ESTIMATION OF SEMIVARYING COEFFICIENT MODELS FOR COUNTING PROCESSES WITH APPLICATIONS

by

Liqiu Deng

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2019

Approved by:

_____

Dr. Yanqing Sun

_____

Dr. Yang Li

_____

Dr. Weihua Zhou

_____

Dr. Weidong Tian

ABSTRACT

LIQIU DENG. Estimation of Semivarying Coefficient Models for Counting Processes with Applications. (Under the direction of DR. YANQING SUN)

Recurrent events are very common in many different fields, including biological, medical, engineering and finance. Existing research have developed methodologies to model constant covariate effects and time-dependent covariate effects. However, in reality, for instance medical cases, covariate effects can be depending on other covariates as well. Therefore, in this dissertation, we investigate a semiparametric model for recurrent events, which incorporates both time-varying covariate effects and covariate-varying effect.

In our model, we use fixed parameters to model constant covariate effects, while we assume both time-dependent effects and covariate-varying effects to be unknown functions. An estimation procedure is proposed to estimate the unknow parameters and functions. Local linear smoothing method is adopted in our estimation procedure. Detailed computation is carried out by using Newton-Raphson iterative method. The asymptoptic properties including asymptotic normality and consistency are established for the proposed estimators.

In order to assess the finite-sample performance of the proposed estimators and estimation procedure, simulation studies are conducted for different cases. The simulation results show that the proposed estimators perform very well with small bias and an empirical coverage probability close to its nominal level 95%.

In addition, the proposed model and methodologies are applied on the dataset from

the Hemodialysis Study (HEMO). The data applications are aiming at examining the treatment effects of two different design in the study and exploring factors that are associated with hemodialysis patients' mortality and hospitalization rate. The results show that both treatments are not significant at neither reducing mortality risk nor hospitalization rate for hemodialysis patients. Some factors, including sex, age, baseline serum albumin level, ICED score and diabetes, are found to be significantly associated with the mortality and hospitalization rate for hemodialysis patients.

ACKNOWLEDGMENTS

It has been a long journey before I get to this point. I would like to acknowledge the roles of all individuals that have helped me complete this dissertation and get through this long journey.

I would like to express my most sincere gratitude to my advisor Dr. Yanqing Sun, who encouraged me to pursue this study and provided me insightful guidance from the beginning to the finishing line. I am deeply grateful for her understanding, patience, enthusiasm to research, encouragement and inspiration, without any of which this dissertation would not have been completed. The weekly meetings we had in the past five years helped me go further in my research, while Friday lunches with Dr. Sun helped me grow to be a better woman. What I have learned from her will benefit me for the rest of my life.

To my committee members, Dr. Weihua Zhou, Dr. Yang Li, and Dr. Weidong Tian, I am deeply grateful for their valuable comments, advise, time and encouragement throughout my dissertation.

I would like to thank Dr. Peter Gilbert at Fred Hutchinson Cancer Research Center and University of Washington for motivating the research of the dissertation and discussions on the future applications to the Malaria vaccine study.

I would also like to thank National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under National Institutes of Health (NIH) for providing the HEMO data set. I am grateful to Dr. Guofen Yan at University of Virginia School of Medicine for her help and inspiring discussion on the HEMO data set. This research was

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Survival analysis aims at modeling time to certain event. It is of interest in many biological, medical, engineering and financial applications. For example, animals life in control and experimental groups, time to relapse of a certain disease after different treatments, first time failure of a mechanical part, or time to bankruptcy of a financial institution. By assuming survival time follows a probability distribution (e.g. Exponential, Gamma, Weibull, or Log-Normal etc.) that depends on covariates, a variety of parametric regression models are developed in order to investigate how survival time distributes given the covariates, as well as exploring how strongly survival time relates to the covariates. Another approach developed by Cox (1972) is a semi-parametric, which assumes a parametric and proportional structure that relates covariates to hazard rate while leaves the non-parametric part completely unspecified. Since Cox model does not require distribution assumption, it soon becomes one of the most popular survival regression methods.

The ordinary Cox regression model suffices in survival analysis in which events are assumed conditionally independent given covariates. However, it is common to observe that the event of interest occurs multiple times on one single subject, which is referred as recurrent events. Typical examples include recurrent episodes of a disease in patients, multiple admissions to hospitals, multiple ear infections on young age children, multiple delinquencies of credit card customers, earthquakes in one city,

repairs in machines and automobiles, etc. It is of interest to identify risk factors that are associated with the frequencies of recurrent events. In this dissertation, we aim at developing new methodologies for analyzing recurrent events data.

1.1   Motivating Examples - HEMO Study

This dissertation is motivated by the Hemodialysis Study (HEMO). In this section, we provide a brief description for the HEMO study, including the background of this clinical trial, experimental design and primary/secondary outcomes of this trial, followed by a short summary of existing research on the HEMO dataset and our research questions.

The Hemodialysis Study (HEMO) is a clinical trial conducted in multi-center, with patients randomized into different treatment groups. The objective of this study is to test if those treatments can reduce the mortality and mobidity of patients with hemodialysis.

A total number of 2677 patients were recruited in this study, but only 1846 were enrolled in this clinical trial after screening process. The following two treatments were considered in this study.

- The dose of dialysis: standard dose (equilibrated Kt/V level of 1.05) or high dose (equilibrated Kt/V level of 1.45).

- The dialyzer with different membrane types: a low-flux membrane or a high-flux membrane.

A 2 x 2 design was considered in this study. A total of 1846 patients were randomly assigned to different treatment groups as shown in Table 1. It is shown that this study

is a balanced design in terms of the number of patients in each treatment group.

The enrollment window for patients ranges from March 1995 and October 2000. Patients were followed up until December 31, 2001 unless loss of follow-up or death happened before the end of study. The primary outcome of this study was death due to different causes, while hospitalizations were also of interest as secondary outcomes. The HEMO study was aiming to check if high-flux membrane and high-dose of dialysis could reduce the mortality and mobidity of patients with hemodialysis.

Many studies have been conducted on the dataset collected from the HEMO study (Sattar et al., 2012; Greene et al., 2000; Eknoyan et al., 2002; Hoenich, 2003; Daugirdas et al., 2004; Lynch et al., 2010; Beddhu et al., 2010; Rocco et al., 2004; Daugirdas et al., 2003; Cheung AK, 2004; Yan and Greene, 2011; Depner et al., 2004; Argyropoulos et al., 2015; Jhamb et al., 2011; Burrowes et al., 2011; Liang KV, 2011; Dwyer JT, 2002). The aforementioned studies have a variety of research focuses, including exploring factors that are associated with hemodialysis patients' mortality, examining the treatment effects on mortality, assessing the association between treatments and secondary outcomes-hospitalizations, evaluating patients' quality of life, etc.

In particular, Greene et al. (2000) provided a thorough description on the HEMO study, including the factorial design of the clinical trial, primary outcomes and secondary outcomes, possible analysis with data collected from the study, etc. In addition, some statistical issues were also discussed by the study. Using the Cox's time varying covariate model, Sattar et al. (2012) found an significant association between diabetes and the risk of mortality in end stage renal disease (ESRD) patients. In addition, their results showed an increasing risk of mortality over time among ESRD

patients. Eknoyan et al. (2002) analyzed the effects of different doses of dialysis and different levels of flux of dialyzer membrane on hemodialysis patients' mortality. Their research found that both treatment effects were not significant, which means hemodialysis patients did not benefit from taking neither high dose of dialysis nor high-flux of dialyzer membrane. Cheung AK (2004) conducted an analysis on cardiac diseases in the HEMO study. This article explored factors associated with cardiac diseases and checked the treatment effects on mortality and cardiac outcomes, for instance, cardiac hospitalizations. The results showed no significant effects of high-flux dialysis on mortality of hemodialysis patients, but significant effects on cardiac outcomes.

In this dissertation, both primary outcomes mortality and the secondary outcomes hospitalizations are of interest. We would like to explore the factors that are associated with mortality and hospitalizations for hemodialysis patients. In particular, it is of our interest to check if the covariate effects vary along with the duration of being on dialysis for such patients, or along with patients' age.

Table 1: Allocation of participants in different treatment groups in Hemo study.

	low-flux	high-flux	Total
standard-KTV	467	459	926
high-KTV	458	462	920
Total	925	921	1846

1.2   Literature Review

In the recent decades, it has become a great interest to model the occurence of recurrent events. Intensive studies have been conducted to develp methodoligies to analyze recurrent event data. For example, the hazard functions of gap times between recerrent events were considered by Prentice and others (1981) and marginal hazard function of each recurrent event by Wei and others (1989). Andersen and Gill (1982) proposed a multiplicative intensity model, in which recurrent events were considered as non-homogeneous Poisson process. Later on, Pepe and Cai (1993), Lawless and Nadeau (1995), Lin et al. (2000, 2001) studied a mean/rate model with the Poisson assumption removed.

$$E\{dN_i^*(t) \mid X_i(t)\} = \exp\{\beta^T X_i(t)\}\lambda_{0i}(t)dt, \tag{1.1}$$

The aforementioned models assume that the covariate coefficients are constant over time. For many applications, however, the covariate effects change over time instead of staying constant. For example, in clinical study, the treatment effects may change over time, and the temporal effects of treatment may be of interest. Chiang and Wang (2007) modified the above the model to allow the coefficients to be time-dependent.

$$E\{dN_i^*(t) \mid X_i(t)\} = \exp\{\beta(t)^T X_i(t)\}\lambda_{0i}(t)dt, \tag{1.2}$$

Amorim et al. (2008) developed a semiparametric rates model allowing covariate effects to be time-dependent. Regression splines techniques are incorporated in its estimation procedure for estimating the time varying covariate coefficients. More

recently, Sun et al. (2011) proposed a marginal rates model, which allowed some covariate coefficients to be time-varying while others constant.

$$E\{dN_i^*(t) \mid X_i(t), Z_i(t)\} = \exp\{\beta_0(t)^T X_i(t) + \gamma_0^T Z_i(t)\} d\mu_0(t), \qquad (1.3)$$

where $\beta_0(t)$ is an unknown p-vactor of time-varying regression coefficients, γ_0 is a q-vector of unknown constant regression coefficients, and $\mu_0(t)$ is an unspecified baseline mean function.

In many applications, for instance the motivating example mentioned above, the covariate effects can not only be time dependent but also vary with other covariates. Thus, a model that can consider such covariate-varying coefficients is desired. However, to our best knowledge, existing models and methodologists have not yet considered such covariate-varying effects on recurrent event data. In this dissertation, we propose a semiparametric model which incorporates both time-varying covariate effects and covariate-varying effects motivated by our motivating example.

The remainder of the dissertation is organized as follows. In Chapter 2, we introduce the proposed semiparametric model, estimation procedure for estimating the unknow parameters and functions in the model followed by a detailed computational algarithm, and bandwidth selection for smoothing. Chapter 3 defines some notations and presents the asymptotic properties of the proposed parameter estimates, the proofs for which are given in Appendix. In Chapter 4, the finite-sample performance of the proposed method is evaluated by two types of simulation studies, one of which is survival analysis with single-event data and the other is for recurrent event data. Chapter 5 is a real data application of the proposed method, followed by a discussion

of our findings on the application.

CHAPTER 2: SEMIPARAMETRIC MODEL WITH NONPARAMETRIC
COVARIATE-VARYING EFFECTS

In Chapter 2, the contents are organized as follows. We introduce the proposed

model in Section 2.1, along with related notations. In Section 2.2, we propose a proce-

dure to estimate the nonparametric parameters and parametric parameters contained

in our model. The detailed algorithm procedure for estimation is given in Section 2.3.

Since bandwidth selection is involved in our estimation procedure, we use cross val-

idation method to choose the optimal bandwidth, details of which are presented in

Section 2.4.

2.1 Model Description

In this section, we start with introducing the notations that are used in our pro-

posed model and methodology.

Let n be the total number of subjects in a random sample, t be the follow up time,

and $t \in [0, \tau]$. Suppose the event of interest can be repeated for the same subject i

and the recurrent event times are recorded by T_{ij}, which means the jth event time for

the ith subject. Let K_i be the total number of recurrent events for subject i, then we

have the event times $0 \leq T_{i1} < T_{i2} < \cdots < T_{iK_i} \leq \tau$ for subject i.

During the follow up period, it is common that subjects could drop out at somepoint

due to a variety of reasons. For instance, in clinical trial, patients could die due to

unrelated causes or lose interest and thus stop following up with the study. Let C_i be

the minimum of end of the study τ and censoring time for subject i. For subject i, only obervations at time points before C_i are possible, which means that the observation for subject i is terminated at C_i. Thus, if we define $N_i^*(t) = \sum_{j=1}^{n_i} I(T_{ij} \leq t)$ to be the number of events recorded for subject i by time t, then the observed event process can be written as $N_i(t) = N_i^*(t \wedge C_i)$. Note that $I(\cdot)$ is the indicator function, and $a \wedge b$ means the minimum of a and b.

Suppose $Q_i(t)$ and $U_i(t)$ are covariates that are associated with subject i. $U_i(t)$ is a scalor process and has support $\mathcal{U}$. $Q_i(t)$ consists three parts $X_i(t)$, $Z_i(t)$ and $W_i(t)$, the dimentions of which are p_1, p_2 and p_3, respectively. Each part of $Q_i(t) = (X_i^T(t), Z_i^T(t), W_i^T(t))^T$ is allowed to be either time-dependent or time-independent over the time interval $[0, \tau]$. Note that A^T means the transpose of matrix/vector A.

We assume that the random processes $\{N_i(\cdot), Q_i(\cdot), U_i(\cdot), i = 1, \cdots, n\}$ are independent identically distributed (iid). In this dissertation, we only consider the case that the censoring is caused by the termination of the study or random loss of follow up, therefore it is reasonable to assume that the censoring time C_i is noninformative and independent of event time T_{ij}. In other words, we have $E\{dN_i^*(t) \mid Q_i(t), U_i(t), C_i \geq t\} = E\{dN_i^*(t) | Q_i(t), U_i(t)\}$. However, the censoring time C_i is allowed to depend on covariates $Q_i(t)$ and $U_i(t)$. Some researchers, for instance Ghosh and Lin (2003), have developed methodologies to deal with recurrent events data in the presence of dependent censoring.

Let $\mathcal{F}_{t-}$ denote the filtration, which is the σ-algebra generated by the observed information. Let $\lambda_i(t) = E\{dN_i^*(t)|\mathcal{F}_{t-}\}$ be the intensity or the mean rate $\lambda_i(t) = E\{dN_i^*(t)|X_i(t), U_i(t)\}$. We propose the following semiparametric varying-coefficients

intensity model:

$$\lambda_i(t) = \exp\{\alpha^T(t)X_i(t) + \beta^T Z_i(t) + \gamma^T(U_i(t))W_i(t)\}, \qquad (2.1)$$

for $0 \le t \le \tau$, where

1. $\alpha(t) = (\alpha_1(t), \alpha_2(t), \cdots, \alpha_{p_1}(t))^T$ is a vector with the same dimension p_1 as $X_i(t)$, each element of which is an unspecified function over the time period $0 \le t \le \tau$. By setting the first element of $X_i(t)$ to be identity vector I, the accoresponding $\alpha_1(t)$ is the nonparametric baseline function.

2. $\beta = (\beta_1, \beta_2, \cdots, \beta_{p_2})^T)$ is a vector of unknown time-independent parameters with the same dimention p_2 as $Z_i(t)$;

3. $\gamma(U_i(t)) = (\gamma_1(U_i(t)), \gamma_2(U_i(t)), \cdots, \gamma_{p_3}(U_i(t)))^T$ is a p_3-dimensional vector of functions of corativate $U_i(t)$.

In the proposed model, we model the time-varying effects by the non-parametric part $\alpha(t)$, constant effects by the parametric part β, and covariate-varying effects by the non-parametric part $\gamma(u)$. $\gamma(u)$ models the effects of covariate $W_i(t)$ at the level u of another covariate $U_i(t)$. Compared with model (1.3) proposed by Sun et al. (2011), our model adds the ability to model the covariate-varying effects, which is common in reality. For example, in the HEMO study, it is of interest to model if and how covariate effects change along with duration time of dialysis. The duration time of dialysis prior to enrollment to the study was recorded as a baseline measurement for each patient. Let $S_i(t)$ denote the baseline measurement of duration time of dialysis prior to enrollment, we can define $U_i(t) = t + S_i(t)$ to be the actual duration time

of diaysis at time t such that we can model how covariate effects change along with duration time of dialysis.

2.2   Estimation Procedure

In this section, we describe the estimating procedure for the parameters in our model. Since the proposed model contains both non-parametric functions $\alpha(t)$ and $\gamma(u)$ and parametric parameters β, we divide our estimating procedure into two parts. First, we estimate the non-parametric part by using Taylor expansion and local linear approximation. Secondly, we estimate the parametric part by using profile likelihood method. Details are as follows.

In order to use Taylor expansion, we first assume that $\alpha(\cdot)$ and $\gamma(\cdot)$ are smooth functions such that their first and second derivatives $\dot{\alpha}(\cdot)$, $\dot{\gamma}(\cdot)$, $\ddot{\alpha}(\cdot)$ and $\ddot{\gamma}(\cdot)$ exist. With this assumption, $\alpha(\cdot)$ and $\gamma(\cdot)$ can be approximated by the first order Taylor expansion as follows.

For any t that belongs to a neighborhood of t_0, $t \in \mathcal{N}_{t_0}$, we have

$$\alpha(t) = \alpha(t_0) + \dot{\alpha}(t_0)(t - t_0) + O((t - t_0)^2),$$

and similarly for any u that belongs to a neighborhood of u_0 $u \in \mathcal{N}_{u_0}$, we have

$$\gamma(u) = \gamma(u_0) + \dot{\gamma}(u_0)(u - u_0) + O((u - u_0)^2).$$

Therefore, for $t \in \mathcal{N}_{t_0}$ and $U_i(t) \in \mathcal{N}_{u_0}$, the proposed model (2.1) can be approximated by the following

$$\lambda_i^*(t, \vartheta^*, \beta; t_0, u_0) = \exp\{\vartheta^{*T}(t_0, u_0)Q_i^*(t, t_0, u_0) + \beta Z_i(t)\}, \qquad (2.2)$$

where $\vartheta^*(t_0, u_0)$ consists a total of four parts, not only the functions themselves but also their first derivatives, as follows.

$$\vartheta^*(t_0, u_0) = (\alpha^T(t_0), \gamma^T(u_0), \dot{\alpha}^T(t_0), \dot{\gamma}^T(u_0))^T$$

And similarly for $Q_i^*(t, t_0, u_0)$, we have

$$Q_i^*(t, t_0, u_0) = (X_i^T(t), W_i^T(t), X_i^T(t) \times (t - t_0), W_i^T(t) \times (U_i(t) - u_0))^T$$

In our estimation procedure, we adopt kernal smoothing method. Suppose

1. $K_1(\cdot)$ and $K_2(\cdot)$ are kernel functions,

2. h_t and h_u are bandwidth parameters,

3. $K_{h_t}(\cdot) = K_1(\cdot/h_t)/h_t$ and $K_{h_u}(\cdot) = K_2(\cdot/h_u)/h_u$.

Then, at each t_0 and u_0, we define $K_{h_t, h_u}(t, U_i(t); t_0, u_0) = K_{h_t}(t - t_0)K_{h_u}(U_i(t) - u_0)$ as a two dimentional product kernel function.

For fixed β, at each t_0 and u_0, by Cook and Lawless (2007), the local log-likelihood function for $\alpha(\cdot)$ and $\gamma(\cdot)$ can be written as follows.

$$\ell_\vartheta(\vartheta^*; \beta, t_0, u_0) = \sum_{i=1}^n \int_0^\tau [\log(\lambda_i^*(t, \vartheta^*, \beta; t_0, u_0)) \, dN_i(t) - \lambda_i^*(t, \vartheta^*, \beta; t_0, u_0) \, dt]$$

$$K_{h_t, h_u}(t, U_i(t); t_0, u_0). \tag{2.3}$$

By taking the first derivative for (2.3) with respect to ϑ^*, at each t_0 and u_0, the local score function for $\alpha(\cdot)$ and $\gamma(\cdot)$ for fixed β is

$$U_\vartheta(\vartheta^*; \beta, t_0, u_0) = \sum_{i=1}^n \int_0^\tau [dN_i(t) - \lambda_i^*(t, \vartheta^*, \beta; t_0, u_0) \, dt] \, Q_i^*(t, t_0, u_0)K_{h_t, h_u}(t, U_i(t); t_0, u_0)$$

$$\tag{2.4}$$

By setting $U_\vartheta(\vartheta^*; \beta, t_0, u_0) = 0$, we can solve it and denote the solution by $\tilde{\vartheta}^*(\beta, t_0, u_0)$.

We define the following notations:

1. $\tilde{\vartheta}(\beta, t, u)$ be the first $p_1 + p_3$ components of $\tilde{\vartheta}^*(\beta, t_0, u_0)$,

2. $\tilde{Q}_i(t)$ be the first $p_1 + p_3$ components of $Q_i^*(t)$, i.e., $\tilde{Q}_i(t) = ((X_i(t))^T, (W_i(t))^T)^T$,

3. $\tilde{\lambda}_i(t, \beta) = \exp\{(\tilde{\vartheta}(\beta, t, U_i(t)))^T \tilde{Q}_i(t) + \beta^T Z_i(t)\}$.

Then the profile likelihood function can be written as follow:

$$\ell_\beta(\beta) = \sum_{i=1}^{n} \int_{t_1}^{t_2} \left[\log(\tilde{\lambda}_i(t, \beta)) \, dN_i(t) - \tilde{\lambda}_i(t, \beta) \, dt \right]. \tag{2.5}$$

By maximizing the profile likelihood function (2.5), we can obtain the profile maximum likelihood estimator $\hat{\beta}$ of β.

By taking derivative of (2.5) with respect to β, the profile estimating equation for β can be obtained by $U_\beta(\beta) = \frac{\partial \ell_\beta(\beta)}{\partial \beta}$ as follows:

$$U_\beta(\beta) = \sum_{i=1}^{n} \int_{t_1}^{t_2} \left[dN_i(t) - \tilde{\lambda}_i(t, \beta) \, dt \right] \left\{ \frac{\partial \tilde{\vartheta}(\beta, t, U_i(t))}{\partial \beta} \tilde{Q}_i(t) + Z_i(t) \right\}, \tag{2.6}$$

where $\frac{\partial \tilde{\vartheta}(\beta, t, U_i(t))}{\partial \beta}$ is the first $p_1 + p_3$ columns of

$$\frac{\partial \tilde{\vartheta}^*(\beta, t, U_i(t))}{\partial \beta} = - \left\{ \frac{\partial U_\vartheta(\vartheta^*; \beta, t, U_i(t))}{\partial \vartheta^*} \right\}^{-1} \frac{\partial U_\vartheta(\vartheta^*; \beta, t, U_i(t))}{\partial \beta} \bigg|_{\vartheta^* = \tilde{\vartheta}^*(\beta, t, U_i(t))}.$$

By setting $U_\beta(\beta) = 0$, we can solve it and denote the solution by $\hat{\beta}$.

In this dissertation, we adopt the Newton-Raphson iterative method to find the estimators of the nonparametric components $\hat{\vartheta}(t_0, u_0)$ and the parametric components $\hat{\beta}$.

We define the following notations.

1. $\hat{\vartheta}(t_0, u_0) = (\hat{\alpha}(t_0, u_0), \hat{\gamma}(t_0, u_0))^T$

2. $\hat{\alpha}(t_0, u_0)$ as the first p_1 elements of $\hat{\vartheta}^*(t_0, u_0)$ corresponding to the position of

   $\alpha(t_0)$ in $\vartheta^*(t_0, u_0)$.

3. $\hat{\gamma}(t_0, u_0)$ as the p_3 elements of of $\hat{\vartheta}^*(t_0, u_0)$ corresponding to the position of $\gamma(u_0)$

   in $\vartheta^*(t_0, u_0)$.

Note that the estimators $\hat{\alpha}^*(t_0, u_0)$ and $\hat{\gamma}^*(t_0, u_0)$ are not efficient, with the rea-

son being that only local observations are utilized during the estimation process.

Therefore, an aggregation through the following formulas are proposed to improve

the efficiency of our estimators for both $\alpha(t_0)$ and $\gamma(u_0)$.

$$\hat{\alpha}(t_0) = \frac{1}{n} \sum_{j=1}^{n} \hat{\alpha}(t_0, U_j(t_0)), \qquad \hat{\gamma}(u_0) = \frac{1}{n_{u_0}} \sum_{j=1}^{n_{u_0}} \hat{\gamma}(t_{u_0,j}, u_0), \qquad (2.7)$$

where $t_{u_0,j} \in U_j^{-1}(u_0) = \{t : U_j(t) = u_0\}$, and n_{u_0} is the number of points in the

union $\cup_{j=1}^{n}\{U_j^{-1}(u_0)\}$.

2.3   Computational Algorithm

The previous section derives the estimators $\hat{\alpha}(t_0)$ and $\hat{\gamma}(u_0)$ for the non-parametric

functions and $\hat{\beta}$ for the parametric parameters in our model. In this section, we

sketch the detailed computational algorithm to accomplish those estimators by using

Newton-Raphson iterative method.

First, we define some notations that will be used in the algorithm as follows.

- Let $\hat{\vartheta}(t, u)^{\{0\}}$ be the initial values of $\hat{\vartheta}(t, u)$ and $\hat{\beta}^{\{0\}}$ be the initial values for $\hat{\beta}$.

- Let $\hat{\vartheta}^{\{k\}}(t, u)$ be the kth step estimator of $\vartheta(t, u)$ and $\hat{\beta}^{\{k\}}$ be the kth step esti-

  mator of β

The steps of computational algorithm are given as follows.

1. Generate equally spaced grid points (t_0, u_0) over t and u.

2. Initialize $\hat{\vartheta}(t, u)^{\{0\}}$ and $\hat{\beta}^{\{0\}}$ by using arbitrary values;

3. For each grid point (t, u), plug $\hat{\beta}^{\{k-1\}}$, the $(k-1)^{th}$ step estimator of β, into the local score function (2.4). Find the root and denote it as the kth step estimator $\hat{\vartheta}^{*\{k\}}(t, u) = \hat{\vartheta}^*(t, u, \hat{\beta}^{\{k-1\}})$. The kth step estimator satisfies that $U_\vartheta(\hat{\vartheta}^{*\{k\}}(t, u); \hat{\beta}^{\{k-1\}}, t, u) = 0$

4. Obtain the estimates $\hat{\alpha}^{\{k\}}(t_0)$ and $\hat{\gamma}^{\{k\}}(u_0)$ by doing aggregation through (2.7) such that the estimated curves are smooth enough.

5. Plug in $\hat{\alpha}^{\{k\}}(t_0)$ and $\hat{\gamma}^{\{k\}}(u_0)$ to (2.5). The kth step estimator $\hat{\beta}^{\{k\}}$ can be obtained by maximizing the profile likelihood (2.5).

6. Repeat step 3, 4, and 5 and update the estimators $\hat{\vartheta}^{*\{k\}}(t, u)$ and $\hat{\beta}^{\{k\}}$ at each iteration until the convergence criteria is met. The estimator $\hat{\beta}$ is $\hat{\beta}^{\{k\}}$ at the convergence.

2.4   Bandwidth Selection

Kernel smoothing method is adopted in our estimation procedure, which involves selecting the optimal bandwidths. Cross-validation method has been commonly used for bandwidth selection. In this dissertation, we choose the optimal bandwidths for estimating the non-parametric functions $\alpha(t)$ and $\gamma(u)$ by using the K-fold cross-validation method, where K represents the number of groups that a given sample is to be split into.

In our estimation procedure, we use a two dimentional product kernel function, which involves two bandwidths h_t and h_u parameters. We use K-fold cross-validation method to choose the optimal bandwidths $h_{t,K}^*$ and $h_{u,K}^*$ parametersfor h_t and h_u, respectively. Briefly, we go through each combination of (h_t, h_u) to see which one results in the least negative log-likelihood define as follows. The optimal bandwidths combination $(h_{t,K}^*, h_{u,K}^*)$ is the one corresponding to the least negative log-likelihood.

Suppose we have a given sample. The detailed computational procedure for carrying out the K-fold cross-validation are given as follows.

1. Create combinations for h_t and h_u.

2. Shuffle the sample randomly, split it into K groups and denote them as $(G_1, G_2, \cdots, G_K)$.

3. For each combination of $(h_{t,K}, h_{u,K})$, do the following:

    3.1. Hold out group G_k $(k = 1, 2, \cdots, K)$ as a test data set and take the re-maining groups as a training data set.

    3.2. Fit the proposed model on the training set and evaluate it on the test data set by calculating the negative log-likelihood.

    3.3. Retain the evaluation score and discard the model.

    3.4. repeat Steps 3.1 - 3.3 until each group gets a turn to be the test data set.

    3.5. Take the simple average of all K evaluation scores from the loop, and denote it as the Score for current testing bandwidth combination.

4. Repeat Step 3 until each bandwidth combination is used to fit the model.

5. Compare all scores corresponding to different bandwidth combinations and denote the one corresponding to the lowest score as the optimal bandwidth combination $(h^*_{t,K}, h^*_{u,K})$

CHAPTER 3: ASYMPTOTIC PROPERTIES

In Chapter 2, we propose estimators $\hat{\alpha}$, $\hat{\gamma}$ and β for the non-parametric functions $\alpha(t)$ and $\gamma(u)$ and the parameter β, respectively. It is naturally of interest to explore the asymptotic properties for those estimators. Therefore, in Chapter 3, we will establish the asymptotic properties for those proposed estimators, including asymptotic normality, consistency, etc. Chapter 3 are organized as follows. In Section 3.1, we define all related notations. Section 2 presents all theorems that we establish regarding the asymptotic properties of our estimators.

3.1   Notations

We define the notations as follows.

- Let $\mathscr{I}_1 = \{\mathscr{I}_{ij}\}_{p_1 \times (p_1 + p_3)}$ be a matrix with elements like the following.

$$
\mathscr{I}_{ij} = \begin{cases} 1 & \text{for } i = 1, \ldots, p_1,\ i = j \\ \\ 0 & \text{otherwise} \end{cases}
$$

- Let $\mathscr{I}_3 = \{\mathscr{I}_{ij}\}_{p_3 \times (p_1 + p_3)}$ be a matrix with elements like the following.

$$
\mathscr{I}_{ij} = \begin{cases} 1 & \text{for } i = 1, \ldots, p_3,\ j = i + p_1 \\ \\ 0 & \text{otherwise} \end{cases}
$$

- Let $\alpha_0(t)$, β_0 and $\gamma_0(u)$ be the true values of $\alpha(t)$, β and $\gamma(u)$ under model (2.1),

respectively.

- Let $\lambda_{i0}(t) = \exp\{\alpha_0(t)X_i(t) + \beta_0 Z_i(t) + \gamma_0(U_i(t))W_i(t)\}$

- Let $\hat{\lambda}_i(t) = \exp\{\hat{\vartheta}^T(t, U_i(t))\tilde{Q}_i(t) + \beta^T Z_i(t)\}$

- Let $f_U(t, u)$ be the density function of $U(t)$ evaluated at u

- Define

$$e_{11}(t, u) = E\left[(-\lambda_{i0}(t)dt)\{\tilde{Q}_i(t)\}^{\otimes 2} \mid U_i(t) = u\right] f_U(t, u)$$

and

$$e_{12}(t, u) = E\left[(-\lambda_{i0}(t)dt) Z_i(t)\{\tilde{Q}_i(t)\}^{\otimes 2} \mid U_i(t) = u\right] f_U(t, u)$$

- Define

$$\hat{E}_{11}(t_0, u_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(t - t_0)K_b(U_i(t) - u_0)\left[-\hat{\lambda}_i(t)dt\right]\{\tilde{Q}_i(t)\}^{\otimes 2},$$

and

$$\hat{E}_{12}(t_0, u_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_h(t - t_0)K_b(U_i(t) - u_0)\left[-\hat{\lambda}_i(t)dt\right]\tilde{Q}_i(t)(Z_i(t))^T.$$

3.2   Asymptotic Properties

In this section, we establish the asymptotic properties for the proposed estimators $\hat{\alpha}(t)$, $\hat{\gamma}(t)$ and $\hat{\beta}$ in this section. Three theorems are given as follows. The proofs for theorems are presented in the Appendix.

Theorem 3.1. *Assuming the conditions given in Appendix are satisfied, then*

$$\sqrt{n}(\hat{\beta} - \beta_0) \longrightarrow N(0, A_\beta^{-1}\Sigma_\beta A_\beta^{-1})$$

in distribution, where

$$A_\beta = E\left[\int_0^\tau \{Z_i(t) - (e_{12}(t, U_i(t)))^T (e_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}^{\otimes 2}\, dt\right]$$

and

$$\Sigma_\beta = E\left[\int_0^\tau \{Z_i(t) - (e_{12}(t, U_i(t)))^T (e_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}\, dM_i(t)\right]^{\otimes 2}.$$

The matrices A_β and Σ_β can be consistently estimated by the following two formulas, respectively.

$$\hat{A}_\beta = \frac{1}{n}\sum_{i=1}^n \int_{t_1}^{t_2} \{Z_i(t) - (\hat{E}_{12}(t, U_i(t)))^T (\hat{E}_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}^{\otimes 2}\, dt$$

and

$$\hat{\Sigma}_\beta = \frac{1}{n}\sum_{i=1}^n \left(\int_{t_1}^{t_2} \{dN_i(t) - \hat{\lambda}_i(t)dt\}\{Z_i(t) - (\hat{E}_{12}(t, U_i(t)))^T (\hat{E}_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}\right)^{\otimes 2},$$

where $0 < t_1 < t_2 < \tau$.

Theorem 3.2. *Assuming the conditions given in Appendix are satisfied, then*

(1) $\displaystyle\sup_{t\in[0,\tau]} |\hat{\alpha}(t) - \alpha_0(t)| = o_p(1)$;

(2) $\displaystyle\sqrt{nh_t}(\hat{\alpha}(t) - \alpha_0(t) - \frac{1}{2}h_t^2\nu_2\ddot{\alpha}(t))\xrightarrow{\mathcal{D}} N(0, \Sigma_\alpha(t))$,

where

$$\Sigma_\alpha(t) = \lim_{n\to\infty} h_t E\left[\int_0^\tau \{dN_i(s) - \lambda_i(s)ds\}\,\mathscr{I}_1 e_{11}(t, U_i(s))^{-1}\tilde{Q}_i(s)K_{h_t}(s - t)\right]^{\otimes 2},$$

which can be consistently estimated by the following formula.

$$\hat{\Sigma}_\alpha(t) = \frac{h_t}{n}\sum_{i=1}^n \left[\int_0^\tau \{dN_i(s) - \hat{\lambda}_i(s)ds\}\,\mathscr{I}_1\hat{E}_{11}(t, U_i(s))^{-1}\tilde{Q}_i(s)K_{h_t}(s - t)\right]^{\otimes 2}.$$

Theorem 3.3. *Assuming the conditions given in Appendix are satisfied, then*

(1) $\displaystyle\sup_{u\in[u_1,u_2]}|\hat{\gamma}(u)-\gamma_0(u)|=o_p(1)$;

(2) $\sqrt{nh_u}(\hat{\gamma}(u)-\gamma_0(u)-\dfrac{1}{2}h_u^2\nu_2\ddot{\gamma}(u))\xrightarrow{\mathcal{D}}N\left(0,\Sigma_\gamma(u)\right).$

where

$$\Sigma_\gamma(u)=\lim_{n\to\infty}h_uE\left[\int_0^\tau\{dN_i(s)-\lambda_i(s)ds\}\,\mathscr{I}_3 e_{11}(u,U_i(s))^{-1}\tilde{Q}_i(s)K_{h_u}(s-u)\right]^{\otimes2},$$

which can be consistently estimated by

$$\hat{\Sigma}_\gamma(u)=\frac{h_u}{n}\sum_{i=1}^n\left[\int_0^\tau\left\{dN_i(s)-\hat{\lambda}_i(s)ds\right\}\mathscr{I}_3\hat{E}_{11}(u,U_i(s))^{-1}\tilde{Q}_i(s)K_{h_u}(s-u)\right]^{\otimes2}.$$

CHAPTER 4: SIMULATION STUDIES

In order to assess the finite sample performance of the proposed model and estimation procedure, we conduct a few simulation studies in this chapter. First, we apply the methodologies on a survival analysis, which is a single-event case. Second, the methodologies are applied to recurrent event case. Details for simulations for both cases will be presented in the following subsections.

4.1 Simulation on Single Event Data - Survival Analysis

In this section, we illustrate our method on single event data and check the finite sample performance of the proposed method. This section first introduces the method that is used to generate single event counting process data, followed by a simulation example with a specific model by using the proposed method.

4.1.1 Generating Single Event Data

Some researchers had studied on how to generate single event data for survival anaysis. For instance, quite many researchers proposed methods to simulate survival data with Cox model proposed by Cox (1972).

In the Cox model, the intensity function is written as follows.

$$\lambda(t|X) = \lambda_0(t) \exp\{\beta^T X\}, \tag{4.1}$$

where $\lambda_0(t)$ is the baseline intensity function, β is the parameter vector and X is

a time independent covariate vector with the same dimention as β. The cumulative intensity function $\Lambda(t|X)$can be written as the integral of $\lambda(t|X)$ from 0 to t as follows.

$$\Lambda(t|X) = \int_0^t \lambda(s|X)ds, \qquad (4.2)$$

The survival function has the following relationship with cumulative intensity function.

$$S(t|X) = \exp\{-\Lambda(t|X)\}, \qquad (4.3)$$

Let $F(t|X) = 1 - S(t|X)$, then we have

$$F(t|X) = 1 - \exp\{-\Lambda(t|X)\}, \qquad (4.4)$$

Supposed Y is a random variable, which follows uniform distribution on $[0, 1]$. By setting $F(t|X) = Y$, we can derive the corresponding T by solving the inverse function as below. T is the desired time to event.

$$T = F^{-1}(Y) = \Lambda^{-1}(-\ln(1 - Y)), \qquad (4.5)$$

where $F^{-1}()$ and $\Lambda^{-1}()$ are the inverse functions of $F()$ and cumulative intensity function $\Lambda()$. In the real world, observations can be censored. For example, patients do not experience death by the end of the study period and thus are censored. In order to take into account censoring, we set the study period to be τ. If the T generated by the equation above is greater than τ, the corresponding subject is marked as censored.

With a given intensity model, a survival dataset with single event can be generated by following the steps.

1. Set the true value for parameter vector β, true function for $\lambda_0(t)$, study period

τ, and the total number of observations n (aka. sample size).

2. Simulate the time independent covariates vector X.

3. Plug in all information from step 1 and 2 to (4.1) and calculate (4.1) through (4.4).

4. Simulate a number y from the distribution $Uniform[0,1]$.

5. Plug y into (4.5), find T.

6. Compare T with τ, if T is less than τ, set the censoring indicator to be 0, otherwise set it to be 1.

7. Repeat step 2 through step 6 for n times to simulate data for n subjects.

4.1.2   Simulation Example

By following the procedure presented in the previous subsection, a single-event sample is generated to be used for evaluating the finite sample performance of our model. In this example, we consider the following hazard model for failure time to illustrate our method.

$$\lambda(t) = exp\left\{\alpha_0(t) + \alpha_1(t)X + \beta Z + \gamma(t - S)W\right\}, \tag{4.6}$$

for $0 \le t \le \tau$ with $\tau = 2$, with the following settings.

- $\alpha_0(t) = -1.5 + 0.8t$, $\alpha_1(t) = t$, $\beta = -0.5$, $\gamma(U_i(t)) = -0.5u$;

- X_i is generated from truncated normal distribution $N(-0.5, 0.5, 0, 1)$, Z_i is an

uniform random variable on $[-0.5, 0.5]$ and W_i is generated from the distribution $Binary(0.5)$;

- $U_i(t) = t - S_i$, where S_i is generated from the uniform distribution $U[0, 0.5]$;

- Censoring time C_i is generated from the uniform distribution $U[1, 3]$.

In the generated sample, about 50 % of subjects are censored. Approximately 50 % of subjects experience an event during the study period $[0, 2]$.

During estimation, cross-validation method described in Chapter 2 is applied for preliminary bandwidth selection. In this example, three sets of bandwidth combination for h_t and h_u are selected to reflect different levels of smoothness, including

- $h_t = h_u = 0.30$

- $h_t = h_u = 0.40$

- $h_t = h_u = 0.50$

Boundaries effect is taking into consideration in all simulation examples in this study, thus we set $t_1 = h_t$ and $t_2 = \tau - h_t$ in the estimating functions in chapter 2. For all simulations examples in this study, we use the Epanechnikov kernel for smoothing, which is given as follows.

$$K(x) = .75(1 - x^2)I(|x| \leq 1)$$

For all simulation examples in this dissertation, we consider some criterias to valuate the performance of the proposed estimators, which are presented as follows.

In order to assess the performance of the estimator $\hat{\beta}$, we measure the following items in all simulation examples.

- Bias = estimate - true value

- the sample standard error of the estimates (SSE)

- the sample mean of the estimated standard errors (ESE)

- the 95% empirical coverage probability (CP)

In order to assess the performance of the estimators $\hat{\alpha}_0(t)$ for baseline, $\hat{\alpha}_1(t)$ for time varying covariate effects and $\hat{\gamma}(u)$ for covariate-varying covariate effect, we measure their pointwise Bias, SSE, ESE, and CP at different fixed time points t and u, respectively.

Besides Bias, SSE, ESE and CP, to better assess the overall performance for those estimators for unknown functions, we calculate the sqaure root of integrated mean square error (RMSE) for each of them respectively.

Suppose

- N is the total number of repetitions;

- $\alpha_{00}(t)$ and $\alpha_{10}(t)$ are the true function values of $\alpha_0(t)$ and $\alpha_1(t)$ at each time point $t \in [0, \tau]$, respectively;

- $\gamma_{10}(u)$ and $\gamma_{20}(u)$ are the true function values of $\gamma_1(u)$ and $\gamma_2(u)$ at each point $u \in [0, \tau]$, respectively;

then, the RMSEs for each of estimators in this chapter are defined as follows.

$$RMSE_{\alpha_0} = \left\{ \frac{1}{N(\tau - 2h_t)} \sum_{j=1}^{N} \int_{h_t}^{\tau - h_t} (\hat{\alpha}_{0j}(t) - \alpha_{00}(t))^2 dt \right\}^{1/2},$$

$$RMSE_{\alpha_1} = \left\{ \frac{1}{N(\tau - 2h_t)} \sum_{j=1}^{N} \int_{h}^{\tau - h} (\hat{\alpha}_{1j}(t) - \alpha_{10}(t))^2 dt \right\}^{1/2},$$

$$RMSE_{\gamma_1} = \left\{ \frac{1}{N(\tau - 2h_u)} \sum_{j=1}^{N} \int_{h_u}^{\tau - h_u} (\hat{\gamma}_{1j}(u) - \gamma_{10}(u))^2 du \right\}^{1/2},$$

$$RMSE_{\gamma_2} = \left\{ \frac{1}{N(\tau - 2h_u)} \sum_{j=1}^{N} \int_{h_u}^{\tau - h_u} (\hat{\gamma}_{2j}(u) - \gamma_{20}(u))^2 du \right\}^{1/2},$$

where $\hat{\alpha}_{0j}(t)$, $\hat{\alpha}_{1j}(t)$, $\hat{\gamma}_{1j}(u)$ and $\hat{\gamma}_{2j}(u)$ are the jth estimate of $\alpha_0(t)$, $\alpha_1(t)$, $\gamma_1(u)$ and $\gamma_2(u)$, respectively, for $j = 1, \cdots, N$.

Sample sizes $n = 400, 600, 800, 1000$ are considered in this example. All results presented below are calculated based on 500 simulation repetitions. Table 2 summarizes the Bias, SSE, ESE and CP for the fixed covariate coefficient estimator $\hat{\beta}$ under model (4.6). The results show the following.

1. The bias for the estimator $\hat{\beta}$ are small among different sample sizes and bandwidth combinations, which indicates that the estimates are unbiased.

2. Both empirical and estimated standard errors presented on Table 2 are reasonably close to each other, and thus the coverage probabilities are close to the nominal level 95%.

Note that for a particular sample size, when the bandwidth gets larger, Bias decreases while both empirical and estimated standard errors increases. Fan and Gijbels (1996) find that bigger bandwidth results in larger variance but smaller bias. Our

results are consistent with their finding. However, the results show that coverage probabilities are not sensitive to bandwidth selection. To conclude, Table 2 indicates that the proposed estimator $\hat{\beta}$ for fixed covariate effect performs well under model (4.6).

Figure 1 and Figure 2 summarize the results produced with the bandwidth $h_t = 0.40$ and $h_u = 0.40$. The results are averaged on 500 simulation repetitions. The red curve corresponds to the results for sample size $n = 400$, while blue curve for $n = 600$, green curve for $n = 800$ and black curve for $n = 1000$, respectively.

The left panel of Figure 1 presents the Bias, SSEs, ESEs and CPs at different fixed time points for $\hat{\alpha}_0(t)$, while the right panel for $\hat{\alpha}_1(t)$. The results show that the bias for both estimators are small. The coverage probability fluctuates around the nominal level 95%. In addition, the plots show that when sample size is 400 or 600, the estimated standard errors are not stable for the latter half of study period. However, when sample size increases ($n = 800, 1000$), SSE and ESE agree to each other very well. This could be due to data sparcity when sample size is not considerably large. The estimators perform better when sample size increases.

Figure 1 show that the pointwise bias for $\hat{\gamma}(u)$ are very small, which indicates that the pointwise estimates are unbiased. An agreement is observed between pointwise SSE and ESE, thus the coverage probability curves slightly fluctuate around the line of 97%, which shows reasonable performance of the estimator $\hat{\gamma}(u)$.

Table 3 summarizes the RMSEs based on 500 simulation repetitions for $\hat{\alpha}_0(t)$, $\hat{\alpha}_1(t)$ and $\hat{\gamma}(u)$ under model 4.6. The results show that the RMSEs for all those four estimators decrease when sample size increases. The same trend is observed for all

three selected bandwidth combinations.

Table 2: Summary of Bias, SSE, ESE and CP for $\hat{\beta}$ under model (4.6).

n	h_t	h_u	Bias	SSE	ESE	CP
400	0.30	0.30	-0.0333	0.3481	0.3511	0.966
	0.40	0.40	-0.0241	0.3630	0.3855	0.966
	0.50	0.50	-0.0097	0.4109	0.4385	0.972
600	0.30	0.30	0.0055	0.4871	0.2822	0.956
	0.40	0.40	-0.0047	0.2939	0.3066	0.958
	0.50	0.50	-0.0014	0.3315	0.3517	0.968
800	0.30	0.30	-0.0159	0.2283	0.2347	0.964
	0.40	0.40	-0.0128	0.2488	0.2622	0.964
	0.50	0.50	-0.0103	0.2902	0.3020	0.964
1000	0.30	0.30	-0.0168	0.1989	0.2079	0.968
	0.40	0.40	-0.0136	0.2180	0.2330	0.966
	0.50	0.50	-0.0146	0.2493	0.2687	0.970

Table 3: Summary of RMSEs for $\hat{\alpha}_0(t)$, $\hat{\alpha}_1(t)$ and $\hat{\gamma}(u)$ under model (4.6).

n	h_t	h_u	RMSE_{α_0}	RMSE_{α_1}	RMSE_{γ}
400	0.30	0.30	0.2701	0.6550	1.0922
	0.40	0.40	0.2134	0.5475	0.3863
	0.50	0.50	0.1849	0.4778	0.3384
600	0.30	0.30	0.2244	0.5154	0.6548
	0.40	0.40	0.1648	0.4418	0.2806
	0.50	0.50	0.1457	0.3874	0.2384
800	0.30	0.30	0.1659	0.4422	1.1642
	0.40	0.40	0.1404	0.3819	0.2240
	0.50	0.50	0.1239	0.3374	0.1934
1000	0.30	0.30	0.1451	0.3996	0.5083
	0.40	0.40	0.1243	0.3488	0.2057
	0.50	0.50	0.1111	0.3114	0.1800
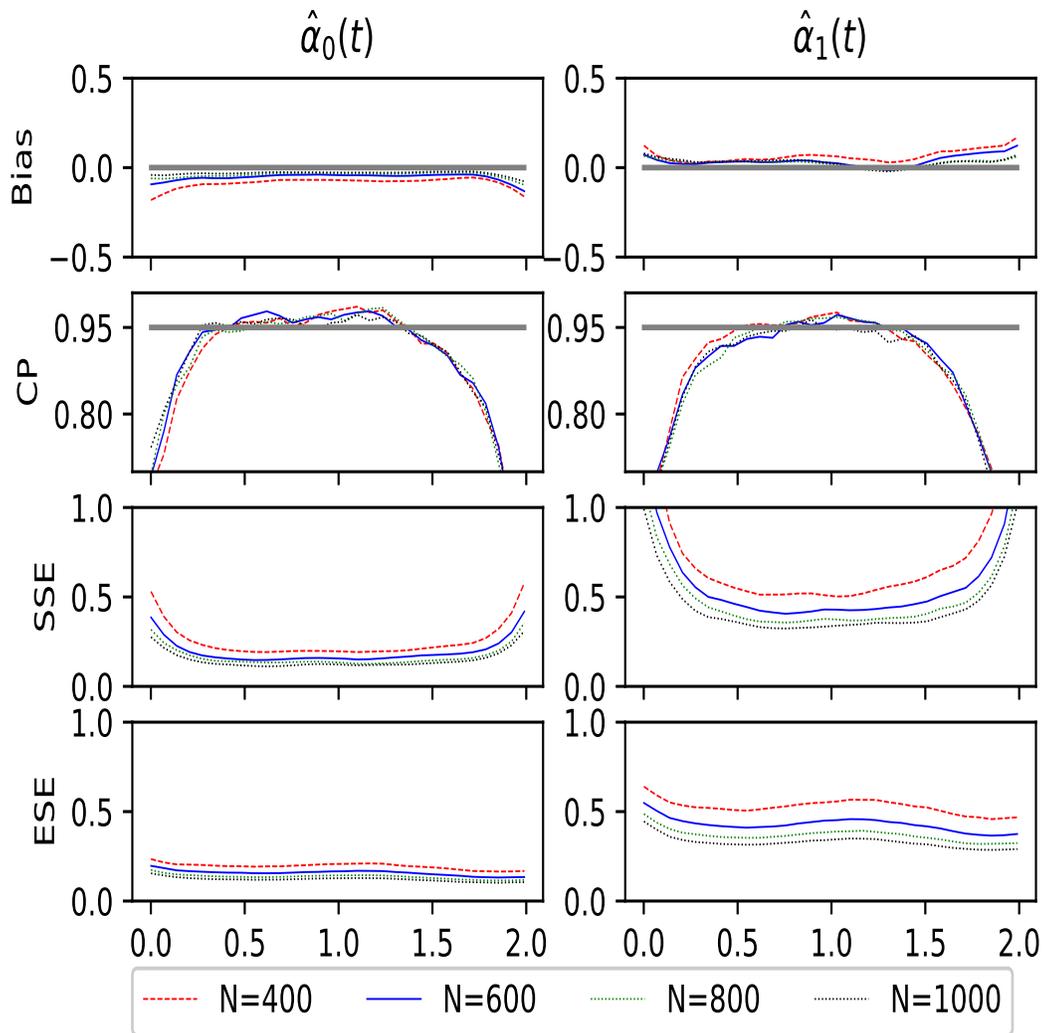
Figure 1: Plots for Bias, CP, SSE and ESE for n=400, 600, 800, and 1000 with $h_t = 0.4, h_u = 0.4$ for $\alpha_0(t) = -1.5 + 0.8t$ and $\alpha_1(t) = t$ for $0 \leq t \leq 2$ under model (4.6).
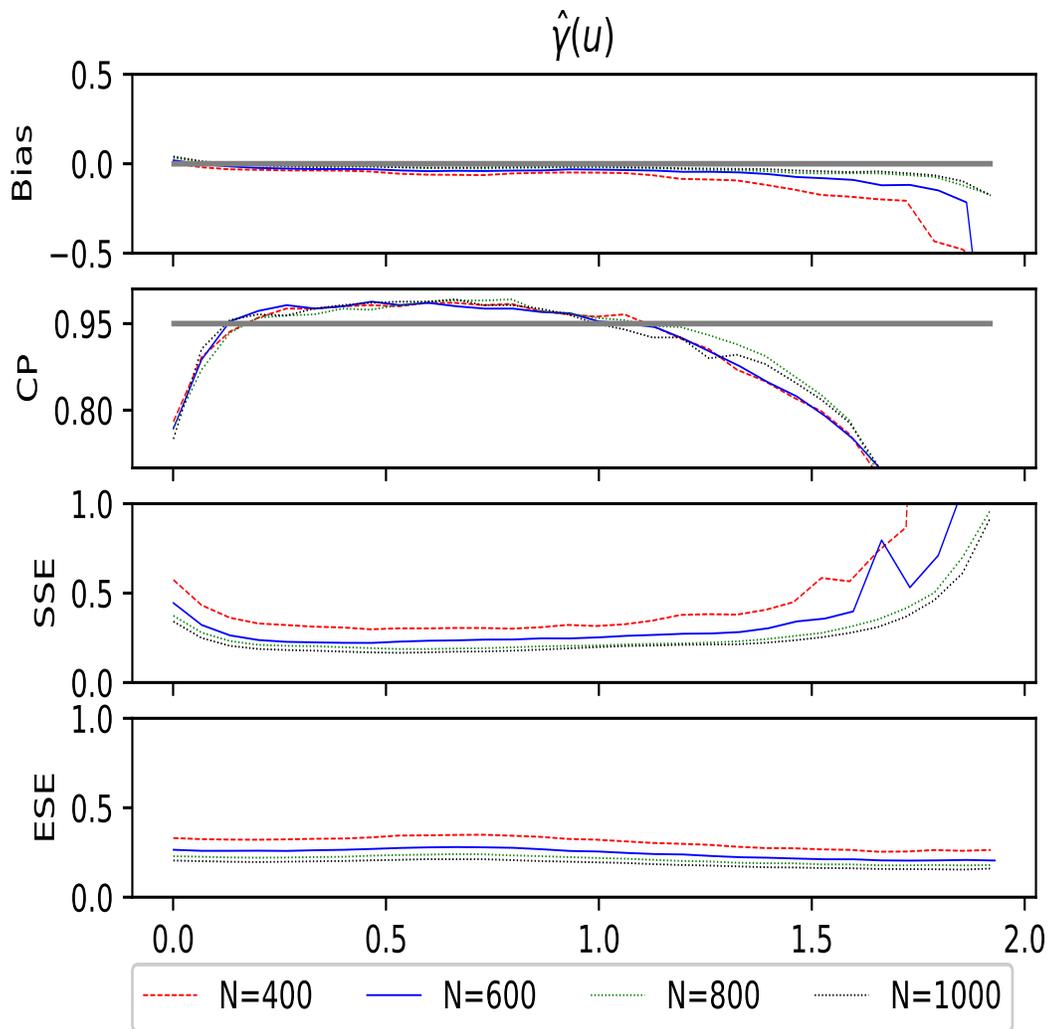
Figure 2: Plots for bias, CP, SSE and ESE for n=400, 600, 800, and 1000 with $h_t = 0.4, h_u = 0.4$ for $\gamma(u) = -0.5u$ under model (4.6).

4.2   Simulation on Recurrent Events Data

The previous section show that the proposed model and estimation procedure work well under the survival analysis framework. In this section, the goal is to check the finite sample performance of the proposed methodology when the event of interest can be repeated for each subject. This section starts with a description on how to generate recurrent events data by thinning method, followed by two simulation examples under different models.

4.2.1   Generating Recurrent Events Data

There are many different existing methods for generating a non-Homogeneous Poisson Process (NHPP). For instance, a NHPP can be generated through time-scale transformation of a Homogeneous Poisson Process (HPP) or it can be generated by using order statistics. Lewis and Shedler (1976) generated a NHPP with log linear rate function. Another approach for generating a NHPP is the thinning method proposed by Lewis and Shedler (1979). Compared with other methods, the thinning method has many advantages. For instance, the thinning method does not require numerical integration of the rate function, ordering of points, or generating Poisson variates. Thus, in this dissertation, we adopt the thinning method for simulating recurrent events data.

Given the intensity function $\lambda(t)$ for $0 \leq t \leq \tau$, choose a constant $\bar{\lambda}$ $s.t.$ $\lambda(t) \leq \bar{\lambda}$ for all t. The detailed procedure for generating recurrent events by using thinning method are given as follows. By repeating the procedure below, we can generate the recurrent event times for each subject i in a sample. The recurrent event times are

recorded by T_{ij} $(i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, K_i)$

1. Set $T_0 = 0$, $T^* = 0$, and $j = 1$.

2. Generate an random variable V from exponential distribution $\exp(1/\bar{\lambda})$

3. Update $T^* = T^* + V$.

4. If $T^* > \tau$, stop; otherwise generate a random variable R from uniform distribution $U(0, 1)$.

5. Compare R with $\lambda(T^*)/\bar{\lambda}$. If $R \leq \lambda(T^*)/\bar{\lambda}$, then accept the arrival time, store it by $T_{ij} = T^*$, and $j = j + 1$; otherwise reject the arrival time and return to Step 2.

In the following subsections, this procedure is used to generate recurrent event samples to conduct simulation studies.

4.2.2   Simulation Example 1

We start with a simple model to illustrate the proposed method and evalute the finite sample performance of the proposed estimators. The model given as follows considers an intercept term as the baseline, time-independent effect and covariate-varying effect.

$$\lambda_i(t) = exp\left\{\alpha_0(t) + \beta Z_i + \gamma(U_i(t))W_i\right\}, \qquad (4.7)$$

for $0 \leq t \leq \tau$ with $\tau = 5$, with the following settings.

- $\alpha_0(t) = 1.5 - \log(1 + t)$, $\beta = 1.5$, $\gamma(U_i(t)) = \sqrt{U_i(t)} - 2$;

- Z_i is an uniform random variable on $[-0.5, 0.5]$ and W_i is generated from truncated normal distribution $N(-0.5, 0.5, 0, 1)$,.

- $U_i(t) = t - S_i$, where S_i is generated from the uniform distribution $U[0, 0.5]$.

- Censoring time C_i for the ith subject is generated from the uniform distribution $U[4, 9]$.

In the generated sample, about 20% of subjects are censored. Approximately a total of 12 recurrent events are observed per subject during the study period $[0, 5]$.

Three sample sizes ($n = 400, 600, 800$) are considered in this example. All results presented below are calculated based on 500 simulation repetitions. Cross-validation method is applied for bandwidth selection. The following three sets of bandwidth combinations for h_t and h_u are selected to reflect different levels of smoothness.

- $h_t = h_u = 0.25$

- $h_t = h_u = 0.30$

- $h_t = h_u = 0.35$

Table 4 summarizes the Bias, SSE, ESE, and CP for the fixed covariate coefficient estimator $\hat{\beta}$ under model (4.7) based on 500 simulation repetitions. The results show the following.

1. The bias for the estimator $\hat{\beta}$ are small among different sample sizes and bandwidth combinations, which indicates that the estimates are unbiased.

2. Both empirical and estimated standard errors presented on Table 4 agree to each other, which results coverage probabilities that are close to the nominal level 95%. In addition, coverage probabilities are found to be not sensitive to bandwidth selection.

3. Bias, empirical standard errors, and estimated standard errors all decrease when sample size increases.

4. For a particular sample size, when the bandwidth gets larger, Bias decreases while both empirical and estimated standard errors increases, which is again consistent with the finding by Fan and Gijbels (1996).

To conclude, Table 4 indicates that the proposed estimator $\hat{\beta}$ for fixed covariate effect performs well under model (4.7).

Figure 3 plots the pointwise bias, empirical standard errors, estimated standard errors and coverage probabilities for the estimate of the baseline function $\alpha_0(t)$, while Figure 4 shows the plots for $\hat{\gamma}(u)$. All plots are generated based on the results from 500 simulation repetitions. The bandwidth combination used to generate these plots is h_t=0.3, h_u=0.3. The red curve represents result for sample size $n = 400$, while blue for $n = 600$ and green for $n = 800$. The plots reveal the following findings.

1. The pointwise bias for $\hat{\alpha}_0(t)$ is reasonably small. When sample size increases, bias decreases. Bias tends to become relatively larger at the end of study period, which could be the boundary effects.

2. The pointwise bias for $\hat{\gamma}(u)$ is reasonably small and the curve fluctuates around

the zero line.

3. For both $\hat{\alpha}_0(t)$ and $\hat{\gamma}(u)$, the pointwise empirical standard error and estimated standard errors are very close to each other, and consequently the curves for pointwise coverage probabilities fluctuate around the nominal level 95%.

Besides pointwise bias, SSE, ESE and CP, RMSEs are also calculated to assess the overall performance of both $\hat{\alpha}_0(t)$ and $\hat{\gamma}(u)$. Table 5 shows the results under model (4.7) with different sample sizes and bandwidth combinations. It appears that RMSEs for both $\hat{\alpha}_0(t)$ and $\hat{\gamma}(u)$ drops when sample size increases, regardless of bandwidth combination.

To conclude, based on the results given by Figure 3, 4 and Table 5, the proposed estimator $\hat{\alpha}_0(t)$ and $\hat{\gamma}(u)$ show satisfied performance under model (4.7).

Table 4: Summary of Bias, SSE, ESE and CP for $\hat{\beta}$ under model (4.7).

n	h_t	h_u	Bias	SSE	ESE	CP
400	0.25	0.25	0.0173	0.0724	0.0697	0.934
	0.30	0.30	-0.0016	0.0736	0.0702	0.938
	0.35	0.35	0.0088	0.0754	0.0722	0.938
600	0.25	0.25	0.0112	0.0580	0.0567	0.944
	0.30	0.30	-0.0051	0.0579	0.0572	0.950
	0.35	0.35	0.0063	0.0596	0.0589	0.954
800	0.25	0.25	0.0070	0.0498	0.0490	0.936
	0.30	0.30	-0.0089	0.0509	0.0495	0.938
	0.35	0.35	0.0031	0.0529	0.0510	0.940

Table 5: Summary of RMSEs for $\hat{\alpha}_0(t)$ and $\hat{\gamma}(u)$ under model (4.7).

n	h_t	h_u	RMSE_{α_0}	RMSE_{γ}
400	0.25	0.25	0.0862	0.2981
	0.30	0.30	0.0781	0.2770
	0.35	0.35	0.0728	0.2643
600	0.25	0.25	0.0683	0.2395
	0.30	0.30	0.0625	0.2230
	0.35	0.35	0.0584	0.2126
800	0.25	0.25	0.0584	0.2089
	0.30	0.30	0.0536	0.1934
	0.35	0.35	0.0502	0.1846
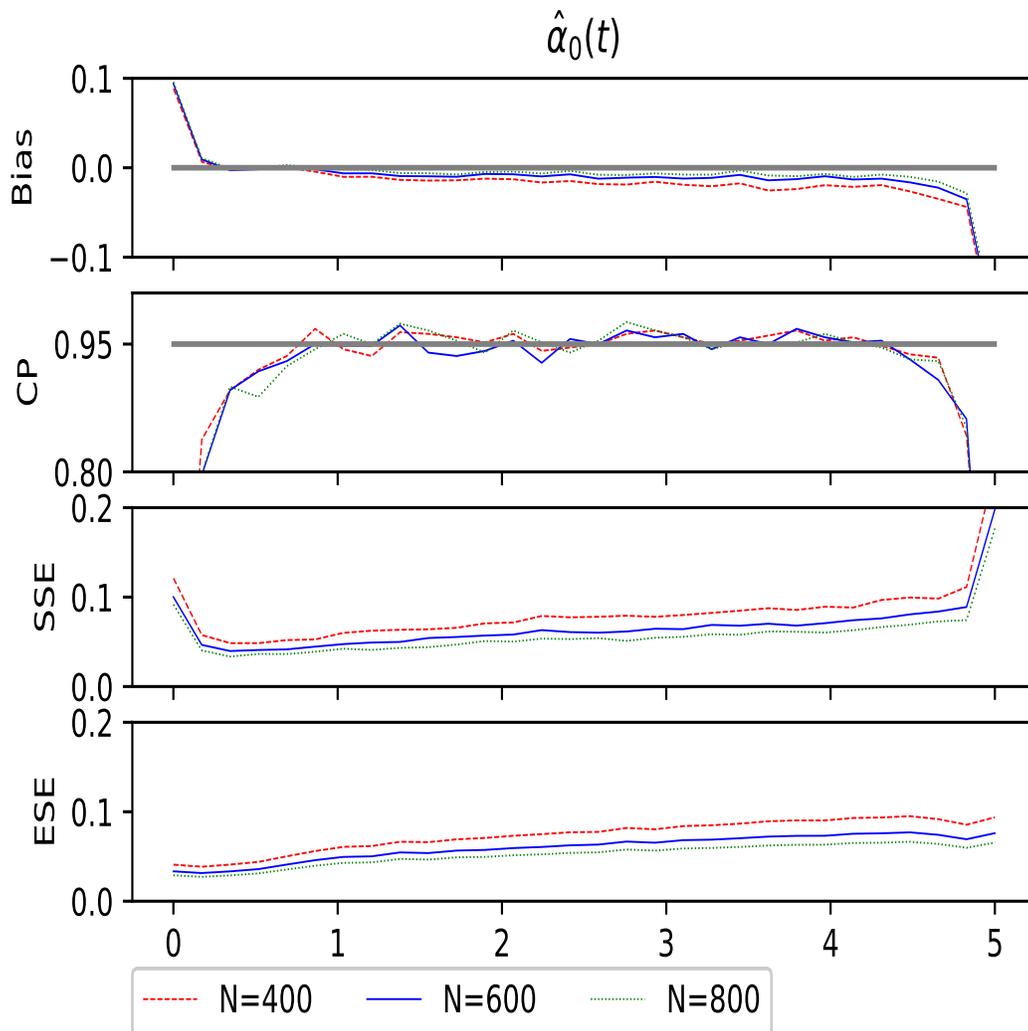
Figure 3: Plots for bias, CP, SSE and ESE for n=400, 600, 800 with h_t=0.3, h_u=0.3 for $\alpha_0(t) = 1.5 - \log(1 + t)$ for $0 \leq t \leq 5$ under model (4.7).
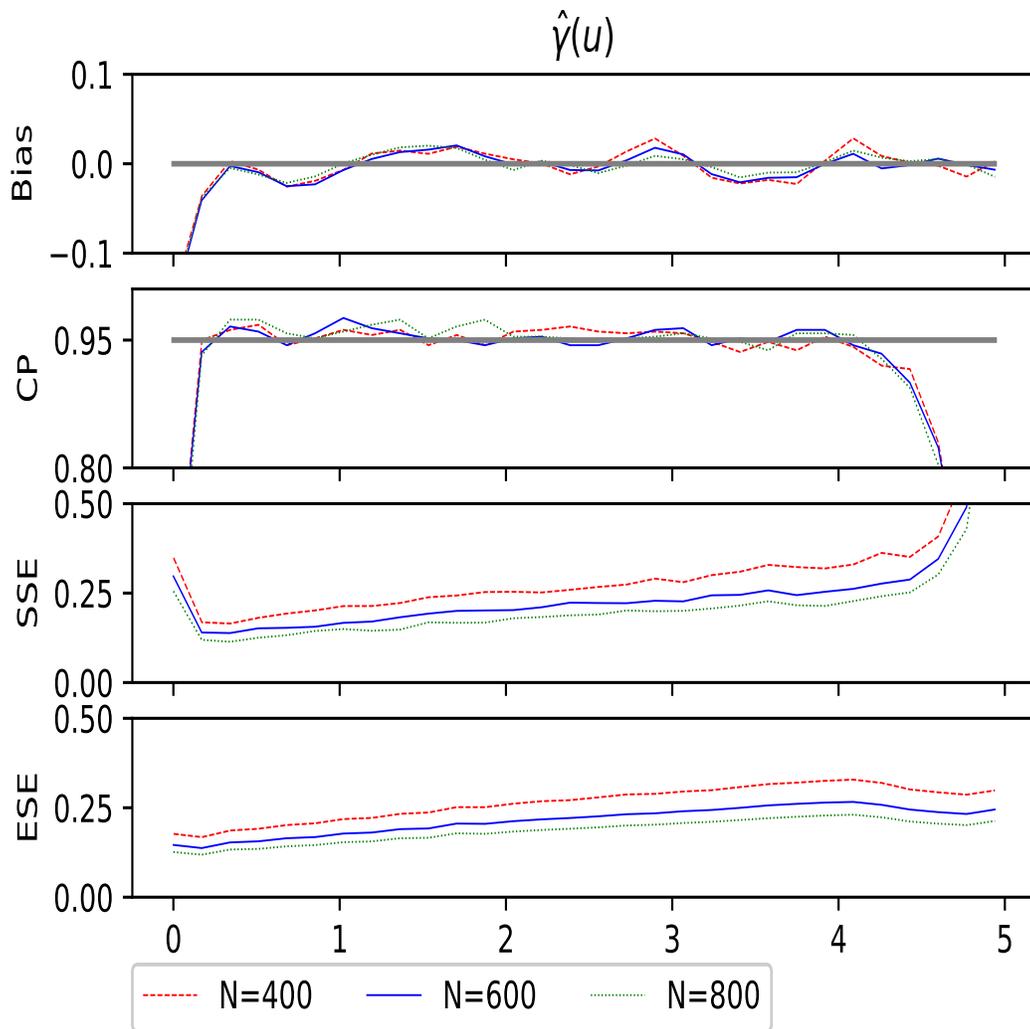
Figure 4: Plots for bias, CP, SSE and ESE for n=400, 600, 800 with h_t=0.3, h_u=0.3 for $\gamma(u) = \sqrt{u} - 2$ under model (4.7).

4.2.3   Simulation Example 2

The previous simulation example considers fixed covariate effects and covariate-varying effects. The results show that the proposed method performs well. In example 2, a more complicated model is considered to illustrate the proposed method, which incorporates fixed covariate effects, time-varying effects and covariate-varying effects. The model is given as follows.

$$\lambda_i(t) = exp\left\{\alpha_1(t) + \alpha_2(t)X_i(t) + \beta Z_i + \gamma_1(U_i(t))W_{1i} + \gamma_2(U_i(t))W_{2i}\right\}, \qquad (4.8)$$

for $0 \leq t \leq \tau$ and $\tau = 5$, with the following settings.

- $\alpha_1(t) = 2 - \log(1+t)$, $\alpha_2(t) = \sin(0.2t)$;

- $\beta = 1.5$;

- $\gamma_1(U_i(t)) = U_i(t) - 1$, $\gamma_2(U_i(t)) = \sqrt{U_i(t)} - 2$;

- Covariate X_i is generated from truncated normal distribution $N(-0.5, 0.5, 0, 1)$;

- Z_i is generated from uniform distribution $U[-0.5, 0.5]$;

- W_{1i} and W_{2i} are generated from truncated bivariate normal with marginal $N(-0.5, 0.5, 0, 1)$ and correlation $\rho = 0.2$;

- Cencoring time C_i is generated from an uniform distribution $U[4, 9]$;

- $U_i(t) = t - S_i$, where S_i is generated from an uniform distribution $U[0, 0.5]$.

With the above settings, about 20% of subjects are cencored. Approximately a total of 12 recurrent events are observed per subject during the study period $[0, 5]$.

Similar to previous examples, we set $t_1 = h_t$ and $t_2 = \tau - h_t$ in the estimating functions to deal with boundary effects. Cross-validation method is applied for bandwidth selection, and the following three sets of bandwidth combination for h_t and h_u are selected to reflect different levels of smoothness.

- $h_t = h_u = 0.25$

- $h_t = h_u = 0.30$

- $h_t = h_u = 0.35$

Three sample sizes $(n = 400, 600, 800)$ are considered in this tudy. All results presented below are calculated based on 500 simulation repetitions.

Table 6 summaries the Bias, SSE, ESE, and CP for the fixed covariate coefficient estimator $\hat{\beta}$ under model (4.8), averaging over 500 simulation repetitions. The results show that the bias for the estimator $\hat{\beta}$ are small among different sample sizes and bandwidth combinations, which indicates that the estimates are unbiased. Both empirical and estimated standard errors presented on Table 6 agree to each other, and thus the coverage probabilities are close to 95%. In addition, Bias, empirical standard errors, and estimated standard errors all decrease when sample size increases. However, the results show that coverage probabilities are not sensitive to bandwidth selection. To conclude, Table 6 indicates that the proposed estimator $\hat{\beta}$ for fixed covariate effect performs well under model (4.8).

To assess the performance of the estimators for time varying covariate effects $\hat{\alpha}_1(t)$ and $\hat{\alpha}_2(t)$ and those for covariate-varying covariate effects $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$, we calculate their Bias, SSE, ESE, and CP at different fixed time points t and u, respectively.

Figure 5 and Figure 6 summarize the results produced with the bandwidth $h_t = 0.30$ and $h_u = 0.30$. The results are averaged on 500 simulation repetitions. The left panel of Figure 5 presents the Bias, SSEs, ESEs and CPs at different fixed time points for $\hat{\alpha}_1(t)$, while the right panel for $\hat{\alpha}_2(t)$. The left panel of Figure 6 shows the Bias, SSEs, ESEs and CPs at different fixed time points for $\hat{\gamma}_1(t)$, while the right panel for $\hat{\gamma}_2(t)$.

The plots from Figure 5 and Figure 6 show the following findings.

1. The pointwise bias for all four sets of estimates $\hat{\alpha}_1(t)$, $\hat{\alpha}_2(t)$, $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$ are very small, thus the pointwise estimates are unbiased. In addition, bias decreases along with increasing sample size.

2. For each of all four estimators, an agreement is observed between pointwise empirical standard error and estiamted standard error. The coverage probability curves slightly fluctuate around the line of nominal level 95%.

To assess the overall performance of $\hat{\alpha}_1(t)$, $\hat{\alpha}_2(t)$, $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$, the RMSE is calculated for each of them. Table 7 summarizes the RMSEs based on 500 simulation repetitions for $\hat{\alpha}_1(t)$, $\hat{\alpha}_2(t)$, $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$ under model 4.8. The results show that the RMSEs for all those four estimators decrease when sample size increases. The same trend is observed for all three selected bandwidths.

To conclude, the proposed estimators $\hat{\alpha}_1(t)$, $\hat{\alpha}_2(t)$, $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$ for unknown non-parametric functions shows strong performance under model 4.8.

In this chapter, in order to illustrate our proposed method and assess the finite sample performance, two simulation examples are conducted on survival analysis framework and another two simulation examples on recurrent events data framework.

All examples show reasonably small bias for our estimates and an agreement between empirical standard errors and estimated standard errors, which result in coverage probabilities that are close to the nominal level 95%. Therefore, the proposed method performs very well with finite samples.

Table 6: Summary of Bias, SSE, ESE and CP for $\hat{\beta}$ under model (4.8).

n	h_t	h_u	Bias	SSE	ESE	CP
400	0.25	0.25	0.0164	0.0532	0.0521	0.930
	0.30	0.30	-0.0018	0.0539	0.0525	0.936
	0.35	0.35	0.0082	0.0559	0.0540	0.936
600	0.25	0.25	0.0099	0.0436	0.0422	0.934
	0.30	0.30	-0.0067	0.0444	0.0426	0.932
	0.35	0.35	0.0046	0.0459	0.0439	0.940
800	0.25	0.25	0.0072	0.0371	0.0364	0.942
	0.30	0.30	-0.0082	0.0380	0.0368	0.938
	0.35	0.35	0.0035	0.0396	0.0379	0.938

Table 7: Summary of RMSEs for $\hat{\alpha}_1(t)$, $\hat{\alpha}_2(t)$, $\hat{\gamma}_1(u)$ and $\hat{\gamma}_2(u)$ under model (4.8).

n	h_t	h_u	RMSE_{α_1}	RMSE_{α_2}	RMSE_{γ_1}	RMSE_{γ_2}
400	0.25	0.25	0.0722	0.2051	0.2190	0.2091
	0.30	0.30	0.0643	0.1891	0.2016	0.1948
	0.35	0.35	0.0597	0.1786	0.1917	0.1855
600	0.25	0.25	0.0567	0.1672	0.1787	0.1685
	0.30	0.30	0.0511	0.1549	0.1654	0.1566
	0.35	0.35	0.0476	0.1463	0.1574	0.1491
800	0.25	0.25	0.0482	0.1451	0.1528	0.1451
	0.30	0.30	0.0437	0.1349	0.1417	0.1353
	0.35	0.35	0.0407	0.1278	0.1349	0.1288
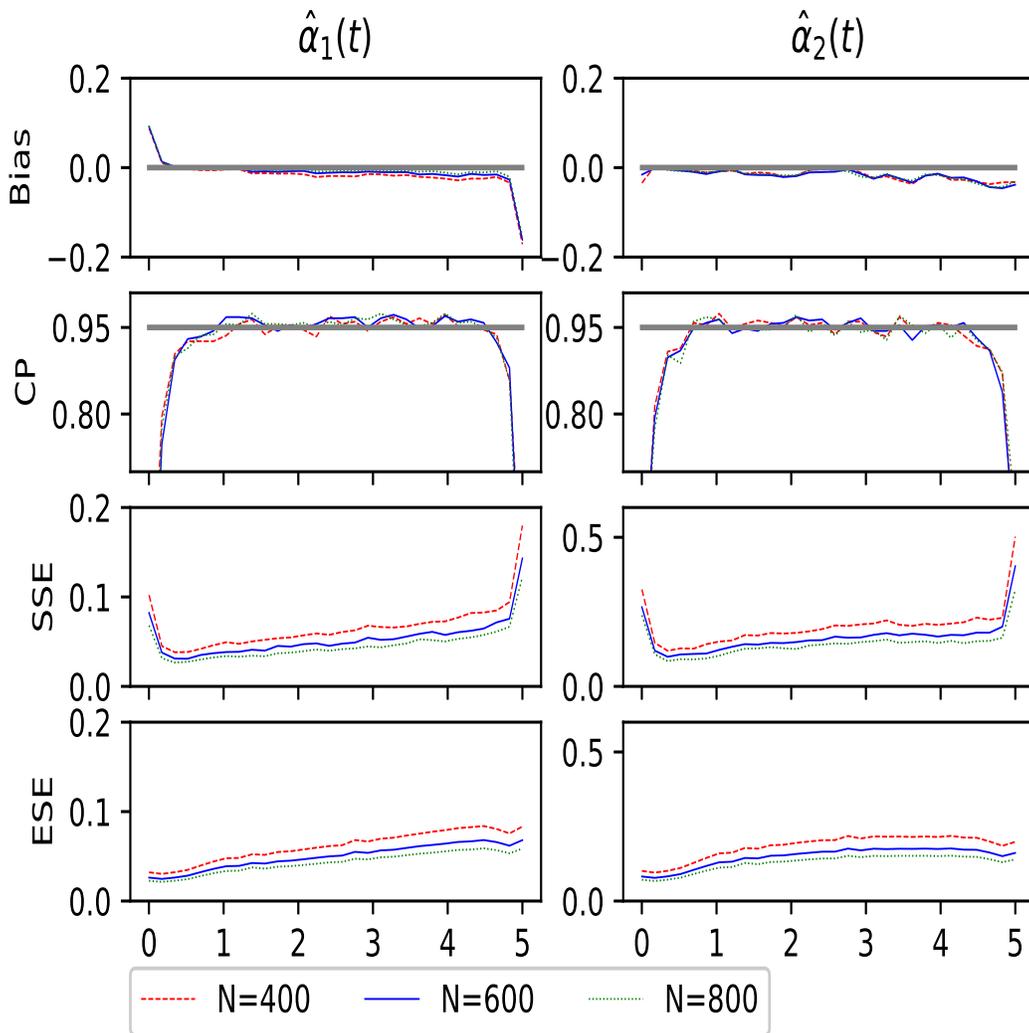
Figure 5: Plots for Bias, CP, SSE and ESE for n=400, 600, 800 with $h_t = 0.3$, $h_u = 0.3$ under model (4.8). Left panel is for $\alpha_1(t) = 2 - log(1 + t)$. Right panel is for $\alpha_2(t) = \sin(0.2t)$.
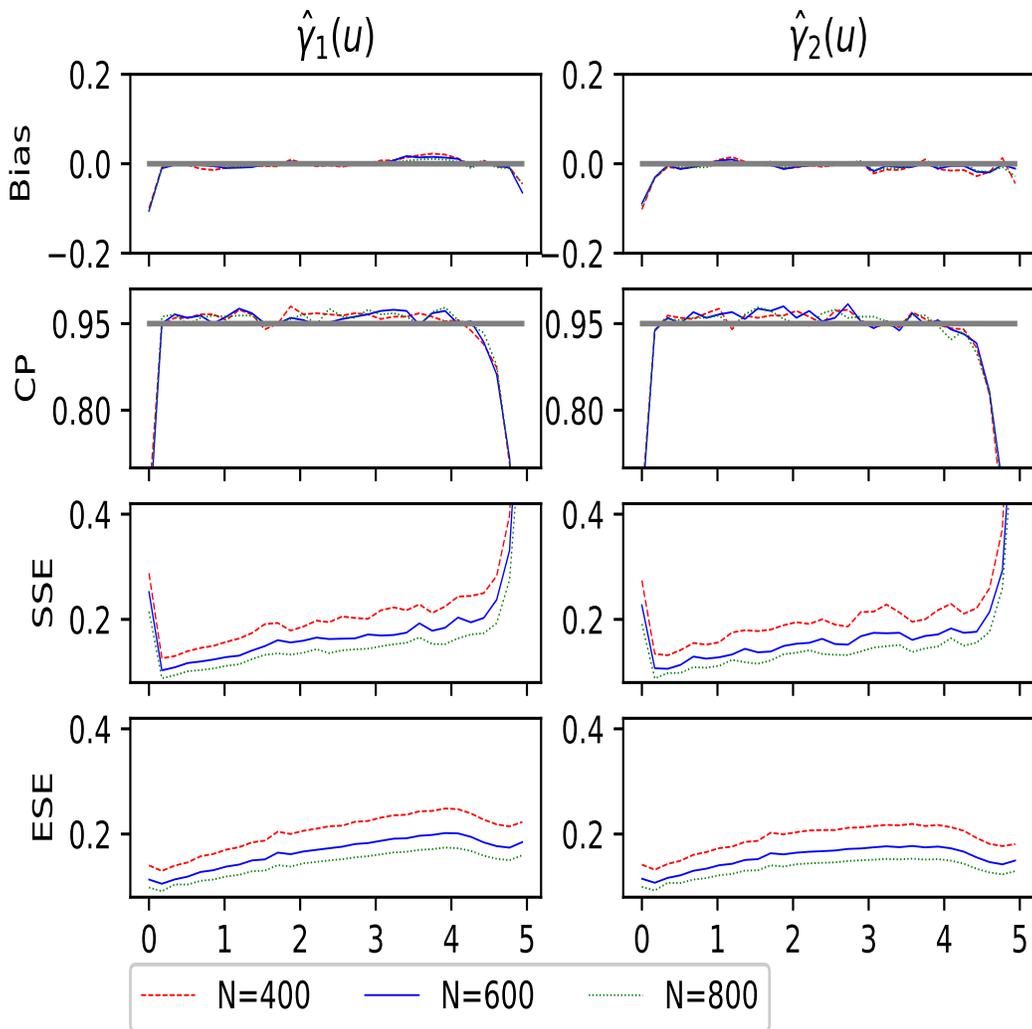
Figure 6: Plots for Bias, CP, SSE and ESE for n=400, 600, 800 with $h_t = 0.3$, $h_u = 0.3$ under model (4.8). Left panel is for $\gamma_1(u) = u - 1$. Right panel is for $\gamma_2(u) = \sqrt{u} - 2$.

CHAPTER 5: DATA APPLICATIONS ON HEMO STUDY

In the HEMO study, the primary outcome is mortality of hemodialysis patients and the secondary outcomes are hospitalizations with different causes, including cardio-vasculor and infection. In this chapter, the proposed model and estimation procedure will be applied to the HEMO dataset.

This chapter is arranged as follows. In Section 5.1, descriptive statistics on the HEMO dataset will be presented to show the distribution of the data. Section 5.2 is an application on time to composite events, combining deaths and transplants. Section 5.3 presents an application on recurrent hospitalizations. Both applications will consider two time dimentions including follow up time and either duration time of dialysis or age of patients.

5.1   Data Description

In this section, we provide a summary of descriptive statistics from the HEMO dataset. A total of 1846 hemodialysis patients enrolled in this clinical trial and were randomized into different treatment groups for different doses/types of dialysis. They were followed up by the study for up to 6.65 years, unless they experienced death, transplant or transfer to non-trial dialysis center, which can be considered as drop out.

In the application of time to composit events of deaths and transplants, patients

are censored by either transfer or end of study, whichever comes first. Figure 7 shows that a total of 796 patients either transfered to non-trial dialysis center or did not experience death or transplant, and therefore are considered censored in the application of time to compsite events of deaths and transplants. A total of 895 deaths and 197 transplants were observed during the HEMO study.

Since hospitalization records were not collected after patients either received transplant or transfered to non-trial dialysis center. Thus, in the application of recurrent hospitalizations, such patients are considered as censored. Patients that did not experience any of transplant, transfer or death were censored by the end of the trial, which is December 31, 2001. Figure 7 summarizes the numbers of patients with different types of censoring.

A total of 1503 patients had at least one or recurrent hospitalizations and 343 patients did not experience any hospitalizations. A total of 7832 hospitalizations were collected for all patients, which produces approximately 4 hospitalizations per patient during the study period.

In this study, the covariates considered in our model include both categorical covariates and continuous covariates. The detailed description for each covariate is listed as follows.

- AGE: baseline age (in years) at randomization date

- BLACK: black race (1=black, 0=others)

- SEX: gender (1=male, 0=female)

- BALB: mean baseline serum albumin (neph)

- DIABET: diabetic (1=diabetic, 0=non-diabetic)

- DURATION: years the patient had been on dialysis prior to randomization date

- FLUX: randomized flux group (1=low-flux, 0=high-flux)

- KTV: randomized Kt/V group (1=standard-ktv, 0=high-ktv)

- ICED: baseline ICED score (=0,1,2,3). The term ICED refers to the Index of Coexisting Disease (Greenfield and Nelson, 1992), which is used to quantify a patient's level of comorbidity. It takes on values 0,1,2,3, with 3 the most severe comorbidity.

Figure 8 and Figure 9 visualize the demographics for patients in HEMO study. The demographics are summarized as follows.

1. About 44% of patients are male, while about 56% are female.

2. About 63% of patients are African American, while about 37% being other races.

3. About 45% of patients are diabetic, while about 55% are non-diabetic. As mentioned in Chapter 1, diabetes were shown to be associated with the mortality of hemodialysis patients by Sattar et al. (2012) using Cox's time varying covariate model. We would like to check such association by using our method. In addition, we would like to check if such association can be found between diabetes and hospitalization rates.

4. The number of patients are roughly even across different ICED scores 1, 2 and 3. However, only 4 patients are with ICED score 0. To avoid the issue of sparse data, we combine the categories ICED score 0 and 1 in our analysis.

5. The number of patients for each treatment group is roughly the same, which means that the HEMO study is a balance design in terms of number of patients.

6. The ages of patients in this trial range from 18 to 80 years, with a mean age 56.6 and a median 59.

7. Before enrolled in this clinical trial, each patient had been on dialysis for different amounts of time, which was recorded by the covariate DURATION (in years). It ranges from 0 to about 30 years. With the baseline duration time available, it is of interest to check if covariate effects vary along with duration time at baseline.

8. Upon enrollment, each patient was measured twice for their serum albumin levels. The baseline serum albumin level (BALB) is the mean of those two measurements. As shown by Figure 9, it ranges from 0.1 to 5.7. We consider the baseline serum albumin level as a covariate, trying to explore if the baseline albumin level is associated with mortality or the rate of hospitalizations.

In order to check if the HEMO study is a balanced design, we further investigate the randomization of patients. Table 8 summarizes the randomization of patients for the combination of treatment groups and different groups of each categorical variables, including ICED, BLACK, DIABET and SEX. As shown by the table, covariate levels

were fully taken into accout in the process of randomization of patients into different treatment groups. Table 9 summarizes the randomization of patients for the combination of treatment groups and different values of each continuous variables, including AGE, BALB and DURATION. Descriptive statistics including mean, standard deviation (std), minimum, 25% quartile, 50% quartile, 75% quartile and maximum for the covariate in each treatment group are summarized by Table 9. It is shown that the randomization process also considered the baseline age of patients, baseline albumin serum level and duration of dialysis. Therefore, the HEMO study is a balanced design.
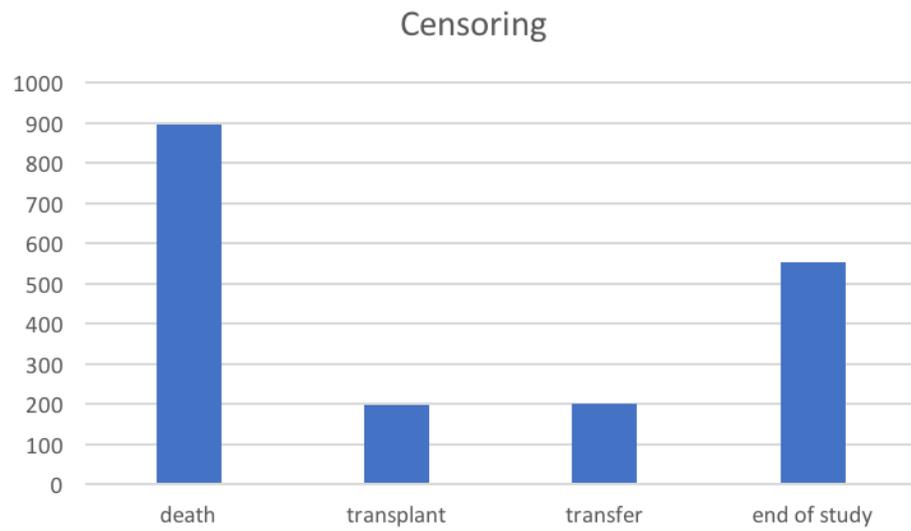
Figure 7: Summary of # of patients for different types of censoring.
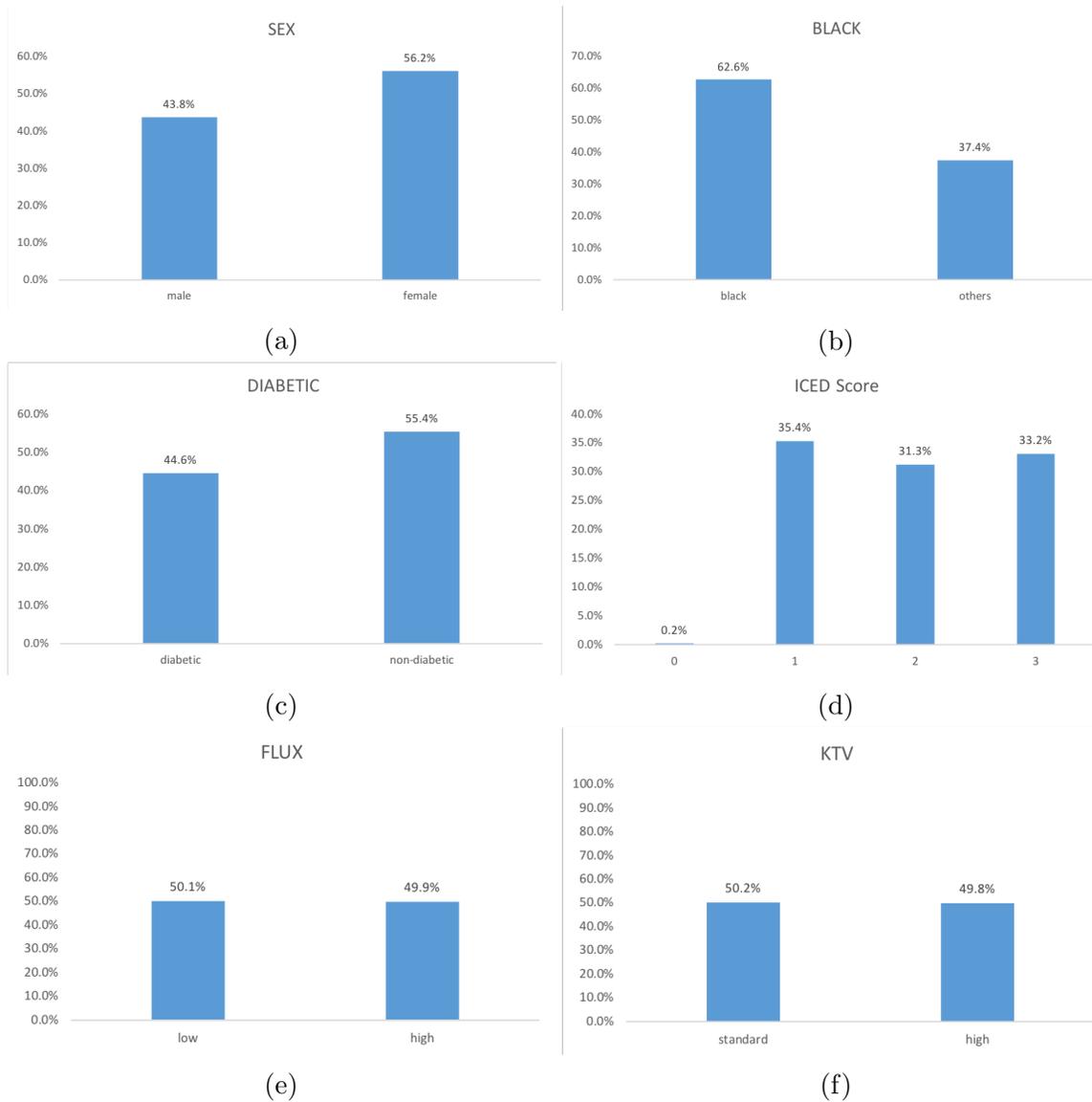
Figure 8: Demographics for patients in HEMO study: (a)-(f) show the percentages of patients for each category of covariates SEX, BLACK, DIABETIC, ICED, FLUX and KTV, respectively.

(a) age

(b) duration of dialysis prior to enrollment

(c) baseline albumin level

(d) distribution of hospitalizations
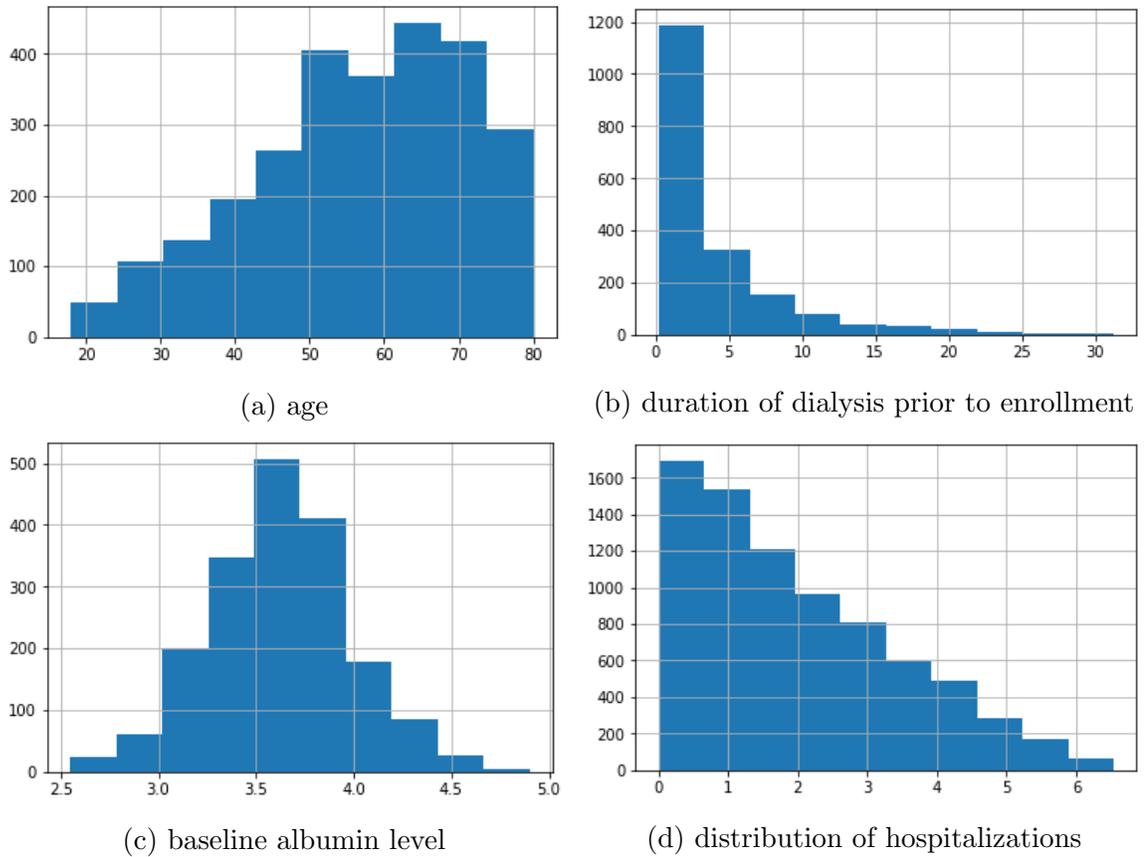
Figure 9: Demographics for patients in HEMO study: (a) is the distribution of age (in years), (b) is the number of years on dialysis prior to enrollment to the study , (c) is the distribution of the baseline serum albumin level (in gm/dL) and (d) is the number of hospitalizations over follow up time.

Table 8: Summary of randomization of patients for the combination of treatment groups and different groups of categorical variables.

Variables		low-flux	high-flux	standard-ktv	high-ktv
ICED	1	338	319	323	334
	2	292	285	286	291
	3	295	317	317	295
BLACK	black	579	577	594	562
	others	346	344	332	358
DIABET	diabetic	411	412	414	409
	non-diabetic	514	509	512	511
SEX	male	409	399	405	403
	female	516	522	521	517

Table 9: Summary of randomization of patients for the combination of treatment groups and different groups of continuous variables.

variable	treatment	mean	std	min	25%	50%	75%	max
AGE	low-flux	57.09	14.18	18.00	48.00	59.00	68.00	80.00
	high-flux	57.19	13.92	18.00	48.00	60.00	68.00	80.00
	standard-ktv	57.34	13.95	18.00	48.00	59.00	68.00	80.00
	high-ktv	920	14.15	18.00	48.00	60.00	68.00	80.00
BALB	low-flux	3.62	0.37	2.55	3.40	3.63	3.87	4.90
	high-flux	3.62	0.35	2.63	3.40	3.64	3.85	4.72
	standard-ktv	3.62	0.36	2.60	3.37	3.63	3.85	4.90
	high-ktv	920	0.36	2.55	3.40	3.64	3.85	4.75
DURATION	low-flux	3.70	4.23	0.19	0.92	2.20	4.68	26.00
	high-flux	3.80	4.48	0.23	0.96	2.11	4.68	31.27
	standard-ktv	3.86	4.46	0.23	1.02	2.27	4.79	31.27
	high-ktv	3.64	4.25	0.19	0.90	2.06	4.43	27.95

5.2   Application on Failure Time to Composite Events

One of the major objectives of the HEMO study was to explore the factors that are associated with the mortality of hemodialysis patients. In this application, we consider death and transplant as composite events. The proposed methodology is applied on the HEMO dataset to explore the factors that are associated with the failure time to composite events to death or transplant.

As a mentioned in Section 5.1, there are a total of 895 deaths and a total of 197 transplants during the HEMO study, with 24 of those patients received transplant before death. Thus, a total of 1068 (57.3%) composite events were detected from the HEMO dataset and 796 (42.7%) patients were censored by either transfer to non-trial clinical or the end of trial.

Before diving into modeling the failure time to composite events, we conduct a bivariate analysis between composite events rate and each covariate considered in this study. For continuous variables, we describe the distribution of each covariate for events group and non-events group seperately. Table 10 shows that the group of patients with composite events are older than those in the non-events group. The baseline serum albumin level for patients with composite events seems to be slightly lower than that for patients without composite events. No noticeable difference is observed for duration of dialysis between the group with composite events and the group without composite events. For categorical covariates, we calculate the rate of composite events for each category of each covariate. Table 11 summarizes the number of composite events in each category for each covariate as well as the rate

of composite events. It is shown that the composite events rate for patients with ICED 1 is much lower than that for patients with ICED score 2 or 3. The composite events rate for patients with ICED score 2 is slightly lower than that for patients with ICED score 3. African American patients has lower composite events rate compared with patients with other races. Patients with diabetes have higher composite events rate than those without diabetes. The composite events rate between male patients and female patients are not noticeable. There are no difference between different treatment groups in terms of composite events rate.

In order to check if and how the treatment interventions and other covariates are associated with time to composite events defined, we consider a full model as follows.

$$\lambda(t) = exp\left\{\alpha_0(t) + \beta_1 FLUX + \beta_2 KTV + \beta_3 AGE + \beta_4 SEX + \beta_5 BALB + \beta_6 DIABET \right.$$
$$\left. + \beta_7 ICED\_1 + \beta_8 ICED\_2 + \beta_9 DURATION + \gamma(t + DURATION)BLACK\right\},$$
$$(5.1)$$

for $0 \leq t \leq 6.65$, with the dummies for ICED groups defined as follows.

- ICED_1 = 1 represents ICED score 1, 0 for others;

- ICED_2 = 1 represents ICED score 2, 0 for others.

In the above model, we have $U(t) = t + DURATION$. Given that DURATION is the baseline duration time prior to study enrollment, u is representing the total duration time since first dialysis. With this setting, we can model the trend of effect difference between African American patients and those with other races over total duration time of dialysis. The total follow up time t is 6.65 years. The range of u is from 0 to 34.3 in years. The left panel of Figure 10 shows the distribution of such events over follow up time t. The right panel of Figure 10 shows the distribution of

such events over duration time of dialysis u. As shown by Figure 10, the number of observed composite events is very low for the tail period of both t and u. Therefore, for the estimations of unkown functions in model (5.1), the tail period $t > 4.5$ and $u > 15$ are not estimated.

Table 12 summarizes the estimates, 95% confidence intervals and p-values for all parameters β_1 to β_9 in model 5.1, respectively. Figure 12 shows the estimated functions for the baseline $\alpha_0(t)$ and $\gamma(u)$. Figure (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively, while Figure (c) and (d) are the estimated functions along with pointwise 95% confidence intervals. The results show the following findings.

1. Both treatment interventions are not significantly associated with the defined composite events to death or transplant for hemodialysis patients. The positive estimates for β_1 indicates that a dialyzer with high-flux membrane compared with low-flux membrane can lower the risk to composite events to death or transplant, while the positive estimates for β_2 suggests that high dose of dialysis delivered can lower the risk to such composite events for hemodialysis patients.

2. The baseline age of patients is very significantly associated with the risk to composite events to death or transplant. In addition, the possitive coefficient suggests that the risk to such composite events is higher for older patients than younger patients.

3. The baseline albumin serum level is significantly associated with the risk to composite events. Such risk increases along with increasing baseline serum

albumin level.

4. Diabetes is significantly associated with the risk to composite events. Compared with non-diabetic patients, diabetic patients have higher risk to composite events to death or transplant, which is consistent with the finding by Sattar et al. (2012) mentioned previously in Section 5.1.

5. The effect difference between ICED score 1 and ICED score 3 and that for ICED score 2 and ICED score 3 are significant. Compared with patients ICED score 3, patients with either ICED score 1 or ICED score 2 have lower risk to composite events to death or transplant.

6. The duration time of dialysis is not significantly associated with the risk to composite events.

7. In Figure 12, Figures (a) and (c) indicate that the baseline function is roughly flat over time. Figures (b) and (d) imply that African American hemodialysis patients tend to have lower risk to composite events to death or transplant compared to patients with other races. In addition, such difference between African American and other races is significant. Figures (b) and (d) also indicate that the difference between African American patients and those with other races decrease when the duration time of dialysis is from 0 to 4 years but then increases when the duration time is from 10 to 14 years.

It is worth mentioned that all above findings are consistent with the descriptive statistics given by Table 10 and Table 11 above.

In order to explore the trend of effect difference between African American and other races over patients' age, another full model with $u(t) = t + AGE$ is fitted as follows.

$$\lambda(t) = exp\{\alpha_0(t) + \beta_1 FLUX + \beta_2 KTV + \beta_3 AGE + \beta_4 SEX + \beta_5 BALB + \beta_6 DIABET \\ + \beta_7 ICED\_1 + \beta_8 ICED\_2 + \beta_9 DURATION + \gamma(t + AGE)BLACK\},$$

$$(5.2)$$

for $0 \le t \le 6.65$.

In this example, we consider $U(t) = t + AGE$, which ranges from 18 to 86.65 years. Given that AGE stands for the baseline age at enrollment, $U(t) = t + AGE$ represents the age of patients at time t. As shown by Figure 11, the number of observed composite events is very low for the right tail of t and left tail of u. Therefore, we do not estimate unkown functions in model (5.2) for the tail period $t > 4.5$ and $u < 40$.

Table 13 presents the coefficient estimates along with the 95% confidence intervals and p-values for all parameters in model 5.2. The results suggest the following findings.

1. Based on the results, baseline serum albumin level (BALB), diabetes (DIABET), ICED score and duration time of dialysis (DURATION) are significantly associated with the risk to composite events to death or transplant for hemodialysis patients, while other covariates are not.

2. Patients with diabetes have higher risk to composite events to death or transplant.

3. The positive coefficient estimates $\hat{\beta}_5$ and $\hat{\beta}_9$ indicate that the risk to composite

events increase along with albumin level and duration time of dialysis. Compared with patients with ICED score 3, those with lower ICED scores (1 or 2) have lower risk to composite events to death or transplant.

4. It is worth to note that given all other covariates in the model, treatment effects and their interaction effects for neither dose of dialysis nor dialyzers with standard or high-flux membrane turn out to be non-significant.

The estimated functions for $\alpha_0(t)$ and $\gamma(u)$ given by Figure 13 show that African American patients have significantly lower risk to composite events compared with patients with other races. Furthermore, such difference increases along with age when patients are between 40 to 60 years old but decreases along with age when patients are between 60 to 80 years old.

Table 10: Distribution of continuous variables in event group and non-event group.

variable	group	mean	std	min	25%	50%	75%	max
AGE	non-event	54.51	14.05	18.00	45.00	55.50	66.00	80.00
	event	59.05	13.74	18.00	50.75	62.00	70.00	80.00
BALB	non-event	3.67	0.35	2.61	3.45	3.68	3.90	4.90
	event	3.59	0.37	2.55	3.35	3.60	3.80	4.75
DURATION	non-event	3.75	4.57	0.23	0.86	1.99	4.44	31.27
	event	3.75	4.20	0.19	1.04	2.26	4.75	29.02

Table 11: Event rate in event group and non-event group for different categories of categorical variables.

Variables		non-event	event	total	event_rate
ICED	1	347	310	657	0.47
	2	231	346	577	0.60
	3	200	412	612	0.67
BLACK	black	523	633	1156	0.55
	others	255	435	690	0.63
DIABET	diabetic	307	516	823	0.63
	non-diabetic	471	552	1023	0.54
SEX	male	334	474	808	0.59
	female	444	594	1038	0.57
FLUX	low-flux	388	537	925	0.58
	high-flux	390	531	921	0.58
KTV	standard-ktv	389	537	926	0.58
	high-ktv	389	531	920	0.58
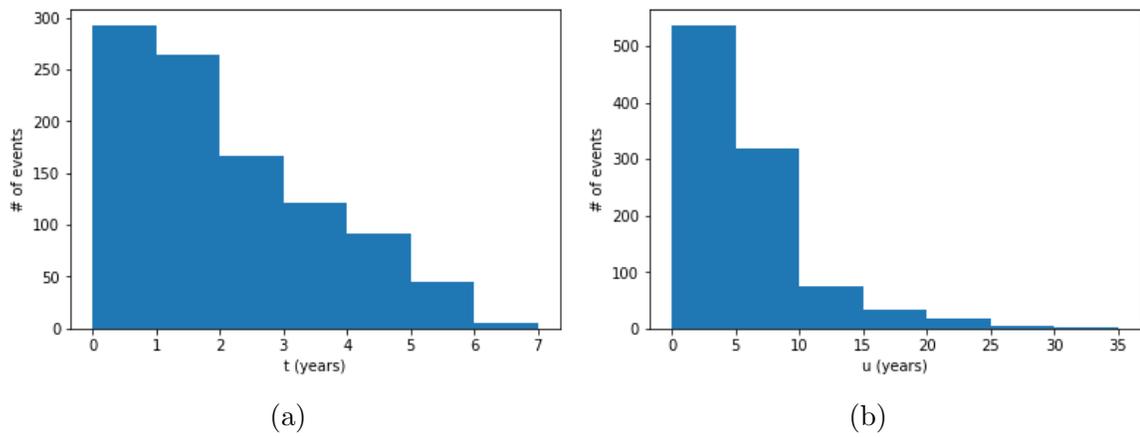
Figure 10: Distribution of composite events over follow up time t and duration time of dialysis u. Figure (a) represents # of composite events over follow up time and (b) represents # of composite events over time of dialysis.
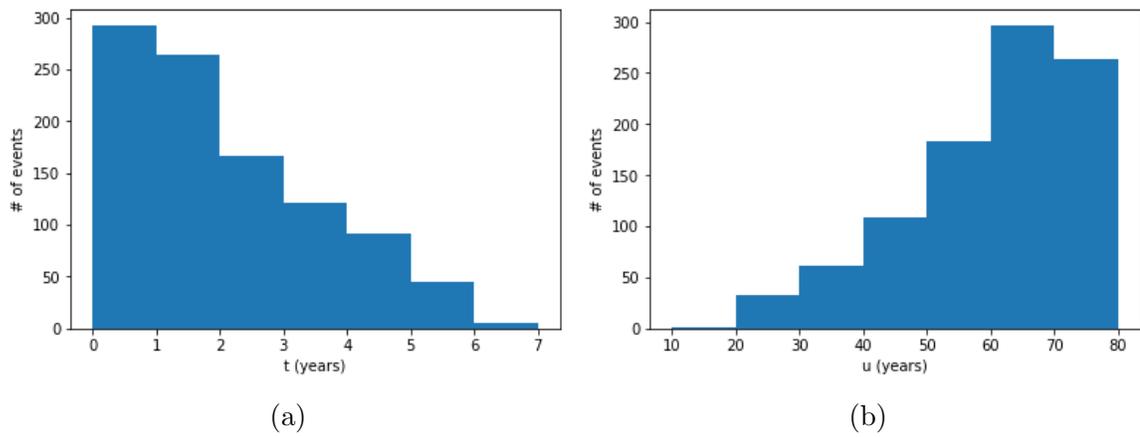
Figure 11: Distribution of composite events over follow up time t and patients' age u. Figure (a) represents # of composite events over follow up time and (b) represents # of composite events over patients' age.

Table 12: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under time to composite events model (5.1) with $u(t) = t + DURATION$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	0.0965	0.0795	-0.0593	0.2524	1.2138	0.2248
KTV	0.0993	0.0799	-0.0572	0.2559	1.2441	0.2135
AGE	0.0245	0.0036	0.0173	0.0316	6.7120	<0.0001
SEX	-0.0036	0.0838	-0.1680	0.1607	-0.0434	0.9654
BALB	0.2974	0.1318	0.0391	0.5557	2.2567	0.0240
DIABET	0.1865	0.0850	0.0199	0.3531	2.1947	0.0282
ICED_1	-0.3082	0.1004	-0.5050	-0.1115	-3.0701	0.0021
ICED_2	-0.2248	0.0946	-0.4101	-0.0395	-2.3773	0.0174
DURATION	0.0394	0.0246	-0.0089	0.0877	1.6002	0.1096

Table 13: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under time to composite events model (5.2) with $u(t) = t + AGE$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	0.1086	0.0656	-0.0200	0.2371	1.6552	0.0979
KTV	0.0922	0.0662	-0.0374	0.2219	1.3941	0.1633
AGE	0.0231	0.0170	-0.0103	0.0565	1.3562	0.1750
SEX	0.0126	0.0694	-0.1235	0.1487	0.1817	0.8558
BALB	0.2567	0.1087	0.0436	0.4698	2.3608	0.0182
DIABET	0.2384	0.0724	0.0966	0.3802	3.2944	0.0010
ICED_1	-0.3232	0.0835	-0.4868	-0.1595	-3.8699	0.0001
ICED_2	-0.1816	0.0789	-0.3362	-0.0269	-2.3006	0.0214
DURATION	0.0227	0.0081	0.0069	0.0385	2.8199	0.0048

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval

(d) $\hat{\gamma}(u)$ with 95% confidence interval

Figure 12: Estimated functions under model (5.1) with $u(t) = t + DURATION$ with bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval

(d) $\hat{\gamma}(u)$ with 95% confidence interval
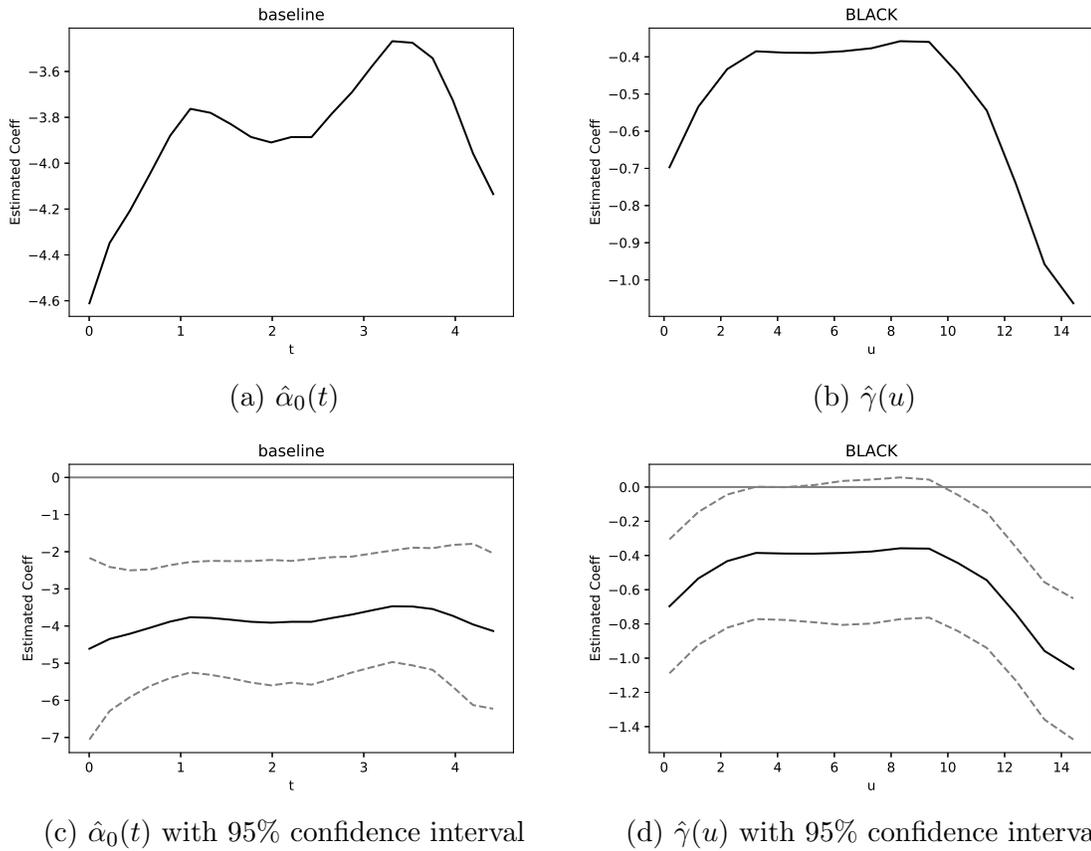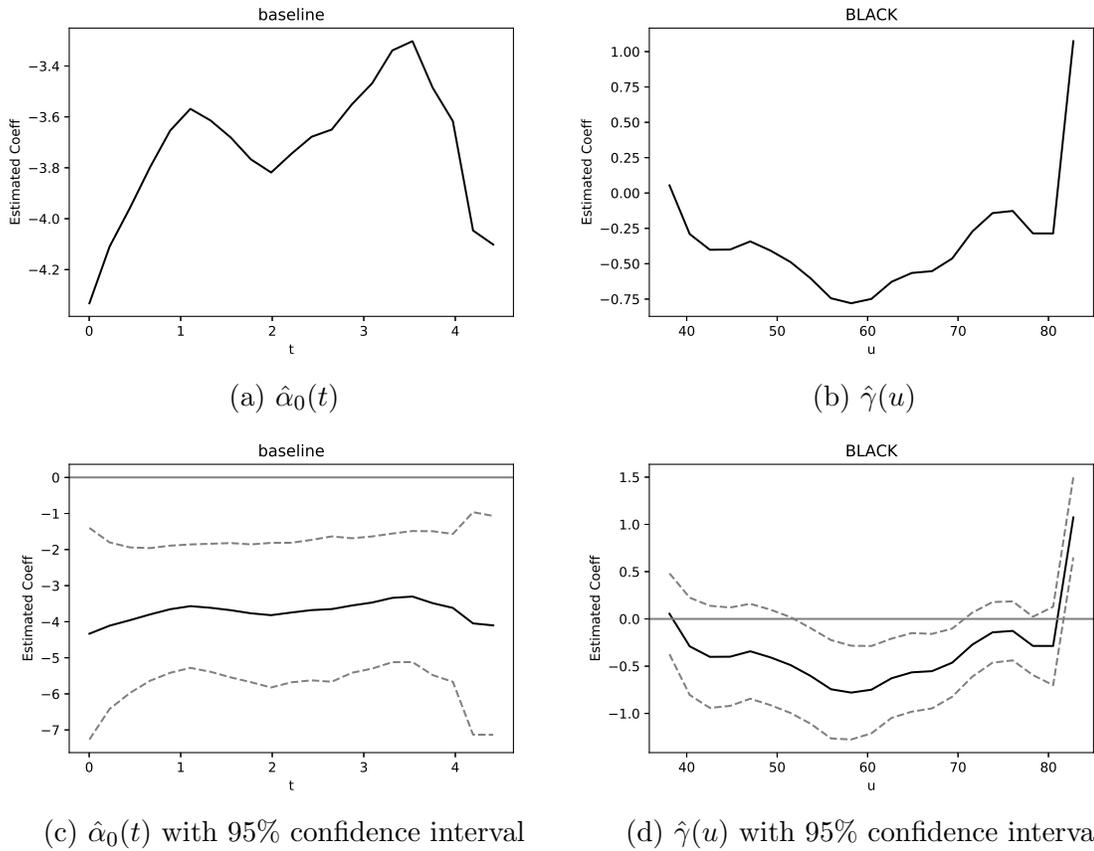
Figure 13: Estimated functions under model (5.2) with $u(t) = t + AGE$ with bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.

5.3   Application on Recurrent Hospitalizations

In the previous application, we focus on modeling time to mortality representing by the composite events to death or transplant. In this section, the goal is to model the rate of recurrent hospitalizations for hemodialysis patients.

Similarly, before we apply our proposed method on the HEMO dataset, we obtain some descriptive statistics regarding the hospitalization rates across different categories of each covariate. Figure 14 and Figure 15 suggest the following findings.

1. There are notable difference across different categories regarding the # of hospitalizations per patient for covariates including sex, age, baseline serum albumin (BALB), black race, ICED score 1 compared with ICED socres 2 and 3, and diabetes.

2. The # of hospitalizations per patient does not differ much when duration time of dialysis prior to enrollment is less than 15 years. However, the # of hospitalizations per patient dramatically decreases when duration time of dialysis is greater than 15 years.

3. The difference of the # of hospitalizations per patient for different treatments or different treatment combinations are not notable.
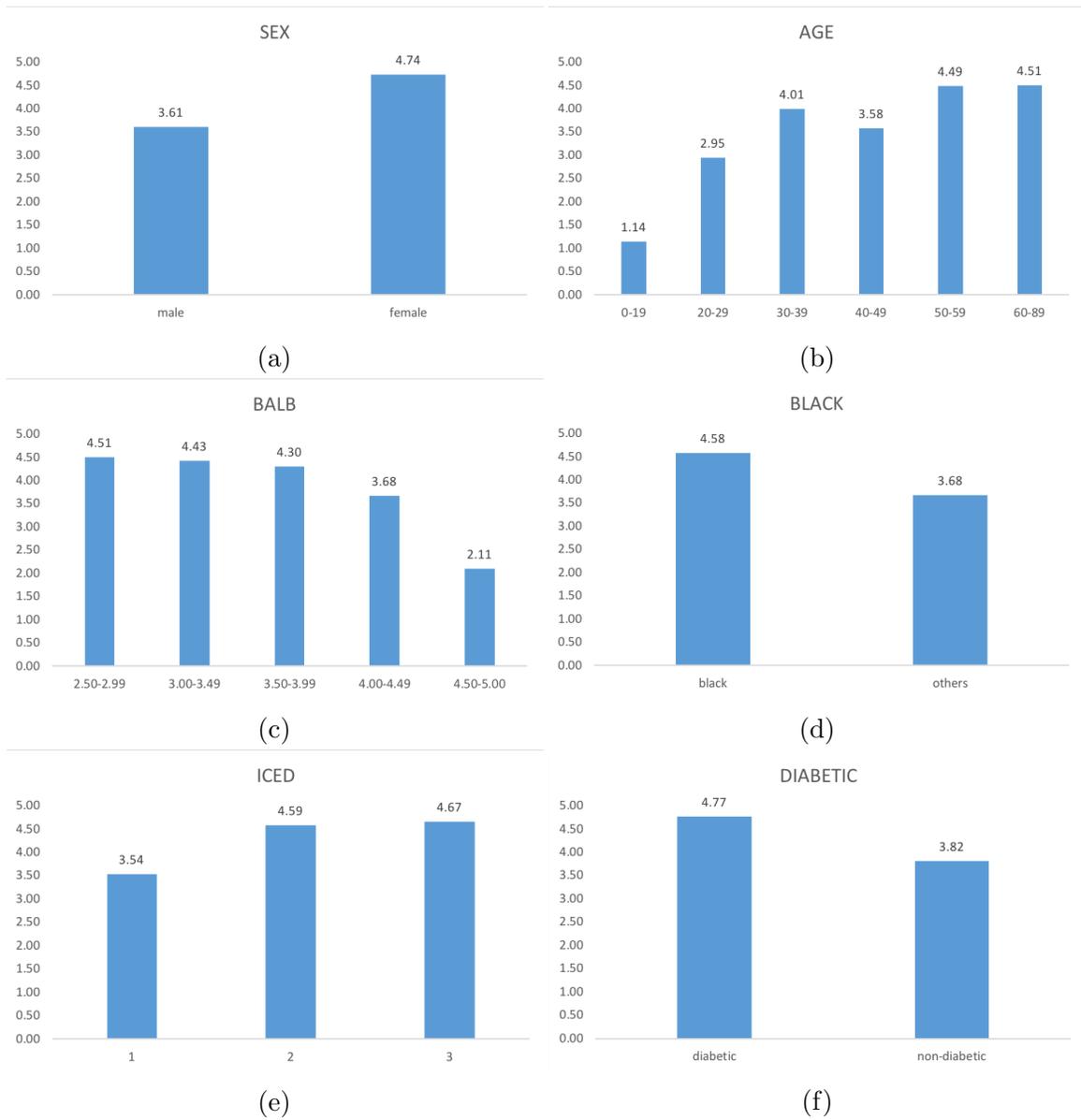
Figure 14: The # of Hospitalization per patient for different categories of covariates: sex, baseline age, baseline serum albumin, black race, ICED score and diabetic.
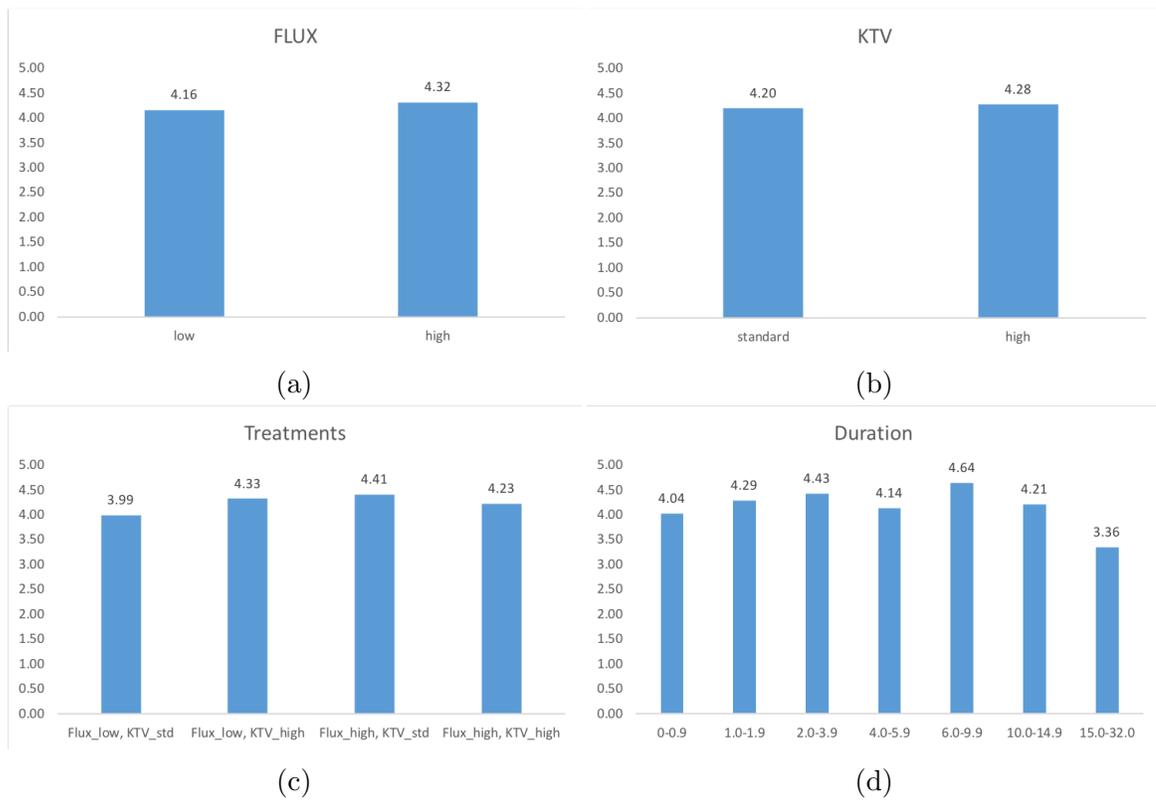
Figure 15: The # of Hospitalization per patient for different categories of covariates: FLUX, KTV, treatment groups of FLUX and KTV combinations and duration of dialysis prior to enrollment.

We start with a simpler model with $U(t) = t + DURATION$ as follows to check if the treatments effects on hospitalization rate are significant for hemodialysis patients.

$$\lambda(t) = exp\{\alpha_0(t) + \beta_1 FLUX \\ + \beta_2 KTV + \beta_3 DURATION + \gamma(t + DURATION)BLACK\}, \quad (5.3)$$

for $0 \le t \le 6.65$.

First, records of hospitalizations are plotted against t and u, respectively, to check how they are distributed. The left panel of Figure 16 shows the # of hospitalization records against follow up time t, while the right panel against duration time of dialysis u. The issue of data sparseness is detected for the same time windows as in Application 1. Thus, we only estimate $\alpha_0(t)$ for $0 \le t \le 4.5$ and $\gamma(u)$ for $0 \le u \le 15$.

The estimated coefficients for β_1 and β_2 are summarized by Table 14 along with their 95% confidence intervals and p-values. Based on the results, no significant association between neither treatments and hospitalization rates are found for hemodialysis patients. The duration time of dialysis prior to enrollment of study is found to be not significantly associated with hospitalization rate neither.

Figure 17 shows the estimated functions for baseline $\alpha_0(t)$ and $\gamma(u)$. It indicates that the baseline intensity for hospitalization is roughly constant over time. Figures (d) shows that he difference between African American patients and patients with other races regarding hospitalization rate is not significant. In addition, the race effect does not change over total duration time of dialysis.

In order to explore the factors that are associated with hospitalization rate for hemodialysis patients, a full model as below is fitted by including all covariates in this study.

$$\lambda(t) = exp\left\{\alpha_0(t) + \beta_1 FLUX + \beta_2 KTV + \beta_3 AGE + \beta_4 SEX + \beta_5 BALB + \beta_6 DIABET \right.$$
$$\left. + \beta_7 ICED\_1 + \beta_8 ICED\_2 + \beta_9 DURATION + \gamma(t + DURATION)BLACK\right\},$$
$$(5.4)$$

for $0 \leq t \leq 6.65$.

Table 15 summarizes the estimated coefficients for all covariates included in model (5.4). The results suggest the following remarkable findings.

1. Covariates that are significantly associated with hospitalization rate for hemodialysis patients include age, sex, diabetes and ICED score.

2. Male patients tend to have lower hospitalization rate compared with female patients.

3. The hospitalization rate for patients increases along with age. Older patients tend to have higher hospitalization rate compared with younger patients.

4. Diabetic hemodialysis patients have higher hospitalization rate compared with non-diabetic patients.

5. Hemodialysis patients with ICED score 1 have significantly lower hospitalization rate compare with those with ICED score 3, however, the difference regarding hospitalization rate between patients with ICED score 2 and ICED score 3 is not significant.

6. The duration time of dialysis prior to enrollment is not significantly associated with the hospitalization rate of hemodialysis patients. Figure 15 (d) shows that the # of hospitalization per patient do not vary across different levels of duration

time, except for the duration interval [15, 32]. However, not many patients in the data set falls into this duration time interval indicated by Figure 9. Thus, it could be the case that the model (5.4) could not capture the difference of hospitalization due to data sparseness in the duration time interval [15, 32].

7. Given other covariates, both treatments effects (FLUX and KTV) are not significant in the full model. In other words, the dose of dialysis delivered to hemodialysis patients and dialyzers with standard or high-flux membrane are not significantly associated with patients' hospitalization rate.

Figure 18 presents the estimated functions for $\alpha_0(t)$ and $\gamma(u)$. The plots show a very similar estimate for baseline $\alpha_0(t)$ as that in the simpler model (5.3). No significant difference regarding hospitalization rate between African American and other races are found when total duration time of dialysis is less than 25 years, while the difference between those groups becomes significant when total duration time of dialysis is between 25 to 30 years.

Table 14 and Table 15 visulize the HEMO dataset and summarize the pairwise correlation between each covariate and hospitalization rate. It is worth mentioned that all the findings from the full model are consistent with those suggested by the pairwise plots in Table 14 and Table 15.
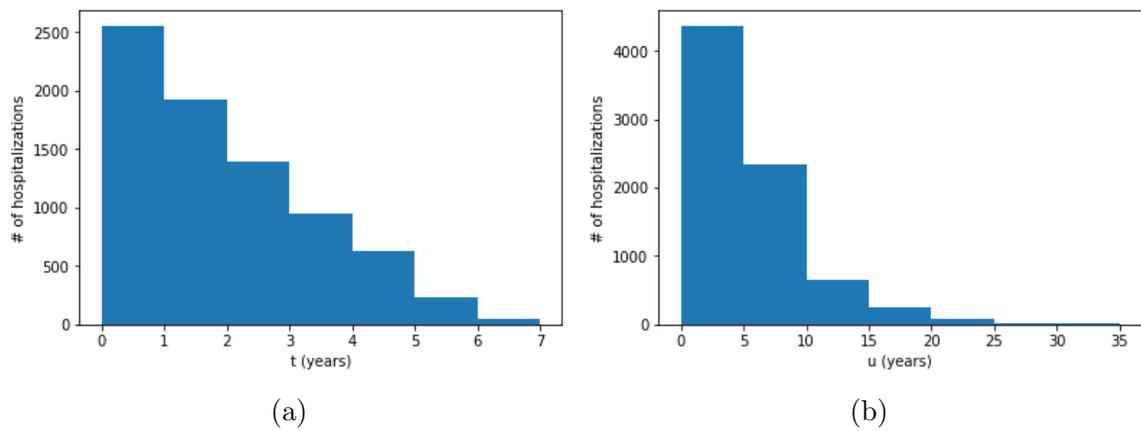
Figure 16: Distribution of hospitalizations over follow up time t and duration time of dialysis u. Figure (a) represents # of hospitalizations over follow up time and (b) represents # of hospitalizations over time of dialysis.

Table 14: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under simpler model (5.3) with $U(t) = t + DURATION$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	0.1165	0.1030	-0.0853	0.3184	1.1316	0.2578
KTV	0.1116	0.1033	-0.0908	0.3140	1.0805	0.2799
DURATION	-0.0139	0.0184	-0.0500	0.0222	-0.7561	0.4496

Table 15: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under full model (5.4) with $U(t) = t + DURATION$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	0.0642	0.0938	-0.1198	0.2481	0.6837	0.4942
KTV	0.0280	0.0941	-0.1564	0.2124	0.2976	0.7660
AGE	0.0073	0.0038	-0.0001	0.0147	1.9401	0.0524
SEX	-0.2184	0.0986	-0.4116	-0.0253	-2.2164	0.0267
BALB	-0.1733	0.1460	-0.4595	0.1129	-1.1866	0.2354
DIABET	0.2203	0.1109	0.0029	0.4376	1.9865	0.0470
ICED_1	-0.3693	0.1285	-0.6212	-0.1173	-2.8730	0.0041
ICED_2	-0.1514	0.1148	-0.3764	0.0737	-1.3182	0.1874
DURATION	-0.0068	0.0171	-0.0404	0.0268	-0.3992	0.6898

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval

(d) $\hat{\gamma}(u)$ with 95% confidence interval

Figure 17: Estimated functions under model (5.3) with $U(t) = t + DURATION$ and bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval
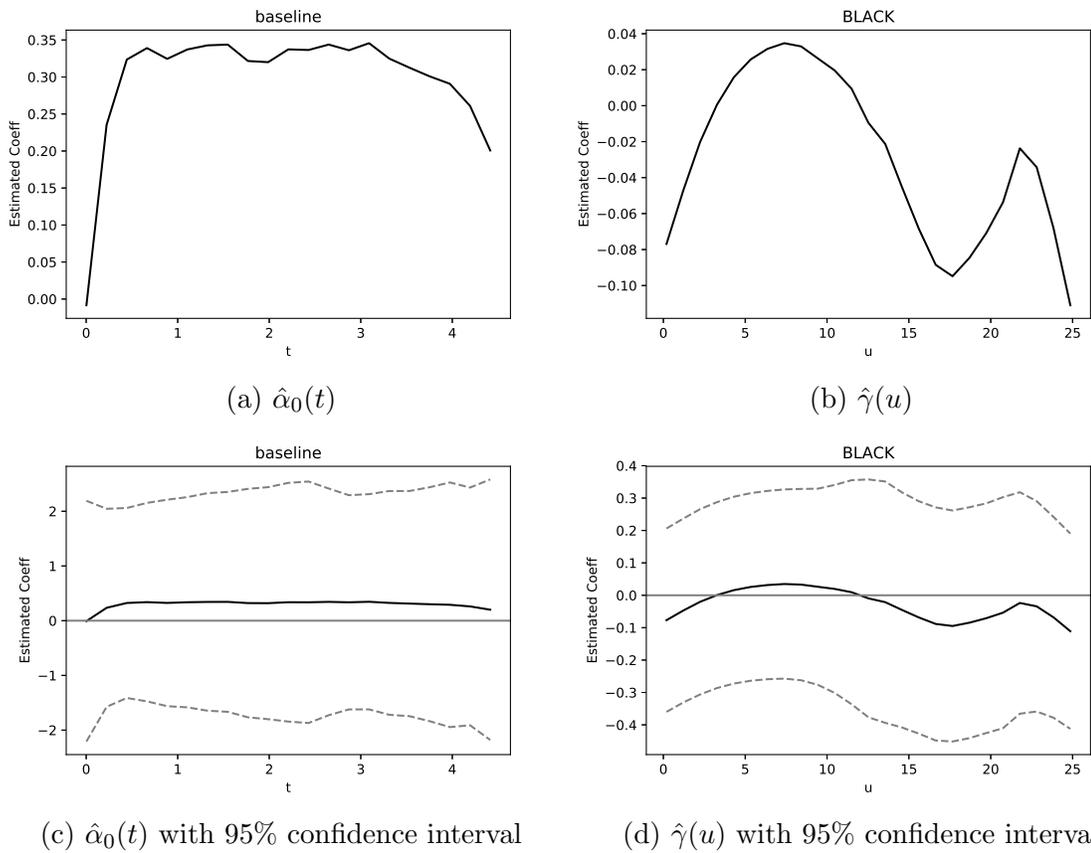
(d) $\hat{\gamma}(u)$ with 95% confidence interval

Figure 18: Estimated functions under model (5.4) with $U(t) = t + DURATION$ and bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.
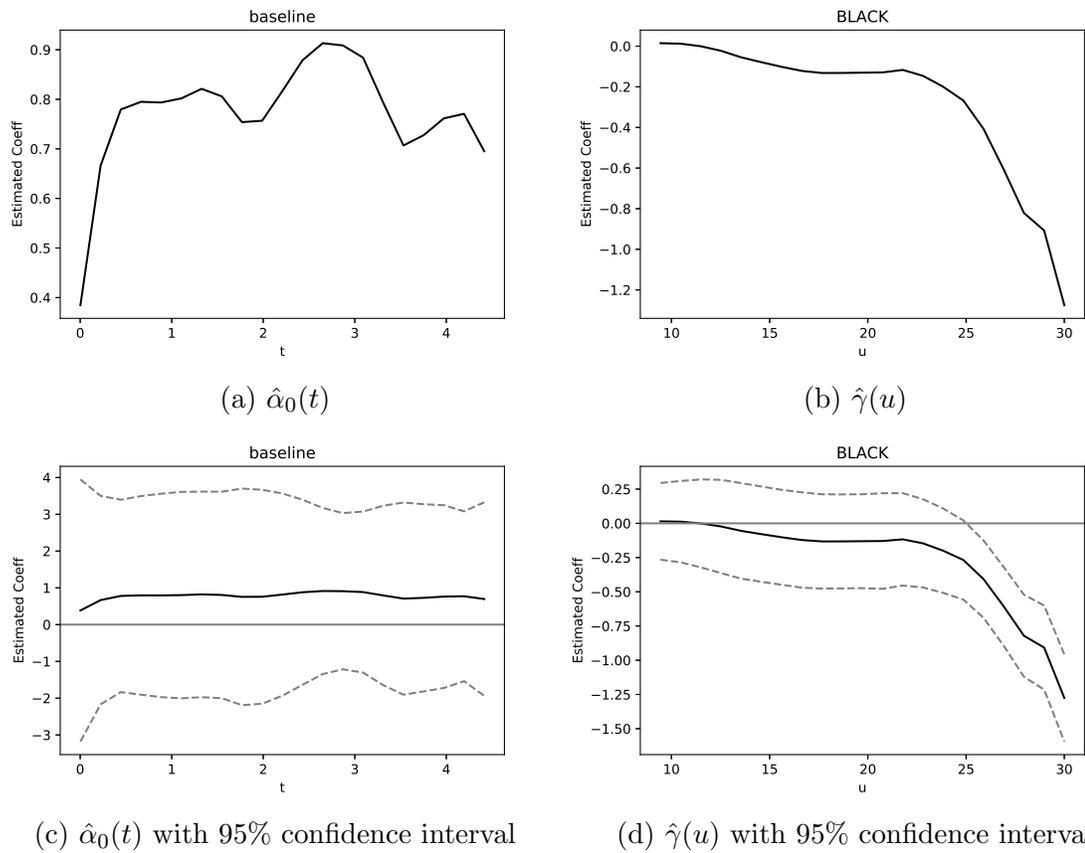
In the last example, we consider $U(t) = t + AGE$ instead of $U(t) = t + DURATION$ in our model. The purpose is to explore if and how the difference of hospitalization rate between African American and other races changes along with patients' age.

As usual, we start with a simpler model as follows to check if the treatments effects on hospitalization rate are significant for hemodialysis patients.

$$\lambda(t) = exp\left\{\alpha_0(t) + \beta_1 FLUX + \beta_2 KTV + AGE + \gamma(t + AGE)BLACK\right\}, \quad (5.5)$$

for $0 \le t \le 6.65$

With the defined $U(t) = t + AGE$, the range of u is $18 \le u \le 86.65$. Figure 19 plots hospitalizations against follow up time t and patients' age u, respectively. It is shown that not many hospitalizations happened in the window $[0, 35]$ for u. Therefore, to avoid the issue of sparse data, $\alpha_0(t)$ is estiamted for $0 \le t \le 4.5$ and $\gamma(u)$ for $35 \le u \le 86.65$.

The results given by Table 16 indicate that both treatment effects are not significant, which means that no significant difference regarding hospitalization rate are detected neither between different dose of dialysis delivered nor different type of dialyzer used for dialysis. The baseline age is not a significant factor to hospitalization rate for hemodialysis patients neither.

The plots from Figure 20 suggest that African American patients, compared with other races, have higher hospitalization rates when they are between 35 to 45 years old or 75 to 85 years old, but lower hospitalization rates when they are 45 to 75 years. However, such difference between African American patients and patients with other races is not significant since the 95% confidence band covers the zero line.

Next, a full model is considered as below to examine the factors that are considered
in this study.

$$\lambda(t) = exp\left\{\alpha_0(t) + \beta_1 FLUX + \beta_2 KTV + \beta_3 AGE + \beta_4 SEX + \beta_5 BALB + \beta_6 DIABET \right.$$
$$\left. + \beta_7 ICED\_1 + \beta_8 ICED\_2 + \beta_9 DURATION + \gamma(t + AGE)BLACK\right\},$$

$$(5.6)$$

for $0 \leq t \leq 6.65$.

Results for estimated coefficients are presented by Table 16. Some notable findings
are listed as follows.

1. The set of covariates found to be significantly associated with hospitaliza-
   tion rate for hemodialysis patients includes sex, baseline serum albumin level
   (BALB), diabetic and ICED score.

2. For each of the significant covariate, similar effects are found compared with
   the previous full model (5.4) with $U(t) = t + DURATION$.

3. Other covariates are not significantly associated with hospitalization rate for
   hemodialysis patients, including age and duration time of dialysis prior to study
   enrollment.

4. Given other covariates, no significant effects are detected for neither treatments.

Figure 21 suggests very similar results as the simpler model (5.5) for the interaction
of black race and age. No significance difference regarding hospitalization rate for
hemodialysis patients are detected by the full model (5.5).

To conclude, the findings in this example are consistent with not only data vi-
sulization of hospitalization records but also the findings in previous example with
$U(t) = t + DURATION$.
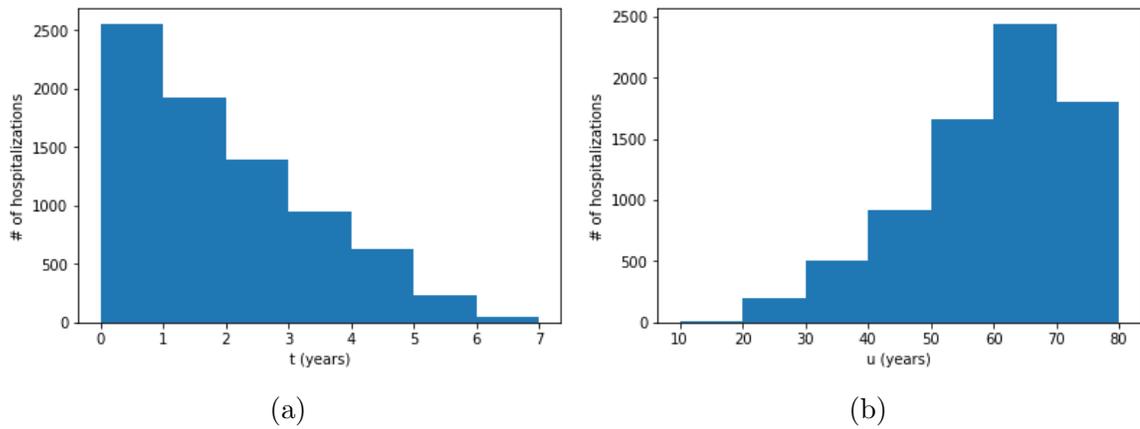
Figure 19: Distribution of hospitalizations over follow up time t and patients' age u. Figure (a) represents # of hospitalizations over follow up time and (b) represents # of composite events over patients' age.

Table 16: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under simpler model (5.5) with $U(t) = t + AGE$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	-0.0783	0.0480	-0.1723	0.0157	-1.6328	0.1025
KTV	-0.0552	0.0480	-0.1493	0.0389	-1.1493	0.2504
AGE	0.0016	0.0060	-0.0102	0.0134	0.2662	0.7901

Table 17: Summary of estimates, standard deviation, 95% confidence interval, Z score and P-value for $\hat{\beta}$s under full model (5.6) with $U(t) = t + AGE$.

Covariate	Coefficient	SD	95% Upper	95%Lower	Z Score	P-value
FLUX	0.0067	0.0473	-0.0861	0.0994	0.1414	0.8876
KTV	0.0094	0.0472	-0.0831	0.1020	0.1997	0.8417
AGE	0.0091	0.0082	-0.0069	0.0252	1.1129	0.2657
SEX	-0.1930	0.0508	-0.2926	-0.0934	-3.7980	0.0001
BALB	-0.3030	0.0755	-0.4509	-0.1551	-4.0143	0.0001
DIABET	0.2066	0.0549	0.0991	0.3141	3.7663	0.0002
ICED_1	-0.3009	0.0649	-0.4280	-0.1737	-4.6379	<0.0001
ICED_2	-0.0347	0.0573	-0.1470	0.0776	-0.6061	0.5444
DURATION	0.0027	0.0055	-0.0081	0.0134	0.4909	0.6235

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval

(d) $\hat{\gamma}(u)$ with 95% confidence interval

Figure 20: Estimated functions under model (5.5) with $U(t) = t + AGE$ and bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.

(a) $\hat{\alpha}_0(t)$

(b) $\hat{\gamma}(u)$

(c) $\hat{\alpha}_0(t)$ with 95% confidence interval

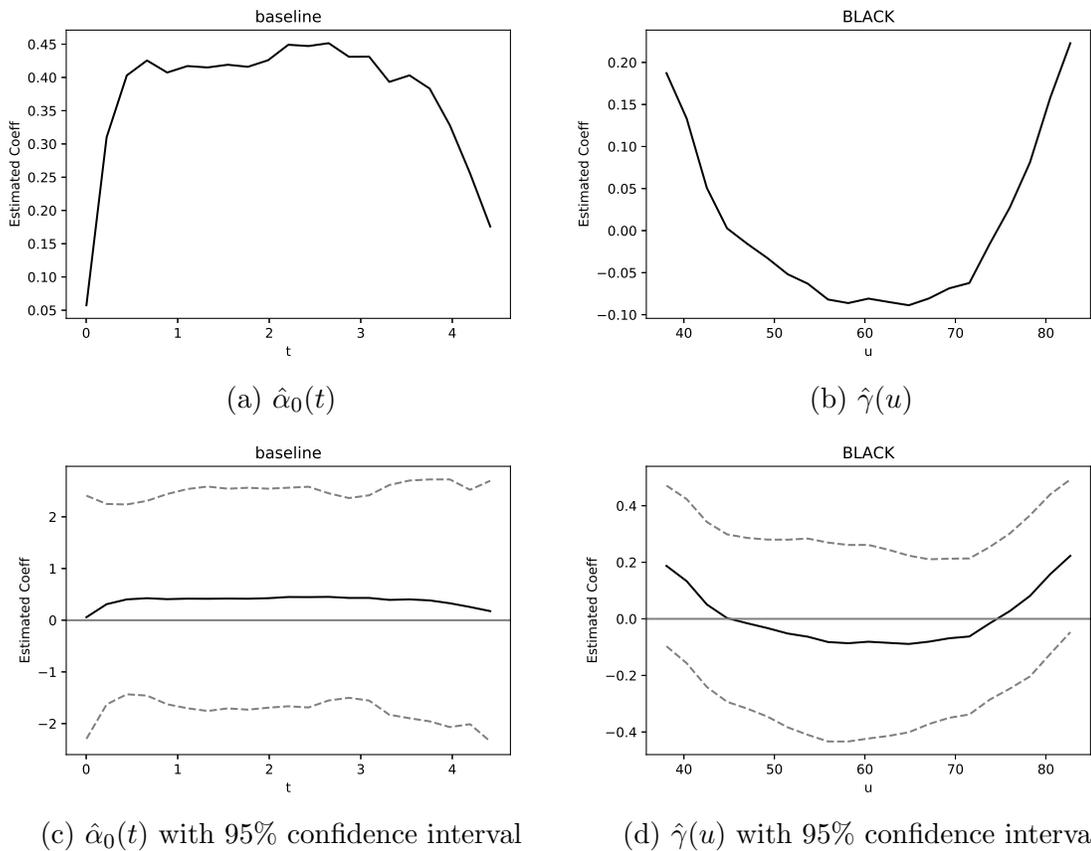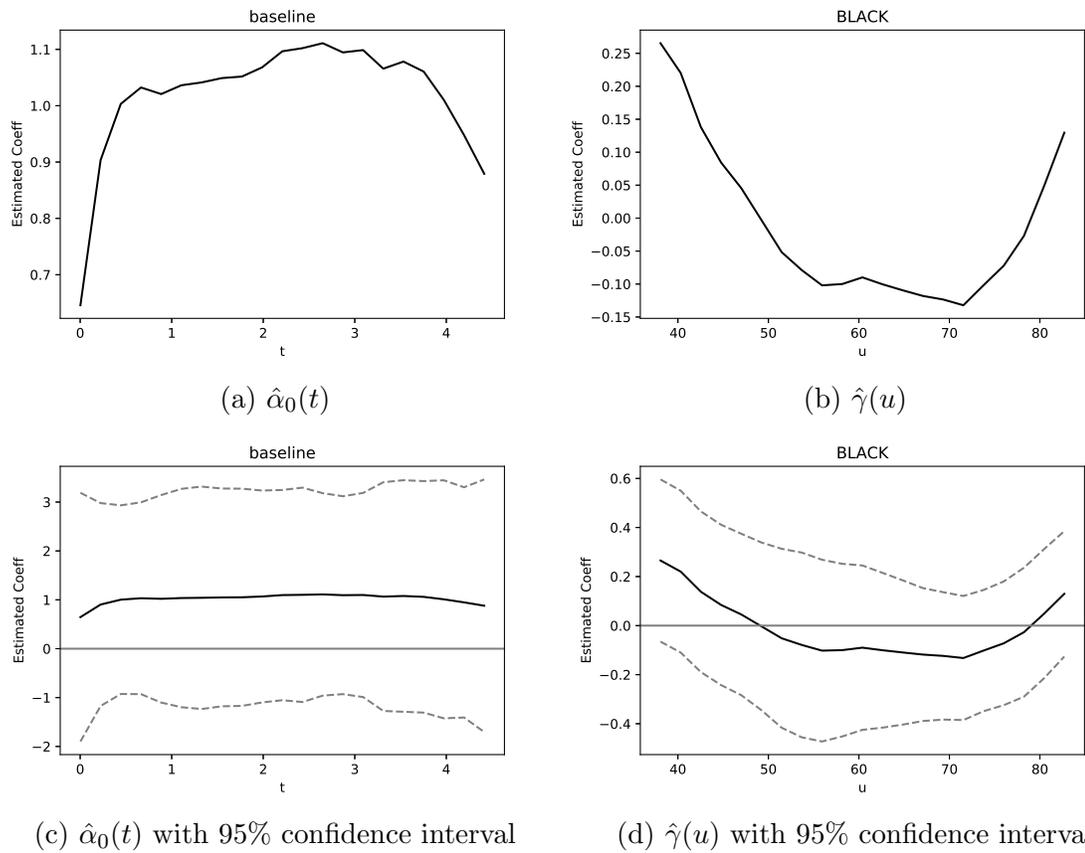(d) $\hat{\gamma}(u)$ with 95% confidence interval

Figure 21: Estimated functions under model (5.6) with $U(t) = t + AGE$ and bandwidth ($h_t = 0.5$, $h_u = 5$): (a) and (b) are the estimated functions for $\alpha_0(t)$ and $\gamma(u)$, respectively; (c) and (d) are the estimated functions with pointwise 95% confidence intervals.

REFERENCES

L. D. Amorim, J. Cai, D. Zeng, and M. L. Barreto. Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistics in Medicine*, 27:5890–5906, 2008.

P. K. Andersen and R. D. Gill. Coxs regression model counting process: A large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

C. Argyropoulos, M.E. Roumelioti, A. Sattar, J.A. Kellum, L. Weissfeld, and M.L. Unruh. Dialyzer reuse and outcomes of high flux dialysis. *PLoS One*, 10(6), 2015.

S. Beddhu, B. Baird, X. Ma, A.K. Cheung, and T. Greene. Serum alkaline phosphatase and mortality in hemodialysis patients. *Clin Nephrol*, 74(2):91–96, 2010.

J.D. Burrowes, G.B. Russell, M. Unruh, and M.V. Rocco. Is nutritional status associated with self-reported sleep quality in the hemo study cohort? *J Ren Nutr*, 22 (5):461–471, 2011.

Yan G Berkoben M Heyka R Kaufman A Lewis J Rocco M Toto R Windus D Ornt D Levey AS HEMO Study Group Cheung AK, Sarnak MJ. Cardiac diseases in maintenance hemodialysis patients: results of the hemo study. *Kidney Int*, 65(6): 2380–2389, 2004.

R.J. Cook and J. Lawless. *The Statistical Analysis of Recurrent Events*. Springer-Verlag New York, 2007.

D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 20:187–220, 1972.

J.T. Daugirdas, T. Greene, T.A. Depner, C. Chumlea, M.J. Rocco, G.M. Chertow, and Hemodialysis (HEMO) Study Group. Anthropometrically estimated total body water volumes are larger than modeled urea volume in chronic hemodialysis patients: effects of age, race, and gender. *Kidney Int*, 64(3):1108–1119, 2003.

J.T. Daugirdas, T. Greene, T.A. Depner, J. Leypoldt, F. Gotch, G. Schulman, R. Star, and Hemodialysis Study Group. Factors that affect postdialysis rebound in serum urea concentration, including the rate of dialysis: results from the hemo study. *J Am Soc Nephrol*, 15(1):194–203, 2004.

T.A. Depner, T. Greene, J.T. Daugirdas, A.K. Cheung, F.A. Gotch, and J.K. Leypoldt. Dialyzer performance in the hemo study: in vivo k0a and true blood flow determined from a model of cross-dialyzer urea extraction. *ASAIO Journal*, 50(1): 85–93, 2004.

Leung J Rocco M Burrowes JD Chumlea WC Frydrych A Kusek JW Uhlin L Dwyer JT, Larive B. Nutritional status affects quality of life in hemodialysis (hemo) study patients at baseline. *J Ren Nutr*, 12(4):213–223, 2002.

G. Eknoyan, G.J. Beck, A.K. Cheung, J.T. Daugirdas, T. Greene, J.W. Kusek, M. Allon, J. Bailey, J.A. Delmez, T.A. Depner, J.T. Dwyer, A.S. Levey, N.W. Levin, E. Milford, D.B. Ornt, M.V. Rocco, G. Schulman, S.J. Schwab, B.P. Teehan, R. Toto, and Hemodialysis (HEMO) Study Group. Effect of dialysis dose and membrane flux in maintenance hemodialysis. *N Engl J Med*, 347(25):2010–2019, 2002.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications: Monographs on statistics and applied probability.* CRC Press, 1996.

D. Ghosh and D.Y. Lin. Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*, 59:877–885, 2003.

T. Greene, G.J. Beck, J.J. Gassman, F.A. Gotch, J.W. Kusek, A.S. Levey, G. Levin, N.W. Schulman, and G. Eknoyan. Design and statistical issues of the hemodialysis (hemo) study. *Control Clin Trials*, 21(5):502–525, 2000.

S. Greenfield and E.C. Nelson. Recent developments and future issues in the use of health status assessment measures in clinical settings. *Medical Care*, 30(5):MS23–MS41, 1992.

N.A. Hoenich. Effect of dialysis dose and membrane flux in maintenance hemodialysis. *N Engl J Med*, 348(15):1491–1494, 2003.

M. Jhamb, F. Pike, S. Ramer, C. Argyropoulos, J. Steel, M.A. Dew, S.D. Weisbord, L. Weissfeld, and M. Unruh. Impact of fatigue on outcomes in the hemodialysis (hemo) study. *Am J Nephrol*, 33(6):515–523, 2011.

J. F. Lawless and C. Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37:158–168, 1995.

P.A.W Lewis and G.S. Shedler. Simulation of nonhomogeneous poisson processes with log linear rate function. *Biometrika*, 63(3):501–505, 1976.

P.A.W Lewis and G.S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

Argyropoulos C Weissfeld L Teuteberg J Dew MA Unruh ML Liang KV, Pike F. Heart failure severity scoring system and medical- and health-related quality-of-life outcomes: the hemo study. *Am J Kidney Dis*, 58(1):84–92, 2011.

D. Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate function of recurrent events. *Journal of the Royal Statistical Society: Series B*, 62:711–730, 2000.

D. Y. Lin, L. J. Wei, and Z. Ying. Semiparametric transformation models for point processes. *Journal of the American Statistical Association*, 96:620–628, 2001.

K.E. Lynch, R. Lynch, G.C. Curhan, and S.M. Brunelli. Prescribed dietary phosphate restriction and survival among hemodialysis patients. *Clin J Am Soc Nephrol*, 6 (3):620–629, 2010.

M. S. Pepe and J. Cai. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88:881–820, 1993.

Li Qi. *Generalized Semiparametric Varying-Coefficient Models For Longitudinal Data*. PhD dissertation, University of North Carolina at Charlotte, 2015.

M.V. Rocco, A.K. Cheung, T. Greene, G. Eknoyan, and Hemodialysis (HEMO) Study Group. The hemo study: applicability and generalizability. *Nephrol Dial Transplant*, 20(2):278–284, 2004.

A. Sattar, C. Argyropoulos, L. Weissfeld, N. Younas, L. Fried, J.A. Kellum, and M. Unruh. All-cause and cause-specific mortality associated with diabetes in prevalent hemodialysis patients. *BMC Nephrol*, 13:130, 2012.

L. Sun, X. Zhou, and S. Guo. Marginal regression models with time-varying coefficients for recurrent event data. *Statistics in Medicine*, 30:2265–2277, 2011.

A.W. Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

G. Yan and T. Greene. Statistical analysis and design for estimating accuracy in clinical-center classification of cause-specific clinical events in clinical trials. *Clinical Trials*, 8:571–580, 2011.

G. Yin, H. Li, and D. Zeng. Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, 103(483), 2008.

APPENDIX: PROOFS OF THE THEOREMS

Conditions

The following conditions are needed for our derivation of asymptotic properties in this study.

- The censoring time C_i is noninformative in the sense that $E\{dN_i^*(t)|Q_i(t), U_i(t), C_i \geq t\} = E\{dN_i^*(t)|Q_i(t), U_i(t)\}$, while the censoring time C_i is allowed to depend on the left continuous covariate process $Q_i(\cdot)$;

- The processes $Q_i(t)$ and $\lambda_i(t)$, $0 \leq t \leq \tau$, are bounded and their total variations are bounded by a constant; $E|N_i(t_2) - N_i(t_1)|^2 \leq L(t_2 - t_1)$ for $0 \leq t_1 \leq t_2 \leq \tau$, where $L > 0$ is a constant; $E|N_i(t+h) - N_i(t-h)|^{2+v} = O(h)$, for some $v > 0$;

- The kernel function $K(\cdot)$ is symmetric with compact support on $[-1, 1]$ and Lipschitz contimuous; Bandwidths $h_t \asymp h_u$; $h_t \to 0$; $nh_t^2 \to \infty$ and nh_t^5 is bounded;

- $\alpha_0(t)$, $\gamma_0(u)$, $e_{11}(t)$ and $e_{12}(t)$ are twice differentiable; $(e_{11}(t))^{-1}$ is bounded over $0 \leq t \leq \tau$; the matrices A_β and Σ_β are positive definite;

- The following two limits exist and are finite.

$$\lim_{n \to \infty} h_t E \left[\int_0^\tau \{dN_i(s) - \lambda_i(s)ds\} \, \mathscr{I}_1 e_{11}(t, U_i(s))^{-1} \tilde{Q}_i(s) K_{h_t}(s-t) \right]^{\otimes 2},$$

and

$$\lim_{n \to \infty} h_u E \left[\int_0^\tau \{dN_i(s) - \lambda_i(s)ds\} \, \mathscr{I}_3 e_{11}(u, U_i(s))^{-1} \tilde{Q}_i(s) K_{h_u}(s-u) \right]^{\otimes 2}.$$

Proof of Theorems

Proof of Theorem 3.1

By Lemma B.1 and Lemma B.3 in Qi (2015) and applying the Glivenko-Cantelli

theorem to the estimating function (2.6), we have the following

$$\frac{1}{n}U_\beta(\beta)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [dN_i(t) - \tilde{\lambda}_i(t,\beta)dt]\{\frac{\partial\tilde{\vartheta}(\beta,t,U_i(t))}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\}]$$

$$\xrightarrow{\mathcal{P}} E\int_0^\tau [dN_i(t) - \lambda\{\vartheta_\beta^T(t,U_i(t))\tilde{Q}_i(t) + \beta^T Z_i(t)\}dt]$$

$$\times \{Z_i(t) - (e_{\beta,12}(t,U_i(t)))^T(e_{\beta,11}(t,U_i(t)))^{-1}\tilde{Q}_i(t)\}$$

$$= E\int_0^\tau [\varphi\{\vartheta_0^T(t,U_i(t))\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}dt - \varphi\{\vartheta_\beta^T(t,U_i(t))\tilde{Q}_i(t) + \beta^T Z_i(t)\}dt]$$

$$\times \{Z_i(t) - (e_{\beta,12}(t,U_i(t)))^T(e_{\beta,11}(t,U_i(t)))^{-1}\tilde{Q}_i(t)\}$$

$$= u(\beta), \tag{1}$$

where β_0 is the unique root of $u(\beta)$. By the Theorem 5.9 of Van Der Vaart (1998),

we have $\hat{\beta}\xrightarrow{\mathcal{P}}\beta_0$.

By the Lemma B.3 in Qi (2015) and the Glivenko-Cantelli theorem, we have the

following

$$-\frac{1}{n}\frac{\partial U_\beta(\beta)}{\partial \beta}\Big|_{\beta=\beta_0}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [dN_i(t) - \lambda\{\tilde{\vartheta}^T(t, U_i, \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}dt]$$

$$\times \left\{\frac{\partial \tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial \beta}\tilde{Q}_i(t) + Z_i(t)\right\}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [-\lambda\{\tilde{\vartheta}^T(t, U_i, \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}]\left\{\frac{\partial \tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial \beta}\tilde{Q}_i(t) + Z_i(t)\right\}^{\otimes 2} dt$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [dN_i(t) - \lambda\{\tilde{\vartheta}^T(t, U_i(t), \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}dt]\left\{\frac{\partial^2 \tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial^2 \beta}\tilde{Q}_i(t)\right\}$$

$$(2)$$

The first and third terms go to zero as $n \to \infty$. Thus, we have

$$-\frac{1}{n}\frac{\partial U_\beta(\beta)}{\partial \beta}\Big|_{\beta=\beta_0}\xrightarrow{\mathcal{P}} E\int_0^\tau \lambda\{\tilde{\vartheta}^T(t, U_i, \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}$$

$$\times \{Z_i(t) - (e_{12}(t, U_i(t)))^T(e_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}^{\otimes 2} dt \equiv A_\beta \qquad (3)$$

Now we show that $n^{-1/2}U_\beta(\beta_0)$ converges in distribution to a normal distribution. By Taylor expansion,

$$\exp\{\tilde{\vartheta}^T(t, U_i(t), \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\} - \exp\{\vartheta_0^T(t, U_i(t))\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}$$

$$= \exp\{\vartheta_0^T(t, U_i(t))\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}[\tilde{\vartheta}^T(t, U_i(t), \beta_0) - \vartheta_0^T(t, U_i(t))]\tilde{Q}_i(t)$$

$$+ O_p(\|\tilde{\vartheta}(t, U_i(t), \beta_0) - \vartheta_0(t, U_i(t))\|^2) \qquad (4)$$

We define $dM_i(t) = dN_i(t) - \exp\{\vartheta_0^T(t, U_i(t))\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}dt$, then we have

$$\frac{1}{\sqrt{n}}U_\beta(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau [dN_i(t) - \exp\{\tilde{\vartheta}^T(t, U_i(t), \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}dt]$$

$$\times \left\{\frac{\partial\tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\right\}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{\frac{\partial\tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\right\} dM_i(t)$$

$$- \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau [\exp\{\tilde{\vartheta}^T(t, U_i(t), \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\} - \varphi\{\vartheta_0^T(t, U_i(t))\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}]$$

$$\times \left\{\frac{\partial\tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\right\} dt$$

$$(5)$$

By Lemma 1 in Lin et al. (2001), the second term equals to the following

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \exp\{\tilde{\vartheta}^T(t, U_i, \beta_0)\tilde{Q}_i(t) + \beta_0^T Z_i(t)\}[\tilde{\vartheta}^T(t, U_i(t), \beta_0) - \vartheta_0^T(t, U_i(t))]\tilde{Q}_i(t)dt$$

$$\times \left\{\frac{\partial\tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\right\} = o_p(1)$$

$$(6)$$

Thus, we have

$$\frac{1}{\sqrt{n}}U_\beta(\beta_0)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{\frac{\partial\tilde{\vartheta}(t, U_i(t), \beta_0)}{\partial\beta}\tilde{Q}_i(t) + Z_i(t)\right\} dM_i(t) + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \{Z_i(t) - (e_{12}(t, U_i(t)))^T(e_{11}(t, U_i(t)))^{-1}\tilde{Q}_i(t)\}dM_i(t)$$

$$(7)$$

which converges in distribution to $N(0, \Sigma_\beta)$ by Central Limit Theorem, where

$$\Sigma_\beta = E\left(\int_0^\tau \{Z_i(t) - (e_{12}(t, U_i(t)))^T (e_{11}(t, U_i(t)))^{-1} \tilde{Q}_i(t)\} dM_i(t)\right)^{\otimes 2}.$$

By Taylor expansion, we have

$$U_\beta(\hat{\beta}) = U_\beta(\beta_0) + \frac{\partial U_\beta(\beta)}{\partial \beta}|_{\beta=\beta_0}(\hat{\beta} - \beta_0) + O_p(\|\hat{\beta} - \beta_0\|^2).$$

Thus,

$$\sqrt{n}(\hat{\beta} - \beta_0) = \left(-\frac{1}{n}\frac{\partial U_\beta(\beta)}{\partial \beta}|_{\beta=\beta_0}\right)^{-1} \times \frac{1}{\sqrt{n}} U_\beta(\beta_0)$$

Hence, by Slutsky Theorem, we have $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, A_\beta^{-1}\Sigma_\beta A_\beta^{-1})$. $\qquad\square$

Proof of Theorem 3.2

(a) Since $\hat{\vartheta}(t_0, u_0) = \tilde{\vartheta}(t_0, u_0, \hat{\beta})$, we have $\hat{\vartheta}(t_0, u_0) \xrightarrow{\mathcal{P}} \vartheta_0(t_0, u_0)$ uniform in $t \in [0, \tau]$ and $u \in [u_1, u_2]$ by Lemma B.1 in Qi (2015) and Theorem 1. Then we have the following

$$\sup_{t_0 \in [t_1, t_2]} |\hat{\vartheta}(t_0) - \vartheta_0(t_0)| = \sup_{t_0 \in [t_1, t_2]} |n^{-1} \sum_{j=1}^n \{\hat{\vartheta}(t_0, U_j(t_0)) - \vartheta_0(t_0, U_j(t_0))\}|$$

$$\leq \sup_{t_0 \in [t_1, t_2], u_0 \in [u_1, u_2]} |\hat{\vartheta}(t_0, u_0) - \vartheta_0(t_0, u_0)| = o_p(1).$$

(b) By Lemma B.4 in Qi (2015), we have

$$\sqrt{nh_t h_u}\{\tilde{\alpha}(t_0, u_0, \beta_0) - \alpha_0(t_0, u_0)\}$$

$$= -\mathscr{I}_1 e_{11}^{-1}(t_0, u_0) \sqrt{\frac{h_t h_u}{n}} \sum_{i=1}^n \int_0^\tau \{dN_i(t) - \lambda_i(t)dt\} \tilde{X}_i(t) K_{h_t}(t - t_0) K_{h_u}(U_i(t) - u_0)$$

$$+ \frac{1}{2}\sqrt{nhbh}h^2 \nu_2 e_{11}^{-1}(t_0, u_0) b_\alpha(t_0, u_0) + \frac{1}{2}\sqrt{nhb}b^2 \nu_2 e_{11}^{-1}(t_0, u_0) b_\gamma(t_0, u_0)$$

$$+ o_p(\sqrt{nhb}(h^2 + b^2))$$

Since $e_{11}^{-1}(t_0, u_0)b_\gamma(t_0, u_0)$ is zero for the first p_1 components and $e_{11}^{-1}(t_0, u_0)b_\alpha(t_0, u_0)$ is $\ddot{\alpha}(t_0)$ for the first p_1 components. Then

$$\sqrt{nh_t}\{\hat{\alpha}(t_0) - \alpha_0(t_0)\}$$

$$= -\sqrt{\frac{h_t}{n}}\sum_{i=1}^{n}\int_0^\tau \{dN_i(t) - \lambda_i(t)dt\}\{\frac{1}{n}\sum_{j=1}^{n}\mathscr{I}_1 e_{11}^{-1}(t_0, U_j(t_0))Q_i(t)K_{h_u}(U_i(t) - U_j(t_0))\}$$

$$\times K_{h_t}(t - t_0)$$

$$+ \sqrt{\frac{h_t}{n}}\frac{1}{n}\sum_{j=1}^{n}\{e_{11}(t_0, U_j(t_0))^{-1}e_{12}(t_0, U_j(t_0))\}(\hat{\beta} - \beta_0) + \frac{1}{2}\sqrt{nh_t}h_t^2\nu_2\ddot{\alpha}(t_0).$$

By Lemma A.1 in Yin et al. (2008),

$$\frac{1}{n}\sum_{j=1}^{n}e_{11}^{-1}(t_0, U_j(t_0))K_{h_u}(u - U_j(t_0)) = e_{11}^{-1}(t_0, u) + O_p(\frac{\log h_u}{\sqrt{nh_u}}) + O(h_u^2)$$

uniformly in $t \in [t_1, t_2]$ and $u \in [u_1, u_2]$. It follows that

$$\sqrt{nh_t}\{\hat{\alpha}(t_0) - \alpha_0(t_0) - \frac{1}{2}h_t^2\nu_2\ddot{\alpha}(t_0)\}$$

$$= \sqrt{\frac{h_t}{n}}\sum_{i=1}^{n}\int_0^\tau \{dN_i(t) - \lambda_i(t)dt\}\mathscr{I}_1 e_{11}^{-1}(t_0, U_i(t))Q_i(t)K_{h_t}(t - t_0) + o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}g_i(t_0) + o_p(1),$$

where $g_i(t_0) = h_t^{1/2}\int_0^\tau \{dN_i(t) - \lambda_i(t)dt\}\mathscr{I}_1 e_{11}^{-1}(t_0, U_i(t))Q_i(t)K_{h_t}(t - t_0)$.

Following the arguments of Lemma 2 of Sun (2010),

$$\sqrt{nh_t}(\hat{\alpha}(t) - \alpha_0(t) - \frac{1}{2}h_t^2\nu_2\ddot{\alpha}(t_0))\overset{\mathcal{D}}{\longrightarrow}N(0, \Sigma_\alpha(t)) \tag{8}$$

where

$$\Sigma_\alpha(t_0) = \lim_{n\to\infty}h_t E\left[\int_0^\tau \{dN_i(t) - \lambda_i(t)dt\}\mathscr{I}_1 e_{11}^{-1}(t_0, U_i(t))Q_i(t)K_{h_t}(t - t_0)\right]^{\otimes 2}.$$

□

Proof of Theorem 3.3

By following the same proofing process of Theorem 3.2, we have the following

$\hat{\gamma}(u) \xrightarrow{\mathcal{P}} \gamma_0(u)$ uniformly in $u \in [u_1, u_2]$, and

$$\sqrt{nh_u}(\hat{\gamma}(u) - \gamma_0(u) - \tfrac{1}{2}h_u^2\nu_2\ddot{\gamma}(u)) \xrightarrow{\mathcal{D}} N\left(0, \Sigma_\gamma(u)\right). \qquad \square$$