint FMCox PH models and its corresponding

Table 2.3: Comparison among the estimation results for FMCox PH models, proposed joint FMCox PH models and the oracle estimators, with the mean of the time-independent covariate $\mu = 0$

Parameter	Scenario $\gamma_{10} = \gamma_{20} = 0, \eta_{10} = 3, \eta_{20} = -3$
$\eta_{FMCox,1}(SD)$	3.45 (0.53)
$\eta_{FMCox,2}(SD)$	-3.46(0.52)
$\gamma_{joint,1}(SD)$	0.01337(0.94)
$\gamma_{joint,2}(SD)$	0.00896(0.89)
$\eta_{joint,1}(SD)$	3.52(0.97)
$\eta_{joint,2}(SD)$	-3.34(0.92)
$\gamma_{oracle,1}(SD)$	0.00956(0.32)
$\gamma_{oracle,2}(SD)$	0.01082(0.36)
$\eta_{oracle,1}(SD)$	3.05(0.34)
$\eta_{oracle,2}(SD)$	-3.02(0.33)
$\pi_1(SD)$	0.5052(0.15)
$\pi_2(SD)$	0.4948(0.18)
Accuracy(range)	0.87(0.81-0.93)
Censoring(range)	0.39(0.32-0.48)

Table 2.4: Comparison among the estimation results for FMCox PH models, proposed joint FMCox PH models and the oracle estimators, with the mean of the time-independent covariate $\mu = 5$

Parameter	Scenario $\gamma_{10} = \gamma_{20} = 0, \eta_{10} = 3, \eta_{20} = -3$
$\eta_{FMCox,1}(SD)$	3.24 (0.58)
$\eta_{FMCox,2}(SD)$	-2.14(0.55)
$\gamma_{joint,1}(SD)$	0.01717(0.85)
$\gamma_{joint,2}(SD)$	0.004952(0.98)
$\eta_{joint,1}(SD)$	3.44(0.56)
$\eta_{joint,2}(SD)$	-3.08(1.09)
$\gamma_{oracle,1}(SD)$	0.00245(0.45)
$\gamma_{oracle,2}(SD)$	0.03984(0.34)
$\eta_{oracle,1}(SD)$	3.03(0.28)
$\eta_{oracle,2}(SD)$	-3.12(0.57)
$\pi_1(SD)$	0.5093(0.14)
$\pi_2(SD)$	0.4907(0.13)
Accuracy(range)	0.89(0.72-0.97)
Censoring(range)	0.39(0.31-0.44)

estimation procedures proposed in thesis could complete the work what the FMCox PH models (Eng and Hanlon, 2014) have done.

2.3.2 Simulation study scenario 2

In this scenario, we evaluate the performance of our proposed joint FMCox PH models and the corresponding estimating procedure when the coefficients for the time-dependent covariate γ are not 0. At first, we need to simulate the time-dependent covariate, which could be observed as the longitudinal process $W(u)$, and could be generated from the linear mixed model:

$$Z_i(u) = \alpha_{0i} + \alpha_{1i}u \quad u = 0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80.$$
$$W_i(u) = Z_i(u) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2 = 1)$$

where the random effects $(\alpha_{0i}, \alpha_{1i})^T$ are generated from the multivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (4.173, -0.0103)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1.24 & -0.001 \\ -0.001 & 0.001 \end{pmatrix}$. Note that this longitudinal process is motivated from the AIDS CLINICAL TRIALS GROUP 175 presented by Tsiatis and Davidian (2001).

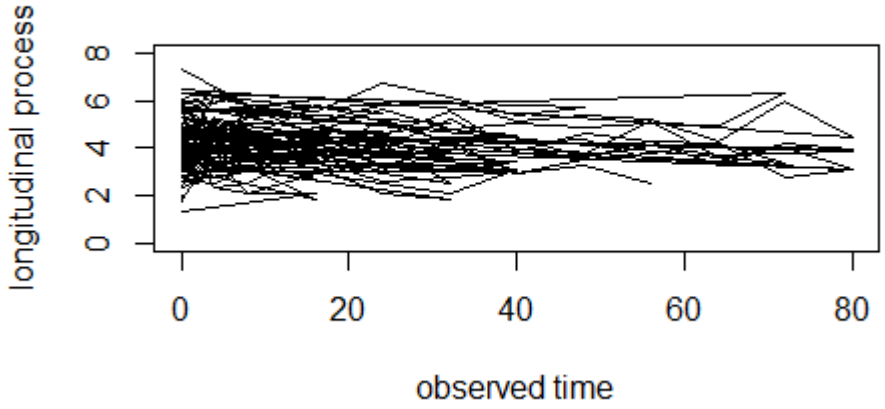

Figure 2.4: 50 samples of the longitudinal processes in simulation study scenario 2

Assume that the number of latent classes $K = 2$, all the $2n$ observations have a single time-independent covariate x, which is obtained from the Bernoulli distribution with the probability 0.5. This simulated time-independent covariate is motivated from the treatment in many medical research. The relationship between the event time

of interest and the time-independent covariate x is controlled by η, where the first class here has $\eta_1 = 0$ and the second class has $\eta_2 = 0.5$. The relationship between the event time and the time-dependent covariate $Z(u)$ is controlled by γ, where the first class here has $\gamma_1 = -1$ and the second class has $\gamma_2 = -0.5$. Let the baseline hazard function for each class be $h_{0k}(u) = 1$. Hence, the survival time $\tilde{T}_i$ for the i-th subject could be generated from the specific formulation proposed by Austin (2012); Bender et al. (2005). The censoring time is generated from $C_i \sim \exp(110)$, where the simulated censored times are bounded by the given observed measurement times in the longitudinal process. The observed event time used in this scenario is $T_i = min(\tilde{T}_i, C_i)$. Set $n = 300$ subjects in each class and set $\eta_1 = 0, \eta_2 = 0.5, \gamma_1 = -1, \gamma_2 = -0.5$, and $C_i \sim \exp(110)$ so that the censoring rate is around 10%. Therefore, the Kaplan-Meier survival curve could be plotted as follow:

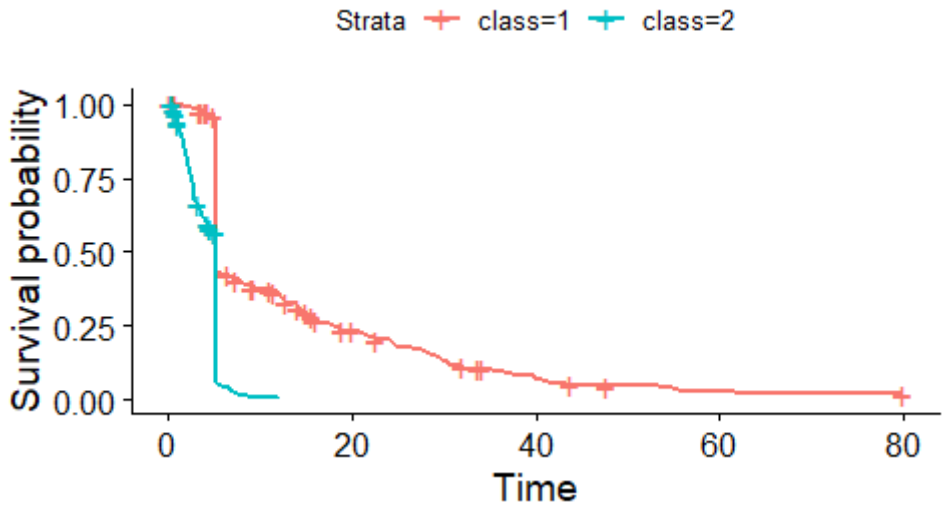

Figure 2.5: the Kaplan-Meier survival curves for the given latent classes in one simulation of the scenario 2.

The standard joint modeling of longitudinal and time-to-event data and the corresponding R package JM (Rizopoulos, 2010) will be used for this scenario at first, and the estimation results are presented in the Table 2.5

From the Table 2.5, the estimation of the longitudinal process is reasonable, the estimation of the covariance matrix for the random effects α and the measurement residual are quite close to the true parameters setting. But in the event process, the coefficient for the time-independent covariate is estimated as $\hat{\eta} = 0.0777$ and

Table 2.5: estimation results from the standard joint modeling of longitudinal and time-to-event data in the scenario 2.

Joint Model Summary:
Longitudinal Process: Linear mixed-effects model
Event Process: Relative risk model with piecewise-constant baseline risk function
Parameterization: Time-dependent

	log.Lik	AIC	BIC
	-6989.349	14006.7	14068

Variance Components:

	StdDev	Corr
(intercept)	1.1227	(Intr)
time	0.0285	0.0602

residual: 1.0092
Coefficients:
Longitudinal Process

	value	Std.Err	z-value	p-value
(intercept)	4.1272	0.0529	78.0335	<0.0001
time	-0.0146	0.0023	-6.3027	<0.0001

Event Process

	value	Std.Err	z-value	p-value
x	0.0773	0.0926	0.8342	0.4042
association	-0.1318	0.0360	-3.6649	0.0002

the coefficient for the time-dependent covariate is estimated as $\hat{\gamma} = -0.1318$, which have large bias compared with the true simulation settings, and could not explain the heterogeneity in this scenario.

We study the same scenario over 500 simulations. In Table 2.6, we report the estimated γ_k and η_k from the proposed joint FMCox PH models and the oracle estimator if the true classification situation is known. Accuracy is the proportion of observations assigned to their correct class.

The results in Table 2.6 imply that the estimation procedures for the longitudinal sub-model perform quite similar to the standard joint modeling work, which will characterize the properties of longitudinal process well. As for the estimation in the survival sub-model, there are non-ignorable but small bias between the estimators of the proposed inference procedures and the true parameters setting. These biases toward larger absolute parameter estimates, which compared with the oracle estimators, that we believes comes from the algorithm greedily reinforcing as discussed in scenario 1. And comparing these results with the estimation results in Table 2.5, the proposed joint FMCox PH models perform better than than the standard joint modeling when

Table 2.6: Comparison between the estimation results for the proposed joint FMCox PH models and the oracle estimators in the scenario 2.

Joint FMCox PH Model Summary:
Longitudinal Process: Linear mixed-effects model
Event Process: Cox PH model with piecewise-constant baseline hazard function
Variance Components:

	StdDev	Corr
(intercept)	1.199504	(Intr)
time	0.2943384	0.53743

residual: 1.021486
Coefficients:
Longitudinal Process

	value	Std.Err
(intercept)	4.106106	0.0453
time	-0.01646654	0.0034

Event Process

	value	Std.Err
x (class 1)	0.03253	0.9445
x (class 2)	0.46726	0.7933
association (class 1)	-0.94352	0.8608
association (class 2)	-0.62669	0.6861

The oracle estimators:

	value	Std.Err
x (class 1)	0.00452	0.4434
x (class 2)	0.53234	0.2313
association (class 1)	-1.03867	0.3935
association (class 2)	-0.57451	0.467

Mixing probability estimation:
π_1 (SD) for class 1: 0.5165 (0.21)
π_2 (SD) for class 2: 0.4835 (0.23)
Accuracy (range): 0.84 (0.78-0.91)
Censoring (range): 0.10 (0.08-0.15)

dealing with the heterogeneous problem in the longitudinal and time-to-event data.

2.4 Real analysis

Among infectious diseases, the AIDS studies are a good example to be used in joint modeling of the longitudinal and survival processes in many literature In the last few decades (Brilleman et al., 2016; Brombin et al., 2016; Farahani et al., 2016; Liu and Huang, 2009). In AIDS studies, CD4 cells are always considered as a sign of disease progression in HIV-infected patients, which are help to coordinate the immune system's response to certain microorganisms e.g. viruses. And based on the medical knowledge (Brombin et al., 2016; Farahani et al., 2016; Song et al., 2017), the lower the CD4 count is, patients are at the higher risk of infection.

In this section, the AIDS dataset from Community Programs for Clinical Research on AIDS (CPCRA) was used (Abrams et al., 1994). And in this study, there are 467 patients infected with HIV. Two outcomes were recorded for this study, the first one is CD4 counts, which were measured at different given time points, i.e. 0,6,12 and 18 months, and several samples of the log CD4 counts processes are presented in Figure 2.6. The other one is the time-to-death outcome, our main interest focuses on how the CD4 process and other time-independent covariate e.g. the given treatment effect the risk of infection or death. In CPCRA study, patients received two different treatments, Zalcitabine (ddC) or Didanosine (ddI), randomly.

Roustaei et al. (2018) proposed an approach to study the latent heterogeneous problem in CPCRA study, compared with joint latent class model (JLCM) proposed by Liu et al. (2015) and separate approach. But these three methods need to model the class membership probability for each patient at fist, and even assume that the longitudinal processes are independent of the event time, conditional on the given classification. For simplicity, some prior knowledge was used for our real data analysis directly, such as there are $K = 2$ latent classes, and the main time-independent covariate of interest is the treatment. Due to the skewed distribution of CD4 cell level, logCD4 was used as the longitudinal outcomes. The baseline hazard functions were assumed to be piecewise-constant functions. At first, we need to plot the longitudinal process, so that we could assume the following linear mixed effects model for this study

based on Figure 2.6.

$$\begin{aligned} Z_i(u) &= \boldsymbol{f}^T(u)\boldsymbol{\alpha}_i \quad \text{e.g. } Z_i(u) = \alpha_{0i} + \alpha_{1i}u \\ W_i(u) &= Z_i(u) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{2.26}$$

where α_i are the random effects, α_{0i} is the random intercept and α_{1i} is the random slope. The observed measurement time points are $u = 0, 6, 12, 18$. It is obvious that the random slope may be negative and its absolute value may be quite small.

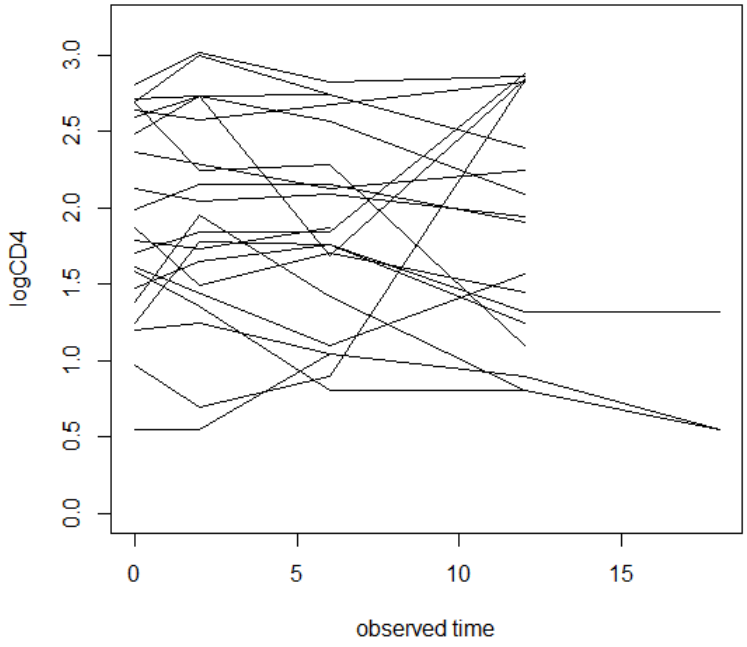

Figure 2.6: 20 samples of the log CD4 counts processes in AIDS study.

Considering the logCD4 as the time-dependent covariate in the survival sub-model, the next objective is estimating the coefficients η_k for time-independent covariate treatment and the coefficients γ_k for time-dependent covariate logCD4. Recall the survival sub-model:

$$\lambda_{ik}(u) = h_{0k}(u) \exp\{\gamma_k \times \log CD4_i + \eta_k \times treatment_i\} \tag{2.27}$$

with the latent classification variable

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ belongs to sub-class 1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbb{P}\{c_{ik} = 1\} = \pi_k \quad \text{and} \quad \sum_{k}^{2} \pi_k = 1.$$

The estimation results were presented in Table 2.7.

Table 2.7: Estimation results for the AIDS study.

Joint FMCox PH Model Summary:
Longitudinal Process: Linear mixed-effects model
Event Process: Cox PH model with piecewise-constant baseline hazard function
Variance Components:

	StdDev	Corr
(intercept)	0.6719533	(Intr)
time	0.1054032	0.3242496

residual: 0.06033
Coefficients:
Longitudinal Process

	value
(intercept)	1.835005
time	-0.0333369

Event Process

	value
x (class 1)	-0.194
x (class 2)	-0.284
association (class 1)	-0.933
association (class 2)	-0.436

Mixing probability estimation:
π_1 for class 1: 0.6843
π_2 for class 2: 0.3157

And the the Kaplan-Meier survival curves for the two latent classes in this AIDS study could be plotted in Figure 2.7

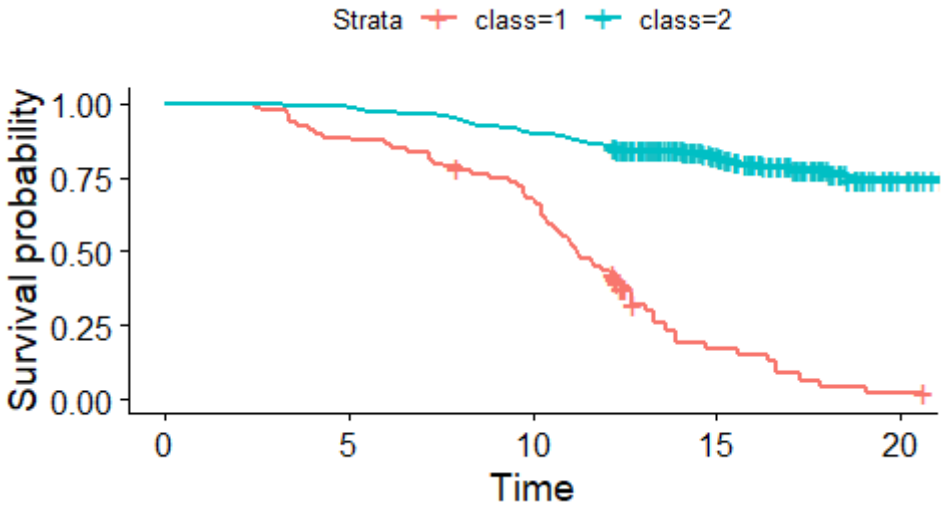

Figure 2.7: the Kaplan-Meier survival curves for the two latent classes in the AIDS study.

Note that the estimations of association parameters between logCD4 and the event time $\gamma_1 = -0.933$ and $\gamma_2 = -0.436$ are both negative, which satisfy the medical fact that the lower the CD4 count is, patients are at the higher risk of infection or death. And the estimations of association parameters between treatment and the event time $\eta_1 = -0.194$ and $\eta_2 = -0.284$ are also negative, which mean that taking the given treatment will decrease the risk of infection or death. And from the Kaplan- Meier survival curves, the CD4 level and the given treatment would have different effects on the patients from different latent classes, the patients in latent class 1 have the lower survival probabilities. For the patient from the latent class 1, the CD4 level would make more effect and treatment would make less effect on the risk of infection or death, compared with the patients from the latent class 2.

2.5 Discussion

In this chapter, we presented the joint FMCox PH models, which could estimate the associations of event times and both time-independent covariates and longitudinal processes, considering the latent heterogeneity in the longitudinal and time-to-event data. To remove the complication which comes from the latent random effects associated two sub-models, the conditional score method which does not need any assumption

on the distribution of the random effects, is proposed to replace the standard partial likelihood weighted estimators in the EM iterations. Based on these ideas, the EM-algorithm with conditional score method was provided as the estimation procedures for the joint FMCox PH models, and the consistency of the estimators have also been represented in this chapter.

To assess the performance of the joint FMCox PH models, two simulation studies have been proposed in this chapter. Under the scenario 1, the performance of the joint FMCox PH models is very similar to the performance of FMCox PH models, which means that the joint FMCox PH models could replace FMCox PH models to handle the heterogeneity in survival data analysis. Under the scenario 2, the performance of the joint FMCox PH models show that this proposed models and the corresponding estimation procedures could give the reasonable estimators, considering the heterogeneity in longitudinal and time-to-event data analysis.

However, there are many limitations in this work, such as the lack of the choice of the number of latent classes, more than one time-independent and one time-dependent covariates simulations. In this chapter, we didn't consider the situations where the number of latent classes is not known, but it is common in practice. Therefore, some choice criteria e.g. AIC, BIC used in standard joint modeling, need to be proposed and assess the corresponding performance in the future work. And the joint FMCox PH models we proposed in section 2.1 consider multiple longitudinal processes and baseline covariates. Therefore, more complex simulation, e.g. including more than one time-independent and time-dependent covariates, will also be needed in the future work.

Chapter 3

FMCox PH Model with Time-varying Coefficients

In the survival analysis, the main objective is to explore the association between the event time T and the observed covariate vector x. The most popular semi-parametric regression model used for time-to-event data analysis is the proportional hazard model (COX PH model), which is proposed by Cox (1972). And the standard COX PH model has the important assumption that the regression coefficients are constant over time.

However, it is often important to characterize the effects over time of the covariates of interest on the event time in many medical research. For example, in the AIDS study which needs to compare a new treatment with an active control, suppose that the time to a clinical event e.g. death is the primary endpoint. The new treatment may work well in the initial clinical period, but may gradually lose the efficiency due to mutation of the virus. If the treatment is indeed time-efficient, then it is necessary to know when and how fast the treatment becomes ineffective. Therefore, the Cox PH model with a time-varying coefficient (Gamerman, 1991; Hastie and Tibshirani, 1993; Martinussen and Scheike, 2002; Marzec and Marzec, 1997; Murphy and Sen, 1991; Zucker and Karr, 1990) is much more flexible and may be needed for medical research and survival data analysis.

To estimate the coefficient functions, Murphy and Sen (1991) assumed the coefficient functions are piecewise constant and proposed the histogram sieve estimation. Based on the similar assumption that both the baseline hazard function $h_0(t)$ and time-varying coefficient functions $\eta(t)$ are piecewise constant functions which are constant

between distinct event times, Gamerman (1991) proposed the dynamic linear model approach for the estimations of time-varying coefficients. This strong assumption are not appropriate in practice, Zucker and Karr (1990) and Hastie and Tibshirani (1993) described the smoothing spline partial likelihood method to overcome the drawback, where the baseline hazard function has the unspecific form and the coefficient functions could be estimated smoothly. King (1997) suggested using a local partial likelihood estimation technique for the hazard regression model with time-varying coefficients. And the local partial likelihood technique used in this chapter has the similar basic idea and is a simple extension of the local linear fitting technique used in the scatterplot smoothing, which is proposed by Cai and Sun (2003).

3.1 The local partial likelihood technique

Let $\tilde{T}$ be the unobserved true event time and C be the right censored time, so that $T = min(\tilde{T}, C)$ be the observed event time, with the indicator variable $\delta = I(\tilde{T} \leq C)$ and the associated covariates x. $\{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^n$ be an i.i.d sample obtained from the population $(\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{x})$ that follows the Cox PH model with time-varying coefficients:

$$\lambda_i(t) = h_0(t) \exp\left(\boldsymbol{\eta}^T(t)\boldsymbol{x}_i\right) \tag{3.1}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\eta}(t) = (\eta_1(t), ..., \eta_p(t))^T$ be the time-varying coefficient functions and x_i is the p-vector covariates. Define $N_i(t) = I\left(T_i \leq t, \delta_i = 1\right)$ be the counting process of observed event time for i-th subject, and $Y_i(t) = I\left(T_i \geq t\right)$ be the 'at risk' indicator process. The logarithm of Cox's partial likelihood based on observations over time interval $[0, \tau]$ for $\tau > 0$ is:

$$\ell(\eta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\eta}(s) - \log \left\{ \sum_{l=1}^{n} Y_l(s) \exp\left(\mathbf{x}_l^{\mathrm{T}} \boldsymbol{\eta}(s)\right) \right\} \right] dN_i(s) \tag{3.2}$$

where the baseline hazard function $h_0(t)$ are assumed to be unspecific, positive and continuous, the coefficient functions $\boldsymbol{\eta}(s) = (\eta_1(s), ..., \eta_p(s))^T$ are assumed to have continuous second derivative in a neighborhood of given time point t. Therefore, for any time s in a neighborhood of given time point t, by Taylor's expression, the time-varying coefficient function could be written as

$$\eta_j(s) \approx \eta_j(t) + b_j(t)(s - t)$$

Let $\boldsymbol{\beta} = (\eta_1(t), \ldots, \eta_p(t), b_1(t), \ldots, b_p(t))^{\mathrm{T}}$ and $\boldsymbol{X}_i(s) = \boldsymbol{x}_i \times (1, s-t)^T$. And let $K(\cdot)$ be a kernel function which weights smoothly down the contribution of remote data points and $h = h_n$ be the bandwidth parameter which controls the size of the local neighborhood of the given time point. Based on the local linear fitting technique and partial likelihood method, the local linear partial likelihood function could be obtained as:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} K_h(s-t) \left[\boldsymbol{X}_i^T(s)\boldsymbol{\beta} - \log \left\{ \sum_{l=1}^{n} Y_l(s) \exp\left(\boldsymbol{X}_i^T(s)\boldsymbol{\beta}\right) \right\} \right] dN_i(s) \quad (3.3)$$

where $K_h(\cdot) = K(\cdot/h)/h$. Therefore, the local linear partial MLE of the time-varying coefficient functions $\boldsymbol{\eta}(t)$ is the vector consisting of the first p components of the estimator $\hat{\boldsymbol{\beta}}$, which maximize (3.3) w.r.t. $\boldsymbol{\beta}$.

3.2 Notations and models

As has been described in section 1.3, FMCox PH model was proposed to address heterogeneity in survival analysis. The extension to time-dependent covariates with measurement error are discussed in chapter 2. Another alternative to make the FMCox PH model (1.13) more flexible is to allow the coefficients η_k to change over time t. For instance, cancer patients may have different time-efficient prognoses to the same new drug. In this case, identifying disease subtypes and applying the more effective treatment are as the same important as characterizing how the effects of the same treatment changes over time. The extension of FMCox PH model to the FMCox PH model with time-varying coefficients may be needed for more flexible use in medical and survival data analysis.

Let T be the observed event time of a subject with a covariate vector x, of length p. Let $\delta = 0$ or 1 indicate whether or not T is censored. And suppose that there are $K \geq 2$ sub-class with the mixing probabilities $\pi = (\pi_1, ..., \pi_K)$, which have the property $\sum_{i=1}^{K} \pi_i = 1$. Assuming that the hazards are proportional within the given sub-classes of subjects, the subject i in class k follows the Cox PH model with time-varying coefficients:

$$\lambda_{ik}(t) = h_{0k}(t) \exp\left[\boldsymbol{\eta}_k^T(t)\boldsymbol{X}_i(t)\right] \quad (3.4)$$

where $\boldsymbol{\eta}_k(t) = (\eta_{k1}(t), \ldots, \eta_{kp}(t))^T$ is the time-varying coefficient functions for sub-class k. The p-vector of covariates $\boldsymbol{X}_i(t)$ could be both time-dependent and time-independent, but there is no measurement error on the covariates. Note that the baseline hazard function $h_{0k}(t)$ are assumed to be unspecific, positive and continuous, the coefficient functions $\boldsymbol{\eta}_k(t) = (\eta_{k1}(t), \ldots, \eta_{kp}(t))^T$ are assumed to have continuous second derivative in a neighborhood of given time point t.

These observations in sub-class k have the following density function:

$$\begin{aligned} f_k(T_i, \delta_i|\boldsymbol{X}_i) &= \lambda_k(T_i|\boldsymbol{X}_i)^{\delta_i} S(T_i|\boldsymbol{X}_i) \\ &= \{h_{0k}(T_i)\exp[\boldsymbol{\eta}_k^T(T_i)\boldsymbol{X}_i(T_i)]\}^{\delta_i} \cdot \exp\{\int_0^{T_i} h_{0k}(t)\exp[\boldsymbol{\eta}_k^T(t)\boldsymbol{X}_i(t)]dt\}. \end{aligned} \tag{3.5}$$

Therefore, the logarithm of the density function is:

$$\log f_k(T_i, \delta_i|\boldsymbol{X}_i) = \sum_{i=1}^{n}\{\delta_i[\log h_{0k}(T_i) + \boldsymbol{\eta}_k^T(T_i)\boldsymbol{X}_i(T_i)] + \int_0^{T_i} h_{0k}(t)\exp[\boldsymbol{\eta}_k^T(t)\boldsymbol{X}_i(t)]dt\}. \tag{3.6}$$

The mixture likelihood based on the observed data, could be written as

$$L(\boldsymbol{T}, \boldsymbol{\delta} \mid \boldsymbol{X}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k f_k\left(T_i, \delta_i \mid \boldsymbol{X}_i\right).$$

Let the latent variable $c_{ik} = 1$ if subject i belongs to sub-class k, otherwise equals 0, with the probability

$$\mathbb{P}\left\{c_{ik} = 1\right\} = \pi_k \quad \text{and} \quad \sum_k \pi_k = 1.$$

The mixture likelihood based on the complete data, if the latent classification variable c_{ik} is known and let $U = (c_{ik})_{n\times K}$, could be written as:

$$L(\boldsymbol{T}, \boldsymbol{\delta}|\boldsymbol{X}) = \prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k f_k(T_i, \delta_i|\boldsymbol{X}_i)]^{c_{ik}}. \tag{3.7}$$

To obtain the estimators of the time-varying coefficient functions of interest, define $N_i(t) = I\left(T_i \le t, \delta_i = 1\right)$ be the counting process of observed event time for i-th subject, and $Y_i(t) = I\left(T_i \ge t\right)$ be the 'at risk' process, the local partial likelihood method will be used for the proposed model, for sub-class k the partial likelihood could be obtained as follow:

$$PL_k = \prod_{i=1}^{n} Pf_k(T_i, \delta_i|\boldsymbol{X}_i) = \prod_{i=1}^{n}\frac{\exp[\boldsymbol{\eta}_k^T(T_i)\boldsymbol{X}_i(T_i)]}{\sum_{l=1}^{n}\exp[\boldsymbol{\eta}_k^T(T_l)\boldsymbol{X}_l(T_l)]Y_l(T_i)}. \tag{3.8}$$

Therefore, the mixture partial likelihood used for estimating time-varying coefficient functions base on the complete data is:

$$PL = \prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k Pf_k(T_i, \delta_i|\boldsymbol{X}_i)]^{c_{ik}}. \tag{3.9}$$

3.3 Estimation procedure and asymptotic results

3.3.1 Proposed algorithm

In order to maximize the complete mixture partial likelihood function (3.9), the EM iteration is needed here. The logarithm of the mixture partial likelihood function w.r.t the mixing probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$, the baseline hazard functions $\boldsymbol{h} = (h_{01}(t), ..., h_{0K}(t))$ and the time-varying coefficient functions $\boldsymbol{\eta}_k = (\eta_{k1}(t), ..., \eta_{kp}(t))$, $k = 1, ..., K$ could be written as

$$pl_c(\boldsymbol{\pi}, \boldsymbol{h}, \boldsymbol{\eta}; \boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{U} \mid \boldsymbol{X}) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log Pf_k\left(T_i, \delta_i \mid \boldsymbol{X}_i\right). \tag{3.10}$$

The objective Q function used in the EM iteration is

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{U}|\hat{\boldsymbol{\theta}}}[pl_c(\boldsymbol{\theta}; \boldsymbol{T}, \boldsymbol{\delta} \mid \boldsymbol{X}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik} \log Pf_k\left(T_i, \delta_i \mid \boldsymbol{X}_i\right), \end{aligned} \tag{3.11}$$

where $\boldsymbol{\theta}$ represents all the unknown parameters of interest, which includes $\boldsymbol{\pi}, \boldsymbol{h}, \boldsymbol{\eta}$. And the $n \times K$ matrix $\boldsymbol{U} = [c_{ik}]_{n\times K}$ has the elements $c_{ik} = 1$ if the subject i belongs to the sub-class k, otherwise $c_{ik} = 0$.

E-step Given the parameters estimates in the m-th iteration, the $(m + 1)$-th E-step calculates the posterior probability:

$$c_{ik}^{(m+1)} := \mathbb{E}[c_{ik}|T_i, \delta_i, \hat{\boldsymbol{\theta}}^{(m)}] = \frac{\pi_k^{(m)} \times Pf_k\left(T_i, \delta_i \mid \boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^{(m)}\right)}{\sum_{r=1}^{K} \pi_r^{(m)} \times Pf_r\left(T_i, \delta_i \mid \boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^{(m)}\right)}, \tag{3.12}$$

after using the Bayes' theorem.

M-step To update the mixing probabilities π, the Lagrange multipliers technique could be used:

$$\hat{\pi}_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} c_{ik}^{(m+1)}. \tag{3.13}$$

To update the time-varying coefficient $\boldsymbol{\eta}(\cdot)$, the local partial likelihood technique described in section 3.1 will be needed. The logarithm of partial likelihood based on observations over the time interval $[0, \tau]$ for $\tau > 0$ is given by:

$$\begin{aligned} Q_2(\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^{(m+1)} log Pf_k(T_i, \delta_i | \boldsymbol{X}_i, \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^{(m+1)} \int_0^{\tau} [\boldsymbol{X}_i(s)^T \boldsymbol{\eta}_k(s) - log\{\sum_{j=1}^{n} Y_j(s) \exp(\boldsymbol{X}_j(s)^T \boldsymbol{\eta}_k(s))\}] dN_i(s). \end{aligned} \tag{3.14}$$

The assumptions of the baseline hazard functions and time-varying coefficient functions described in section 3.2 will be applied in the following procedures. For any time s in a neighborhood of the given time point t, using the Taylor's expression, the time-varying coefficient function could be re-written as:

$$\eta_{kj} \approx \eta_{kj}(t) + b_{kj}(t)(s - t) \tag{3.15}$$

Let $\boldsymbol{\beta}_k = (\eta_{k1}(t), ..., \eta_{kp}(t), b_{k1}(t), ..., b_{kp}(t))^T$ and $\tilde{\boldsymbol{X}}_i(s, s-t) = \boldsymbol{X}_i(s) \bigotimes (1, s-t)^T$ with $\bigotimes$ being the Kronecker product. And let $h = h_n > 0$ be the bandwidth parameter that controls the size of a local neighborhood and let $K(\cdot)$ be the kernel function which is a symmetric probability density function with support $[1, 1]$, mean 0, and bounded first derivative. The logarithm of local linear partial likelihood function in the $(m + 1)$-th M-step, is:

$$\begin{aligned} \ell(\beta_k) = & \\ & \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^{(m+1)} \int_0^{\tau} K_h(s - t)[\tilde{\boldsymbol{X}}_i(s, s - t)^T \boldsymbol{\beta}_k \\ & - log\{\sum_{j=1}^{n} Y_j(s) \exp(\tilde{\boldsymbol{X}}_i(s, s - t)^T \boldsymbol{\beta}_k)\}] dN_i(s), \end{aligned} \tag{3.16}$$

where $K_h(\cdot) = K(\cdot/h)/h$. Computing such an implicit estimator requires an iterative algorithm such as Newton-Raphson method or Fisher's scoring method. Even worse, for certain given t, there does not exist a local partial likelihood estimator due to the limited amount of data around t. Because of these drawbacks, the one-step estimator (Fan and Chen, 1999) was proposed as a viable alternative.

The local partial likelihood estimator $\hat{\boldsymbol{\beta}}_k$ is found via solving the likelihood equation $\ell'(\boldsymbol{\beta}_k, \tau) = 0$. To facilitate notation, from now on the dependence of $\ell(\boldsymbol{\beta}_k, \tau)$ on τ will be dropped. For a given initial estimator $\hat{\boldsymbol{\beta}}_{0k}$, by Taylor expansion:

$$\ell'(\hat{\boldsymbol{\beta}}_{0k}) + \ell''(\hat{\boldsymbol{\beta}}_{0k})(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_{0k}) \approx 0. \tag{3.17}$$

Thus, the one-step estimator $\hat{\boldsymbol{\beta}}_{OSk}$ is defined as

$$\hat{\boldsymbol{\beta}}_{OSk} = \hat{\boldsymbol{\beta}}_{0k} - \{\ell''(\hat{\boldsymbol{\beta}}_{0k})\}^{-1}\ell'(\hat{\boldsymbol{\beta}}_{0k}). \tag{3.18}$$

However, the choice of given initial estimator $\hat{\boldsymbol{\beta}}_{0k}$ plays an important role on calculating the one-step estimator $\hat{\boldsymbol{\beta}}_{OSk}$, and Cai et al. (2000) provided a useful idea for solving this problem. The basic idea is to compute the local partial likelihood estimates at several given points at first, e.g. computing the local partial likehood estimates at specific grid time points $t_{10}, t_{30}, t_{50}, t_{70}, t_{90}$. Then using these estimators as the initial values of their nearest grid points and calculating the one-step estimators at these grid points. These obtained one-step estimators (e.g. at time points $t_9, t_{11}, t_{29}, t_{31}, \ldots$) will be used as the initial values of their nearest grid points to compute the one-step estimators and so on, until all the one-step estimators at the given time points are obtained.

Therefore, this one-step estimator $\hat{\boldsymbol{\beta}}_{OSk}^{(m+1)}$ is the final update of the time-varying coefficient functions in the $(m+1)$-th iteration.

3.3.2 Estimation of the baseline hazard function

With estimators of time-varying coefficient functions $\eta_k(\cdot)$ obtained from the proposed algorithm, the following estimate for the cumulative baseline hazard function $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u)du$:

$$\hat{\Lambda}_{0k}(t) = \int_0^t \frac{1}{\hat{S}^*_{n,0k}(u)} d\bar{N}_k(u), \tag{3.19}$$

where $\hat{S}^*_{n,0k}(t) = n^{-1}\sum_{i=1}^n \hat{c}_{ik}Y_i(t)\exp(\boldsymbol{X}_i^T(t)\hat{\boldsymbol{\eta}}_k(t))$ and $\bar{N}_k(u) = n^{-1}\sum_{i=1}^n \hat{c}_{ik}I(T_i \leq t, \delta_i = 1)$, for some consistent estimators $\hat{\boldsymbol{\eta}}_k(t)$ of $\boldsymbol{\eta}_k(t)$, $\hat{c}_{ik}$ of c_{ik}. This estimator is an analogue to the estimator (Breslow, 1972) commonly used to estimate the cumulative baseline hazard function in the ordinary Cox PH model. To estimate $\lambda_{0k}(t)$ itself, a

kernel smoothing technique can then be employed here to obtain an estimate via

$$\hat{\lambda}_{0k}(t) = \int K_{h_\lambda}(u-t)d\hat{\Lambda}_{0k}(u) = \frac{1}{nh_\lambda}\sum_{i=1}^{n}\frac{K((T_i-t)/h_\lambda)\delta_i \cdot \hat{c}_{ik}}{\hat{S}^*_{n,0k}(T_i)}, \tag{3.20}$$

where $K_{h_\lambda}(\cdot) = K(\cdot/h_\lambda)/h_\lambda$, $K(\cdot)$ is a given kernel function and h_λ is a given bandwidth.

3.3.3 Concavity of the local linear partial likelihood

In the most of the parametric likelihood theories (Lehmann and Casella, 2006), it is only known that there exists a consistent solution to the local partial likelihood equation. But if there are multiple roots, the consistent estimator may not be found. However, if $\ell(\boldsymbol{\beta}_k)$ is strictly concave, then the solution to (3.16) is unique and must be consistent. To obtain the consistent estimators in every iteration of the proposed algorithm, the concavity of the local linear partial likelihood for each sub-class k is need to be proved.

Theorem 3.3.1. *The estimators in every iteration of the proposed algorithm are the consistent estimators, i.e. the local linear partial likelihood for each sub-class k is convex.*

Proof. Let $G_i(s,\boldsymbol{\beta}_k) = \exp\left(\tilde{\mathbf{X}}_i(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)$ and $G(s,\boldsymbol{\beta}_k) = \sum_{i=1}^{n} Y_i(s)G_i(s,\boldsymbol{\beta}_k)$. Then (3.16) can be re-expressed as follows:

$$\ell(\beta_k) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\int_0^\tau K_h(s-t)[logG_i(s,\beta_k) - logG(s,\beta_k)]dN_i(s),$$

and the first order derivative of $\ell(\boldsymbol{\beta}_k)$ is

$$\ell'(\boldsymbol{\beta}_k) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\int_0^\tau K_h(s-t)[\tilde{\boldsymbol{X}}_i(s,s-t) - \frac{G'(s,\boldsymbol{\beta}_k)}{G(s,\boldsymbol{\beta}_k)}]dN_i(s),$$

where $G'(s,\boldsymbol{\beta}_k) = \sum_{i=1}^{n} Y_i(s)G_i(s,\boldsymbol{\beta}_k)\tilde{\boldsymbol{X}}_i(s,s-t)$. Therefore, the Hessian matrix of

$\ell(\boldsymbol{\beta}_k)$ is given by

$$\begin{aligned}\ell''(\boldsymbol{\beta}_k) &= -\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\int_0^\tau \frac{K_h(s-t)}{G^2(s,\beta_k)}[G(s,\boldsymbol{\beta}_k)\sum_{j=1}^{n} Y_j(s)G_j(s,\boldsymbol{\beta}_k)\tilde{\boldsymbol{X}}_j(s,s-t)^{\otimes 2} \\ &\quad - G'(s,\boldsymbol{\beta}_k)^{\otimes 2}]dN_i(s) \\ &= -\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\int_0^\tau \frac{K_h(s-t)}{G^2(s,\boldsymbol{\beta}_k)}[\sum_{j<l} Y_j(s)Y_l(s)G_j(s,\boldsymbol{\beta}_k)G_l(s,\boldsymbol{\beta}_k) \\ &\quad \cdot (\boldsymbol{X}_j(s)-\boldsymbol{X}_l(s))^{\otimes 2}\otimes\begin{pmatrix} 1 & s-t \\ s-t & (s-t)^2 \end{pmatrix}]dN_i(s),\end{aligned} \tag{3.21}$$

where $\mathbf{A}^{\otimes 2}$ denotes $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ for a vector or matrix $\mathbf{A}$. To prove the right-hand side of (3.21) is negatively definite as $n\to\infty$, we need to prove

$$(\boldsymbol{X}_j(s)-\boldsymbol{X}_l(s))^{\otimes 2}\otimes\begin{pmatrix} 1 & s-t \\ s-t & (s-t)^2 \end{pmatrix}$$

is positive defined. Let any $2p$-vector $d=(d_1^T,d_2^T)^T$ and $\boldsymbol{X}_j(s)-\boldsymbol{X}_l(s)=B$, we have

$$\begin{aligned}&tr[\boldsymbol{d}^T(\boldsymbol{X}_j(s)-\boldsymbol{X}_l(s))^{\otimes 2}\otimes\begin{pmatrix} 1 & s-t \\ s-t & (s-t)^2 \end{pmatrix}\boldsymbol{d}] \\ &= tr[\boldsymbol{d_1^T B B^T d_1}+\boldsymbol{d_2^T B B^T}(s-t)\boldsymbol{d_1}+\boldsymbol{d_1^T B B^T}(s-t)\boldsymbol{d_2}+\boldsymbol{d_2^T B B^T}(s-t)^2\boldsymbol{d_2}].\end{aligned}$$

Let $\boldsymbol{d_1^T B}=e_1$ and $\boldsymbol{d_2^T B}(s-t)=e_2$, it can be written as

$$tr[e_1^2+2e_1e_2+e_2^2]\geq 0.$$

Note that $\boldsymbol{X}_j(s)\neq\boldsymbol{X}_l(s)$, therefore the proof is completed. □

3.3.4 Asymptotic theory

In this section, the large sample properties of the one-step local linear partial likelihood estimators $\hat{\boldsymbol{\eta}}_k$ for each sub-class k would be presented in the following theorems.

Conditions A

(A.1) The kernel function $K(\cdot)$ is a bounded and symmetric density with a bounded support, say, $[-1,1]$.

(A.2) There exists a random vector Y such that $\sup_{s\in\mathcal{N}(t,\epsilon)}|\mathbf{X}(s)|\leq Y$ and

$$\mathbb{E}[\exp\left\{2\left(\sup_{s\in\mathcal{N}(t,\epsilon)}|\boldsymbol{\eta}_k(s)|+\boldsymbol{\eta}_k'(t)+3\right)Y\right\}]<\infty.$$

(A.3) Let $P_k(t \mid \mathbf{x}) = P(T \geq t \mid \mathbf{X}(t) = \mathbf{x}, \mathbf{U} = k)$, define

$$Q_{0k}(t) = \mathbb{E}[P_k(t \mid \mathbf{X}(t))\lambda_k(t \mid \mathbf{X}(t))];$$

$$Q_{1k}(t) = \mathbb{E}[P_k(t \mid \mathbf{X}(t))\lambda_k(t \mid \mathbf{X}(t))\mathbf{X}(t)];$$

$$Q_{2k}(t) = \mathbb{E}[P_k(t \mid \mathbf{X}(t))\lambda_k(t \mid \mathbf{X}(t))\mathbf{X}(t)^{\otimes 2}];$$

$Q_{0k}(s) > 0$, $Q_{1k}(s)$ and $Q_{2k}(s)$ are continuous in the neighborhood $\mathcal{N}(t, \epsilon)$.

(A.4) The sequence $h \to 0$ and $nh \to \infty$ as $n \to \infty$ and $nh^5 = O(1)$

(A.5) The classification indicator variables c_{ik} are unbiased estimated in every iteration of the proposed algorithm.

(A.6) Assume that in a neighbourhood of t, $\lambda_{0k}(s)$ is positive and continuous, $P_k(s|x) > 0$, and coefficient functions $\{\eta_{kj}(s)\}$ have a continuous second derivative.

(A.7) Denote $\mu_j = \int s^j K(s)ds$ and $v_j = \int s^j K^2(s)ds$ for $0 \leq j \leq 2$.

(A.8) Denote $\mathbf{\Sigma}_k(t) = \mathbf{Q}_{2k}(t) - \mathbf{Q}_{1k}(t)\mathbf{Q}_{1k}(t)^{\mathrm{T}}/Q_{0k}(t)$.

Theorem 3.3.2. *under conditions A1 − A8,* $\widehat{\boldsymbol{\eta}}_k(t) \xrightarrow{P} \boldsymbol{\eta}_k(t)$ *as* $n \to \infty$

Lemma 3.3.3. *Let* $c_{nk}(s) = n^{-1}\sum_{i=1}^{n} Y_i(s)c_{ik}g_k(s, X_i(s))$ *and* $c_k(s) = \mathbb{E}[P_k(s|X(s)g_k(s, X(s)))]$ *If* $\sup_{s\in\mathcal{N}(t,\varepsilon)} E\left[g_k^2(s, \mathbf{X}(s))\right] < \infty$, *then*

$$\sup_{s\in\mathcal{N}(t,\varepsilon)} |c_{nk}(s) - c_k(s)| = O_p\left(n^{-1/2}\right)$$

Proof. Let $\boldsymbol{H} = diag\{\boldsymbol{I}_p, h\boldsymbol{I}_p\}$ and $\tilde{\boldsymbol{U}}_i(s, s-t) = \boldsymbol{H}^{-1}\tilde{\boldsymbol{X}}_i(s, s-t)$. Let $\tilde{\boldsymbol{\beta}}_k$ be the running parameter in local linear partial likelihood function (3.16), for the true parameter $\boldsymbol{\beta}_k$, $\hat{\boldsymbol{\beta}}_k$ be the MLE maximizing (3.16). Let $\boldsymbol{\alpha}_k = H(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ and $\hat{\boldsymbol{\alpha}}_k = \boldsymbol{H}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$. Then, by the local linear partial likelihood function (3.16), $\hat{\boldsymbol{\alpha}}_k$ maximizes

$$\begin{aligned}\ell_n(\boldsymbol{\alpha}_k, \tau) = &\int_0^\tau K_h(s-t)n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left[\widetilde{\boldsymbol{X}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k + \tilde{\boldsymbol{U}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k\right] dN_i(s) \\ &- \int_0^\tau K_h(s-t)n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\log\{nS_{n,0k}(\boldsymbol{\alpha}_k, s)\}dN_i(s)\end{aligned} \tag{3.22}$$

w.r.t. α_k, where $S_{n,0k}(\boldsymbol{\alpha}_k, s) = n^{-1}\sum_{i=1}^{n} Y_i(s)\exp\left(\widetilde{\boldsymbol{X}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k + \widetilde{\boldsymbol{U}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k\right)$. It is easy to see that

$$\begin{aligned}\ell_n(\boldsymbol{\alpha}_k, \tau) - \ell_n(0, \tau) = &\int_0^{\tau} K_h(s-t)n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\tilde{\boldsymbol{U}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k dN_i(s) \\ &- \int_0^{\tau} K_h(s-t)n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\log\{\frac{S_{n,0k}(\boldsymbol{\alpha}_k, s)}{S_{n,0k}(0, s)}\}dN_i(s).\end{aligned} \tag{3.23}$$

Let the filtration $\mathscr{F}_{nt}$ be the statistical information accruing during the time $[0, t]$, namely,

$$\mathscr{F}_{nt} = \sigma\left\{X_i(s), N_i(s), Y_i(s), i = 1, \ldots, n, 0 \le s \le t\right\}.$$

Then, under the independent censoring scheme,

$$M_{ik}(t) = N_i(t) - \int_0^t Y_i(s)\lambda_k(s|X_i(s))ds \tag{3.24}$$

is an $\mathscr{F}_{nt}$-martingale. A substitution of (3.24) into (3.23) gives

$$\ell_n(\boldsymbol{\alpha}_k, \tau) - \ell_n(0, \tau) = A_n(\boldsymbol{\alpha}_k, \tau) + R_n(\boldsymbol{\alpha}_k, \tau), \tag{3.25}$$

where

$$A_n(\boldsymbol{\alpha}_k, \tau) = \int_0^{\tau} K_h(s-t)\sum_{k=1}^{K}[S^*_{n,k,1}(s)^T\boldsymbol{\alpha}_k - \log\{\frac{S_{n,0k}(\boldsymbol{\alpha}_k, s)}{S_{n,0k}(0, s)}\}S^*_{n,k,0}(s)]\lambda_{0k}(s)ds, \tag{3.26}$$

where $S^*_{n,k,j}(s) = n^{-1}\sum_{i=1}^{n} c_{ik}Y_i(s)\exp(\boldsymbol{X}_i^T(s)\boldsymbol{\eta}_k(s))\tilde{\boldsymbol{U}}_i^j(s, s-t)$, and

$$R_n(\boldsymbol{\alpha}_k, \tau) = \int_0^{\tau} K_h(s-t)n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}[\tilde{\boldsymbol{U}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k - \log\{\frac{S_{n,0k}(\boldsymbol{\alpha}_k, s)}{S_{n,0k}(0, s)}\}]dM_{ik}(s). \tag{3.27}$$

The proof is straightforward, and omitted, for details, see the proof of lemma 6.1 in Masry and Tjøstheim (1997).

Under condition A1-A8, by lemma 3.3.3, we have:

$$\begin{aligned}A_n(\boldsymbol{\alpha}_k, \tau) &= \int_0^{\tau} K_h(s-t)\sum_{i=1}^{K}[\boldsymbol{S}^*_{k,1}(s)^T\boldsymbol{\alpha}_k - \log\{\frac{S_{0k}(\boldsymbol{\alpha}_k, s)}{S_{0k}(s)}\}S^*_{k,0}(s)]\lambda_{0k}(s)ds + o_p(1) \\ &= \sum_{k=1}^{K}\{\boldsymbol{Q}_{1k}(t)^T \otimes (1, \mu_1)\boldsymbol{\alpha}_k - Q_{0k}(t)\int\log\{\frac{S_k(\boldsymbol{\alpha}_k, t, v)}{Q_{0k}(t)}\}K(v)dv\} + o_p(1)\} \\ &= \sum_{k=1}^{K} A(\boldsymbol{\alpha}_k, \tau) + o_p(1),\end{aligned} \tag{3.28}$$

where

$$S^*_{k,j}(s) = \mathbb{E}[P_k(s|X(s)) \exp(\boldsymbol{X}^T(s)\boldsymbol{a}_k(s))\tilde{\boldsymbol{U}}(s, s-t)^{\otimes j}];$$

$$S_{0k}(\boldsymbol{\alpha}_k, s) = \mathbb{E}[P_k(s|X(s)) \exp(\tilde{\boldsymbol{X}}(s, s-t)^T)\boldsymbol{\beta}_k + \tilde{\boldsymbol{U}}(s, s-t)^T\boldsymbol{\alpha}_k)];$$

$$S_k(\boldsymbol{\alpha}_k, t, v) = \mathbb{E}[P_k(t|X(t))\lambda_k(t|X(t)) \exp(\tilde{\boldsymbol{X}}(t, v))^T\boldsymbol{\alpha}_k].$$

Using the facts that $E\left(\boldsymbol{Y^2Z}\right) - E(\boldsymbol{YZ})^{\boldsymbol{2}} = E\left[(\boldsymbol{Y} - \boldsymbol{E}(\boldsymbol{YZ}))^{\boldsymbol{2}}\boldsymbol{Z}\right] \geq 0$ for $\boldsymbol{Z} \geq 0$ and a matrix $\boldsymbol{B}$ being positive definite is equivalent to $\boldsymbol{a^TBa} \geq 0$ for any column vector $\boldsymbol{a}$ with an appropriate dimension, it is not difficult to check that $A(\boldsymbol{\alpha}_k, \tau)$ is strictly concave, with a maximum at the point $\boldsymbol{\alpha}_k = 0$. The process $R_n(\boldsymbol{\alpha}_k, \cdot)$ is a locally square integrable martingale with the predictable variation process

$$\begin{aligned} C_n(v) &\equiv \langle R_n(\boldsymbol{\alpha}_k, \cdot), R_n(\boldsymbol{\alpha}_k, \cdot)\rangle(v) \\ &= n^{-2}\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^2 \int_0^v K_h^2(s-t)[\tilde{U}_i(s, s-t)^{\mathrm{T}}\alpha_k - \log\{\frac{S_{n,0k}(\alpha_k, s)}{S_{n,0k}(0, s)}\}]^2 Y_i(s)\lambda_k(s|X_i(s))ds. \end{aligned}$$

By condition $A1 - A8$ and lemma 3.3.3, one can show that for any $0 \leq v \leq \tau$ and $0 \leq c_{ik} \leq 1$,

$$E\left[R_n(\boldsymbol{\alpha}_k, v)\right]^2 = E\left[C_n(v)\right] = O\left((nh)^{-1}\right) = o(1).$$

This, in conjunction with (3.25) and (3.28), implies that

$$l_n(\boldsymbol{\alpha}_k, \tau) - l_n(\mathbf{0}, \tau) = \sum_{k=1}^{K} A(\boldsymbol{\alpha}_k, \tau) + o_p(1).$$

Since $\hat{\alpha}_k$ maximizes the concave function $l_n(\boldsymbol{\alpha}_k, \tau) - l_n(\mathbf{0}, \tau)$, by the concavity lemma in appendix II of Andersen and Gill (1982), we have

$$\widehat{\boldsymbol{\alpha}}_k = \mathbf{H}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \xrightarrow{P} 0.$$

This completes the proof of theorem. □

The following notations are needed to construct the asymptotic theorem:

Using the same notation as in the proof of theorem 1, let

$$S_{n,0k}(\boldsymbol{\alpha}_k, s) = n^{-1}\sum_{i=1}^{n} Y_i(s) \exp\left(\widetilde{\mathbf{X}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k + \widetilde{\mathbf{U}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k\right)$$

$$S_{0k}(\boldsymbol{\alpha}_k, s) = E\left[P_k(s|\mathbf{X}(s)) \exp\left(\widetilde{\mathbf{X}}(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k + \widetilde{\mathbf{U}}(s, s-t)^{\mathrm{T}}\boldsymbol{\alpha}_k\right)\right]$$

For $j = 0$ and 1, set

$$\boldsymbol{S}^*_{n,jk}(s) = n^{-1}\sum_{i=1}^{n} Y_i(s)\exp\left(\mathbf{X}_i(s)^{\mathrm{T}}\eta_k(s)\right)\widetilde{\mathbf{U}}_i^j(s, s-t),$$

and

$$\boldsymbol{S}^*_{jk}(s) = E\left[P_k(s|\mathbf{X}(s))\exp\left(\mathbf{X}(s)^{\mathrm{T}}\eta_k(s)\right)\widetilde{\mathbf{U}}(s, s-t)^{\otimes j}\right].$$

For $0 \le j \le 2$, let

$$\boldsymbol{S}_{n,jk}(s) = n^{-1}\sum_{i=1}^{n} Y_i(s)\exp\left(\widetilde{\mathbf{X}}_i(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\widetilde{\mathbf{U}}_i(s, s-t)^{\otimes j},$$

and

$$\boldsymbol{S}_{jk}(s) = E\left[P_k(s|\mathbf{X}(s))\exp\left(\widetilde{\mathbf{X}}(s, s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\tilde{\mathbf{U}}(s, s-t)^{\otimes j}\right].$$

Note that $\boldsymbol{S^*_{n,1k}(s)}, \boldsymbol{S^*_{1k}(s)}, \boldsymbol{S_{n,1k}(s)}$ and $\boldsymbol{S_{1k}(s)}$ are $2p$-vectors, $\boldsymbol{S_{n,2k}(s)}$ and $\boldsymbol{S_{2k}(s)}$ are $2p \times 2p$ matrices and the rest are scalar.

Motivated from the theorem 2 proposed by Cai and Sun (2003), the asymptotical optimal bandwidth depends on the unknown parameters $\boldsymbol{\Sigma}_k(t)$ and $\boldsymbol{\eta}''_k(t)$, we propose the following theorem:

Theorem 3.3.4. *(Asymptotic normality) under conditions $A1 - A8$, when t is the interior point of $[0, \tau]$, we have,*

$$\sqrt{nh}\left[\widehat{\boldsymbol{\eta}}_k(t) - \boldsymbol{\eta}_k(t) - \frac{h^2}{2}\mu_2\boldsymbol{\eta}''_k(t)\right] \xrightarrow{D} N\left\{0, v_0\boldsymbol{\Sigma}_k^{-1}(t)\right\}$$

as $n \to \infty$.

Proof. Let $\gamma_n = (nh)^{(-1/2)}$, $\tilde{\boldsymbol{\beta}}_k$ be the running parameter in local linear partial likelihood function (3.16), for the true parameter $\boldsymbol{\beta}_k$, $\hat{\boldsymbol{\beta}}_k$ be the MLE maximizing (3.16). Define $\boldsymbol{\alpha}_k == \gamma_n^{-1}\mathbf{H}(\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$. Then $\tilde{\boldsymbol{\beta}}_k = \gamma_n\mathbf{H}^{-1}\boldsymbol{\alpha}_k + \boldsymbol{\beta}_k$, and by (3.25), it is easy to obtain that

$$l_n\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) - l_n(\mathbf{0}, \tau) = A_n\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) + R_n\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right),$$

where $A_n(\cdot, \tau)$ and $R_n(\cdot, \tau)$ are defined in (3.26) and (3.27), respectively. By Taylor's expansion at $\alpha_k = 0$, it follows that

$$\log\left\{\frac{S_{n,0k}\left(\gamma_n\boldsymbol{\alpha}_k, s\right)}{S_{n,0k}(0, s)}\right\} = \frac{S_{n,1k}(s)^{\mathrm{T}}\gamma_n\boldsymbol{\alpha}_k}{S_{n,0k}(0, s)} + \frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}\left[\frac{\mathbf{S}_{n,2k}(s)}{S_{n,0k}(0, s)} - \frac{S_{n,1k}(s)^{\otimes 2}}{S^2_{n,0k}(0, s)}\right]\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right) \tag{3.29}$$

Under condition A, by lemma 3.3.2, for $|s-t| < ch$, (3.29) becomes

$$\log\left\{\frac{S_{n,0k}\left(\gamma_n\boldsymbol{\alpha}_k, s\right)}{S_{n,0k}(0,s)}\right\} = \frac{S_{1k}(s)^{\mathrm{T}}\gamma_n\boldsymbol{\alpha}_k}{S_{0k}(0,s)} + \frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}\left[\frac{\mathbf{S}_{2k}(s)}{S_{0k}(0,s)} - \frac{S_{1k}(s)^{\otimes 2}}{S_{0k}^2(0,s)}\right]\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right) \tag{3.30}$$

as $n \to \infty$. Substituting (3.30) into $A_n\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right)$ given in (3.26) and applying lemma 3.3.2 to $S_{n,jk}^*(s)$ for $j = 0, 1$, we can obtain

$$A_{nk}\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) = \gamma_n A_{n,1k}(\tau)^{\mathrm{T}}\boldsymbol{\alpha}_k - \frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}F_{n,1k}(\tau)\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right),$$

where

$$A_{n,1k}(\tau) = \int_0^\tau K_h(s-t)\left[S_{1k}^*(s) - \frac{S_{1k}(s)}{S_{0k}(s)}S_{0k}^*(s)\right]\lambda_{0k}(s)ds,$$

and

$$F_{n,1k}(\tau) = \int_0^\tau K_h(s-t)\left[\frac{S_{2k}(s)}{S_{0k}(s)} - \frac{S_{1k}(s)^{\otimes 2}}{S_{0k}^2(s)}\right]S_{0k}^*(s)\lambda_{0k}(s)ds.$$

It follows from condition A1-A8 and theorem 1 in Sun (1984) that, for each $\tau > 0$, as $n \to \infty$,

$$F_{n,1k}(\tau) - \boldsymbol{\Sigma}_k(t) \otimes \Omega = o_p(1)$$

where $\boldsymbol{\Omega} = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}$. Therefore,

$$A_{nk}\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) = \gamma_n A_{n,1k}(\tau)^{\mathrm{T}}\boldsymbol{\alpha}_k - \frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k(t) \otimes \Omega\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right), \tag{3.31}$$

Similarly, substituting (3.30) into $R_{nk}\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right)$ given in (3.27), we have

$$R_{nk}\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) = \gamma_n R_{n,1k}(\tau)^{\mathrm{T}}\boldsymbol{\alpha}_k - \frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}F_{n,2k}(\tau)\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right),$$

where

$$R_{n,1k}(\tau) = \int_0^\tau K_h(s-t)n^{-1}\sum_{i=1}^n\left[\widetilde{\mathbf{U}}_i(s, s-t) - \frac{S_{n,1k}(s)}{S_{n,0k}(s)}\right]dM_{ik}(s),$$

and

$$F_{n,2k}(\tau) = \int_0^\tau K_h(s-t)\left[\frac{S_{2k}(s)}{S_{0k}(s)} - \frac{S_{1k}(s)^{\otimes 2}}{S_{0k}^2(s)}\right]d\bar{M}_k(s)$$

with $\bar{M}_k(t) = (1/n)\sum_{i=1}^n M_{ik}(t)$. By considering the second moment of $F_{n,2k}(\tau)$ and some simple analysis, we have $F_{n,2k}(\tau) = O_p\left(\gamma_n\right)$. Therefore,

$$R_{nk}\left(\gamma_n\boldsymbol{\alpha}_k, \tau\right) = \gamma_n R_{n,1k}(\tau)^{\mathrm{T}}\boldsymbol{\alpha}_k + o_p\left(\gamma_n^2\right).$$

This, in conjunction with (3.25) and (3.31), implies that

$$l_n\left(\gamma_n\boldsymbol{\alpha}_k,\tau\right)-l_n(\mathbf{0},\tau)=\left[A_{n,1k}(\tau)+R_{n,1k}(\tau)\right]^{\mathrm{T}}\gamma_n\boldsymbol{\alpha}_k-\frac{1}{2}\gamma_n^2\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}_k(t)\otimes\Omega\boldsymbol{\alpha}_k+o_p\left(\gamma_n^2\right).$$

Now, let $\widehat{\boldsymbol{\alpha}}_k=\gamma_n^{-1}\mathbf{H}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)$. Then $\widehat{\boldsymbol{\alpha}}_k$ maximizes $l_n\left(\gamma_n\boldsymbol{\alpha}_k,\tau\right)$ w.r.t α_k. By the quadratic approximation lemma (Fan and Gijbels, 2018), we can obtain

$$\widehat{\boldsymbol{\alpha}}_k=\gamma_n^{-1}(\boldsymbol{\Sigma}_k(t)\otimes\Omega)^{-1}\left[A_{n,1k}(\tau)+R_{n,1k}(\tau)\right]+o_p(1) \tag{3.32}$$

Define

$$\widetilde{S}_{1k}(s)=E\left[P(s|\mathbf{X}(s))\exp\left(\widetilde{\mathbf{X}}(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\mathbf{X}(s)\right],$$

and

$$\widetilde{S}_{n,1k}(s)=n^{-1}\sum_{i=1}^{n}Y_i(s)\exp\left(\widetilde{\mathbf{X}}_i(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\mathbf{X}_i(s).$$

Let

$$A_{n,1k}^*(\tau)=\int_0^\tau K_h(s-t)\left[\mathbf{Q}_{1k}(s)-\widetilde{S}_{1k}(s)\lambda_{0k}(s)\frac{S_{0k}^*(s)}{S_{0k}(s)}\right]ds, \tag{3.33}$$

and

$$R_{n,1k}^*(\tau)=\int_0^\tau K_h(s-t)n^{-1}\sum_{i=1}^{n}\left[\mathbf{X}_i(s)-\frac{\widetilde{S}_{n,1k}(s)}{S_{n,0k}(s)}\right]dM_{ik}(s).$$

Since $(\boldsymbol{\Sigma}_k(t)\otimes\Omega)^{-1}=\boldsymbol{\Sigma}_k(t)^{-1}\otimes\Omega^{-1}$, the first p-components of (3.32) yields

$$\gamma_n^{-1}(\widehat{\mathbf{a}}_k(t)-\mathbf{a}_k(t))=\gamma_n^{-1}\boldsymbol{\Sigma}_k^{-1}(t)\left[A_{n,1k}^*(\tau)+R_{n,1k}^*(\tau)\right]+o_p(1). \tag{3.34}$$

Next, we apply the Taylor expansion to the term $\mathbf{Q}_{1k}(s)-\widetilde{S}_{1k}(s)\lambda_{0k}(s)S_{0k}^*(s)/S_{0k}(s)$ in (3.33) around t to calculate the bias for $\widehat{\mathbf{a}}_k(t)$. Note that $\tilde{\mathbf{X}}(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k=\eta_k(t)^{\mathrm{T}}\mathbf{X}(s)+(s-t)\eta_k'(t)^{\mathrm{T}}\mathbf{X}(s)$ and

$$\mathbf{Q}_{1k}(s)-\widetilde{S}_{1k}(s)\lambda_{0k}(s)=E\left[P_k(s|\mathbf{X}(s))\lambda_{0k}(s)\mathbf{X}(s)\left(\exp\left(\eta_k(s)^{\mathrm{T}}\mathbf{X}(s)\right)-\exp\left(\tilde{\mathbf{X}}(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right)\right].$$

For $|u-t|<h$,

$$\eta_k(s)^{\mathrm{T}}\mathbf{X}(s)=\eta_k(t)^{\mathrm{T}}\mathbf{X}(s)+(s-t)\eta_k'(t)^{\mathrm{T}}\mathbf{X}(s)+\frac{1}{2}(s-t)^2\eta_k''(t)^{\mathrm{T}}\mathbf{X}(s)+o_p\left((s-t)^2\right).$$

Thus, by condition A1-A8,

$$\begin{aligned}&\mathbf{Q}_{1k}(s)-\widetilde{S}_{1k}(s)\lambda_{0k}(s)\\&=E\left[P_k(s|\mathbf{X}(s))\lambda_{0k}(s)\exp\left(\eta_k^{\mathrm{T}}(s)\mathbf{X}(s)\right)\tfrac{1}{2}(s-t)^2\mathbf{X}(s)\mathbf{X}(s)^{\mathrm{T}}\eta_k''(t)\right]+o\left((s-t)^2\right)\\&=\tfrac{1}{2}(s-t)^2\mathbf{Q}_{2k}(s)\eta_k''(t)+o\left(h^2\right)\end{aligned}$$

Similarly,

$$\begin{aligned} S_{0k}^*(s) - S_{0k}(s) =& E\left[P_k(s|\mathbf{X}(s))\left(\exp\left(\eta_k^{\mathrm{T}}(s)\mathbf{X}(s)\right) - \exp\left(\widetilde{\mathbf{X}}(s,s-t)^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right)\right], \\ =& \frac{1}{2}(s-t)^2\mathbf{Q}_{1k}(s)^{\mathrm{T}}\eta_k''(t)/\lambda_{0k}(s) + o\left(h^2\right) \end{aligned}$$

and

$$\left(S_{0k}^*(s) - S_{0k}(s)\right)/S_{0k}(s) = \frac{1}{2}(s-t)^2\mathbf{Q}_{1k}(s)^{\mathrm{T}}\eta_k''(t)/Q_{0k}(s) + o\left(h^2\right).$$

Therefore,

$$\mathbf{Q}_{1k}(s) - \widetilde{S}_{1k}(s)\lambda_{0k}(s)\frac{S_{0k}^*(s)}{S_{0k}(s)} = \frac{1}{2}(s-t)^2\boldsymbol{\Sigma}_k(s)\eta_k''(t) + o_p\left(h^2\right). \tag{3.35}$$

Plugging (3.35) in the expression for $A_{n,1k}^*(\tau)$ given in (3.33), (3.34) becomes

$$\gamma_n^{-1}\left[\widehat{\eta}_k(t) - \eta_k(t) - \frac{h^2}{2}\mu_2\eta_k''(t)\right] = \gamma_n^{-1}\boldsymbol{\Sigma}_k^{-1}(t)X_{n,1k}^*(\tau) + o_p(1),$$

so that

$$\sqrt{nh}\left[\widehat{\eta}_k(t) - \eta_k(t) - \frac{h^2}{2}\mu_2\eta_k''(t)\right] = \boldsymbol{\Sigma}_k^{-1}(t)\sqrt{nh}X_{n,1k}^*(\tau) + o_p(1).$$

Finally, the process $U_{nk}^*(v) = \sqrt{nh}X_{n,1k}^*(v)$ is a locally square integrable martingale with the predictable variation process

$$\left\langle U_{nk}^*, U_{nk}^*\right\rangle(v) = n^{-1}h\sum_{i=1}^n\int_0^v K_h^2(s-t)\left[\mathbf{X}_i(s) - \frac{\widetilde{S}_{n,1k}(s)}{S_{n,0k}(s)}\right]^{\otimes 2} Y_i(s)\lambda_k\left(s|\mathbf{X}_i(s)\right)ds.$$

By lemma 3.3.3, we can show that

$$\left\langle U_{nk}^*, U_{nk}^*\right\rangle(v) = \int K^2(s)ds\left[Q_{0k}(t)\mathbf{Q}_{2k}(t) - \mathbf{Q}_{1k}(t)^{\otimes 2}\right]/Q_{0k}(t) + o_p(1) = v_0\boldsymbol{\Sigma}_k(t) + o_p(1).$$

Write the lth element of the vector $U_{nk}^*(v)$ as

$$\frac{\sqrt{nh}}{n}\sum_{i=1}^n\int_0^t K_h(s-t)H_{n,i,l}(s)dM_{ik}(s),$$

To prove the asymptotic normality, we need to check the Lindeberg condition

$$\sum_{i=1}^n\int_0^v n^{-1}hK_h^2(s-t)H_{n,i,l}^2(s)I\left\{\sqrt{h/n}K_h(s-t)\left|H_{n,i,l}(s)\right| > \varepsilon\right\}Y_i(s)\lambda_k\left(s|\mathbf{X}_i(s)\right)ds \xrightarrow{P} 0$$

for all $\varepsilon > 0$. The last statement is valid by condition A1-A8 and lemma 3.3.3. This establishes that

$$\sqrt{nh}X_{n,1k}^*(v) \xrightarrow{D} N\left\{0, v_0\boldsymbol{\Sigma}_k(t)\right\}, \quad 0 \le v \le \tau.$$

Therefore,

$$\sqrt{nh}\left[\widehat{\eta}_k(t) - \eta_k(t) - \frac{h^2}{2}\mu_2\eta_k''(t)\right] \xrightarrow{D} N\left\{0, v_0\boldsymbol{\Sigma}_k^{-1}(t)\right\}.$$

The proof of the theorem is complete. □

3.4 Simulation

In this section, the simulation studies show that the proposed estimation procedures for FMCox PH models with time-varying coefficients are reliable and useful. The Epanechnikov kernel function $K(s) = 0.75(1 - s^2)_+$ is used for all the following examples. And the performance of the estimated time-varying coefficients $\hat{\eta}_{jk}(\cdot)$ could be assessed by the mean absolute deviation (MAD)

$$\text{MAD}_{jk} = n^{-1} \sum_{l=1}^{n} \left| \widehat{\eta}_{jk}(t_l) - \eta_{jk}(t_l) \right|, \quad j = 1, \ldots, p,$$

where $\{t_l, l = 1, \ldots, n\}$ are the grid points at which $\eta_{jk}(\cdot)$ are estimated.

3.4.1 Simulation study scenario 1

Motivated by the simulation study proposed by Eng and Hanlon (2014), the main objective is estimating the effect of the time-independent covariate x. To compare the performance of the proposed FMCox PH models with time-varying coefficients and the FMCox PH models in Eng and Hanlon (2014), scenario 1 focuses on the situation of a single time-independent covariate x with the constant coefficients $\eta_k(t) = constant$ of interest.

Assume that the number of latent classes $K = 2$, all the $2n$ observations have a single covariates $(x_1, ..., x_{2n}) \sim N(\mu \cdot \mathbf{1}_{2n}, \boldsymbol{I}_{2n})$. For the two different latent classes, we consider the models:

$$\lambda_{ik}(t) = h_{0k}(t) exp\{\eta_k(t) x_i\},$$

where $h_{0k}(t) = 1$, the first class has $\eta_1(t) = 3$ and the second class has $\eta_2(t) = -3$. Hence, the survival time and censoring time for the ith subject is generated through the same technique described in Section 2.3. Set $n = 500$ subjects in each class and to target 40% censoring rate, the mean value of covariates is $\mu = 0$ and censoring event time is generated from $U(0, exp(0.99))$. Therefore, the Kaplan-Meier survival curve for the simulated event time could be plotted in Figure 3.1:

Table 3.1: Simulation results for MAD based on 100 replicates for scenario 1

40% censoring					
n	h	$Av.MAD_1$	$Std.MAD_1$	$Av.MAD_2$	$Std.MAD_2$
500	0.6	0.6732	0.5315	0.5348	0.4628
750	0.6	0.5861	0.4127	0.4359	0.3376
1000	0.6	0.5076	0.3894	0.4318	0.3148
Av. classification accuracy: 87%					

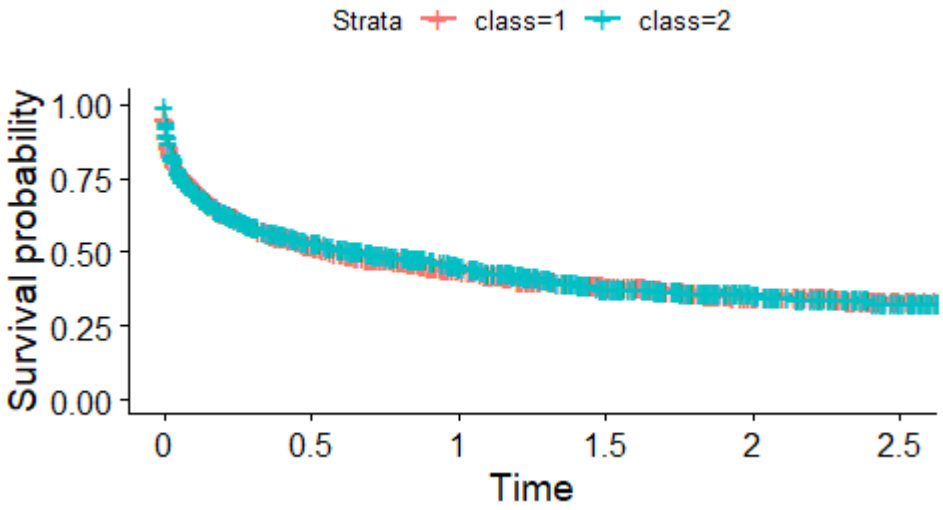

Figure 3.1: the Kaplan-Meier survival curves for the given latent classes in the simplest simulation study with the mean of the time-independent covariate $\mu = 0$.

In both models, we restrict to the estimation of the time-varying coefficients on the time interval $[0, 2.5]$. In this simulation study, the choice of bandwidth is $h = 0.6$ and the choice of the grid points are $0.005 + 0.25l, l = 0, 1, ..., 9$. The simulation results for MAD are showed in Table 3.1, and the average classification accuracy is 87%. These results show that the proposed procedure is reliable. Figure 3.2 and Figure 3.3 display the estimate of the time-varying coefficients for the two given latent class $\eta_1(t)$ and $\eta_2(t)$ based on the sample data.

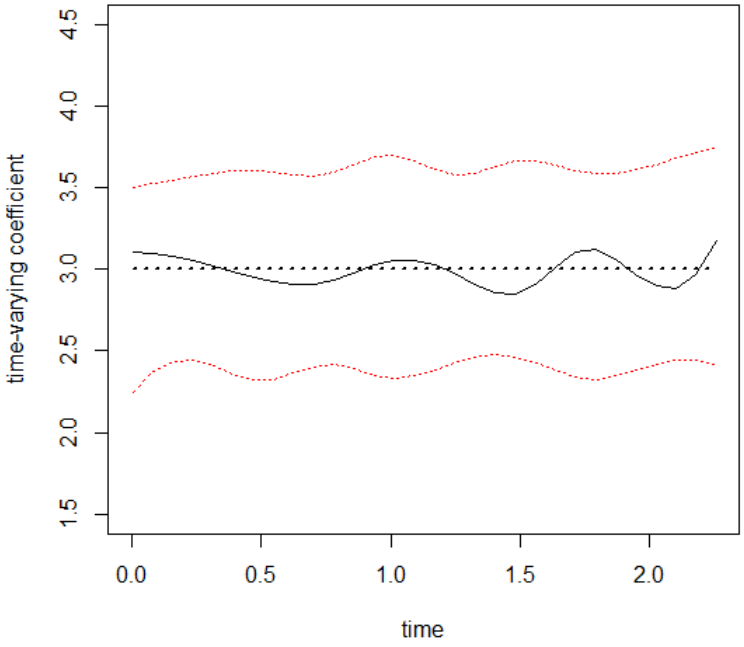

Figure 3.2: estimation of $\eta_1(t)$ for the simulated models

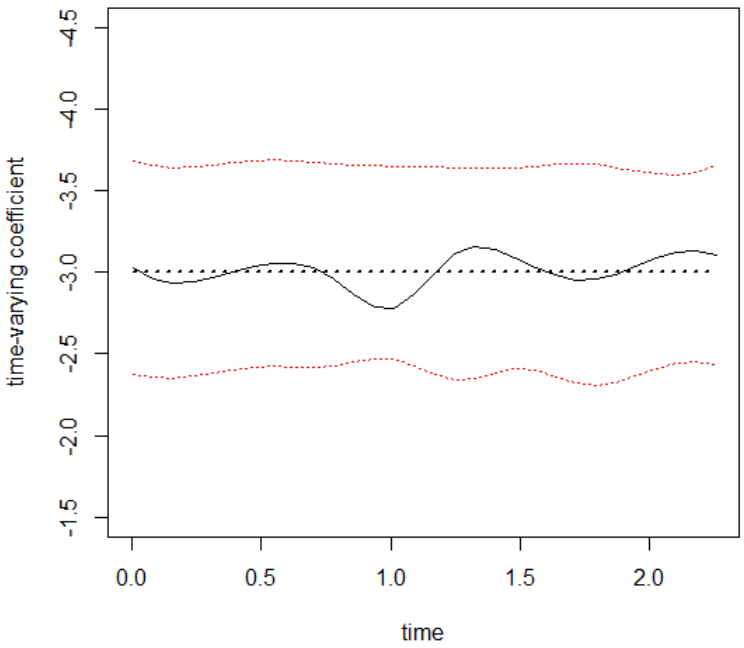

Figure 3.3: estimation of $\eta_2(t)$ for the simulated models

3.4.2 Simulation study scenario 2

In this scenario, we evaluate the performance of the proposed FMCox PH model with time-varying for the non-constant coefficients. Assume that the number of latent classes $K = 2$, all the $2n$ observations have a single covariates $(x_1, ..., x_{2n}) \sim N(\mu \cdot \mathbf{1}_{2n}, \boldsymbol{I}_{2n})$. For the two different latent classes, we consider the models:

$$\lambda_{ik}(t) = h_{0k}(t)exp\{\eta_k(t)x_i\},$$

where $h_{0k}(t) = 1$, the first class has $\eta_1(t) = 4-0.5t$ and the second class has $\eta_2(t) = -1$. Hence, the survival time and censoring time for the ith subject is generated through the same technique described in Section 2.3. Set $n = 500$ subjects in each class and to target 30% censoring rate, the mean value of covariates is $\mu = -0.1$ and censoring event time is generated from $U(0, exp(1.6))$. Therefore, the Kaplan-Meier survival curve for the simulated event time could be plotted as follow:

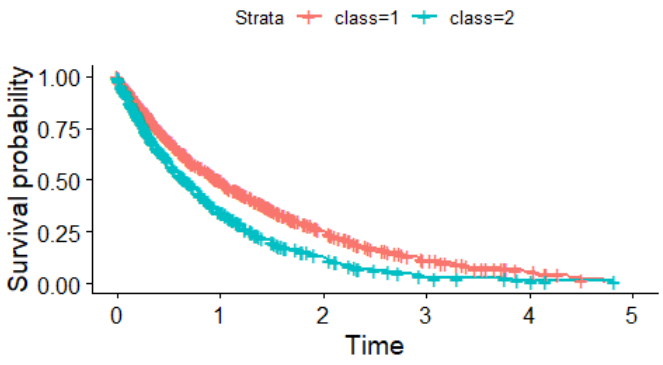

Figure 3.4: the Kaplan-Meier survival curves for the given latent classes in the given scenario with the mean of the time-independent covariate $\mu = -0.1$.

In both models, we restrict to the estimation of the time-varying coefficients on the time interval $[0, 5]$. In this simulation study, the choice of bandwidth is $h = 1$ and the choice of the grid points are $0.005 + 0.5l, l = 0, 1, ..., 9$. The simulation results for MAD are presented in Table 3.2, and the average classification accuracy is 82%. These results show that the proposed procedure for the proposed FMCox PH model with time-varying for non-constant coefficients, is reliable. Figure 3.5 and Figure 3.6 display the estimate of the time-varying coefficients for the two latent class η_1 and η_2 based on the sample data for this simulation study.

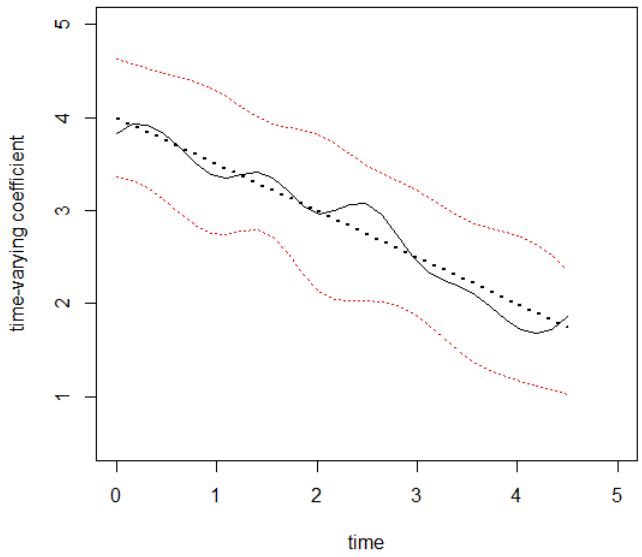

Figure 3.5: estimation of $\eta_1(t)$ for the simulated models

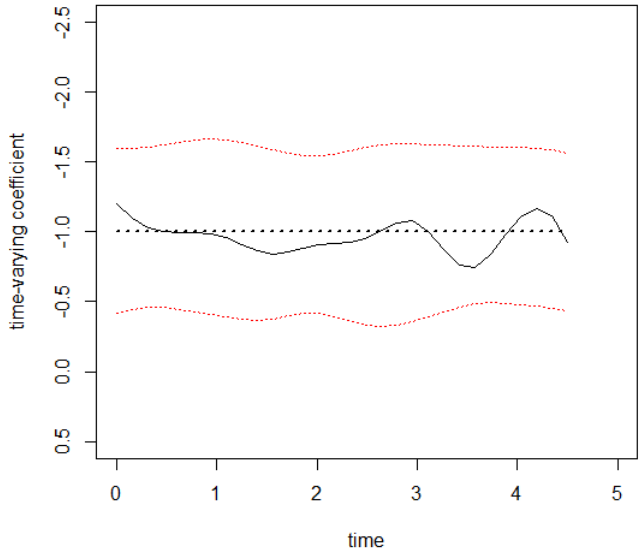

Figure 3.6: estimation of $\eta_2(t)$ for the simulated models

Table 3.2: Simulation results for MAD based on 100 replicates for scenario 2

25% censoring					
n	h	$Av.MAD_1$	$Std.MAD_1$	$Av.MAD_2$	$Std.MAD_2$
500	1	0.7459	0.6568	0.5734	0.6973
750	1	0.7045	0.5676	0.5475	0.5475
1000	1	0.6784	0.4865	0.4967	0.4367
Av. classification accuracy: 82%					

3.5 Real data analysis

We will illustrate the proposed method by applying the procedure to analysis a real data set from the study of AIDS (Abrams et al., 1994). And in this study, there are 467 patients infected with HIV, and they have received two different treatments, Zalcitabine (ddC) or Didanosine (ddI), randomly. The main interest focuses on how the treatment effects the risk of infection or death.

For simplicity, some prior knowledge (Liu et al., 2015; Roustaei et al., 2018) were used for our real data analysis directly, such as there are $K = 2$ latent classes, and the main time-independent covariate of interest is the treatment. The baseline hazard functions were assumed to be piecewise-constant functions. The objective is estimating the coefficients $\eta_k(\cdot)$ for time-independent covariate treatment. Recall the survival sub-model:

$$\lambda_{ik}(u) = h_{0k}(u) \exp\{\eta_k(u) \times treatment_i\}, \tag{3.36}$$

with the latent classification variable

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ belongs to sub-class 1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbb{P}\{c_{ik} = 1\} = \pi_k \quad \text{and} \quad \sum_{k}^{2} \pi_k = 1.$$

The estimation results of the time-varying coefficients for the two latent class $\eta_1(t)$ and $\eta_2(t)$ are presented in Figure 3.7 and Figure 3.8. And the Kaplan-Meier survival curves for the two latent classes in this AIDS study could be plotted in Figure 3.9. Note that the estimation of the treatment effects in the two latent classes are negative at any time point, which means that taking the given treatment will decrease the risk of infection or death. And the given treatment is more effective for the patients in the

second latent class. From the Figure 3.9, the similar results could be obtained, i.e. the patients in the second latent class have the higher survival probability.

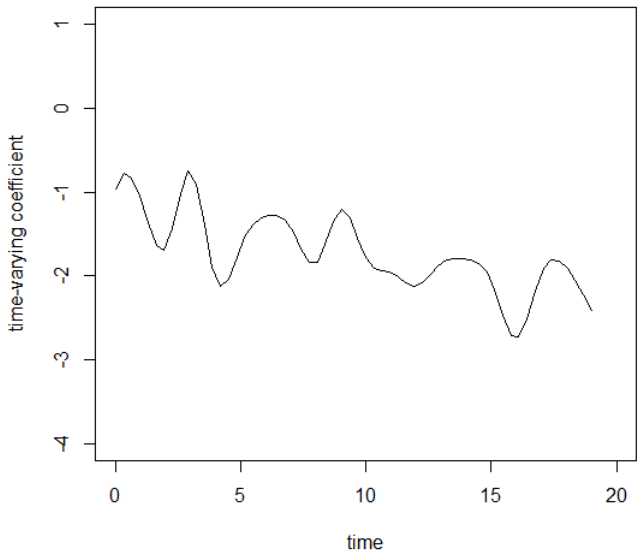

Figure 3.7: estimation of $\eta_1(t)$ for AIDS data

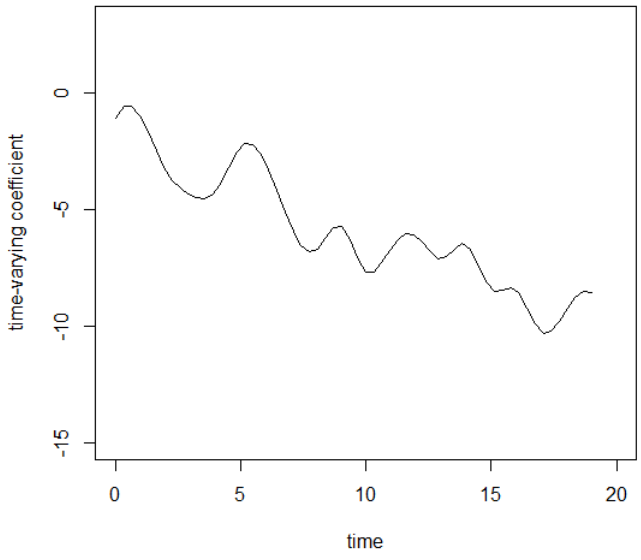

Figure 3.8: estimation of $\eta_2(t)$ for AIDS data

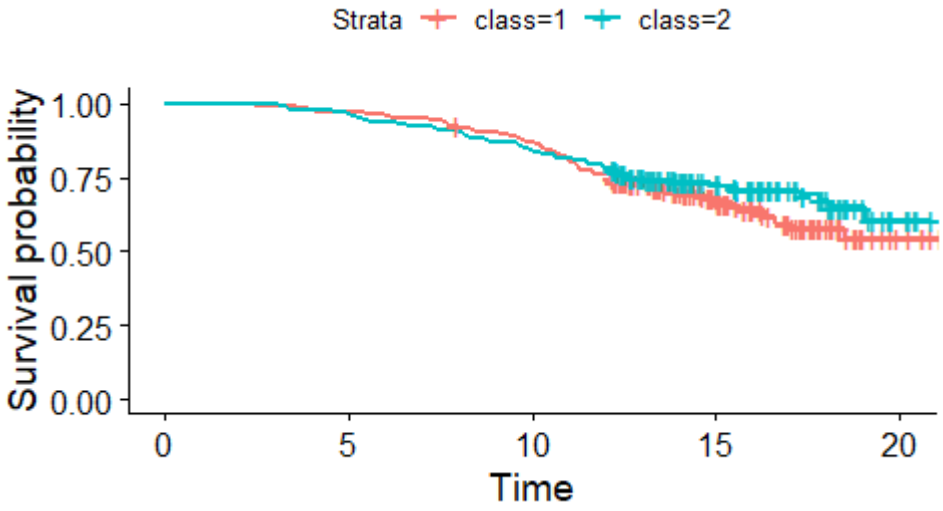

Figure 3.9: the Kaplan-Meier survival curves for the two latent classes in the AIDS study.

3.6 Discussion

In this chapter, we extend the FMCox PH models to characterize the effects over time of the covariates of interest, and proposed the FMCox PH models with time-varying coefficients. The local partial likelihood technique has been used for the estimation procedures, and the drawbacks of a local partial likelihood estimator due to the limited amount of data around the given time point t, could be overcome through the one-step estimator used in the EM iterations. The concavity of the local linear partial likelihood and the asymptotic normality of the local partial likelihood estimators across every sub-class have also been provided in this chapter.

To assess the performance of the FMCox PH models with time-varying coefficients and the corresponding proposed estimation procedures, two simulation studies have been proposed in this chapter. Under the scenario 1, the performance of the FMCox PH models with time-varying coefficients is quite reliable, compared with the performance of FMCox PH models. Under the scenario 2, the performance of the FMCox PH models with time-varying coefficients show that this proposed models and the corresponding proposed estimation procedures could provide the reasonable estimators, considering non-constant coefficients. And we also finish the real data analysis, which

means the proposed models could handle the clinical trial data analysis and provide the explanatory estimators.

However, there are many limitations in this work, such as the lack of the choice of the number of latent classes, more than one time-independent covariates simulations and time-dependent covariates without measurement error simulations. In this chapter, we didn't consider the situations where the number of latent classes is not known, but it is common in practice. Therefore, some choice criteria e.g. AIC, BIC used in standard joint modeling, need to be proposed and assess the corresponding performance in the future work. Therefore, more complex simulation, e.g. including more than one time-independent covariates or time-dependent covariates without measurement error, will also be needed in the future work.

Chapter 4

Conclusion and Discussion

In medical studies, most common diseases including cancer are heterogeneous that they vary in etiology, pathogensis, and prognosis, which we have limited knowledge. Some longitudinal biomarkers include important information from the past history and provide feedback to the future events. It is frequent to collect both repeated measures of longitudinal processes and the time to an event of interest simultaneously. The existing literature considers heterogeneity of survival data analysis under the FMCox PH models, joint analysis of longitudinal and time-to-event data under the standard joint models.

It is natural to extend the FMCox PH model to the case with multiple longitudinal covariates measured with error and to the case which could characterize the effects over time of the covariates of interest. Therefore, the joint FMCox PH models and FMCox PH models with time-varying coefficients have been proposed in chapter 2 and 3, respectively. And the conditional score method is proposed in EM iteration for the joint FMCox PH models, which does not need any assumption on the distribution of the random effects, the estimation procedures lead to consistent estimators for the parameters of interest. To assess the performance of the joint FMCox PH models, two simulation studies have been proposed in chapter 2. Under the scenario 1, the performance of the joint FMCox PH models is similar to the performance of FMCox PH models, which means that the joint FMCox PH models could replace FMCox PH models to handle the heterogeneity in survival data analysis. Under the scenario 2, the performance of the joint FMCox PH models show that this proposed models and the corresponding estimation procedures could give the reasonable estimators,

considering the heterogeneity in longitudinal and time-to-event data analysis. And the local partial likelihood technique has been used for the estimation procedures for FMCox PH models with time-varying coefficients, and the drawbacks of a local partial likelihood estimator due to the limited amount of data around the given time point, could be overcome through the one-step estimator used in the EM iterations. The concavity of the local linear partial likelihood and the asymptotic normality of the local partial likelihood estimators across every sub-class have also been provided in chapter 3. To assess the performance of the FMCox PH models with time-varying coefficients and the corresponding proposed estimation procedures, two simulation studies have been proposed in chapter 3. Under the scenario 1, the performance of the FMCox PH models with time-varying coefficients is quite reliable, compared with the performance of FMCox PH models. Under the scenario 2, the performance of the FMCox PH models with time-varying coefficients show that this proposed models and the corresponding proposed estimation procedures could provide the reasonable estimators, considering non-constant coefficients. And we also finish the real data analysis, which means the proposed models could handle the clinical trial data analysis and provide the explanatory estimators.

However, there are many limitations in this thesis. In these two works, we didn't consider the situation where the number of latent classes is not known, but it is common in practice. Therefore, some choice criteria e.g. AIC, BIC used in standard joint modeling, need to be proposed and assess the corresponding performance in the future work. And in both two projects, the simulation studies are limited, e.g. just including one time-independent covariate and time-dependent covariate in chapter 2, no time-dependent covariate without measurement error in chapter 3. Therefore, more simulation studies and real data analysis will be finished in the future work.

References

Odd Aalen. Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, pages 15–27, 1976.

Donald I Abrams, Anne I Goldman, Cynthia Launer, Joyce A Korvick, James D Neaton, Lawrence R Crane, Michael Grodesky, Steven Wakefield, Katherine Muth, Sandra Kornegay, et al. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, 330(10):657–662, 1994.

Bernard Altshuler. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11, 1970.

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*, 31(29):3946–3958, 2012.

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

N Breslow. Discussion of dr. cox's paper" regression models and life tables.". *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.

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

Samuel L Brilleman, Michael J Crowther, Margaret T May, Mark Gompels, and Keith R Abrams. Joint longitudinal hurdle and time-to-event models: an application

related to viral load and duration of the first treatment regimen in patients with hiv initiating therapy. *Statistics in medicine*, 35(20):3583–3594, 2016.

Chiara Brombin, Clelia Di Serio, and Paola MV Rancoita. Joint modeling of hiv data in multicenter observational studies: A comparison among different approaches. *Statistical methods in medical research*, 25(6):2472–2487, 2016.

Zongwu Cai and Yanqing Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003.

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

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Jean-françois Dupuy and Mounir Mesbah. Joint modeling of event time and nonignorable missing longitudinal data. *Lifetime data analysis*, 8(2):99–115, 2002.

Kevin H Eng and Bret M Hanlon. Discrete mixture modeling to address genetic heterogeneity in time-to-event regression. *Bioinformatics*, 30(12):1690–1697, 2014.

Jianqing Fan and Jianwei Chen. One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):927–943, 1999.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

Mansour Farahani, Vladimir Novitsky, Rui Wang, Hermann Bussmann, Sikhulile Moyo, Rosemary M Musonda, Themba Moeti, Joseph M Makhema, Max Essex, and Richard Marlink. Prognostic value of hiv-1 rna on cd4 trajectories and disease progression among antiretroviral-naive hiv-infected adults in botswana: a joint modeling analysis. *AIDS research and human retroviruses*, 32(6):573–578, 2016.

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

Lloyd D Fisher and Danyu Y Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458): 611–631, 2002.

Dani Gamerman. Dynamic bayesian models for survival data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 40(1):63–79, 1991.

Major Greenwood et al. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

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

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

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

Martin S King. *Local likelihood and local partial likelihood in hazard regression.* The University of North Carolina at Chapel Hill, 1997.

DCFR Koboldt, Robert Fulton, Michael McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua McMichael, Lucinda Fulton, David Dooling, Li Ding, Elaine Mardis, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490 (7418):61–70, 2012.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Erich L Lehmann and George Casella. *Theory of point estimation.* Springer Science & Business Media, 2006.

Lei Liu and Xuelin Huang. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):65–81, 2009.

Yue Liu, Lei Liu, and Jianhui Zhou. Joint latent class model of survival and longitudinal data: An application to cpcra study. *Computational Statistics & Data Analysis*, 91:40–50, 2015.

Torben Martinussen and Thomas H Scheike. A flexible additive multiplicative hazard model. *Biometrika*, 89(2):283–298, 2002.

Leszek Marzec and Pawel Marzec. On fitting cox's regression model with time-dependent coefficients. *Biometrika*, 84(4):901–908, 1997.

Elias Masry and Dag Tjøstheim. Additive nonlinear arx time series and projection estimates. *Econometric Theory*, 13(2):214–252, 1997.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. M. Dekker New York, 1988.

Susan Allbritton Murphy and Pranab Kumar Sen. Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1): 153–180, 1991.

Bengt Muthén and Katherine Masyn. Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics*, 30(1):27–58, 2005.

Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

Li-Xuan Qin and Steven G Self. The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):526–533, 2006.

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

Narges Roustaei, Seyyed Mohammad Taghi Ayatollahi, and Najaf Zare. A proposed approach for joint modeling of the longitudinal and time-to-event data in heterogeneous populations: An application to hiv/aidss disease. *BioMed research international*, 2018, 2018.

Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC medical genomics*, 5(1): 66, 2012.

Ruiguang Song, H Irene Hall, Timothy A Green, Célia L Szwarcwald, and Nikos Pantazis. Using cd4 data to estimate hiv incidence, prevalence, and percent of undiagnosed infections in the united states. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 74(1):3–9, 2017.

Z Sun. Asymptotic unbiased and strong consistency for density function estimator. *Acta Mathematica Sinica*, 27:769–782, 1984.

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

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

Anastasios A Tsiatis, Victor Degruttola, and Michael S Wulfsohn. 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, 1995.

Edward F Vonesh, Tom Greene, and Mark D Schluchter. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in medicine*, 25(1):143–163, 2006.

CY Wang. Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*, pages 235–253, 2006.

Michel Wedel and Wayne S DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of classification*, 12(1):21–55, 1995.

Lang Wu, Wei Liu, Grace Y Yi, and Yangxin Huang. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012, 2012.

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

David M Zucker and Alan F Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1):329–353, 1990.