SEMIPARAMETRIC ANALYSIS OF COMPLEX LONGITUDINAL DATA

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Dayu Sun

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

May 2020

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

SEMIPARAMETRIC ANALYSIS OF COMPLEX LONGITUDINAL DATA

presented by Dayu Sun,

a candidate for the degree of Doctor of Philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. (Tony) Jianguo Sun

Dr. Tieming Ji

Dr. Shih-Kang Chao

Dr. Shawn Ni

ACKNOWLEDGMENTS

First and foremost, I owe my deepest gratitude to my esteemed advisor Dr. (Tony) Jianguo Sun for his wonderful direction and extensive support throughout the development of this dissertation. His critical encouragement and inspiring instruction were invaluable for my doctoral studies and guaranteed the accomplishment of this work.

I extend my gratitude to my advisory committee members, Drs. Shih-Kang Chao, Tieming Ji, and Xiaoguang Ni, for their insightful comments and advice on my work. Special thanks are due to Drs. Hui Zhao and Drs. Hongyuan Cao for her generous academic support and great collaboration throughout my research.

I also owe a debt of gratitude to all faculty in the Department of Statistics, especially those who have taught me, for their inspiration and encouragement in my graduate studies. I am grateful to Dr. Larry Ries, who provided me with helpful guidance and precious advice for being a better instructor. I also appreciate the plenty of help by our great staff Judy Dooley and Abbie Van Nice Booher. In addition, I want to take the opportunity to thank all my classmates and friends I met here. Because of them, my life in Mizzou is so enjoyable and memorable.

In the end, I especially owe my gratitude to my parents for their love ans support throught my life. Thanks you both for giving me strength to pursue an academic career. I also want to express my appreciation for my fiancée Yuanyuan, for her endless love, best care, and generous support, which make my life so wonderful and happy.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

SEMIPARAMETRIC ANALYSIS OF COMPLEX LONGITUDINAL DATA

Dayu Sun

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Event history data consist of the longitudinal records of event occurrence times. Recurrent event data and panel count data are two common types of event history data that occur in many areas, such as medical studies and social sciences. A great deal of literature has been established for their analyses. Nevertheless, only limited research exists on the variable selection for recurrent event data and panel count data. The existing methods can be seen as direct generalizations of the available penalized procedures for linear models, but may not perform as well as expected due to the complex structure of event history data. The first and second parts of this dissertation then discuss simultaneous parameter estimation and variable selection for event history data. We present a new variable selection method with a new penalty function, which will be referred to as the broken adaptive ridge regression approach. In addition to the establishment of the oracle property, we also show that the proposed variable selection method has the clustering or grouping effect when covariates are highly correlated. Furthermore, the numerical studies are performed and indicate that the method works well for practical situations and can outperform the existing methods. Applications to real data are provided.

Most of the existing studies of longitudinal data assume that covariates can be observed at the same observation times for the response variable, and the observation process is independent of the response variable completely or given covariates. In practice, the response variables and covariates are sometimes observed intermittently at different time points, leading to sparse asynchronous longitudinal data. The observation process may also be related to the response variable even given covariates and sometimes both issues can even occur at the same time. Although each of the two issues has been developed to address in literature, it does not seem to exist an established approach that can deal with both together. To address both issues simultaneously, the third part of this dissertation proposes a flexible semiparametric transformation conditional model and a kernel-weighted estimating equation based approach. The proposed estimators of regression parameters are shown to be consistent and asymptotically follow the normal distribution. For the assessment of the finite sample performance of the proposed method, an extensive simulation study is carried out and suggests that it performs well for practical situations. The approach is applied to a prospective HIV study that motivated this investigation.

Chapter 1

Introduction

1.1 Introduction to Event History Data

Event history data consist of the longitudinal records of event occurrence times from a sample of individuals. Two follow-up schemes are typically used for event history studies. One is to follow all study subjects continuously, leading to recurrent event data. Recurrent event data comprise all occurrence times of events for each individual during the follow-up (Andersen, 1997; Cai and Schaubel, 2004; Cook and Lawless, 2007). Recurrent event data occur in many areas such as medical studies and social sciences, and a great deal of literature has been established for their analysis (Andersen, 1997; Cook and Lawless, 2007; Lawless and Nadeau, 1995; Lin et al., 2000; Schaubel et al., 2006). In particular, Cook and Lawless (2007) gave a comprehensive review of the literature on the analysis of such data. An example is the study of the hospitalization rate for certain patient groups, and another example is given by

investigating the occurrence rate of some disease symptoms or infections, as discussed in Section 2.6.

Another scheme is to observe all study subjects only at discrete time points and thus only incomplete data, which are often referred to as panel count data, are available for inference. In panel count data, we only know the number of event occurrences between two consecutive observation times points. The panel count data can then be regarded as interval-censored recurrent event data. One complicated factor for panel count data is that both observation and follow-up times usually vary from subject to subject. Fields in which panel count data are common include demographical studies, epidemiological studies, medical periodic follow-up studies and tumorigenicity experiments (Balakrishnan and Zhao, 2009; Hu et al., 2009). A most recent comprehensive review for panel count data is Sun and Zhao (2013). An example is the skin cancer study, discussed in Section 3.5, investigating if a treatment can reduce the rates of two types of skin tumor recurrence (Li et al., 2010; Sun and Zhao, 2013). In this skin cancer study, each patient was observed at a sequence of discrete observation times and the numbers of occurrences of two types of skin tumors between the observation times were recorded. However, as expected, these real observation times differed from patient to patient and so as the follow-up times.

1.2 Variable Selection for Recurrent Event Data and Panel Count Data

Variable selection is an important topic and has been discussed in many areas such as linear regression and failure time data analysis (Fan and Li, 2002; Tibshirani, 1996).

Among the available methods, the most commonly used is perhaps the penalized approach that maximizes or minimizes an objective function minus or plus a penalty function. It is well known that a natural approach is the L_0-penalized regression that directly penalizes the cardinality of the variables in the model and seeks the most parsimonious model explaining the data. However, solving an exact L_0-penalized nonconvex optimization problem involves exhaustive combinatorial best subset search, which is NP-hard and computationally infeasible, especially for high dimensional data. To overcome this, a popular approach is to replace the nonconvex L_0-norm by the L_1-norm, which is known as the closest convex relaxation of the L_0-norm. In addition, the L_1-penalized optimization problem can be solved exactly with efficient algorithms and the method became popular since introducing the least absolute shrinkage and selection operator (LASSO) method (Tibshirani, 1996). Nevertheless, it is known that the LASSO does not have the oracle property as it tends to select too many small noise features and is biased for large parameters. In addition, it cannot accommodate the grouping effect when covariates are highly correlated.

To address these, many penalty functions have been proposed such as the smoothly clipped absolute derivation (SCAD) (Fan and Li, 2001) and the adaptive LASSO (ALASSO) (Zou, 2006) functions. However, none of them have both the oracle property and grouping effect at the same. In this dissertation, we propose a broken adaptive ridge (BAR) regression approach that approximates the L_0-penalized regression using an iteratively reweighted L_2-penalized algorithm for variable selection. The advantage of the BAR is that it has both the grouping effect and oracle property

However, these methods cannot be directly applied to the event history data because they have a much more complicated and different structure. To our knowledge,

two penalized procedures have been developed in Tong et al. (2009a) and Zhang et al. (2013), and both considered some commonly used penalty functions, namely, the smoothly clipped absolute derivation (SCAD) and the seamless-L_0 (SELO) functions (Dicker et al., 2013). Also they both discussed the multiplicative mean model, which assumes that covariates affect the mean function of the process in a multiplicative way, for the underlying recurrent event process (Sun and Zhao, 2013).

As a useful alternative to the multiplicative model, the additive rate model assumes that the rate function of a recurrent event process $N^*(t)$ has the form

$$\mathrm{E}\{dN^*(t)|\boldsymbol{Z}^*(t)\} = d\mu_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}^*(t)dt; \tag{1.1}$$

see (Sun and Zhao, 2013; Zhao et al., 2013). In the above, $\boldsymbol{Z}^*(t)$ denotes a p-dimensional vector of possibly time-dependent covariates, $\mu_0(t)$ is an unspecified non-descreasing function and $\boldsymbol{\beta}$ denotes a p-dimensional vector of regression parameters. The model above provides a characterization of the regression effects different from the multiplicative model and has some remarkable features that are not shared by the latter. In particular, model (1.1) pertains to the risk difference or excess risk, a measure that is especially relevant and informative in epidemiological and clinical studies. Very limited literature has discussed the variable selection for event history data under model (1.1). To our best knowledge, only Chen and Wang (2013) addressed this issue for recurrent event data by ALASSO and SCAD method.

To fill the research gap, we will propose a simultaneous estimation and variable selection method for recurrent event data and panel count data under model (1.1) by the novel BAR approach in Chapters 2 and 3, respectively. In particular, unlike the two existing methods given in Tong et al. (2009a) and Zhang et al. (2013), the

resulting estimator from the proposed method enjoys the oracle property and grouping effect at the same time.

1.3 Regression Analysis of Asynchronous Longitudinal Data

A great deal of literature has been established for regression analysis of longitudinal data (Diggle et al., 1994; Hand and Crowder, 1996). For most of the existing methods, two basic assumptions are that 1) covariates can be observed completely or at the same observation times for the response variable, and 2) the observation process is independent of the response variable completely or given covariates. However, as pointed out by many authors, sometimes the response variables and covariates may be observed intermittently at different time points, leading to sparse asynchronous longitudinal data. The sparsity here means that only a few observations are available at discrete time points for each subject. In addition, the observation process may be related to the response variable even given covariates, resulting in an informative observation process. In this case, the observation processes can be modeled and studied as event history data.

One example of sparse asynchronous longitudinal data is given in Wohl et al. (2005), which discussed the analysis of a prospective observational cohort study of 190 HIV-infected subjects. The subjects were followed for up to five years and during the study, their HIV viral load and CD4 cell counts were measured repeatedly. Figure 1.1 presents the observation times for both viral load and CD4 cell counts for all the subjects studied by Wohl et al. (2005) and it is clear that they are not

matched over time because the two variables were measured on different days. Some

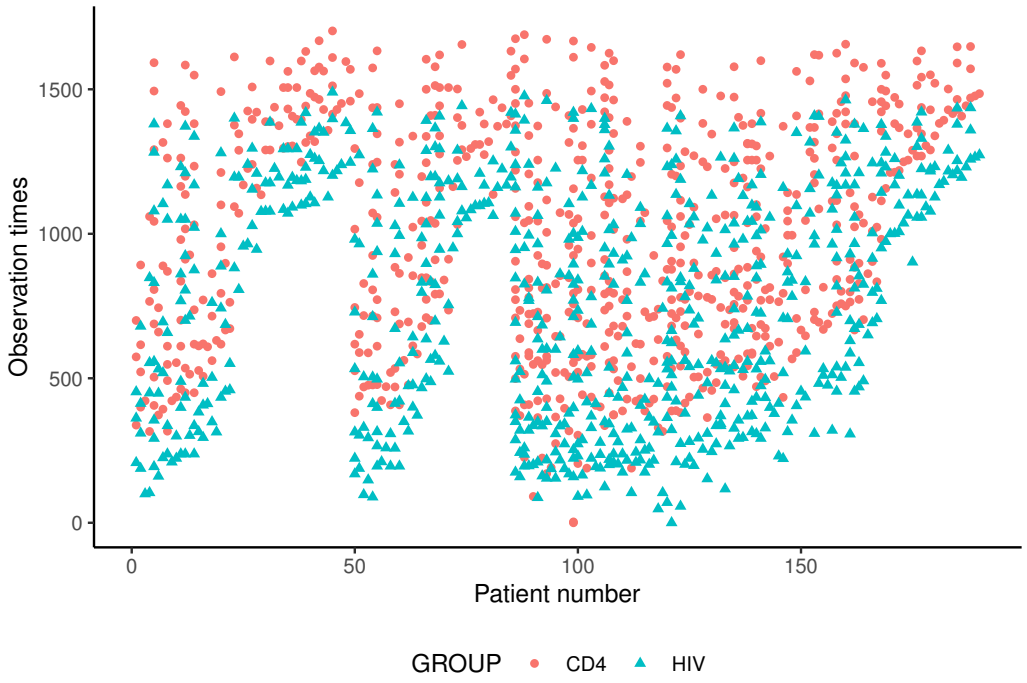

Figure 1.1: Observation times of CD4 cell counts and HIV viral load by each patient.

patients had additional visits because of random HIV-related infection occurrence, implying possible informative observation times. The existence of informative observation times or processes can often occur too in other longitudinal studies. For example, Sun et al. (2005) discussed such an example arising from a bladder cancer follow-up study conducted by the Veterans Administration Cooperative Urological Research Group. It followed the patients with superficial bladder tumors and some patients had significantly more clinical visits than others. Medical cost data also give an example of informative observation processes when some patients may make more visits to clinics or hospitals (Sun et al., 2005). To deal with the asynchronicity between response observation times and covariates measurement times, one ad hoc but

commonly used method is the last value carried forward approach, which imputes the values of covariates by their most recent observed values. However, it is easy to see that it can yield biased estimators (Cao et al., 2016).

To properly take into account the asynchronicity between response observation times and covariates measurement times, Cao et al. (2015b) proposed a generalized estimation equation approach using kernel weighting under the generalized linear model. Following them, Chen and Cao (2017) generalized the method to the partial linear model situation. In addition, Cao et al. (2015a) considered regression analysis of recurrent event data with sparse longitudinal covariates and developed a similar kernel weighted method.

Some literature has been developed for regression analysis of synchronous longitudinal data with informative observation processes and among others, early work includes Lin et al. (2004) and Sun et al. (2005). The former proposed a class of inverse intensity-of-visit process-weighted estimators under the framework of typical marginal regression models, while the latter presented some conditional models and developed an estimating equation-based estimation procedure. In addtion, Sun et al. (2007), Song et al. (2012) and Qu et al. (2018) gave some joint modeling methods, and Han et al. (2014) provided an estimating equation-based method when the follow-up time may be informative too. In addition, among others, He et al. (2009), Li et al. (2010), and Li et al. (2013) discussed the same problem where the longitudinal variable of interest represents some counts, which can be regarded as panel count data. However, it does not seem that there exists an established approach that can allow both the asynchronicity and the informative observation process. As pointed out by many authors, when they exist, the analysis that ignores either the asynchronicity

or the informative observation process could result in biased results and misleading conclusions. In Chapter 4, we will present a class of general and flexible models and an estimating equation-based approach that can deal with both issues.

1.4 Outline of the Dissertation

The remainder of this dissertation will be organized as follows. In Chapter 2, we consider variable selection for the recurrent event data with the additive model (1.1). We first introduce the existing estimating method for the recurrent event data with the additive model by the estimating equation method. Then we will introduce the motivation and details of the BAR method for variable selection. As discussed above, most variable selection methods cannot be directly applied to the event history data. We therefore propose a pseudo loss function method to enable variable selection for recurrent event data by the penalized method. We rigorously established the oracle properties and grouping effects of the proposed method. A simulation study shows the performance of the variable selection and compares it with other popular variable selection methods, namely, LASSO, ALASSO and SCAD. Another numerical study also demonstrates the grouping effect of the method proposed in Chapter 2. An application to the chronic granulomatous disease (CGD) study illustrates the proposed method in practice.

Chapter 3 considers the BAR variable selection method for panel count data under the additive model (1.1). Similarly to the method for the recurrent event data, a pseudo loss function is constructed from the estimating equation for panel count used to infer the panel count data. The oracle property and grouping effect of the

proposed method are proved. Besides, numerical studies are carried out to compare the finite-sample properties of the proposed method with other methods and illustrate its grouping effect. We then discuss the application of the proposed method to the skin cancer trial.

Chapter 4 discusses regression analysis for asynchronous longitudinal data in the presence of informative observation times. We will present a class of flexible semiparametric transformation models and a kernel-based estimating equation method for inference. We develop the asymptotic properties of the estimators from the proposed method. An extensive simulation shows the finite-sample properties of the estimates. An application to the HIV study (Wohl et al., 2005) demonstrates the proposed method in practice.

Chapter 2

Variable Selection for Recurrent Event Data with Broken Adaptive Ridge Regression

2.1 Introduction

As described in Section 1.2, recurrent event data are common in many areas and many methods have been developed for analyzing recurrent event data. However, there exists little research on the variable selection in the context of recurrent event data except Tong et al. (2009b) and Chen and Wang (2013). The former considered the data arising from a multiplicative rate model and proposed a penalized estimating equation-based procedure with the SCAD penalty function (Fan and Li, 2001). In contrast, the latter considered the additive rate model and developed a penalized least-squares loss function-based method with the use of the SCAD penalty function and the ALASSO proposed by Zou (2006). As will be seen below, the two meth-

ods may not perform well sometimes and do not have the grouping effect, which is especially important when variables are highly correlated.

In this chapter, we elaborate on the BAR regression approach mentioned in Section 1.2 for variable selection of recurrent event data. To describe the BAR regression approach, we will first introduce some notation and the model to be used throughout this chapter and also briefly describe the existing estimation procedure in Section 2.2. The proposed method is presented in Section 2.3, and as mentioned above, it approximates the L_0-penalized regression using an iteratively reweighted L_2-penalized algorithm and takes the limit of the algorithm as the BAR estimator. The approach has the advantages of performing simultaneous variable selection and parameter estimation and accommodating clustering effects. Also in Section 2.3, the asymptotic properties of the proposed BAR estimator including the oracle property are established. Section 2.4 presents the results of extensive simulation studies to assess the performance of the proposed methodology, and they suggest that the proposed method works well and can outperform the existing methods for practical situations. An application is provided in Section 2.5 and Section 2.6 contains some discussion and concluding remarks.

2.2 Model and the Existing Estimation Procedure

Consider an event history study consisting of n independent subjects which concerns the occurrence of a recurrent event of interest. For subject i, let $N_i^*(t)$ denote the underlying recurrent event process indicating the total number of the occurrences of the event over the time interval $[0, t]$. Let C_i denote the follow-up time on subject

i and let $N_i(t) = N_i^*(t \wedge C_i)$ be the observed recurrent event process. Suppose that there exists a p-dimensional vector of external covariate process, denoted by $\boldsymbol{Z}_i^*(t) = (Z_{i1}^*(t), \ldots, Z_{ip}^*(t))'$, and the main objective is to perform regression analysis with the focus on parameter estimation and covariate selection.

To describe the covariate effect, we assume that given $\boldsymbol{Z}_i^*(t)$, $N_i^*(t)$ satisfies the model (1.1). As introduced in Section 1.2, (1.1) is usually referred to as the additive rate model and has been studied by many authors for analyzing recurrent event data. Cook and Lawless (2007), Lin et al. (2000) and Schaubel et al. (2006) discussed the validity and usefulness of the model and Schaubel et al. (2006) provided an approach to measure the validity of the model given the observed recurrent event data. But there does not seem to exist an established method that can be used for covariate selection except Chen and Wang (2013). As opposed to its commonly used alternative, the proportional rate model, where the regression coefficient reflects relative effects, model (1.1) characterizes the absolute covariate effects, which are often of direct interest in epidemiological and clinical studies.

For the time being, we are only interested in the estimation of the regression parameter $\boldsymbol{\beta}$. For this, define

$$dM_i(t; \boldsymbol{\beta}) = dN_i(t) - I(C_i > t)\{d\mu_0(t) + \boldsymbol{\beta}' \boldsymbol{Z}_i^*(t) dt\}.$$

One can easily show that $M_i(t; \boldsymbol{\beta})$ is a zero-mean stochastic process at the true value, say $\boldsymbol{\beta}_0$, of the regression parameter, and this motivates the estimating equations

$$\sum_{i=1}^{n} \int_0^t dM_i(s; \boldsymbol{\beta}) = 0 \tag{2.1}$$

and

$$\sum_{i=1}^{n} \int_0^{\tau} \boldsymbol{Z}_i^*(s) dM_i(s;\boldsymbol{\beta}) = 0 \tag{2.2}$$

for estimation of $\boldsymbol{\beta}_0$ and $\mu_0(t)$, where τ is a pre-specified constant such that $P(C_i \geq \tau) > 0$ for $i = 1, \ldots, n$. Solving (2.1) for $\mu_0(t)$ with $\boldsymbol{\beta}$ fixed and plugging the solution into the (2.2), we obtain the estimating equation

$$U_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{Z}_i^*(s) - \bar{\mathbf{Z}}^*(s) \right\} dM_i(s;\boldsymbol{\beta}) \; = \; 0 \, . \tag{2.3}$$

Here $\bar{\boldsymbol{Z}}^*(t) = \boldsymbol{S}^{(1)}(t)/S^{(0)}(t)$ and $\boldsymbol{S}^{(k)}(t) = n^{-1} \sum_{i=1}^{n} I(C_i \geq t) \boldsymbol{Z}_i^{*\otimes k}(t)$ with $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = a$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$ for a vector $\boldsymbol{a}$, $k = 0, 1, 2$. Such equations were also given by Schaubel et al. (2006).

Solving (2.3) gives an explicit estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{b}} = \Big[\sum_{i=1}^{n} \int_0^{\tau} I(C_i \geq s) \big\{ \boldsymbol{Z}_i^*(s) - \bar{\boldsymbol{Z}}^*(s) \big\}^{\otimes 2} ds \Big]^{-1} \Big[\sum_{i=1}^{n} \int_0^{\tau} \big\{ \boldsymbol{Z}_i^*(s) - \bar{\boldsymbol{Z}}^*(s) \big\} dN_i(s) \Big] .$$

Schaubel et al. (2006) showed that under some regularity conditions, $\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\beta}$ and $\sqrt{n}(\hat{\boldsymbol{b}} - \boldsymbol{\beta})$ has an asymptotic normal distribution.

2.3 BAR Regression Estimation Procedure

Now we discuss the covariate selection problem and for this, let

$$\boldsymbol{P}_n = \sum_{i=1}^{n} \int_0^{\tau} \big\{ \boldsymbol{Z}_i^*(s) - \bar{\boldsymbol{Z}}^*(s) \big\} dN_i(s)$$

and

$$\boldsymbol{\Omega}_n = \sum_{i=1}^{n} \int_0^{\tau} I(C_i \geq s)\{\boldsymbol{Z}_i^*(s) - \bar{\boldsymbol{Z}}^*(s)\}^{\otimes 2} ds\,,$$

and define a pseudo loss function

$$l(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Omega}_n\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{P}_n\,.$$

One can easily show that $\hat{\boldsymbol{b}}$ is the minimizer of $l(\boldsymbol{\beta})$. Also let $\boldsymbol{X}$ denote the $p \times p$ upper triangular matrix given by the Cholesky decomposition of $\boldsymbol{\Omega}_n$ in $\boldsymbol{\Omega}_n = \boldsymbol{X}'\boldsymbol{X}$, and $\boldsymbol{y} = (\boldsymbol{X}')^{-1}\boldsymbol{P}_n$. Then we can see that the minimization of $l(\boldsymbol{\beta})$ is equivalent to minimizing the least-squares loss function $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ up to a constant. This suggests that by borrowing the idea behind the penalized least squares, we define the L_0-penalized least-squares estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}(L_0) = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p} I(\beta_j \neq 0)\}\,, \tag{2.4}$$

where $\lambda_n > 0$ is a tuning parameter.

Although with some good properties, the determination of $\hat{\boldsymbol{\beta}}(L_0)$ is very difficult or computationally infeasible. To deal with this, we propose to approximate (2.4) by

$$g(\tilde{\boldsymbol{\beta}}) \equiv \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p} \frac{\beta_j^2}{\tilde{\beta}_j^2}\}\,,$$

which can be rewritten as $g(\tilde{\boldsymbol{\beta}}) = \{\boldsymbol{X}'\boldsymbol{X} + \lambda_n \boldsymbol{D}(\tilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{X}'\boldsymbol{y}$, or

$$g(\tilde{\boldsymbol{\beta}}) = \{\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\tilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{P}_n\,,$$

where $\boldsymbol{D}(\tilde{\boldsymbol{\beta}}) = \text{diag}\{\tilde{\beta}_1^{-2}, \ldots, \tilde{\beta}_p^{-2}\}$ with $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)'$ denoting a reasonable or consistent estimator of $\boldsymbol{\beta}_0$ such that there is no zero components in $\tilde{\boldsymbol{\beta}}$. More discussion on this will be given below.

To see why the proposed idea works, we note that the quadratic function is an approximation to the L_0 penalized regression defined in (2.4) and that L_0 penalty is a good tool for variable selection except for its difficult computation feature. Also note that the approximation idea of iteratively reweighted quadratic penalization actually has its roots in the well-known Lawsons algorithm in the classical approximation theory and the sparse signal reconstruction theory (Gorodnitsky and Rao, 1997; Lawson, 1961). In the approximation above, the weighted penalty for a zero component will iteratively become large and as a consequence, a zero coefficient estimate is expected to decrease and converge to zero. In contrast, the weighted penalty for a non-zero component is expected to converge to a constant.

We define the BAR estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_R^* = \lim_{k \to \infty} \hat{\boldsymbol{\beta}}_R^{(k)}$$

based on the iterative formula $\hat{\boldsymbol{\beta}}_R^{(k)} = g(\hat{\boldsymbol{\beta}}_R^{(k-1)})$. For the selection of the initial values for the iteration procedure above, we suggest $\hat{\boldsymbol{\beta}}_R^{(0)} = (\boldsymbol{\Omega}_n + \xi_n \boldsymbol{I})^{-1} \boldsymbol{P}_n$, where $\xi_n \geq 0$. When $\xi_n > 0$, $\hat{\boldsymbol{\beta}}_R^{(0)}$ is the ridge estimator; if $\xi_n = 0$, $\hat{\boldsymbol{\beta}}_R^{(0)}$ reduces to the unpenalized estimator $\hat{\boldsymbol{b}} = \boldsymbol{\Omega}_n^{-1} \boldsymbol{P}_n$. The idea discussed above has been considered under different contexts by Frommlet and Nuel (2016) and Liu and Li (2016), who demonstrated empirically that as an automatic variable selection and parameter estimation procedure, the BAR method could provide substantial improvements over some existing methods. However, no theoretical justification was provided.

To establish the asymptotic properties of $\hat{\boldsymbol{\beta}}_R^*$, we write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ with $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$ being the first q and remaining $p-q$ components of $\boldsymbol{\beta}_0$, respectively, and without loss of generality assume that $\boldsymbol{\beta}_{02} = 0$. Let $\hat{\boldsymbol{\beta}}_R^{(k)} = (\hat{\boldsymbol{\beta}}_{R1}^{(k)'}, \hat{\boldsymbol{\beta}}_{R2}^{(k)'})'$ and $\hat{\boldsymbol{\beta}}_R^* = (\hat{\boldsymbol{\beta}}_{R1}^{*'}, \hat{\boldsymbol{\beta}}_{R2}^{*'})'$ denote the corresponding decomposition of $\hat{\boldsymbol{\beta}}_R^{(k)}$ and $\hat{\boldsymbol{\beta}}_R^*$, respectively. Let $\bar{\boldsymbol{z}}^*(t) = \lim_{n\to\infty} \bar{\boldsymbol{Z}}^*(t)$, $\boldsymbol{\Omega} = E[\int_0^\tau I(C_i \geq t)\{\boldsymbol{Z}_i^*(t) - \bar{\boldsymbol{z}}^*(t)\}^{\otimes 2} ds]$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = E[\int_0^\tau \{\boldsymbol{Z}_i^*(t) - \bar{\boldsymbol{z}}^*(t)\} dM_i(t;\boldsymbol{\beta})]^{\otimes 2}$, and let $\boldsymbol{\Omega}_{(1)}$, $\boldsymbol{\Sigma}_{(1)}$, $\boldsymbol{P}_n^{(1)}$ and $\boldsymbol{\Omega}_n^{(1)}$ be the leading $q \times q$ submatrix of $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$, $\boldsymbol{P}_n$ and $\boldsymbol{\Omega}_n$, respectively.

The following are the regularity conditions needed for the asymptotic properties of $\hat{\boldsymbol{\beta}}_R^*$.

(C1) $\{N_i^*(\cdot), C_i, \boldsymbol{Z}_i^*\}$, $i = 1, \ldots, n$, are independent and identically distributed.

(C2) $P(C_i \geq \tau) > 0$, for $i = 1, \ldots, n$.

(C3) $N_i(\tau)$ is bounded by a constant.

(C4) The matrix $\boldsymbol{\Omega}$ is positive definite.

(C5) The $\boldsymbol{Z}_i^*(\cdot)$'s have bounded total variations, i.e., $\{\|\boldsymbol{Z}_i^*(0)\| + \int_0^\tau \|d\boldsymbol{Z}_i^*(t)\|\}$ is bounded for all i, where $\|\boldsymbol{Z}_i^*(\cdot)\|$ is the Euclidean metric of the vector $\boldsymbol{Z}_i^*(\cdot)$.

(C6) $c^{-1} < \lambda_{min}(\boldsymbol{\Omega}) \leq \lambda_{max}(\boldsymbol{\Omega}) < c$, for all $n > 0$ and some large constant $c > 1$, where $\lambda(\boldsymbol{Q})$ stands for the eigenvalues of the matrix $\boldsymbol{Q}$.

(C7) $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$ as $n \to \infty$.

Theorem 1. *Assume that the regularity conditions (C1)-(C7) given above hold. Suppose that the initial estimator satisfies* $\hat{\boldsymbol{\beta}}_R^{(0)} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$. *Then with probability tending to 1, the BAR estimator* $\hat{\boldsymbol{\beta}}_R^* = (\hat{\boldsymbol{\beta}}_{R1}^{*'}, \hat{\boldsymbol{\beta}}_{R2}^{*'})'$ *has the following properties:*

1. $\hat{\boldsymbol{\beta}}^*_{R1}$ *exists and is the unique fixed point of the equation* $\boldsymbol{\beta}_1 = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1))^{-1} \boldsymbol{P}_n^{(1)}$, *where* $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = diag\{\beta_1^{-2}, \ldots, \beta_q^{-2}\}$;

2. $\hat{\boldsymbol{\beta}}^*_{R2} = 0$;

3. $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{R1} - \boldsymbol{\beta}_{01}) \xrightarrow{D} N(0, \boldsymbol{\Omega}_{(1)}^{-1} \boldsymbol{\Sigma}_{(1)} \boldsymbol{\Omega}_{(1)}^{-1})$ *as* $n \to \infty$.

For a variable selection procedure, in addition to the oracle property, another property that is often desired is the grouping effect, meaning that the highly correlated covariates should have similar regression coefficients and be selected or deleted simultaneously. For the proposed BAR approach, it follows from $\boldsymbol{\Omega}_n = \boldsymbol{X}'\boldsymbol{X}$ that the (j, k) elements of the two matrices are equal, giving

$$\sum_{i=1}^{n} \int_0^{\tau} I(C_i \geq s)\{Z_{ij}^*(s) - \bar{Z}_j^*(s)\}\{Z_{ik}^*(s) - \bar{Z}_k^*(s)\} ds = \boldsymbol{x}_j' \boldsymbol{x}_k \,,$$

where $\boldsymbol{x}_j$ denotes the jth p-dimensional column vector of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $j, k = 1, \ldots, p$. This implies that the correlation between the original covariates $\boldsymbol{Z}_j^*$ and $\boldsymbol{Z}_k^*$ can be described by that between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ and leads to the following grouping effect property.

Theorem 2. *Assume that* $\boldsymbol{X}$ *is standardized in the sense that* $\Sigma_{j=1}^n x_{ij} = 0$ *and* $\Sigma_{j=1}^n x_{ij}^2 = 1$. *Then with probability tending to one, the BAR estimator* $\hat{\boldsymbol{\beta}}^*_R = (\hat{\beta}^*_{R1}, \ldots, \hat{\beta}^*_{Rp})'$ *satisfies*

$$\Big| \frac{1}{\hat{\beta}^*_{Ri}} - \frac{1}{\hat{\beta}^*_{Rj}} \Big| \leq \frac{1}{\lambda_n} \|\boldsymbol{y}\| \sqrt{2(1 - \rho_{ij})}$$

for those nonzero components $\hat{\beta}^*_{Ri}$ *and* $\hat{\beta}^*_{Rj}$, *where* ρ_{ij} *denotes the sample correlation coefficient of* $\boldsymbol{x}_i$ *and* $\boldsymbol{x}_j$.

The proofs of Theorems 1 and 2 are sketched in Appendix A.1. To implement the procedure, one needs to choose the tuning parameter λ_n as well as ξ_n. To reduce the computational burden, we suggest employing the K-fold cross-validation to chooose those tuning parameters, where K is a positive integer.

2.4 A Simulation Study

An extensive simulation study was conducted to investigate the performance of the proposed BAR regression estimation procedure. In the study, recurrent event data were generated from the mixed Poisson processes with the rate function given in (1.1) multiplied by the Gamma random variables with mean 1 and variance σ^2. The follow-up times C_i's were generated from the uniform distribution over $(0, 4.5)$. For the covariate $\boldsymbol{Z}^*$, we considered several settings with different p values, different locations for non-zero components, and different types of distributions, continuous or discrete, with the components being independent or correlated. For each scenario, we considered the sample size n =100, 300 or 500 with 500 replications.

For comparison, we also considered the approach of Chen and Wang (2013) with the LASSO, ALASSO and SCAD penalty functions. The LASSO and ALASSO were implemented by using the R package *glmnet* and the SCAD was realized by the R package *ncvreg*. For the selection of the tuning parameter, both packages use similar methods to select 50 candidates, and the largest candidates are $\max_l |\boldsymbol{x}_l'\boldsymbol{y}|$, $\max_l |w_l\boldsymbol{x}_l'\boldsymbol{y}|$, and $\max_l |\boldsymbol{x}_l'\boldsymbol{y}|$ for the LASSO, ALASSO, and SCAD, respectively. For the ALASSO, the weights $w_i = 1/\left|\beta_i^{ols}\right|$ and $\gamma = 1$ were used as suggested by Zou (2006), where $\boldsymbol{\beta}^{ols}$ denotes the ordinary least-squares estimator of $\boldsymbol{\beta}$. For the selection

of the tuning parameter λ_n as well as ξ_n, we used the 5-fold cross-validation and considered five candidate values for ξ_n chosen to be the equally spaced points over $[0.01, 10]$ in a logarithmic scale. For each ξ_n, 50 candidate values were considered for λ_n that are equally spaced over $[\epsilon\lambda_{max}, \lambda_{max}]$ also in a logarithmic scale, where ϵ was set to be 0.001 and $\lambda_{max} = \max_l 4\left(\boldsymbol{x}_l'\boldsymbol{y}\right)^2/(\boldsymbol{x}_l'\boldsymbol{x}_l)$ with $\boldsymbol{x}_l$ being the lth column of $\boldsymbol{X}$.

Table 2.1 presents the the covariate selection results with $p = 9$, where the true value of $\boldsymbol{\beta}_0$ is $(1, 0, -1, 0, 0, 0, 0, 0, 0)'$, $\mu_0(t) = 0.45t$, and $\sigma^2 = 0.5$ or 1. Here the covariates were generated from the normal distribution and the correlation between Z_i^* and Z_j^* was set as $\rho^{|j-i|}$ with $\rho = 0.1$, 0.5 or 0.95. The table includes the average of the empirical values of the mean squared error (MSE) defined as $(\hat{\boldsymbol{\beta}}_R^* - \boldsymbol{\beta}_0)' E(\boldsymbol{Z}^*\boldsymbol{Z}^{*\prime})(\hat{\boldsymbol{\beta}}_R^* - \boldsymbol{\beta}_0)$, the average percent of the numbers of correctly selected zero coefficients (Corr), and the average percent of the numbers of incorrectly selected zero coefficients (Inco). It is seen that with respect to the correctly selected zero coefficients, the BAR always gave the best performance and the BAR and SCAD seemed to significantly outperform LASSO and ALASSO. On the incorrectly selected zero coefficients, no method gave overall better performance than others. With respect to the MSE, the BAR yielded the least MSE and as expected, all methods gave better performance as the sample size increased.

Tables 2.2 and 2.3 give the results for the same setup as Table 2.1 but different values for p and $\boldsymbol{\beta}_0$. In Table 2.2, $p = 50$ and the first six components of $\boldsymbol{\beta}_0$ were $(1, 0, -1, 1, 0, -1)'$ with the other components being zero, while in Table 2.3, $p = 100$ and the first thirty components of $\boldsymbol{\beta}_0$ were replicators of $(1, 0, -1)'$ with the other components being zero. Also all components were assumed to be independent of each other and the other setup is the same as Table 2.1 except that the recurrent event

processes followed either the homogeneous Poisson process ($\sigma^2 = 0$) or the mixed Poisson process with the variance of the Gamma random variables being $1(\sigma^2 = 1)$. It is apparent that all tables gave similar conclusions to those given by Table 2.1 and again suggested that the proposed BAR procedure seems to give better performance than the other procedures on the covariate selection in general.

To assess the grouping ability of the methods discussed above, we performed a simulation study with $p = 12$, $\boldsymbol{\beta}_0 = (1, 1, 1, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0)$, and $n = 10000$; other settings are the same as before. A large sample size was used to order to reveal the grouping effect. For the covariate generation, we assumed that

$$Z_j^* = x_1 + \epsilon_j, \quad x_1 \sim N\left(0, 50^2\right), \quad j = 1, 2, 3;$$
$$Z_j^* = x_2 + \epsilon_j, \quad x_2 \sim N\left(0, 50^2\right), \quad j = 4, 5, 6;$$

$$Z_j^* \sim N(0, 50^2), \quad \text{independent with each other, } j = 7, \ldots, 12;$$

where the ϵ_j are i.i.d. $N(0, 0.025^2)$. That is, the 12 covariates were from three different groups with the within-group correlations being nearly 1 and the between-group correlations being close to 0. Figure 2.1 shows the solution paths of the estimates given by the four methods. One can see the obvious grouping effect of the proposed BAR estimator when the tuning parameters are in the proper range. The BAR method clearly selected the six relevant covariates into their two separate groups with almost the same coefficients within each group when $log(\lambda_n)$ is roughly between -4.3 and 2.5. Although the estimate paths were unstable when λ is small, the cross-validation criterion can successfully find the tuning parameters inducing the grouping effect. In contrast, all three other methods showed little grouping effects as they selected only

β_1 and β_4.

2.5 An Application to the Chronic Granulomatous Disease Study

We apply the proposed BAR regression estimation procedure to the recurrent event data arising from the chronic granulomatous disease (CGD) study discussed by Chen and Wang (2013), Fleming and Harrington (2005) and Tong et al. (2009a), among others. The CGD is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections that are usually present in early life and may lead to death in childhood. The CGD study consists of 128 patients with the chronic granulomatous disease between October 1988 and March 1989. For each subject, the collected information includes the occurrence times of all recurrent serious infections during the study period, and the study involves two treatments, placebo (65) and gamma interferon (63). During the study period, 20 placebo patients and 7 treated patients had experienced at least one serious infection. A goal of the study is to investigate the ability of gamma interferon to reduce the occurrence rate of serious infections.

We use Z_1^* to denote the treatment with $Z_1^* = 0$ for the patients on gamma interferon and $Z_1^* = 1$ otherwise. In addition to the treatment, the data set includes information on eight other covariates. They are the pattern of inheritance ($Z_2^* = 0$ for X-linked patients and $Z_2^* = 1$ for autosomal recessive patients), the age, height and weight of the patient (Z_3^*, Z_4^*, Z_5^*), the use of corticosteroids at time of study entry ($Z_6^* = 0$ if yes and $Z_6^* = 1$ if no), the use of prophylactics at time of study

entry ($Z_7^* = 0$ if yes and $Z_7^* = 1$ if no), the patient's gender ($Z_8^* = 1$ if female and $Z_8^* = 0$ if male), and the hospital category (US-NIH, US-other, Europe-Amsterdam and Europe-other). For the last covariate, following Tong et al. (2009a) and Chen and Wang (2013), we describe it using three dummy variables as $Z_9^* = 1$ for US-NIH and $Z_9^* = 0$ otherwise, $Z_{10}^* = 1$ for US-other and $Z_{10}^* = 0$ otherwise, $Z_{11}^* = 1$ for Europe-Amsterdam and $Z_{11}^* = 0$ otherwise.

For the analysis, we consider both the selection of covariates or predictive factors and the estimation of their effects simultaneously. The results are given in Table 2.4 and here for comparison, we also included the results given by the LASSO, ALASSO, SCAD and ordinary least-squares procedure. Table 2.4 includes the selected covariates with their estimated effects and the estimated standard errors (in parentheses) for each method. It is seen that all procedures selected the treatment indicator, indicating that gamma interferon had a significant effect on reducing the rate of serious infections. In addition, the BAR procedure suggested that the pattern of inheritance and the use of prophylactics at time of study entry seem to have some effects on the rate of serious infections, and also the infection rate appears to be different between the patients in the hospitals in the US-other group and other types of hospitals. Based on the ALASSO procedure, the patient's age and the use of corticosteroids at the time of study entry could affect the infection rate. It is interesting to note that as LASSO and SCAD, the method given by Chen and Wang (2013) selected only the treatment indicator Z_1^*, while the results obtained here are similar to those given by Tong et al. (2009a) under the proportional rate model.

To assess the appropriateness of the selected models by each of the four methods

given above, we performed the model checking by using the statistic

$$D = \left\{ \sum_{l=1}^{L} \sum_{i=1}^{n} I_i(t_l < C_i) \right\}^{-1} \sum_{l=1}^{L} \sum_{i=1}^{n} \left\{ \hat{M}_i(t_l) \right\}^2$$

based on the sum of the mean squared residuals. Here $\hat{M}_i(t)$ is defined as $M_i(t, \boldsymbol{\beta})$ with the unknowns replaced by their estimates and the $\{t_l\}_{l=1}^{L}$ denote all time points where the $N_i(t)$ have jumps. Lin et al. (2001) employed a similar statistic for checking transformation models with the focus on estimation. The application of this statistic to the data yielded $D = 0.4598$, 0.4670, 0.4588 and 0.4662 for the models selected by the four methods in Table 2.4, respectively, and indicates that the four models are not significantly different.

2.6 Discussion and Concluding Remarks

This chapter discussed simultaneous covariate selection and estimation of covariate effects for the event history study that yields recurrent event data. As mentioned above, only limited research exists for covariate selection in Tong et al. (2009a) and Chen and Wang (2013) due to the special data structures and the difficulties involved. We presented a BAR regression estimation procedure that can not only allow for estimation and variable selection simultaneously but also accommodate the clustering effect when covariates are highly correlated. The oracle property of the proposed approach was established, and the simulation study indicated that the proposed method has better performance than the existing procedures.

Table 2.1: Result on the selection of the normal covariates with $p = 9$.

			$n = 100$			$n = 300$			$n = 500$		
σ^2	ρ	Method	MSE	Corr	Inco	MSE	Corr.	Inco.	MSE	Corr.	Inco.
0.5	0.1	BAR	0.586	0.829	0.084	0.136	0.922	0.000	0.082	0.919	0.000
		LASSO	0.722	0.545	0.071	0.230	0.518	0.000	0.149	0.531	0.000
		ALASSO	0.557	0.771	0.040	0.157	0.819	0.000	0.096	0.826	0.000
		SCAD	0.723	0.563	0.072	0.124	0.667	0.000	0.072	0.789	0.000
	0.5	BAR	0.493	0.823	0.098	0.125	0.913	0.001	0.070	0.909	0.000
		LASSO	0.594	0.557	0.104	0.195	0.525	0.000	0.117	0.539	0.000
		ALASSO	0.464	0.734	0.060	0.140	0.827	0.000	0.079	0.827	0.000
		SCAD	0.600	0.617	0.117	0.111	0.750	0.000	0.060	0.830	0.000
	0.9	BAR	0.196	0.759	0.179	0.051	0.869	0.012	0.030	0.881	0.000
		LASSO	0.202	0.615	0.176	0.071	0.535	0.000	0.047	0.561	0.000
		ALASSO	0.187	0.701	0.136	0.059	0.743	0.003	0.035	0.789	0.000
		SCAD	0.225	0.722	0.236	0.049	0.804	0.010	0.027	0.870	0.001
1	0.1	BAR	0.564	0.797	0.078	0.147	0.908	0.000	0.088	0.923	0.000
		LASSO	0.704	0.523	0.070	0.241	0.482	0.000	0.152	0.504	0.000
		ALASSO	0.547	0.758	0.041	0.167	0.797	0.000	0.100	0.836	0.000
		SCAD	0.715	0.550	0.075	0.136	0.651	0.000	0.076	0.787	0.000
	0.5	BAR	0.443	0.803	0.069	0.125	0.895	0.000	0.071	0.900	0.000
		LASSO	0.539	0.524	0.072	0.201	0.501	0.000	0.125	0.503	0.000
		ALASSO	0.430	0.736	0.042	0.144	0.790	0.001	0.085	0.791	0.000
		SCAD	0.558	0.602	0.085	0.113	0.739	0.001	0.064	0.812	0.000
	0.9	BAR	0.207	0.771	0.209	0.053	0.843	0.016	0.027	0.893	0.002
		LASSO	0.209	0.610	0.188	0.072	0.566	0.004	0.045	0.568	0.000
		ALASSO	0.190	0.697	0.144	0.059	0.749	0.005	0.033	0.796	0.000
		SCAD	0.234	0.718	0.241	0.050	0.797	0.020	0.023	0.886	0.000

Table 2.2: Result on the selection of the normal covariates with $p = 50$.

			$n = 100$			$n = 300$			$n = 500$		
σ^2	ρ	Method	MSE	Corr	Inco	MSE	Corr.	Inco.	MSE	Corr.	Inco.
0.5	0.1	BAR	3.199	0.947	0.540	0.559	0.966	0.020	0.279	0.981	0.001
		LASSO	3.299	0.843	0.516	1.132	0.693	0.014	0.701	0.679	0.000
		ALASSO	3.309	0.747	0.219	0.680	0.817	0.003	0.384	0.834	0.000
		SCAD	3.347	0.857	0.519	0.652	0.685	0.013	0.240	0.752	0.000
	0.5	BAR	2.059	0.949	0.576	0.403	0.965	0.031	0.184	0.982	0.001
		LASSO	2.135	0.858	0.597	0.863	0.660	0.046	0.481	0.636	0.001
		ALASSO	2.198	0.759	0.281	0.500	0.804	0.005	0.265	0.829	0.000
		SCAD	2.146	0.874	0.607	0.441	0.728	0.037	0.141	0.813	0.000
	0.9	BAR	0.682	0.950	0.660	0.212	0.965	0.253	0.098	0.970	0.060
		LASSO	0.638	0.883	0.691	0.316	0.763	0.262	0.193	0.679	0.070
		ALASSO	0.908	0.758	0.462	0.251	0.771	0.093	0.138	0.779	0.014
		SCAD	0.658	0.918	0.737	0.262	0.865	0.295	0.127	0.853	0.113
1	0.1	BAR	3.119	0.945	0.501	0.575	0.967	0.017	0.285	0.980	0.001
		LASSO	3.294	0.839	0.492	1.152	0.675	0.009	0.714	0.690	0.000
		ALASSO	3.226	0.739	0.197	0.715	0.802	0.001	0.385	0.837	0.000
		SCAD	3.305	0.850	0.493	0.653	0.677	0.010	0.250	0.758	0.000
	0.5	BAR	2.025	0.942	0.522	0.379	0.964	0.017	0.201	0.981	0.002
		LASSO	2.130	0.838	0.565	0.819	0.648	0.033	0.502	0.638	0.001
		ALASSO	2.206	0.744	0.262	0.490	0.799	0.001	0.272	0.831	0.000
		SCAD	2.134	0.860	0.572	0.390	0.719	0.020	0.153	0.806	0.000
	0.9	BAR	0.691	0.950	0.687	0.216	0.961	0.253	0.100	0.965	0.056
		LASSO	0.645	0.891	0.711	0.311	0.751	0.240	0.191	0.679	0.079
		ALASSO	0.883	0.759	0.474	0.249	0.766	0.092	0.135	0.770	0.013
		SCAD	0.664	0.926	0.766	0.270	0.864	0.300	0.126	0.851	0.118

Table 2.3: Result on the selection of the normal covariates with $p = 100$.

			$n = 100$			$n = 300$			$n = 500$		
σ^2	ρ	Method	MSE	Corr	Inco	MSE	Corr.	Inco.	MSE	Corr.	Inco.
0.5	0.1	BAR	22.903	0.970	0.931	16.480	0.954	0.713	7.110	0.889	0.182
		LASSO	20.336	0.917	0.867	16.222	0.813	0.584	8.877	0.541	0.124
		ALASSO	63.584	0.569	0.493	11.013	0.704	0.172	5.469	0.691	0.022
		SCAD	19.821	0.936	0.887	16.626	0.847	0.610	7.493	0.606	0.131
	0.5	BAR	10.771	0.979	0.957	8.305	0.960	0.755	3.307	0.900	0.153
		LASSO	9.660	0.941	0.917	8.678	0.877	0.765	5.325	0.567	0.261
		ALASSO	27.797	0.620	0.556	5.748	0.704	0.180	2.804	0.686	0.017
		SCAD	9.518	0.947	0.922	8.609	0.893	0.754	3.509	0.666	0.175
	0.9	BAR	2.566	0.973	0.946	1.848	0.971	0.850	1.307	0.934	0.551
		LASSO	2.282	0.940	0.912	1.899	0.896	0.800	1.492	0.739	0.513
		ALASSO	5.791	0.679	0.611	1.703	0.722	0.378	1.003	0.665	0.121
		SCAD	2.205	0.957	0.934	1.929	0.935	0.851	1.616	0.876	0.652
1	0.1	BAR	23.857	0.969	0.935	16.213	0.953	0.699	7.618	0.899	0.214
		LASSO	20.412	0.928	0.880	16.051	0.810	0.580	9.441	0.567	0.156
		ALASSO	63.446	0.578	0.500	10.684	0.705	0.160	5.585	0.698	0.025
		SCAD	19.897	0.939	0.891	16.395	0.848	0.606	8.117	0.625	0.159
	0.5	BAR	10.768	0.976	0.955	8.225	0.943	0.719	3.489	0.902	0.172
		LASSO	9.611	0.942	0.918	8.667	0.864	0.753	5.536	0.575	0.280
		ALASSO	26.281	0.651	0.581	5.724	0.680	0.164	2.919	0.681	0.019
		SCAD	9.470	0.951	0.928	8.614	0.894	0.762	3.611	0.666	0.180
	0.9	BAR	2.571	0.973	0.951	1.846	0.974	0.859	1.324	0.936	0.571
		LASSO	2.262	0.944	0.918	1.895	0.900	0.802	1.505	0.743	0.519
		ALASSO	5.757	0.692	0.640	1.695	0.736	0.394	0.993	0.663	0.121
		SCAD	2.232	0.958	0.933	1.938	0.939	0.859	1.643	0.884	0.681

Table 2.4: Results on covariate selection and their estimated effects for the CGD study.

	BAR	LASSO	ALASSO	SCAD	OLS
Z_1^*	0.127(0.043)	0.080(0.038)	0.104(0.016)	0.088(0.036)	0.170(0.050)
Z_2^*	0.018(0.004)	0	0	0	0.152(0.113)
Z_3^*	0	0	-0.138(0.020)	0	-0.409(0.152)
Z_4^*	0	0	0	0	0.081(0.397)
Z_5^*	0	0	0	0	0.105(0.266)
Z_6^*	0	0	-0.105(0.016)	0	-0.343(0.269)
Z_7^*	0.027(0.007)	0	0	0	0.140(0.087)
Z_8^*	0	0	0	0	-0.134(0.111)
Z_9^*	0	0	0	0	0.059(0.065)
Z_{10}^*	0.018(0.004)	0	0	0	0.094(0.069)
Z_{11}^*	0	0	0	0	-0.051(0.076)

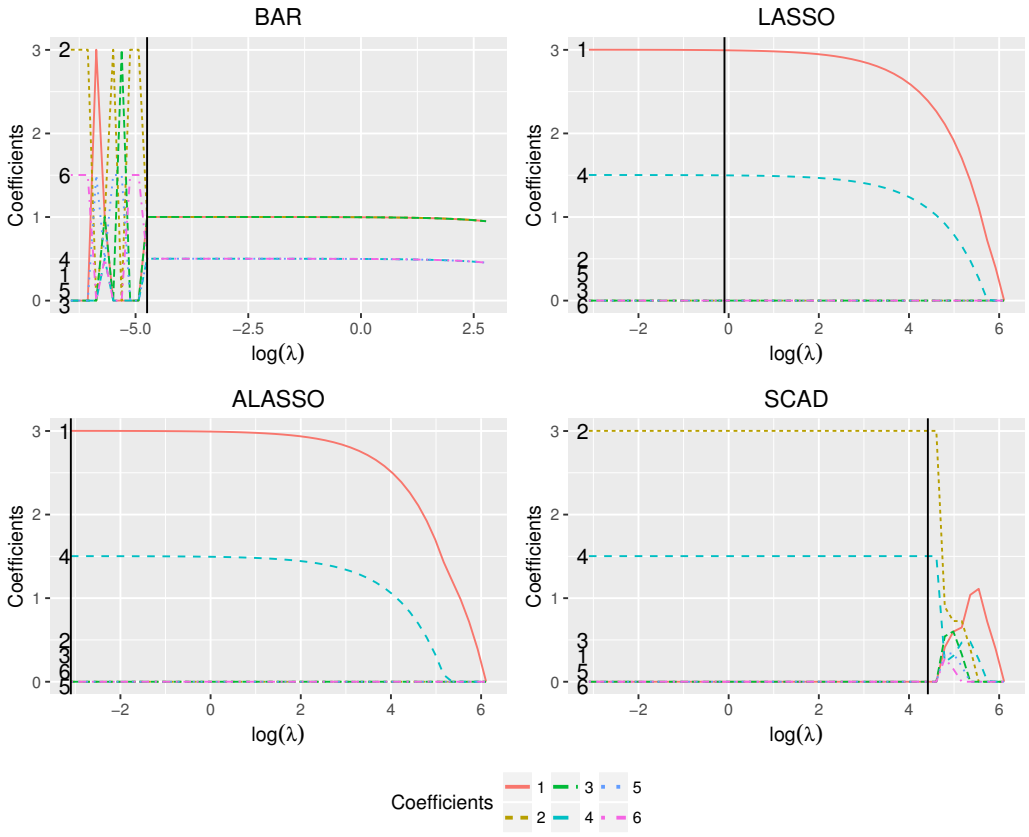

Figure 2.1: Path plots of the four estimates with the black solid vertical lines corresponding to the optimal tuning parameters based on the 5-fold cross-validation.

Chapter 3

Simultaneous Estimation and Variable Selection for Panel Count Data

3.1 Introduction

In this chapter, we will discuss simultaneous estimation and variable selection for panel count data under model (1.1) using the BAR method. In particular, unlike the two existing methods given in Tong et al. (2009a) and Zhang et al. (2013), the resulting estimator from the proposed method has the oracle property and grouping effect.

To present the new approach, we first introduce some notation and assumptions in Section 3.2; also briefly describe there is an estimation procedure if one is only interested in estimation. We then discuss in Section 3.3 the proposed method, which can be seen as a combination of the nonconcave penalized likelihood approach (Fan

and Li, 2001) and the estimating equation approach (Lin and Ying, 1994). The proposed method is similar to that described in Section 2.3 but more complicated due to the incompleteness of the observed data. In Section 3.3, we establish the asymptotic properties of the proposed method, including the oracle property and grouping effect. Section 3.4 reports the results from simulation studies conducted to assess the performance of the proposed method and they indicate that it works well in practical situations. An application to the skin cancer study is provided in Section 3.5. Some discussion and concluding remarks can be found in Section 3.6.

3.2 Notation, Assumptions and Estimation Procedure

Consider an event history study on a recurrent event and suppose that it consists of n independent subjects. For subject i, let $N_i^*(t)$ and $\boldsymbol{Z}_i^*(t)$ be defined as above but corresponding to the subject with $N_i^*(t)$ denoting the total number of the occurrences of the event until time t from the subject. Also for subject i, assume that $N_i^*(t)$ can only be observed at the discrete time points $t_{i1} \leq \cdots \leq t_{im_i}$ and define a counting process

$$H_i^*(t) = \sum_{j=1}^{m_i} \mathbf{1}(t_{ij} \leq t),$$

where m_i denotes the potential or scheduled number of observations on subject i. That is, we only have panel count data on the $N_i^*(t)$ and $N_i^*(t)$ can be observed only at the time points where $H_i^*(t)$ jumps. In addition, suppose that there exists a follow-up time, denoted by C_i, on subject i and the observed recurrent event and observation processes are $N_i(t) = N_i^*(t \wedge C_i)$ and $H_i(t) = H_i^*(t \wedge C_i)$, respectively. Then the

observed data have the form $\{N_i(t)dH_i(t), H_i(t), C_i, \boldsymbol{Z}_i^*(t)\}$ for all $i \in \{1, \ldots, n\}$. In the following, we will focus on the situation where p can diverge to infinity but $p < n$.

We assume the covariate effect to be described by model (1.1). We also assume that the covariates are external (Kalbfleisch and Prentice, 2002) and as Lin et al. (2000) pointed out, in this case, model (1.1) is equivalent to the additive mean model

$$\mathrm{E}\{N_i^*(t)|\boldsymbol{Z}_i(t)\} = \mu_0(t) + \boldsymbol{\beta}'\boldsymbol{Z}_i(t),$$

where $\boldsymbol{Z}_i(t) = \int_0^t \boldsymbol{Z}_i^*(u)du$. Furthermore, it will be assumed that the observation process $H_i^*(t)$ satisfies the following marginal rate model

$$\mathrm{E}\{dH_i^*(t)|\boldsymbol{Z}_i^*(t)\} = d\Lambda_0(t), \tag{3.1}$$

and $H_i^*(t)$ and C_i are mutually independent and also independent of $N_i^*(t)$ and $\boldsymbol{Z}_i^*(t)$, where Λ_0 is an unspecified positive nondecreasing function.

For the time being, suppose that one is only interested in the estimation of covariate effects. For this, define

$$dM_i(t; \boldsymbol{\beta}, \mu, \Lambda) = N_i(t)dH_i(t) - \mathbf{1}(C_i \geq t)\{\mu(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}d\Lambda(t),$$

and

$$\bar{\boldsymbol{Z}}(t) = \sum_{j=1}^n \mathbf{1}(C_j \geq t)\boldsymbol{Z}_j(t) \Big/ \sum_{j=1}^n \mathbf{1}(C_j \geq t), \quad \hat{\Lambda}_0(t) = \int_0^t \sum_{i=1}^n dH_i(s) \Big/ \sum_{j=1}^n \mathbf{1}(C_j \geq s),$$

which is usually referred to as the Aalen–Breslow-type estimator of $\Lambda_0(t)$. Then under

models (1.1) and (3.1), one can show that

$$\mathrm{E}\{dM_i(t;\boldsymbol{\beta}_0,\mu_0,\Lambda_0)\}=0.$$

This motivates the estimating equations

$$\sum_{i=1}^{n} dM_i(t;\boldsymbol{\beta},\mu_0,\hat{\Lambda}_0)=0 \tag{3.2}$$

and

$$U_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\boldsymbol{Z}_i(t)dM_i(t;\boldsymbol{\beta},\hat{\mu}_0,\hat{\Lambda}_0)=0 \tag{3.3}$$

for $\mu_0(t)$ and $\boldsymbol{\beta}$, where τ is the longest follow-up time. A similar idea was used in Lin and Ying (1994) for the estimation of regression parameters in the additive hazards model based on right-censored failure time data.

By solving (3.2), one can obtain

$$\hat{\mu}_0(t)=\sum_{i=1}^{n}N_i(t)dH_i(t)\Big/\sum_{i=1}^{n}dH_i(t)-\boldsymbol{\beta}^{\top}\bar{\boldsymbol{Z}}(t). \tag{3.4}$$

Then by substituting (3.4) into (3.3) and solving (3.3), we can obtain an estimator of $\boldsymbol{\beta}$ given by $\hat{\boldsymbol{b}}=\boldsymbol{\Omega}_n^{-1}\mathbf{v}_n$, where

$$\mathbf{v}_n=\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i(t)-\bar{\boldsymbol{Z}}(t)\}N_i(t)dH_i(t)$$

and

$$\boldsymbol{\Omega}_n=\sum_{i=1}^{n}\int_0^{\tau}\mathbf{1}(C_i\geq t)\{\boldsymbol{Z}_i(t)-\bar{\boldsymbol{Z}}(t)\}^{\otimes 2}d\hat{\Lambda}_0(t).$$

Note that it is easy to see that $\hat{\boldsymbol{b}}$ is the minimizer of the loss function

$$\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\Omega}_n \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{v}_n.$$

Also note that by applying the Cholesky decomposition to $\boldsymbol{\Omega}_n$, we have $\boldsymbol{\Omega}_n = \boldsymbol{X}^\top \boldsymbol{X}$, where $\boldsymbol{X}$ is a $p \times p$ upper triangular matrix. Define $\boldsymbol{y} = (\boldsymbol{X}^\top)^{-1}\mathbf{v}_n$. Then one can easily show that the minimization of $\ell(\boldsymbol{\beta})$ is equivalent to minimizing the least-square loss function $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. In other words, $\hat{\boldsymbol{b}}$ can be seen as the least-squares estimator of $\boldsymbol{\beta}$ in the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

3.3 Simultaneous Estimation and Variable Selection

Now suppose that we are interested in simultaneous estimation and variable selection on covariate effects. We will first develop a general penalized estimation procedure similar to the method in Section 2.3. The oracle property and grouping effect of the proposed approach are then established.

3.3.1 Broken Adaptive Ridge Regression

To develop a penalized procedure for $\boldsymbol{\beta}$ based on the estimation procedure presented above, it would be natural to consider and minimize the penalized function

$$\ell_0(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0)$$

based on the L_0 penalty function, where $\lambda_n > 0$ is a tuning parameter. By contrast, it is well known that this minimization process is usually computationally difficult. Corresponding to this, we propose to consider the approximation

$$\ell_2(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) = \ell(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} \beta_j^2/\tilde{\beta}_j^2, \tag{3.5}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_{p_n})^\top$ denotes a good estimator of $\boldsymbol{\beta}_0$ with no zero component to be discussed below. Note that the same idea has been discussed by Frommlet and Nuel (2016) and Liu and Li (2016) under the contexts of linear models and generalized linear models, respectively, but they only considered the empirical properties of the developed methods. It will be seen that the penalty function in (3.5) inherits some appealing properties associated with both L_0 and L_2 penalty functions.

Let $g(\tilde{\boldsymbol{\beta}})$ denote the minimizer of $\ell_2(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ given $\tilde{\boldsymbol{\beta}}$. Then we have

$$g(\tilde{\boldsymbol{\beta}}) = \{\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\tilde{\boldsymbol{\beta}})\}^{-1}\mathbf{v}_n, \tag{3.6}$$

where $\boldsymbol{D}(\tilde{\boldsymbol{\beta}}) = \mathrm{diag}(\tilde{\beta}_1^{-2}, \ldots, \tilde{\beta}_p^{-2})$. This suggests that we can estimate $\boldsymbol{\beta}$ by the broken adaptive ridge estimator for panel count data defined as

$$\hat{\boldsymbol{\beta}}_P^* = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}_P^{(k)}$$

based the iterated formula $\hat{\boldsymbol{\beta}}_P^{(k)} = g(\hat{\boldsymbol{\beta}}_P^{(k-1)})$.

For the initial value of the iteration, we suggest using $\hat{\beta}_P^{(0)} = (\boldsymbol{\Omega}_n + \xi_n \mathbf{I})^{-1}\mathbf{v}_n$ with $\xi_n \geq 0$. When $\xi_n > 0$, $\hat{\boldsymbol{\beta}}_P^{(0)}$ is the ridge estimator. If $\xi_n = 0$, $\hat{\beta}_P^{(0)}$ reduces to the unpenalized estimator $\hat{\boldsymbol{b}} = \boldsymbol{\Omega}_n^{-1}\mathbf{v}_n$. In general, the update in each iteration is well

defined since the initial $\boldsymbol{\beta}_P^{(0)}$ and its subsequent updates $\boldsymbol{\beta}_P^{(k)}$ do not yield any zero coefficient.

As the iterated values converge to the limit, $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_P^{(k-1)})$ will eventually involve some divisions by very small nonzero values that can lead to an arithmetic overflow. To avoid this, we can rewrite (3.6) as

$$g(\tilde{\boldsymbol{\beta}}) = \Gamma(\tilde{\boldsymbol{\beta}})\{\Gamma(\tilde{\boldsymbol{\beta}})\boldsymbol{\Omega}_n\Gamma(\tilde{\boldsymbol{\beta}}) + \lambda_n I_p\}^{-1}\Gamma(\tilde{\boldsymbol{\beta}})\mathbf{v}_n, \tag{3.7}$$

where $\Gamma(\tilde{\boldsymbol{\beta}}) = \text{diag}(\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$. Note that the right-hand side of (3.7) only involves multiplication by $\tilde{\beta}_j$ and hence avoids the computational instability.

Note that for the function $\ell_2(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$ defined in (3.5), it is apparent that one can define or consider the function with replacing the penalty function there by some commonly used penalty functions, such as the LASSO, ALASSO and SCAD. It will be seen that the proposed approach has good properties, namely, the oracle property and grouping effect and gives better empirical performance as discussed below. The idea used above is also similar to that behind the local quadratic approximation used in Fan and Li (2001), but the resulting iterative equations are actually quite different.

3.3.2 Asymptotic Properties of $\hat{\boldsymbol{\beta}}_P^*$

Now we discuss the asymptotic properties of the proposed BAR estimate $\hat{\boldsymbol{\beta}}_P^*$. For this, let $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p_n})^\top$ denote the true value of $\boldsymbol{\beta}$ and without loss of generality, assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top)^\top$, where $\boldsymbol{\beta}_{01}$ consists of all q_n nonzero components and $\boldsymbol{\beta}_{02}$ the remaining zero components. Correspondingly, we divide $\boldsymbol{\beta}$ and $\hat{\boldsymbol{b}}$ in the same way and also in the following, we will denote p by p_n to emphasize the dependence of p

on n.

To establish the asymptotic properties, we need the following regularity conditions.

(D1) The observations $\{N_i^*(t), H_i^*(t), C_i, \boldsymbol{Z}_i(t) : 0 \le t \le \tau\}$ with $i \in \{1, \ldots, n\}$ are independent and identically distributed.

(D2) We have $\Pr(C_i \ge \tau) > 0$ for all $i \in \{1, \ldots, n\}$.

(D3) For all $i \in \{1, \ldots, n\}$, $N_i^*(\tau)$ is bounded by a constant with $\mu_0(\tau) < \infty$.

(D4) The $\boldsymbol{Z}_i$'s have bounded total variations, i.e., $\|\boldsymbol{Z}_i(0)\| + \int_0^\tau \|d\boldsymbol{Z}_i(t)\|$ is bounded for all i, where $\|\boldsymbol{Z}_i\|$ is the Euclidean metric of the vector $\boldsymbol{Z}_i$.

(D5) For all $n > 0$, we have $1/c < \lambda_{\min}(\boldsymbol{\Omega}_n/n) \le \lambda_{\max}(\boldsymbol{\Omega}_n/n) < c$, where $c > 1$ is some large constant and $\lambda(\boldsymbol{Q})$ stands for the eigenvalues of the matrix $\boldsymbol{Q}$.

(D6) As $n \to \infty$, $p_n q_n/\sqrt{n} \to 0$, $\lambda_n/\sqrt{n} \to 0$, $\xi_n/\sqrt{n} \to 0$, $\lambda_n\sqrt{q_n/n} \to 0$ and $\lambda_n^2/(p_n\sqrt{n}) \to \infty$ as $n \to \infty$.

(D7) There exist positive constants a_0 and a_1 such that $a_0 \le |\beta_{0,j}| \le a_1$ for all $j \in \{1, \ldots, q_n\}$.

(D8) The initial estimator satisfies $\|\hat{\boldsymbol{\beta}}_P^{(0)} - \boldsymbol{\beta}_0\| = O_p(\sqrt{p_n/n})$.

Define $\hat{\Xi}_n = (\hat{U}_1\hat{U}_1^\top + \cdots + \hat{U}_n\hat{U}_n^\top)/n$, where $\hat{U}_i = \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} d\hat{M}_i(t, \hat{\boldsymbol{b}}, \hat{\mu}_0, \hat{\Lambda}_0)$. First we will describe the oracle property.

Theorem 3. *Assume that the regularity conditions* (D1)–(D8) *given above hold. Then with probability tending to* 1*, the BAR estimator* $\hat{\boldsymbol{\beta}}_P^* = (\hat{\boldsymbol{\beta}}_{P1}^{*\top}, \hat{\boldsymbol{\beta}}_{P2}^{*\top})^\top$ *has the following properties:*

(i) $\hat{\boldsymbol{\beta}}^*_{P1}$ *exists and is the unique fixed point of the equation* $\boldsymbol{\beta}_1 = \{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1)\}^{-1}\mathbf{v}_n^{(1)}$, *where* $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = \mathrm{diag}\{\beta_1^{-2}, \ldots, \beta_{q_n}^{-2}\}$.

(ii) $\hat{\boldsymbol{\beta}}^*_{P2} = 0$.

(iii) $\sqrt{n}\,(\hat{\boldsymbol{\beta}}^*_{P1} - \boldsymbol{\beta}_{01})$ *converges in distribution to a mean-zero multivariate normal distribution whose covariance matrix can be consistently estimated by* $\{n^{-1}\boldsymbol{\Omega}_n^{(1)}\}^{-1}\,\hat{\Xi}_n^{(1)}\,\{n^{-1}\boldsymbol{\Omega}_n^{(1)}\}^{-1}$, *where* $\boldsymbol{\Omega}_n^{(1)}$ *and* $\hat{\Xi}_n^{(1)}$ *denote the* $q_n \times q_n$ *left-up submatrices of* $\boldsymbol{\Omega}_n$ *and* $\hat{\Xi}_n$, *respectively.*

By the grouping effect, we usually mean that when the true model or covariates have a natural group structure, a selection approach can have all coefficients within a group clustered or selected together. This is clearly a desirable property and one example of this is given by the gene network relationship where some of genes are strongly correlated and are often referred to as grouped genes. To describe the grouping effect of the BAR regression estimator proposed above, first note that based on $\boldsymbol{\Omega}_n = \boldsymbol{X}^\top \boldsymbol{X}$, we have that, for all $j, k \in \{1, \ldots, p_n\}$,

$$\sum_{i=1}^{n} \int_0^\tau \mathbf{1}(C_i \geq t)\{Z_{ij}(t) - \bar{Z}_j(t)\}\{Z_{ik}(t) - \bar{Z}_k(t)\} d\hat{\Lambda}_0(t) = \boldsymbol{x}_j^\top \boldsymbol{x}_k, \tag{3.8}$$

which are the (j, k) elements of the two matrices, where $\boldsymbol{x}_j$ denotes the jth p_n-dimensional column vector of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n})$. This implies that the correlation between the original covariates $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_k$ can be described by that between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ and thus we have the following grouping effect property.

Theorem 4. *Assume that* $\boldsymbol{X}$ *has been standardized. Then with probability tending*

to 1 *as* $n \to \infty$, *the BAR estimator* $\hat{\boldsymbol{\beta}}_P^* = (\hat{\beta}_{P1}^*, \ldots, \hat{\beta}_{Pp_n}^*)^\top$ *satisfies the inequality*

$$|1/\hat{\beta}_{Pi}^* - 1/\hat{\beta}_{Pj}^*| \leq \|\boldsymbol{y}\|\sqrt{2(1-\rho_{ij})}/\lambda_n,$$

where $\hat{\beta}_{Pi}^* \times \hat{\beta}_{Pj}^* \neq 0$ *and* ρ_{ij} *denotes the sample correlation coefficient of* $\boldsymbol{x}_i$ *and* $\boldsymbol{x}_j$.

The proof of the asymptotic properties above is sketched in Appendix A.2. Based on (3.8), the result given in Theorem 4 suggests that if the correlation between z_i and z_j is strong, the estimators $\hat{\beta}_{Pi}^*$ and $\hat{\beta}_{Pj}^*$ will be very close. For the determination of the proposed BAR estimator, we need to choose the tuning parameters ξ_n and λ_n. To reduce the computational burden, we suggest employing the K-fold cross-validation, where K is an integer.

More specifically, let O denote the full dataset and randomly divide O into K equal sub-datasets $O^1, \ldots, O^K$. Also let $\boldsymbol{\theta} = (\xi_n, \lambda_n)$ and $\hat{\boldsymbol{\beta}}_P^{-k}(\boldsymbol{\theta})$ denote the proposed BAR estimator of $\boldsymbol{\beta}$ based on the dataset $O - O^k$ for given $\boldsymbol{\theta}$ and k. Then one can choose $\boldsymbol{\theta}$ that minimizes the cross-validation error given by

$$CV(\boldsymbol{\theta}) = \sum_{k=1}^{K} \ell^k\{\hat{\boldsymbol{\beta}}_P^{-k}(\boldsymbol{\theta})\},$$

where ℓ^k denotes the loss function $\ell(\boldsymbol{\beta})$ based on the sub-dataset O^k.

3.4 Simulation Studies

For the assessment or evaluation of the proposed BAR regression estimation procedure, we conducted simulation studies. In addition to the BAR estimator, we considered and compared the BAR estimator to the estimators given by (3.5) with

replacing the BAR penalty function by the LASSO, ALASSO and SCAD penalty functions. In the tables below, the resulting estimators for the four methods are denoted by BAR, LASSO, ALASSO and SCAD, respectively. Note that although the same penalty functions are used here as in Fan and Li (2001), Fan and Li (2002), and Tibshirani (1996), the estimation procedures here are quite different from that given in these papers as both models and data structures are quite different.

To generate the incomplete event history data, we first generated the observation time points by simulating the follow-up times C_is from the uniform distribution over $(0, 8)$ and assuming that the observation processes H_i^*s are homogeneous Poisson processes with $\Lambda_0(t) = t$. The observed panel count data were then generated by assuming that the $N_i^*(t)$s are either homogeneous or mixed Poisson processes with $\mu_0(t) = t$ given covariates. For the results given below, we generated the covariates $\boldsymbol{Z}_i^*$s from standard normal distributions with correlation between the j_1 and j_2 components given by $\rho^{|j_1-j_2|}$, where $\rho = 0$ or 0.5. The results given below are based on $n = 100$ or 500 with 500 replications.

Table 3.1 presents the results obtained by the four methods, the proposed BAR, LASSO, ALASSO and SCAD, on the estimation and covariate selection with $p_n = 10$, the first $q_n = 3$ components of $\boldsymbol{\beta}_0$ being 1 and the remaining $p_n - q_n$ components being zero. Here the 5-fold cross-validation was used for the determination of the tuning parameters for all four methods. The results include the average of L_1 prediction errors (PE), the average percent of the numbers of the zero components that were correctly identified as zero (Corr), and the average percent of the numbers of the non-zero components that were incorrectly identified as zero (Incr).

One can clearly see from Table 3.1 that the proposed BAR method always has

a higher or significantly higher Corr than the other methods, meaning that it has the least probability to identify non-relevant or non-important factors to be relevant or important. Also the proposed BAR method tends to have lower prediction errors than the other three methods, although the differences may not be significant. On the Incr, it seems that all methods are close to one another, especially when the sample size is large. The same conclusions can be seen from Table 3.2, where $p_n = 30$, the first $q_n = 5$ components of $\boldsymbol{\beta}_0$ being 1 and the remaining $p_n - q_n$ components being zero. We also considered other set-ups and obtained similar conclusions.

To assess and compare the grouping effects of the four methods above, we performed a simulation study as above but with $\boldsymbol{\beta}_0 = (1, 1, 1, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0)$ and $n = 10{,}000$, meaning $p_n = 12$ and $q_n = 6$. Here the large sample size was used since the grouping effect, as an asymptotic property of variable selection methods, can only be revealed with a large sample. For the covariate generation, we assumed that, for $j \in \{1, 2, 3\}$,

$$Z_j = x_1 + \epsilon_j, \quad x_1 \sim \mathcal{N}(0, 50^2),$$

and, for $j \in \{4, 5, 6\}$,

$$Z_j = x_2 + \epsilon_j, \quad x_2 \sim \mathcal{N}(0, 50^2),$$

where the ϵ_js are iid $\mathcal{N}(0, 0.025^2)$. For $j \in \{7, \ldots, 12\}$, $Z_j \sim \mathcal{N}(0, 50^2)$ are independent with each other. This way, the 12 covariates are from three different groups with the within-group correlations being nearly 1 and the between-group correlations being close to 0.

Table 3.3 gives the estimated coefficients and the corresponding optimal tuning parameters chosen again by the 5-fold cross-validation criterion, while Figure 3.1

shows the solution paths of the four estimators. One can see the obvious grouping effect of the proposed BAR estimator from both Table 3.3 and Figure 3.1 when the tuning parameters are in the proper range. In particular, the method clearly selected the six relevant covariates into their two separate groups with almost the same coefficients within each group when $\ln(\lambda_n)$ is roughly between 0.7 and 5.6.

Note that although the estimator paths are unstable when λ_n is small, the cross-validation criterion can successfully find the tuning parameters inducing the grouping effect and the true values for the two group parameters are 0.5 and 1, respectively. Also note that Figure 3.1 indicates that sometimes the BAR estimator paths may be unstable too when λ_n is large than the optimal tuning parameter but will go to zero when λ_n increases. In contrast, all three other methods showed little grouping effects as only β_1 and β_4 were estimated as nonzero or relevant by them.

3.5 An Application to the Skin Cancer Study

In this section, we apply the BAR estimation and variable selection method proposed in the previous sections to a set of panel count data arising from a skin cancer trial conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Sun and Zhao, 2013; Zhang et al., 2013). The study is a double-blinded and placebo-controlled randomized Phase III clinical trial with the primary goal being to evaluate the effectiveness of 0.5 g/m2/day PO difluoromethylornithine (DFMO) in reducing the recurrence rate of skin cancers in a population of the patients with a history of non-melanoma skin cancers: basal cell carcinoma and squamous cell carcinoma. The data set consists of 290 skin cancer patients who were randomized

into either the placebo (147) or DFMO (143) group. For each patient, in addition to the observed numbers of occurrences of two types of skin cancers and the treatment indicator, the data also include the information on three baseline covariates, gender, age at the diagnosis and the number of prior skin cancers.

For the analysis, define $Z_{i1} = 1$ if patient i was in the DFMO group and 0 otherwise, Z_{i2} and Z_{i3} to represent the number of prior skin cancers and the age of the patient, respectively, and $Z_{i4} = 1$ if patient i is male and 0 otherwise. Table 4 presents the estimation and variable selection results given by the four methods discussed in the previous section on the squamous cell carcinoma based on the 10-fold cross-validation and grid search for the tuning parameter selection. In addition to the estimated effects for the selected covariates, we also obtained and included in the table the estimated standard errors (SD) given in Theorem 3.

One can see that the results are quite similar for the proposed BAR, LASSO and SCAD methods, which selected two covariates. In contrast, the ALASSO only selected one covariate. All four methods indicated that the DFMO treatment had no effect on reducing the occurrence rate of new skin cancers and the occurrence rate did not seem to be significantly related to the gender of the patient. Instead, the occurrence rate seems to be positively or negatively related to the number of prior skin cancers or the age, respectively. These results are consistent with those obtained by Zhang et al. (2013) based on the multiplicative model in general but the latter indicated that the occurrence rate and the age did not seem to be related.

3.6 Discussion and Concluding Remarks

In this chapter, we have discussed simultaneous estimation and variable selection for panel count data and proposed a BAR regression approach. In the method, instead of using the commonly used penalty functions such as LASSO, ALASSO and SCAD, we discussed a new penalty function that iteratively approximates the L_0 penalty function by using L_2 type of penalty functions. One main advantage of the proposed method is that unlike the methods given in Tong et al. (2009a) and Zhang et al. (2013) or based on the commonly used penalty functions, the resulting estimator has both the oracle property and clustering effect. Also the numerical studies suggested that it performs well in practical situations.

Table 3.1: The simulation results with $q_n = 3$, $p_n = 10$ and the numbers in the parentheses denoting the sample standard deviations.

Poisson	n	ρ	Methods	PE		Corr	Incr
Homogeneous	100	0	BAR	1.044	(0.635)	0.845	0.022
			LASSO	1.508	(0.566)	0.406	0.019
			ALASSO	1.129	(0.549)	0.741	0.007
			SCAD	1.240	(0.681)	0.496	0.017
		0.5	BAR	1.469	(0.893)	0.835	0.097
			LASSO	1.578	(0.714)	0.560	0.017
			ALASSO	1.471	(0.741)	0.750	0.045
			SCAD	1.818	(0.829)	0.548	0.023
	500	0	BAR	0.390	(0.240)	0.930	0.000
			LASSO	0.688	(0.247)	0.396	0.000
			ALASSO	0.454	(0.232)	0.815	0.000
			SCAD	0.373	(0.220)	0.818	0.000
		0.5	BAR	0.496	(0.316)	0.922	0.000
			LASSO	0.739	(0.325)	0.536	0.000
			ALASSO	0.591	(0.315)	0.793	0.000
			SCAD	0.527	(0.323)	0.702	0.000
Mixed	100	0	BAR	2.353	(1.140)	0.739	0.259
			LASSO	2.609	(0.859)	0.573	0.282
			ALASSO	2.209	(0.884)	0.717	0.193
			SCAD	2.689	(0.906)	0.592	0.269
		0.5	BAR	2.918	(1.466)	0.807	0.398
			LASSO	2.668	(1.072)	0.668	0.269
			ALASSO	2.766	(1.308)	0.746	0.288
			SCAD	2.763	(1.246)	0.677	0.257
	500	0	BAR	0.811	(0.469)	0.884	0.001
			LASSO	1.201	(0.430)	0.423	0.000
			ALASSO	0.893	(0.422)	0.774	0.000
			SCAD	0.856	(0.441)	0.580	0.000
		0.5	BAR	1.209	(0.751)	0.836	0.037
			LASSO	1.366	(0.598)	0.561	0.003
			ALASSO	1.223	(0.609)	0.759	0.016
			SCAD	1.463	(0.776)	0.523	0.006

Table 3.2: The simulation results with $q_n = 5$, $p_n = 30$ and the numbers in the parentheses representing the sample standard deviations.

Poisson	n	ρ	Methods	PE		Corr	Incr
Homogeneous	100	0	BAR	2.781	(1.555)	0.905	0.172
			LASSO	4.053	(1.077)	0.602	0.157
			ALASSO	3.090	(1.269)	0.725	0.033
			SCAD	3.665	(1.319)	0.645	0.148
		0.5	BAR	3.829	(1.912)	0.928	0.329
			LASSO	3.812	(1.479)	0.720	0.107
			ALASSO	4.606	(2.126)	0.734	0.148
			SCAD	4.173	(1.413)	0.754	0.122
	500	0	BAR	0.733	(0.355)	0.979	0.000
			LASSO	1.729	(0.478)	0.537	0.000
			ALASSO	1.051	(0.433)	0.821	0.000
			SCAD	0.759	(0.308)	0.827	0.000
		0.5	BAR	1.067	(0.584)	0.970	0.003
			LASSO	1.676	(0.564)	0.697	0.000
			ALASSO	1.473	(0.678)	0.833	0.000
			SCAD	1.462	(0.615)	0.675	0.000
Mixed	100	0	BAR	5.103	(2.037)	0.923	0.663
			LASSO	5.249	(1.292)	0.847	0.648
			ALASSO	5.326	(2.128)	0.790	0.386
			SCAD	5.255	(1.199)	0.842	0.632
		0.5	BAR	5.912	(3.096)	0.945	0.685
			LASSO	5.246	(1.699)	0.833	0.516
			ALASSO	7.357	(3.937)	0.768	0.462
			SCAD	5.018	(1.805)	0.831	0.481
	500	0	BAR	1.679	(0.953)	0.942	0.020
			LASSO	3.102	(0.842)	0.562	0.014
			ALASSO	2.090	(0.836)	0.796	0.002
			SCAD	2.243	(0.946)	0.577	0.013
		0.5	BAR	2.925	(1.440)	0.939	0.197
			LASSO	3.036	(0.982)	0.723	0.020
			ALASSO	3.217	(1.319)	0.795	0.062
			SCAD	3.711	(1.289)	0.715	0.034

Table 3.3: Estimated coefficients by the four methods and the corresponding optimal tuning parameters.

Parameters	BAR	LASSO	ALASSO	SCAD
β_1	0.990	2.952	2.969	2.973
β_2	0.990	0	3.971e-09	0
β_3	0.990	0	0	0
β_4	0.494	1.471	1.482	1.485
β_5	0.494	0	0	0
β_6	0.494	0	0	0
β_7	0	−0.002	0	0
β_8	0	−0.014	0	0
β_9	0	0.014	0	0
β_{10}	0	0	0	0
β_{11}	0	−0.009	0	0
β_{12}	0	−0.015	0	0
Tuning parameters	$\lambda_n = 5.777$ $\xi_n = 0.001778$	$\lambda_n = 56.049$	$\lambda_n = 8.555$	$\lambda_n = 1368.746$

Table 3.4: The estimation and variable selection results for the skin cancer data on squamous cell carcinoma with SD in the parentheses representing the estimated standard errors.

Method	$\hat{\beta}_1$	$\hat{\beta}_2$ (SD)	$\hat{\beta}_3$ (SD)	$\hat{\beta}_4$
BAR	0 (—)	1.431e-04 (6.306e-05)	1.609e-05 (8.525e-06)	0 (—)
LASSO	0 (—)	1.410e-04 (6.294e-05)	1.737e-05 (8.521e-06)	0 (—)
ALASSO	0 (—)	4.617e-05 (7.157e-05)	0 (—)	0 (—)
SCAD	0 (—)	1.433e-04 (6.289e-05)	1.782e-05 (8.552e-06)	0 (—)

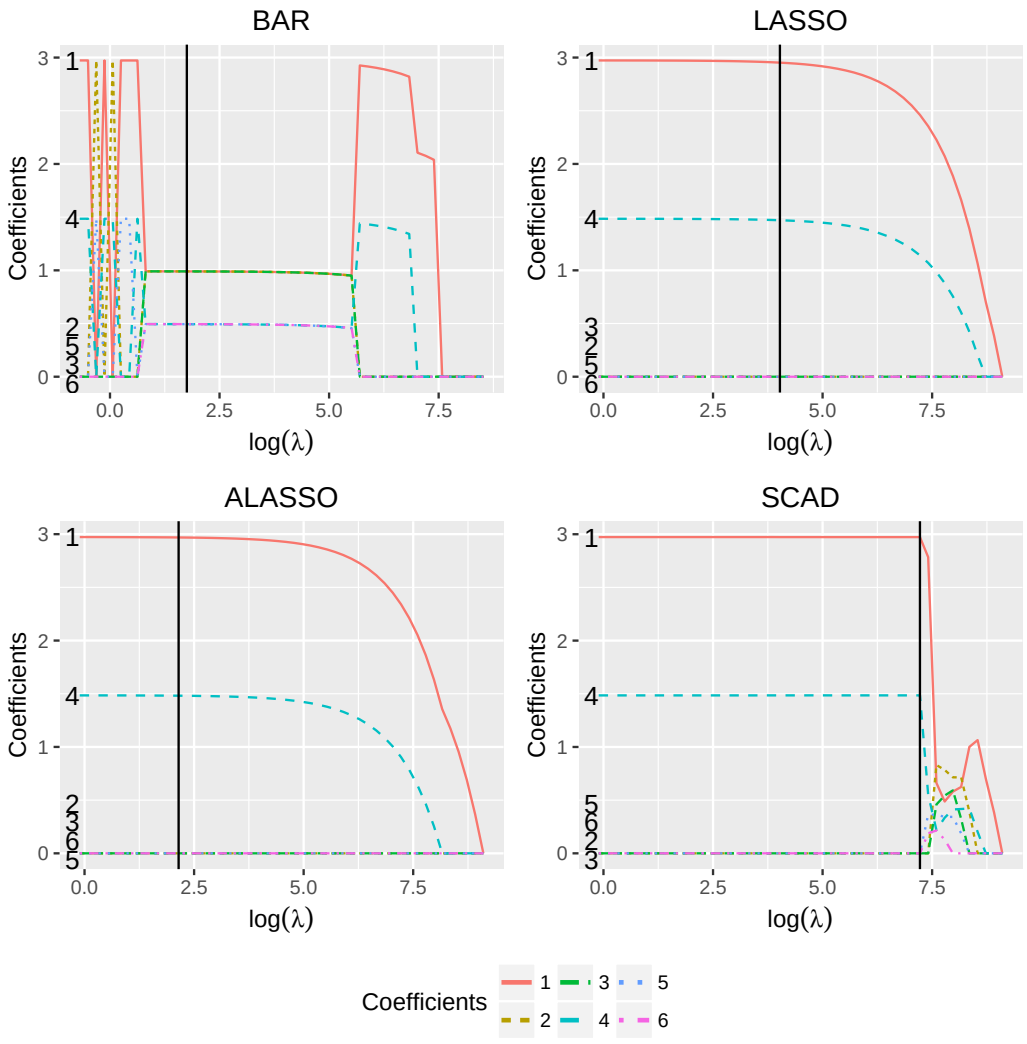

Figure 3.1: Path plots of the four estimators with the black solid vertical lines corresponding to the optimal tuning parameters based on the 5-fold cross-validation.

Chapter 4

Regression Analysis of Asynchronous Longitudinal Data with Informative Observation Processes

4.1 Introduction

As introduced in Section 1.3, sparse asynchronous longitudinal data with informative observation times are not uncommon in medical studies but, to our best knowledge, no previous studies addressed this issue. This chapter will describe a kernel-based generalized estimating equation method that can deal with asynchronicity and informative observation times simultaneously. Since panel count data are also a type of longitudinal data, the proposed method can be naturally applied to panel count data too.

In Section 4.2, we will begin with introducing some notation, assumptions and

models that will be used throughout the chapter. In particular, we will present a class of flexible semiparametric transformation models for the longitudinal process of interest. Then we will first discuss the simple situation where one observes synchronous longitudinal data. In Section 4.3, the estimating equation-based approach described in Section 4.2 for the synchronous case will be generalized to asynchronous situations with the use of the kernel weighting technique similar to that used in Cao et al. (2015b). The asymptotic distribution of the proposed estimators will be established in Section 4.4 and also in this section, the bandwidth selection for kernel weighting is discussed. Results obtained from an extensive simulation study are presented in Section 4.5 and indicate that the proposed method works well in practical situations. In Section 4.6, we apply the proposed approach to the HIV longitudinal data described in Section 1.3 and Section 4.7 contains some discussion and concluding remarks.

4.2 Estimation with Synchronous Longitudinal Data

Consider a longitudinal study that consists of n independent subjects. Let $Y(t)$ denote the longitudinal response variable of interest and suppose that there exists a $p-$dimensional vector of covariates, denoted by $Z(t)$, that are external and may depend on time t. Also suppose that for subject i, there exist two sequences of observation times $T_{i1} < T_{i2} < \cdots < T_{i,J_i}$ and $R_{i1} < R_{i2} < \cdots < R_{i,K_i}$ for the observation of $Y_i(t)$ and $Z_i(t)$, respectively, and a follow-up or censoring time denoted by C_i. Here J_i and K_i denote the potential numbers of the observations on $Y_i(t)$ and $Z_i(t)$, respectively. Let $N_i^*(t)$ and $O_i^*(s)$ be the counting processes describing the observation

processes on $Y_i(t)$ and $Z_i(t)$, respectively, $i = 1, ..., n$. Then it is easy to see that the actual observations processes for them are $N_i(t) = N_i^*(\min(t, C_i)) = \sum_{j=1}^{J_i} I(T_{ij} \le t)$ and $O_i(s) = O_i^*(\min(s, C_i)) = \sum_{j=1}^{K_i} I(R_{ij} \le s)$, respectively. In other words, $Y_i(t)$ and $Z_i(t)$ are observed only at the jump time points of $N_i(t)$ and $O_i(t)$, respectively. It is apparent that one observes synchronous longitudinal data if $N_i(t) = O_i(t)$ for all i, and the observed data have the form $\left\{ Y_i\left(T_{ij}\right)' s, Z_i\left(R_{ik}\right)' s, C_i\,,\ i = 1, \ldots, n \right\}$.

To present the regression model, define $\mathcal{F}_{it} = \{N_i(s),\, 0 \le s < t\}$, the history or filtration of the observation process $N_i\left(t\right)$ up to time t^-. In the following, we will assume that given $Z_i(t)$ and $\mathcal{F}_{it}$, the conditional mean function of $Y_i(t)$ has the form

$$E\{Y_i(t)|Z_i(t), \mathcal{F}_{it}\} = g\{\mu(t)e^{\beta' Z_i(t) + \alpha' H(\mathcal{F}_{it})}\}\,. \tag{4.1}$$

Here $g(\cdot)$ is a known twice continuously differentiable function, $\mu(t)$ is an unspecified smooth function of t, α and β are vectors of unknown parameters, and $H(\cdot)$ is a vector of known functions of $\mathcal{F}_t$. In the above, the function $g(\cdot)$ can take many forms to account for various types of dependence of $Y_i(t)$ on $Z_i(t)$ and $\mathcal{F}_{it}$, and two simple choices are $g(x) = x$ and $g(x) = \log x$. Especially, the latter yields the additive mean model

$$E\{\, Y_i(t)|Z_i(t), \mathcal{F}_{it}\,\} = \log\{\mu(t)\} + \beta' Z_i(t) + \alpha' H(\mathcal{F}_{it})$$

discussed in Sun et al. (2005). A more general choice is the so-called Box-Cox transformation $g(x) = \{(x+1)^a - 1\}/a$ with $g(x) = \log(x+1)$ for $a = 0$, where a is a constant. For $H(\cdot)$, also various forms can be taken and one is $H(\mathcal{F}_{it}) = N_i(t - t_0)$, meaning that $Y_i(t)$ may depend on the number of the observations during the period from $t - t_0$ to the current time, where t_0 is a constant. It is apparent that the model

above is a conditional model and alternatively one could consider the marginal or joint model approach. More comments on this are given in Section 4.7.

In reality, of course, the observation process $N_i^*(t)$ may depend on covariates too. In the following, it will be assumed that it is a nonhomogeneous Poisson process satisfying the proportional rate model

$$E\{dN_i^*(t)|Z_i(t)\} = e^{\gamma' Z_i(t)} d\Lambda(t)\,, \tag{4.2}$$

where γ is a vector of unknown parameters and $\Lambda(\cdot)$ is an unspecified baseline mean function. Furthermore, we will assume that $E\{\, dO_i^*(t)|Z_i(t)\,\} = \nu\,(t)\,dt$, meaning that the observation process for covariates does not depend on $\{N_i(t), Y_i(t), C_i\}$ given $Z_i\,(t)$, where $\nu(t)$ is an unspecific baseline rate function. More comments on these will be given in Section 4.7. Also it will be assumed that conditional on $Z_i(t)$, C_i is independent of $\{N_i(t), O_i(t), Y_i(t)\}$ or we have independent and non-informative censoring times.

In the remaining of this section, it will be assumed that one observes synchronous longitudinal data or we have $J_i = K_i$ and $N_i(t) = O_i(t)$ for all i. To present the estimation procedure, let $\Lambda_0\,(\cdot)$, $\mu_0\,(\cdot)$, $\nu_0\,(\cdot)$, β_0, α_0 and γ_0 denote the true values of $\Lambda\,(\cdot)$, $\mu\,(\cdot)$, $\nu\,(\cdot)$, β, α and γ, respectively, and define $\theta = (\alpha', \beta')'$ and $\theta_0 = (\alpha_0', \beta_0')'$ for convenience. Also define $X_i(t,s) = (H(\mathcal{F}_{it})', Z_i(s)')'$ and $\Delta_i(t) = I(C_i > t)$, $i = 1, ..., n$. By following Li et al. (2013), we will first consider the estimation of γ and $\Lambda(t)$. More specifically, for the estimation of γ, we can consider the estimating equations

$$\tilde{U}_\gamma(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \Delta_i(t) \left\{ Z_i(t) - \tilde{Z}(t;\gamma) \right\} dN_i(t) = 0\,. \tag{4.3}$$

In the above, τ is the longest follow-up time and $\tilde{Z}(t;\gamma) = \tilde{S}^{(1)}(t;\gamma)/\tilde{S}^{(0)}(t;\gamma)$, where

$$\tilde{S}^{(k)}(t;\gamma) = n^{-1}\sum_{i=1}^{n}\Delta_i(t)Z_i(t)^{\otimes k}e^{\gamma' Z_i(t)},\ k = 0, 1, 2,$$

with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$ for any vector a. Let $\tilde{\gamma}$ denote the solution to the estimating equation (4.3). Then $\Lambda(t)$ can be estimated by

$$\tilde{\Lambda}(t;\tilde{\gamma}) = \sum_{i=1}^{n}\int_0^t \frac{\Delta_i(u)dN_i(u)}{n\tilde{S}^{(0)}(u;\tilde{\gamma})} = \sum_{i=1}^{n}\sum_{j:T_{ij}\le t}\frac{1}{n\tilde{S}^{(0)}(T_{ij};\tilde{\gamma})}.$$

For estimating θ and $\mu(t)$, define

$$d\tilde{M}_i(t;\theta,\gamma,\Lambda,\mu) = \Delta_i(t)\Big[Y_i(t)dN_i(t) - g\{\mu(t)e^{\theta' X_i(t,t)}\}e^{\gamma' Z_i(t)}d\Lambda(t)\Big], \tag{4.4}$$

and

$$\tilde{U}_\theta(\theta;\gamma,\Lambda,\mu) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau X_i(t)d\tilde{M}_i(t;\theta,\gamma,\Lambda,\mu). \tag{4.5}$$

Then under models (4.1) and (4.2) and the assumptions above, one can easily show that $E\{\tilde{M}_i(t;\theta_0,\gamma_0,\Lambda_0)\} = 0$. This suggests that one can estimate θ and $\mu(t)$ based on the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} d\tilde{M}_i(t;\theta,\gamma,\Lambda,\mu(\cdot)) = 0 \quad \text{and} \quad \tilde{U}_\theta(\theta;\gamma,\Lambda,\mu(\cdot)) = 0$$

with replacing γ and $\Lambda(t)$ by $\tilde{\gamma}$ and $\tilde{\Lambda}(t,\tilde{\gamma})$, respectively. Furthermore, by following the arguments similar to those used in Lin et al. (2001), Li et al. (2010) and Sun et al. (2005), one can show that for large n, the estimators defined above always exist and are unique and consistent. In the next section, we will generalize the estimation

approach above to the case of sparse asynchronous longitudinal data.

4.3 Estimation with Asynchronous Longitudinal Data

Now we discuss estimation of models (4.1) and (4.2) or unknown parameters for sparse asynchronous longitudinal data. For this, it is easy to see that the estimating equations or the estimators given in the previous section are not available or cannot be used anymore. To generalize them, we will employ the kernel weighting technique.

First as before, we will consider the estimation of model (4.2), that is, the estimation of γ and $\Lambda(t)$. For this, based on the estimating function $\tilde{U}_\gamma(\gamma)$ and by following Cao et al. (2015a), we can consider the estimating function

$$U_\gamma(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i} K_h(T_{ij} - R_{ik})\{Z_i(R_{ik}) - \bar{Z}(T_{ij};\gamma)\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\{Z_i(r) - \bar{Z}(t;\gamma)\}\Delta_i(t) dO_i(r)\, dN_i(t)\,,$$

where $K_h(t) = K(t/h)/h$ is a symmetric kernel function with the bandwidth h, and $\bar{Z}(t;\gamma) = S_n^{(1)}(t;\gamma)/S_n^{(0)}(t;\gamma)$ with

$$S_n^{(l)}(t;\gamma) = n^{-1}\sum_{i=1}^{n}\int_0^\infty K_h(t-r)\Delta_i(t)Z_i(r)^{\otimes l}\exp\{\gamma' Z_i(r)\}\, dO_i(r)\,,\ l = 0,1,2.$$

Let $\hat{\gamma}$ denote the estimator of γ given by the solution to $U_\gamma(\gamma) = 0$. Then a natural

estimator of $\Lambda(t)$ is given by

$$\hat{\Lambda}(t;\hat{\gamma}) = \sum_{i=1}^{n} \int_0^t \frac{\Delta_i(t) \sum_{k=1}^{K_i} K_h\left(t - R_{ik}\right)}{nS^{(0)}(t;\hat{\gamma})} dN_i\left(t\right) .$$

For estimation of θ and μ, based on the estimating functions given in (4.4) and (4.5) and similarly as above, we can consider the following kernel weighted estimating equations

$$U_\theta(\theta;\hat{\gamma},\mu\left(\cdot\right)) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\infty \int_0^\tau K_h\left(t-r\right)\Delta_i(t) X_i\left(t,r\right)$$

$$\times \left[Y_i\left(t\right) dN_i\left(t\right) - g\left\{\mu\left(t\right) e^{\theta' X_i(t,r)}\right\} e^{\hat{\gamma}' Z_i(r)} d\hat{\Lambda}\left(t\right)\right] dO_i\left(r\right) = 0\,, \tag{4.6}$$

and

$$U_\mu\left(\mu\left(\cdot\right);\hat{\gamma}\right) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\infty K_h\left(t-r\right)\Delta_i(t)$$

$$\times \left[Y_i\left(t\right) dN_i\left(t\right) - g\left\{\mu\left(t\right) e^{\theta' X_i(t,r)}\right\} e^{\hat{\gamma}' Z_i(r)} d\hat{\Lambda}\left(t\right)\right] dO_i\left(r\right) = 0\,. \tag{4.7}$$

It is easy to see that the key difference between the estimating equations (4.4)-(4.5) and (4.6)-(4.7) is that $K_h(T_{ij} - R_{ik})$ is added as the weight for each pair of T_{ij} and R_{ik}. If T_{ij} and R_{ik} are close to each other, then the weight $K_h(T_{ij} - R_{ik})$ is close to 1, while if they are distant from each other, $K_h(T_{ij} - R_{ik})$ will be nearly or exactly 0.

Let $\hat{\theta}$ and $\hat{\mu}$ denote the respective estimators of θ and μ given by the solutions to the estimating equations (4.6) and (4.7), and $t_1 < t_2 < \ldots < t_m$ the distinct ordered observation times of $\{T_{ij}, i = 1,\ldots,n; j = 1,\ldots,J_i\}$. It is apparent that the function $\mu(t)$ can be estimated only at the time points t_l's. More specifically, after some algebra and for given θ, the estimated value of $\mu(t)$ at t_l based on equation

(4.7), denoted by $\hat{\mu}(t_l;\theta,\hat{\gamma})$, can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i} I\left(t_l = T_{ij}\right) K_h\left(t_l - R_{ik}\right)\Delta_i(t_l)\left[Y_i\left(t_l\right) - \bar{g}\left(t_l;\mu(t_l),\theta,\hat{\gamma}\right)\right] = 0\,, \qquad (4.8)$$

$l = 1,\ldots,m$, where

$$\bar{g}\left(t;\mu\left(\cdot\right),\theta,\gamma\right) = \frac{\sum_{i=1}^{n}\int_0^\tau K_h\left(t-r\right)\Delta_i(t) g\left\{\mu\left(t\right)\exp\left(\theta' X_i\left(t,r\right)\right)\right\}\exp\left(\gamma' Z_i\left(r\right)\right) dO_i\left(r\right)}{\sum_{i=1}^{n}\int_0^\tau K_h\left(t-r\right)\Delta_i(t)\exp\left(\gamma' Z_i\left(r\right)\right) dO_i\left(r\right)}\,.$$

Also the estimating equation (4.6) can be rewritten as

$$U_\theta(\theta;\hat{\gamma},\mu\left(\cdot\right)) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h\left(t-r\right)\Delta_i(t)\left[X_i\left(t,r\right)Y_i\left(t\right)\right.$$

$$\left. - \frac{\sum_{j=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_j(t) X_j\left(t,s\right) g\left\{\mu\left(t\right)\exp\left(\theta' X_j\left(t,s\right)\right)\right\}\exp\left(\hat{\gamma}' Z_j\left(s\right)\right) dO_j\left(s\right)}{\sum_{j=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_j(t)\exp\left(\hat{\gamma}' Z_j\left(s\right)\right) dO_j\left(s\right)}\right]$$

$$\times\, dO_i\left(r\right) dN_i\left(t\right) = 0\,. \qquad (4.9)$$

It is easy to see that in general, there are no closed-forms for $\hat{\theta}$ and $\hat{\mu}\left(\cdot;\hat{\theta},\hat{\gamma}\right)$ and some iterative algorithms need to be used for their determination. However, for some special cases, the estimators $\hat{\theta}$ and $\hat{\mu}\left(\cdot;\hat{\theta},\hat{\gamma}\right)$ can have closed-forms. One such situation is $g(x) = \log(x)$, under which model (4.1) would reduce to the linear model discussed in Sun et al. (2005). For the case, one can easily show that

$$\hat{\mu}\left(t;\theta,\hat{\gamma}\right) = \exp\left\{\frac{\sum_{i=1}^{n}\Delta_i(t)Y_i\left(t\right)dN_i\left(t\right)}{\sum_{i=1}^{n}\Delta_i(t)dN_i\left(t\right)} - \theta'\bar{X}\left(t;\hat{\gamma}\right)\right\}\,, \qquad (4.10)$$

where

$$\bar{X}\left(t;\hat{\gamma}\right)=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K_i}K_h\left(t-R_{ik}\right)\Delta_i(t)X_i\left(t,R_{ik}\right)\exp\left\{\hat{\gamma}'Z_i\left(R_{ik}\right)\right\}}{\sum_{i=1}^{n}\sum_{k=1}^{K_i}K_h\left(t-R_{ik}\right)\Delta_i(t)\exp\left\{\hat{\gamma}'Z_i\left(R_{ik}\right)\right\}}.$$

By plugging (4.10) into (4.9), one can obtain

$$U_\theta\left(\theta;\hat{\gamma}\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})Y_i\left(T_{ij}\right)\left[X_i\left(T_{ij},R_{ik}\right)-\bar{X}\left(T_{ij};\hat{\gamma}\right)\right]$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})\left\{\overline{X\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}}-\bar{X}\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}\right\}\theta\,,$$

where

$$\overline{X\left(t;\hat{\gamma}\right)^{\otimes 2}}=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K_i}K_h\left(t-R_{ik}\right)I\left(C_i\geq t\right)X_i\left(t,R_{ik}\right)^{\otimes 2}\exp\left(\hat{\gamma}'Z_i\left(R_{ik}\right)\right)}{\sum_{i=1}^{n}\sum_{k=1}^{K_i}K_h\left(t-R_{ik}\right)I\left(C_i\geq t\right)\exp\left(\hat{\gamma}'Z_i\left(R_{ik}\right)\right)}\,.$$

This yields

$$\hat{\theta}=\left[\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})\left\{\overline{X\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}}-\bar{X}\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}\right\}\right]^{-1}$$

$$\times\left[\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})Y_i\left(T_{ij}\right)\left\{X_i\left(T_{ij},R_{ik}\right)-\bar{X}\left(T_{ij};\hat{\gamma}\right)\right\}\right].$$

Another case where the proposed estimator $\hat{\mu}$ has a closed-form is $g(x)=x$, corresponding to the proportional mean model. For the situation, we have

$$\hat{\mu}\left(t;\theta,\hat{\gamma}\right)=\frac{\sum_{i=1}^{n}\Delta_i(t)Y_i\left(t\right)dN_i\left(t\right)}{\sum_{i=1}^{n}\Delta_i(t)dN_i\left(t\right)}$$

$$\times \frac{\sum_{i=1}^{n} \int_0^\tau K_h(t-r)\,\Delta_j(t)\exp\left(\hat{\gamma}' Z_j(r)\right) dO_j(r)}{\sum_{j=1}^{n} \int_0^\tau K_h(t-r)\,\Delta_j(t)\exp\left(\theta' X_j(t,r) + \hat{\gamma}' Z_j(r)\right) dO_j(r)} . \tag{4.11}$$

By plugging (4.11) into (4.9), the estimating equation (4.6) or (4.9) becomes

$$\tilde{U}_\theta(\theta;\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \int_0^\infty K_h(t-r)\,\Delta_i(t) Y_i(t) \left[X_i(t,r)\right.$$

$$\left. - \frac{\sum_{j=1}^{n} \int_0^\tau K_h(t-u)\,\Delta_j(t) X_j(t,u)\exp\left(\theta' X_j(t,u) + \hat{\gamma}' Z_j(u)\right) dO_j(u)}{\sum_{j=1}^{n} \int_0^\tau K_h(t-u)\,\Delta_j(t)\exp\left(\theta' X_j(t,u) + \hat{\gamma}' Z_j(u)\right) dO_j(u)}\right] dO_i(r)\, dN_i(t) .$$

In the next section, we will establish the asymptotic properties of the estimators proposed above.

4.4 Asymptotic Properties and Bandwidth Selection

In this section, we will derive the asymptotic distribution of $\hat{\theta}$. For $k = 0, 1, 2$, define

$$Q_n^{(k)}(t;\theta,\gamma) = \frac{1}{n}\sum_{l=1}^{n} \int_0^\tau K_h(t-s)\,\Delta_l(t)\dot{g}\{\hat{\mu}(t;\theta,\gamma)\exp\left(\theta' X_l(t,s)\right)\}$$

$$\times X_l^{\otimes k}(t,s)\exp\left(\theta' X_l(t,s) + \gamma' Z_l(s)\right) dO_l(s) ,$$

and suppose that $S_n^{(k)}(t;\gamma)$ and $Q_n^{(k)}(t;\theta,\gamma)$ converge with the limits

$$s^{(k)}(\gamma,t) = E\left[\Delta(t)\, Z(t)^{\otimes k}\exp\left\{\gamma^T Z(t)\right\}\right]\nu_0(t)$$

and $q^{(k)}(t;\theta,\gamma) = E\left[Q_n^{(k)}(t;\theta,\gamma)\right]$, respectively. For a function $g(\cdot)$, let $\dot{g}(\cdot)$ denote its first order derivative and define $\tilde{x}(t;\theta,\gamma) = q^{(1)}(t;\theta,\gamma)/q^{(0)}(t;\theta,\gamma)$,

$$A(\gamma_0) = -\int_0^\tau \left\{ s^{(2)}(t;\gamma_0) - \frac{s^{(1)}(t;\gamma_0)^{\otimes 2}}{s^{(0)}(t;\gamma_0)} \right\} d\Lambda_0(t) ,$$

$$\Sigma_\gamma(\gamma_0) = \int_{-\infty}^{\infty} K(z)^2 dz \int_0^\tau \left\{ s^{(2)}(t;\gamma_0) - \frac{s^{(1)}(t;\gamma_0)^{\otimes 2}}{s^{(0)}(t;\gamma_0)} \right\} d\Lambda_0(t) ,$$

$$V(t,\theta,\gamma) = E\left[\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta' X(t,t))\}X(t,t)\right.$$

$$\left.\times \{X(t,t) - \tilde{x}(t;\theta,\gamma)\}' \exp(\theta' X(t,t) + \gamma' Z(t))\right] \nu_0(t) ,$$

$$B(\theta_0,\gamma_0) = E\int_0^\tau \Delta(t)\frac{V(t,\theta_0,\gamma_0)}{s^{(0)}(t;\gamma_0)}\mu_0(t)\exp(\gamma_0' Z(t))\, d\Lambda_0(t) ,$$

$$D(\theta_0,\gamma_0) = -E\int_0^\tau \Delta(t)\frac{\nu_0(t)}{s_n^{(0)}(t;\gamma)}\left\{X(t,t) - \frac{q^{(1)}(t,\theta_0,\gamma_0)}{q^{(0)}(t,\theta_0,\gamma_0)}\right\} g\{\mu_0(t)\exp(\theta_0' X(t,t))\}$$

$$\times \left\{ Z(t) - \frac{s^{(1)}(t;\gamma_0)}{s^{(0)}(t;\gamma_0)} \right\} \exp(2\gamma_0' Z(t))\, d\Lambda_0(t) ,$$

$$I(t,s) = X(t,s)Y(t) - \frac{Q_n^{(1)}(t;\theta_0,\gamma_0)}{S_n^{(0)}(t,\gamma_0)} + D(\theta_0,\gamma_0)A^{-1}(\gamma_0)\left\{ Z(s) - \frac{S_n^{(1)}(t,\gamma_0)}{S_n^{(0)}(t,\gamma_0)} \right\} \exp(\gamma_0' Z(t)) ,$$

$$\Sigma_\theta(\theta_0) = \int_z K^2(z)\, dz \int_0^\tau \nu_0(t) E\left[\Delta(t)\left[X(t,t)Y(t) - \frac{q^{(1)}(t;\theta_0,\gamma_0)}{s^{(0)}(t,\gamma_0)}\right.\right.$$

$$\left.\left. + D(\theta_0,\gamma_0)A^{-1}(\gamma_0)\left\{ Z(t) - \frac{s^{(1)}(t,\gamma_0)}{s^{(0)}(t,\gamma_0)} \right\}\right] \exp(\gamma_0' Z(t))\right] d\Lambda_0(t) ,$$

$$G_1(t,s) = E\left[\dot{g}\{\mu_0(t)\exp(\theta_0' X(t,s))\}X(t,s)\right.$$

$$\left.\times \{X(t,s) - \tilde{x}(t;\theta_0,\gamma_0)\}' \exp(\theta_0' X(t,s) + \gamma_0' Z(s))\lambda(t)\nu_0(s)\right] ,$$

$$G_2(t,s) = E\left[\left\{ X(t,s) - \frac{Q_n^{(1)}(t,\theta_0,\gamma_0)}{Q_n^{(0)}(t,\theta_0,\gamma_0)} \right\} g\left\{\hat{\mu}(t;\theta_0,\gamma_0)\exp(\theta_0' X(t,s))\right\}\right.$$

$$\times \left\{ Z(s) - \frac{S_n^{(1)}(t;\gamma_0)}{S_n^{(0)}(t;\gamma_0)} \right\} \exp\left(\gamma_0' Z(t)\right) \lambda_0(t) \nu_0(s) \Bigg] ,$$

and

$$G_3(t_1, t_2, s_1, s_2) = E[\Delta(t_1)\Delta(t_2) I(t_1, s_1) I'(t_2, s_2) \nu_0(s_1) \lambda_0(t_1) \nu_0(s_2) \lambda_0(t_2)] .$$

For the asymptotic properties, we need the following regularity conditions.

(E1) $\Pr(C \geq \tau) > 0$, where τ is a predetermined constant.

(E2) $N(\tau)$ and $O(\tau)$ are bounded by finite constants, and $\lambda_0(t)$ and $\nu_0(t)$ are twice continuously differentiable for $t \in [0, \tau]$.

(E3) For $i = 1 \ldots n$, the $Z_i(t)$'s have bounded variation. In addition, $E[Z(s)\Delta(t) \exp\{\beta_0' Z(t)\}]$ is twice continuously differentiable for $s, t \in [0, \tau]^{\otimes 2}$.

(E4) If there is a vector θ such that $\theta' X(t, s) = 0$, then $\theta = 0$ for any t and s.

(E5) The functions $g(\cdot)$ and $\mu_0(\cdot)$ are twice continuously differentiable for $t \in [0, \tau]$, and $H(\cdot)$ is twice continuously differentiable too. Furthermore, $G_1(t, s)$ and $G_2(t, s)$ are bounded for $(t, s) \in [0, \tau]^{\otimes 2}$ and twice continuously differentiable for $s \in [0, \tau]$, and $G_3(t_1, t_2, s_1, s_2)$ is bounded for $(t_1, t_2, s_1, s_2) \in [0, \tau]^{\otimes 4}$ and continuously differentiable for $(s_1, s_2) \in [0, \tau]^{\otimes 2}$. Also

$$E\left[\int_0^\tau \int_0^\infty \Delta(t) \left[X(t, s) Y(s) - \frac{Q_n^{(1)}(t;\theta_0,\gamma_0)}{S_n^{(1)}(t;\gamma_0)} \right] \exp\left(\gamma_0' Z(t)\right) \lambda_0(t) \nu_0(s) \right] < \infty ,$$

and

$$E\left[\int_0^\tau \Delta(t) \frac{V(t,\theta_0,\gamma_0)}{s^{(0)}(t;\gamma_0)} \mu_0(t) \exp\left(\gamma_0' Z(t)\right) d\Lambda_0(t) \right] < \infty .$$

(E6) The function $K(z)$ is a symmetric density function satisfying $\int_{-\infty}^{\infty} K(z)\,dz = 1$, $\int_{-\infty}^{\infty} zK(z)\,dz = 0$, $\int_{-\infty}^{\infty} z^2K(z)\,dz < \infty$, and $\int_{-\infty}^{\infty} K(z)^2\,dz < \infty$.

(E7) For the bandwidth, we have that $h \to 0$, $nh \to \infty$ and $nh^5 \to 0$ as $n \to \infty$.

Before describing the asymptotic distribution of $\hat{\theta}$, note that for the estimator $\hat{\gamma}$, Cao et al. (2015a) showed that under the Conditions (E1)-(E3) and (E6)-(E7), we have that

$$(nh)^{1/2} A(\gamma_0)(\hat{\gamma} - \gamma_0) \to_d N(0, \Sigma_\gamma(\gamma_0)) .$$

Also $A(\gamma_0)$ and $\Sigma_\gamma(\gamma_0)$ can be consistently estimated by

$$\hat{A}(\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \sum_{k=1}^{K_i} K_h(t - R_{ik})\Delta_i(t)\left\{\frac{S^{(2)}(t,\hat{\gamma})}{S^{(0)}(t,\hat{\gamma})} - \left[\frac{S^{(1)}(t,\hat{\gamma})}{S^{(0)}(t,\hat{\gamma})}\right]^{\otimes 2}\right\} dN_i(t),$$

and

$$\hat{\Sigma}_\gamma(\hat{\gamma}) = \frac{1}{n^2}\sum_{i=1}^{n}\left[\int_0^\tau\int_0^\infty K_h(t-r)\left\{Z_i(r) - \bar{Z}(t,\hat{\gamma})\right\} dO_i(r)\,dN_i(t)\right]^{\otimes 2},$$

respectively. Now we are ready to establish the asymptotic distribution of $\hat{\theta}$.

Theorem 5. *Suppose that Conditions (E1) - (E7) hold. Then as $n \to \infty$, we have that $(nh)^{1/2}\left\{B(\theta_0,\gamma_0)\left(\hat{\theta} - \theta_0\right)\right\} \to N(0, \Sigma_\theta(\theta_0))$ in distribution. Furthermore, $B(\theta_0,\gamma_0)$ and $\Sigma_\theta(\theta_0)$ can be consistently estimated by their empirical counterparts*

$$\hat{B}\left(\hat{\theta},\hat{\gamma}\right) = -\left.\frac{\partial U_\theta(\theta;\hat{\gamma},\hat{\mu}(\cdot;\theta,\hat{\gamma}))}{\partial\theta}\right|_{\theta=\hat{\theta}},$$

and

$$\hat{\Sigma}_\theta\left(\hat{\theta}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left[\int_0^\tau\int_0^\infty\left\{K_h\left(t-r\right)\Delta_i(t)\left[X_i\left(t,r\right)Y_i\left(t\right) - \frac{Q_n^{(1)}\left(t;\theta,\hat{\gamma}\right)}{S_n^{(0)}\left(t;\hat{\gamma}\right)}\right]\right.$$

$$\left.-\hat{D}\left(\hat{\theta},\hat{\gamma}\right)\hat{A}^{-1}\left(\hat{\gamma}\right)\left\{Z_i(r)-\bar{Z}(t;\hat{\gamma})\right\}\right\}dO_i\left(r\right)dN_i(t)\Bigg]^{\otimes 2},$$

respectively, where

$$\hat{D}\left(\hat{\theta},\hat{\gamma}\right) = \left.\frac{\partial U_\theta\left(\hat{\theta};\gamma,\hat{\mu}\left(\cdot;\hat{\theta},\gamma\right)\right)}{\partial\gamma}\right|_{\gamma=\hat{\gamma}}.$$

The proof of Theorem 5 is sketched in Appendix A.3. To make use of the result given above, it is apparent that one needs to choose a kernel function $K(\cdot)$ and the bandwidth h. For the former, there exist many choices and a commonly used one is the Epanechnikov kernel to be used below. For the selection of bandwidth h, it is well-known that a smaller h usually leads to a smaller bias but a larger variance and a proper h should be chosen based on their trade-off. For this, note that the optimal bandwidth may depend on the total number of the observations on covariates of all subjects, and the second term in the square bracket of (4.9) can be regarded as the Nadaraya-Watson estimator of a certain function at time point t given the observations at the R_{ik}'s, $i = 1\ldots,n$, $k = 1,\ldots,K_i$. Also it is well-known that the optimal bandwidth of the Nadaraya-Watson estimator in such context is $O\left(\tilde{n}^{-0.2}\right)$, where $\tilde{n} = \sum_i K_i$. These suggest that we can consider the bandwidth $h = \tilde{n}^{-a}$ with $0.2 < a < 1$. The simulation study below indicates that the value of a between 0.6 and 0.7 gives a good balance between the bias and variance.

Instead of the approach considered above, sometimes one may prefer a data-driven

method to select the bandwidth through the grid search. For this, by following Cao et al. (2015a) and Cao et al. (2015b), we suggest choosing h that minimizes the estimated mean square error (MSE) $\hat{C}^2h^4+\hat{V}(h)$ over the range of bandwidth candidates, where $\hat{V}(h)$ denotes the estimated asymptotic variance of $\hat{\beta}$ for a given h and $\hat{C}$ is defined below. To define the range, it is apparent that a natural lower bound is given by $\max_i \min_{k=(1,\ldots,K_i),j=(1,\ldots,J_i)} |T_{ij} - R_{ik}|$. The upper bound of the range is usually taken to be M times the lower bound with the simulation study indicating $M = 5$ being a good choice. Given the range, one can regress $\hat{\beta}(h)$ to h^2 with 25 equally spaced h over the range to obtain the slope $\hat{C}$, which can be used to estimate the bias by the slope $\hat{C}$. Note that Cao et al. (2015a) and Cao et al. (2015b) used a different method to estimate the asymptotic variance of $\hat{\beta}$ and the approach given here can be more reliable and faster.

4.5 A Simulation Study

Now we report some results obtained from an extensive simulation study conducted to assess the finite sample performance of the method proposed in the previous sections. In the study, we assumed that the covariate process $Z(t)$ is a piecewise constant function given by

$$Z(t) = \sum_{i=1}^{20} I\{(i-1)/20 \le t < i/20\} z_i,$$

where $\{z_i\}_{i=1}^{20}$ follows the multivariate normal distribution with mean zero and the covariance matrix $\{e^{-|i-j|/20}\}_{i,j=1,\ldots,20}$. For the response process, it was assumed that

$$Y_i(t) = g[\mu_0(t) \exp\{\beta_0' Z_i(t) + \alpha_0' N_i(t-)\}] + \epsilon_i(t)$$

with $g(t) = \log(t)$ or $g(t) = t$, where $\epsilon_i(t)$ is a Gaussian process with mean zero and covariance $2^{-|t-s|}$ between $\epsilon_i(t)$ and $\epsilon_i(s)$. On the observation processes, we assumed that both $N_i^*(t)$ and $O_i^*(t)$ are Poisson processes with the intensity functions $\exp(\gamma_0' Z(t))\,\lambda_0(t)$ and $\nu_0(t)$, respectively. In the following, we considered the set-up with $\lambda_0(t) = 3$ and $\nu_0(t) = 6$ or $\lambda_0(t) = 4$ and $\nu_0(t) = 8$. Also the follow-up times C_i's were generated from the uniform distribution over $(0.8, 1)$. The results given below are based on $n = 100$ or 500 with 1000 replications.

Table 4.1 presents the results on the estimation of β and α given by the proposed method with $\beta_0 = -2$, $\alpha_0 = 0.1$, $\gamma_0 = 1.5$, $\mu_0(t) = \exp(2)$ and $g(x) = \log(x)$. In the table, we calculated the estimated bias (Bias) given by the average of the estimates minus the true value, the sample standard errors of the estimates (SSD), the average of the estimated standard errors (ESE), and the 95% empirical coverage probabilities (CP). For the results here, we used the Epanechnikov kernel $K(t) = 0.75\,(1 - t^2)_+$ and set the bandwidth h to be $\tilde{n}^{-a}$ with $a = 0.6$ or 0.7 or selected by the data-driven grid search approach (grid) described in the previous section. For comparison, we also include the results on the estimation of β given by the method proposed in Cao et al. (2015b), denoted by $\hat{\beta}_{Cao}$, which assumes that the observation process on the response process is independent of or non-informative above the response process.

One can see from Table 4.1 that the proposed estimators seem to be unbiased, and the variance estimation and the normal approximation to the distribution of the proposed estimator appear to be appropriate. On the bandwidth selection, all three methods seem to give reasonable results and as expected, the results became better when the sample size increased. In addition, the results indicate that the method given in Cao et al. (2015b) gave biased results or in other words, in the presence

of dependent or informative observation processes, the proposed approach should be used. Table 4.2 gives the results on the estimation of β and α given by the proposed method with $\mu_0(t) = \exp(\sin(2\pi t))$ and the other set-up being the same as in Table 4.1, and Tables 4.3 and 4.4 present the results as those given in Tables 4.1 and 4.2, respectively, except $g(x) = x$. Again they all suggest that the proposed methodology seems to work well and furthermore, it can be seen that the bias and standard error became smaller when the number of observations increased. We also considered some other set-ups, including different kernel and link functions, namely, $K(t)$ and $g(x)$, and obtained similar results.

4.6 An Application to the HIV study

In this section, we will apply the methodology proposed in the previous sections to the cohort study of HIV-infected subjects discussed above and in Cao et al. (2015b) and Wohl et al. (2005) among others. As mentioned before, the study consisted of 190 patients who were followed for about five years from July 1997 to September 2002, and among others, the patient's HIV viral load and CD4 cell counts were measured at different time points during the study. In other words, both the measurement times and the numbers of observations on the HIV viral load and CD4 cell counts differ and we have sparse and asynchronous longitudinal data on the two variables. In the analysis here, by following Cao et al. (2015b), we will focus on the 181 patients who had at least one measurement on both the HIV viral load and CD4 cell counts and infer the relationship between the two variables.

To apply the proposed approach, again by following Cao et al. (2015b), we will

treat the HIV viral load as the response variable and the CD4 cell count as the covariate with the use of the log transformation on both. Also we transfer the scale to make the follow-up period to be $(0, 1]$. For the data, we have $\tilde{n} = 723$, the total number of observations on the covariate, and $\max_i \min_{k=(1,\ldots,K_i),j=(1,\ldots,J_i)} |T_{ij} - R_{ik}| = 0.0180$, the smallest possible bandwidth. For the results given below, we used the bandwidths $h = \tilde{n}^{-a}$ with $a = -0.4, -0.5, -0.6$ as well as the choice given by the data-driven bandwidth selection method described above. Note that $\tilde{n}^{-0.7}$ is smaller than the smallest possible bandwidth.

Table 4.5 presents the estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\alpha}$ given by the proposed method based on the Epanechnikov kernel along with their estimated standard errors and the p-values for testing the corresponding parameter equal to zero. Here we took $g(x) = \log(x)$ and set $H_i(\mathcal{F}_{it})$ to be the total number of the observations on the HIV viral load of the ith patient during the last three months. For comparison, the estimate of β given by the method proposed by Cao et al. (2015b), denoted by $\hat{\beta}_{Cao}$, was also obtained and included in the table as well as their estimated standard errors and the p-values for testing β being equal to zero. One can see from Table 4.5 that both methods suggest that the HIV viral load and CD4 cell counts were significantly and negatively correlated, which is consistent with the literature (Cao et al., 2015b), but the observation process on the HIV viral load did not seem to be related to the CD4 cell counts. Furthermore, the results indicate that the HIV viral load seems to be significantly related to the number of the observations during the last three months, and as seen in the simulation study, the method that ignores this relationship tended to overestimate the covariate effect.

4.7 Discussion and Concluding Remarks

This chapter discussed regression analysis of asynchronous longitudinal data in the presence of informative observation times, and for the problem, we presented a class of flexible semiparametric transformation models, which include many commonly used models as special cases. For estimation, an estimating equation-based approach was developed with the use of the kernel weighting technique, and the asymptotic distribution of the proposed estimator of regression parameters was derived. The numerical studies were carried out for the assessment of the finite sample performance of the proposed methodology and suggested that it seems to work well for practical situations. In particular, they indicated that as expected, the use of the method that ignores the dependent or informative observation process could yield biased results or conclusions.

As mentioned above, the focus here is on the longitudinal process $Y_i(t)$ and with respect to $Y_i(t)$, the proposed estimation approach is essentially a conditional one. With respect to $Y_i(t)$ and $N_i^*(t)$ together or given their correlation, instead of the conditional model (4.1), one could take a joint modeling approach or a marginal modeling method. The former models them together by using, for example, some latent variables to describe their correlation, while the latter models $Y_i(t)$ marginally and then $N_i^*(t)$ conditional on $Y_i(t)$. It is apparent that the latter method would be a natural choice if the counting process $N_i^*(t)$ is of primary interest. One advantage of the proposed conditional model (4.1) is that in addition to the evaluation of covariate effects on the longitudinal process, it also allows the prediction and the independence testing between the response and observation processes.

Table 4.1: The simulation results based on $\mu_0(t) = \exp(2)$ and $g(x) = \log(x)$.

		$n = 100$				$n = 500$			
h	para	Bias	ESE	SSD	CP(%)	Bias	ESE	SSD	CP(%)
$\lambda_0(t) = 3$, $\nu_0(t) = 6$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.072	0.579	0.548	94.8	0.053	0.345	0.336	95.7
	$\hat{\beta}_{Cao}$	0.371	0.459	0.478	86.0	0.380	0.278	0.284	72.1
	$\hat{\alpha}$	-0.012	0.098	0.089	94.9	-0.009	0.058	0.053	96.2
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.061	0.656	0.657	94.5	0.044	0.423	0.414	95.9
	$\hat{\beta}_{Cao}$	0.359	0.529	0.555	86.5	0.369	0.367	0.386	81.2
	$\hat{\alpha}$	-0.011	0.111	0.109	94.6	-0.008	0.070	0.068	95.3
grid	$\hat{\beta}$	0.143	0.507	0.484	94.3	0.075	0.276	0.267	94.2
	$\hat{\beta}_{Cao}$	0.416	0.398	0.441	78.8	0.394	0.166	0.176	34.4
	$\hat{\alpha}$	-0.015	0.088	0.075	95.5	-0.007	0.047	0.039	97.6
$\lambda_0(t) = 4$, $\nu_0(t) = 8$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.051	0.561	0.507	96.1	0.027	0.320	0.313	95.5
	$\hat{\beta}_{Cao}$	0.492	0.426	0.415	77.2	0.486	0.247	0.253	47.4
	$\hat{\alpha}$	-0.006	0.080	0.065	97.3	-0.003	0.045	0.041	96.7
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.044	0.625	0.600	95.3	0.026	0.383	0.398	95.3
	$\hat{\beta}_{Cao}$	0.469	0.488	0.489	81.9	0.479	0.323	0.344	65.2
	$\hat{\alpha}$	-0.008	0.089	0.077	97.0	-0.003	0.054	0.052	96.1
grid	$\hat{\beta}$	0.106	0.506	0.460	96.0	0.049	0.262	0.243	95.9
	$\hat{\beta}_{Cao}$	0.529	0.380	0.385	68.8	0.514	0.196	0.200	26.1
	$\hat{\alpha}$	-0.008	0.074	0.056	97.2	-0.003	0.038	0.030	97.9

Table 4.2: The simulation results based on $\mu_0(t) = \exp(\sin(2\pi t))$ and $g(x) = \log(x)$.

		$n = 100$				$n = 500$			
h	para	Bias	ESE	SSD	CP(%)	Bias	ESE	SSD	CP(%)
$\lambda_0(t) = 3$, $\nu_0(t) = 6$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.021	0.615	0.557	96.7	0.019	0.366	0.358	95.8
	$\hat{\alpha}$	0.004	0.092	0.082	97.5	0.004	0.054	0.051	96.5
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.009	0.694	0.651	96.1	0.006	0.449	0.458	94.8
	$\hat{\alpha}$	0.005	0.104	0.098	96.1	0.004	0.065	0.067	95.6
grid	$\hat{\beta}$	0.102	0.536	0.492	95.2	0.051	0.292	0.286	95.1
	$\hat{\alpha}$	-0.003	0.083	0.074	97.2	0.001	0.043	0.039	97.6
$\lambda_0(t) = 4$, $\nu_0(t) = 8$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.022	0.549	0.490	96.2	0.007	0.316	0.300	96.1
	$\hat{\alpha}$	0.000	0.063	0.056	96.1	0.001	0.036	0.033	97.3
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.001	0.613	0.582	95.2	0.003	0.380	0.387	94.9
	$\hat{\alpha}$	0.001	0.070	0.066	95.8	0.001	0.043	0.041	96.2
grid	$\hat{\beta}$	0.097	0.491	0.447	94.5	0.037	0.256	0.239	96.1
	$\hat{\alpha}$	-0.005	0.058	0.050	97.2	-0.001	0.029	0.026	97.4

Table 4.3: The simulation results based on $\mu_0(t) = \exp(2)$ and $g(x) = x$.

		$n = 100$				$n = 500$			
h	para	Bias	ESE	SSD	CP(%)	Bias	ESE	SSD	CP(%)
$\lambda_0(t) = 3$, $\nu_0(t) = 6$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.071	0.381	0.399	95.2	0.040	0.255	0.263	95.2
	$\hat{\alpha}$	-0.013	0.070	0.075	96.2	-0.009	0.046	0.048	95.1
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.076	0.301	0.355	96.4	0.045	0.199	0.217	94.8
	$\hat{\alpha}$	-0.013	0.054	0.067	96.5	-0.009	0.035	0.039	96.0
grid	$\hat{\beta}$	0.135	0.312	0.264	94.7	0.063	0.174	0.150	95.5
	$\hat{\alpha}$	-0.017	0.060	0.047	97.5	-0.008	0.032	0.025	97.2
$\lambda_0(t) = 4$, $\nu_0(t) = 8$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	0.060	0.275	0.343	96.5	0.030	0.184	0.200	96.0
	$\hat{\alpha}$	-0.007	0.035	0.050	98.5	-0.003	0.024	0.029	98.4
$\tilde{n}^{-0.7}$	$\hat{\beta}$	0.063	0.333	0.379	96.0	0.030	0.234	0.236	95.4
	$\hat{\alpha}$	-0.008	0.044	0.055	98.2	-0.004	0.030	0.034	97.8
grid	$\hat{\beta}$	0.107	0.245	0.309	95.4	0.053	0.139	0.164	96.7
	$\hat{\alpha}$	-0.009	0.031	0.046	98.0	-0.004	0.018	0.024	98.9

Table 4.4: The simulation results based on $\mu_0(t) = \exp(\sin(2\pi t))$ and $g(x) = x$.

		$n = 100$				$n = 500$			
h	para	Bias	ESE	SSD	CP(%)	Bias	ESE	SSD	CP(%)
$\lambda_0(t) = 3$, $\nu_0(t) = 6$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	-0.143	1.128	1.016	95.6	-0.042	0.606	0.613	95.4
	$\hat{\alpha}$	0.061	0.173	0.125	99.6	0.016	0.092	0.072	98.6
$\tilde{n}^{-0.7}$	$\hat{\beta}$	-0.191	1.311	1.175	96.0	-0.125	0.776	0.770	96.9
	$\hat{\alpha}$	0.084	0.215	0.164	99.5	0.038	0.122	0.097	99.0
grid	$\hat{\beta}$	0.007	1.725	1.311	93.0	-0.002	0.701	0.735	95.3
	$\hat{\alpha}$	-0.024	0.253	0.197	97.9	-0.011	0.104	0.113	97.8
$\lambda_0(t) = 4$, $\nu_0(t) = 8$									
$\tilde{n}^{-0.6}$	$\hat{\beta}$	-0.083	1.075	1.047	94.4	-0.002	0.487	0.513	94.6
	$\hat{\alpha}$	-0.012	0.198	0.193	98.3	-0.005	0.060	0.062	96.0
$\tilde{n}^{-0.7}$	$\hat{\beta}$	-0.109	1.314	1.241	95.2	-0.020	0.601	0.669	94.6
	$\hat{\alpha}$	-0.016	0.284	0.252	98.8	-0.004	0.076	0.084	95.5
grid	$\hat{\beta}$	0.085	0.907	0.868	94.1	0.002	0.552	0.558	94.6
	$\hat{\alpha}$	-0.014	0.127	0.128	97.9	0.000	0.060	0.060	96.1

Table 4.5: The analysis results for the HIV study, including the estimated parameters (Est.), the estimated standard errors (SE) and the p-values (p-val).

h		Est.	SE	p-val
0.019 ($\tilde{n}^{-0.6}$)	$\hat{\gamma}$	-0.009	0.103	0.930
	$\hat{\alpha}$	-3.031	1.743	0.082
	$\hat{\beta}$	-0.945	0.219	0.000
	$\hat{\beta}_{Cao}$	-1.08	0.226	0.000
0.037 ($\tilde{n}^{-0.5}$)	$\hat{\gamma}$	-0.039	0.105	0.712
	$\hat{\alpha}$	-2.944	1.394	0.035
	$\hat{\beta}$	-1.026	0.200	0.000
	$\hat{\beta}_{Cao}$	-1.103	0.191	0.000
0.072 ($\tilde{n}^{-0.4}$)	$\hat{\gamma}$	-0.036	0.107	0.734
	$\hat{\alpha}$	-3.298	1.173	0.005
	$\hat{\beta}$	-1.072	0.204	0.000
	$\hat{\beta}_{Cao}$	-1.147	0.167	0.000
grid	$\hat{\gamma}$	-0.011	0.103	0.913
	$\hat{\alpha}$	-3.036	1.732	0.080
	$\hat{\beta}$	-0.950	0.217	0.000
	$\hat{\beta}_{Cao}$	-1.139	0.228	0.000

Chapter 5

Future Research

5.1 Research Topics for Variable Selection of Event History Data

In Chapter 2 and 3, we have focused on the situation where the number of variables or predictors is fixed or can increase with the sample size n but smaller than n, for which little literature exists, although the problem discussed can occur often. It is apparent that one direction for future research is to generalize the idea or approach discussed above to high-dimensional situations where p_n is larger than n. One example of this could be that the event of interest is related to the occurrence of some diseases or the symptoms associated with some diseases and an objective of the study is to establish the relationship between the disease and genes or markers, which can be hundreds of thousands or even larger. Although the idea described here may still apply, the implementation procedure would not work due to the irregularity of some needed

matrices.

The BAR method may have a high computational burden if the dimension of parameters is large, due to the matrix inversion in the algorithm. A more efficient algorithm needs to be developed to ease the computational burden. For example, Dai et al. (2018) proposed to use the iterative hard-thresholding algorithm to approximately solve the ridge regression iteration for the linear regression case. Another possible direction is to apply the coordinate descent (Friedman et al., 2010; Mazumder et al., 2011) to compute each iteration. However, since the BAR penalty is essentially non-convex, one should theoretically show the coordinate descent can achieve the same local optimum as the original BAR algorithm.

In addition, we have focused on the underlying recurrent event process that follows the additive rate model (1.1). It would be useful to generalize the proposed method to the situation where the event process may follow other models, such as the proportional rate model or the semiparametric transformation model (Li et al., 2010).

Another extension is to consider a terminal event, which corresponds to the situation where the follow-up time C_i may be related to $N_i^*(t)$. In this case, it is clear that the proposed method may not be valid and could give biased results (Cook and Lawless, 2007).

Last but not least, as discussed in the application in Section 3.5 , instead of one recurrent event of interest, one may face more than one correlated recurrent events of interest and it would be useful to generalize the proposed method to these multivariate cases.

5.2 Research Topics for Asynchronous Longitudinal Data

In Chapter 4, for simplicity, we have used the same kernel function and bandwidth for the estimation of γ and θ. The proposed method is still valid if different kernel functions and bandwidths are used instead. However, the derivation of the asymptotic distribution of the proposed estimator $\hat{\theta}$ will be different. Also note that in the proposed method, we have focused on the situation where the observation process on the response process may be informative but the observation process on the covariate process is non-informative. It is apparent that sometimes both observation processes could be informative and thus it would be useful to generalize the proposed method to this latter situation.

Another direction for the generalization of the proposed approach is that throughout Chapter 4, we have assumed that the follow-up times C_i's are independent of both the response process and the observation process. In reality, this may not be true as, for example, the C_i's may be caused by a terminal event that is correlated to the response process, and such situations have been discussed by many authors when one observes synchronous longitudinal data or the covariate process can be completely observed. However, it does not seem to exist any literature for the situation of sparse asynchronous longitudinal data.

Appendix A

Theoretical Proofs

A.1 Proofs of Theorem 1 and 2

In this section, we will sketch the proofs of the oracle and grouping effect properties of the proposed BAR estimator $\widehat{\boldsymbol{\beta}}_R^*$ described in Theorems 1 and 2 of Chapter 2.

A.1.1 Preliminaries and Lemmas

For the proof, define

$$\begin{pmatrix} \boldsymbol{\alpha}^*(\boldsymbol{\beta}) \\ \boldsymbol{\gamma}^*(\boldsymbol{\beta}) \end{pmatrix} \equiv g(\boldsymbol{\beta}) = (\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta}))^{-1} \boldsymbol{P}_n \,, \tag{A.1}$$

and the partition of matrix $(n^{-1}\boldsymbol{\Omega}_n)^{-1}$ as

$$\left(n^{-1}\boldsymbol{\Omega}_n\right)^{-1} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{G} \end{pmatrix},$$

where $\boldsymbol{A}$ is a $q \times q$ matrix. Note that since $\boldsymbol{\Omega}_n$ is nonsingular, by multiplying $\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta}))$ and subtracting $\boldsymbol{\beta}_0$ on both sides of Equation (A.1), we can obtain

$$\begin{pmatrix} \boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} \\ \boldsymbol{\gamma}^* \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} A\boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^* + B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^* \\ B'\boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^* + G\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^* \end{pmatrix} = \hat{\boldsymbol{b}} - \boldsymbol{\beta}_0 = O_p(n^{-1/2}), \quad \text{(A.2)}$$

where $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = \text{diag}(\beta_1^{-2}, \ldots, \beta_q^{-2})$ and $\boldsymbol{D}_2(\boldsymbol{\beta}_2) = \text{diag}(\beta_{q+1}^{-2}, \ldots, \beta_p^{-2})$.

To prove Theorem 1, we need the following two lemmas.

Lemma 1. *Let $\{\delta_n\}$ be a sequence of positive real numbers satisfying $\delta_n \to \infty$ and $\delta_n^2/\lambda_n \to 0$. Define $H \equiv \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' : \boldsymbol{\beta}_1 \in [1/K_0, K_0]^q, \|\boldsymbol{\beta}_2\| \le \delta_n/\sqrt{n}\}$, where $K_0 > 1$ is a constant such that $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^q$. Then under the regular conditions (C1)–(C7) and with probability tending to 1, we have*

(i) $\sup_{\boldsymbol{\beta} \in H} \dfrac{\|\boldsymbol{\gamma}^*(\boldsymbol{\beta})\|}{\|\boldsymbol{\beta}_2\|} < \dfrac{1}{c_0}$ *for some constant* $c_0 > 1$.

(ii) $g(\cdot)$ *is a mapping from* H *to itself.*

Proof. Based on conditions (C6) and (C7) and note that $\boldsymbol{\beta}_1 \in [1/K_0, K_0]^q$ and $\|\boldsymbol{\alpha}^*\| \le \|g(\boldsymbol{\beta})\| < K$ for some constant $K > 0$, we have

$$\sup_{\boldsymbol{\beta} \in H} \left\| \frac{\lambda_n}{n} B'\boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^* \right\| = o_p(n^{-1/2}).$$

Since $\lambda_{min}(\boldsymbol{G}) > c^{-1}$, it follows from Equation (A.2) that, with probability tending to 1,

$$c^{-1}\left\|\frac{\lambda_n}{n}\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*\right\| - \|\boldsymbol{\gamma}^*\| \leq \sup_{\boldsymbol{\beta}\in H}\left\|\boldsymbol{\gamma}^* + \frac{\lambda_n}{n}G\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*\right\| \leq \frac{\delta_n}{\sqrt{n}}. \tag{A.3}$$

Let $m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2} = (\gamma_1^*/\beta_{q+1}, \gamma_2^*/\beta_{q+2}, \ldots, \gamma_{p-q}^*/\beta_p)'$, then from the Cauchy-Schwarz inequality and the assumption $\|\boldsymbol{\beta}_2\| \leq \delta_n/\sqrt{n}$, we have

$$\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \|\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*\|\delta_n/\sqrt{n},$$

and

$$\|\boldsymbol{\gamma}^*\| = \|(\boldsymbol{D}_2(\boldsymbol{\beta}_2))^{-1/2}m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \cdot \|\boldsymbol{\beta}_2\| \leq \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\|\delta_n/\sqrt{n}, \tag{A.4}$$

for all large n. Thus

$$\frac{\lambda_n}{nC}\frac{\sqrt{n}}{\delta_n}\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| - \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\|\frac{\delta_n}{\sqrt{n}} \leq \frac{\delta_n}{\sqrt{n}}.$$

Immediately from $\delta_n^2/\lambda_n \to 0$, we have

$$\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \frac{1}{\frac{\lambda_n}{\delta_n^2 c} - 1} < \frac{1}{c_0}, \tag{A.5}$$

with probability tending to one. Hence it follows from inequality (A.4) and (A.5) that

$$\|\boldsymbol{\gamma}^*\| < \|\boldsymbol{\beta}_2\| \leq \delta_n/\sqrt{n} \to 0 \ \text{ as } \ n \to \infty, \tag{A.6}$$

which implies that conclusion (i) holds.

To prove (ii), we only need to verify that $\boldsymbol{\alpha}^* \in [1/K_0, K_0]^q$ with probability tending to 1. Analogously, given conditions (C6), $\boldsymbol{\beta}_1 \in [1/K_0, K_0]^q$ and $\|\boldsymbol{\alpha}^*\| < K$,

$$\sup_{\beta \in H} \left\| \frac{\lambda_n}{n} A \boldsymbol{D}_1(\boldsymbol{\beta}_1) \boldsymbol{\alpha}^* \right\| = o_p(n^{-1/2}).$$

Then from Equation (A.2), we have

$$\sup_{\boldsymbol{\beta} \in H} \left\| \boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} + \frac{\lambda_n}{n} B \boldsymbol{D}_2(\boldsymbol{\beta}_2) \boldsymbol{\gamma}^* \right\| = O_p(n^{-1/2}) \leq \frac{\delta_n}{\sqrt{n}}. \tag{A.7}$$

Also according to inequality (A.3) and Condition (C6), we know that as $n \to \infty$ and with probability tending to one,

$$\sup_{\beta \in H} \left\| \frac{\lambda_n}{n} \boldsymbol{B} \boldsymbol{D}_2(\boldsymbol{\beta}_2) \boldsymbol{\gamma}^* \right\| \leq \frac{\lambda_n}{n} \|\boldsymbol{B}\| \sup_{\boldsymbol{\beta} \in H} \|\boldsymbol{D}_2(\boldsymbol{\beta}_2) \boldsymbol{\gamma}^*\| \leq \frac{2c^2 \delta_n}{\sqrt{n}}. \tag{A.8}$$

Therefore from inequality (A.7) and (A.8), we can get

$$\sup_{\boldsymbol{\beta} \in H} \|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01}\| \leq \frac{(2c^2 + 1)\delta_n}{\sqrt{n}} \to 0$$

with probability tending to one, which implies that for any $\epsilon > 0$, $P(\|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01}\| \leq \epsilon) \to 1$. Since $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^q$, thus $\boldsymbol{\alpha}^* \in [1/K_0, K_0]^q$ holds for large n. Then with the fact that $\|\boldsymbol{\gamma}^*\| \leq \delta_n/\sqrt{n}$, we proved that (ii) is true. □

Lemma 2. *Under the regular conditions (C1)−(C7) and with probability tending to 1, the equation* $\boldsymbol{\alpha} = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\alpha}))^{-1} \boldsymbol{P}_n^{(1)}$ *has a unique fixed-point* $\hat{\boldsymbol{\alpha}}^*$ *in the domain* $[1/K_0, K_0]^q$.

Proof. Since $\boldsymbol{\beta}_{02} = 0$, we define

$$f(\boldsymbol{\alpha}) = (f_1(\boldsymbol{\alpha}), f_2(\boldsymbol{\alpha}), \ldots, f_q(\boldsymbol{\alpha}))' \equiv (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\alpha))^{-1} \boldsymbol{P}_n^{(1)}, \tag{A.9}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$. It is obviously that $(f(\boldsymbol{\alpha})', 0')' = g(\boldsymbol{\alpha}, 0)$ and $f(\boldsymbol{\alpha})$ is a map from $[1/K_0, K_0]$ to itself. Multiplying $\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\alpha})$ and taking derivative with respect to $\boldsymbol{\alpha}$ on both sides of Equation (A.9), we have

$$\Big(\frac{1}{n}\boldsymbol{\Omega}_n^{(1)} + \frac{\lambda_n}{n}\boldsymbol{D}_1(\boldsymbol{\alpha})\Big)\dot{f}(\boldsymbol{\alpha}) + \frac{\lambda_n}{n}\mathrm{diag}\Big(\frac{-2f_1(\alpha)}{\alpha_1^3}, \ldots, \frac{-2f_q(\alpha)}{\alpha_q^3}\Big) = 0,$$

where $\dot{f}(\boldsymbol{\alpha}) = \partial f(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}'$. Then

$$\sup_{\alpha \in [1/K_0, K_0]^q} \Big\| \Big(\frac{1}{n}\boldsymbol{\Omega}_n^{(1)} + \frac{\lambda_n}{n}\boldsymbol{D}_1(\boldsymbol{\alpha})\Big)\dot{f}(\boldsymbol{\alpha}) \Big\| = \sup_{\alpha \in [1/K_0, K_0]^q} \frac{2\lambda_n}{n} \Big\| \mathrm{diag}\Big(\frac{f_1(\boldsymbol{\alpha})}{\alpha_1^3}, \ldots, \frac{f_q(\boldsymbol{\alpha})}{\alpha_q^3}\Big) \Big\| = o_p(1).$$

According to Condition (C6) and the fact that $\boldsymbol{\alpha} \in [1/K_0, K_0]^q$, we can derive

$$\Big\| \Big(\frac{1}{n}\boldsymbol{\Omega}_n^{(1)} + \frac{\lambda_n}{n}\boldsymbol{D}_1(\boldsymbol{\alpha})\Big)\dot{f}(\boldsymbol{\alpha}) \Big\| \geq \Big\| \frac{1}{n}\boldsymbol{\Omega}_n^{(1)}\dot{f}(\boldsymbol{\alpha}) \Big\| - \Big\| \frac{\lambda_n}{n}\boldsymbol{D}_1(\boldsymbol{\alpha})\dot{f}(\boldsymbol{\alpha}) \Big\| \geq \Big(\frac{1}{c} - \frac{\lambda_n}{n}K_0^2\Big)\|\dot{f}(\boldsymbol{\alpha})\|.$$

Thus, $\sup_{\boldsymbol{\alpha} \in [1/K_0, K_0]^q} \|\dot{f}(\boldsymbol{\alpha})\| \to 0$, which implies that $f(\cdot)$ is a contraction mapping from $[1/K_0, K_0]^q$ to itself with probability tending to one. Hence, according to the contraction mapping theorem there exists one unique fixed-point $\hat{\boldsymbol{\alpha}}^* \in [1/K_0, K_0]^q$, such that

$$\hat{\boldsymbol{\alpha}}^* = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*))^{-1} \boldsymbol{P}_n^{(1)}. \tag{A.10}$$

□

A.1.2 Proof of Theorem 1

We prove conclusion (2) of Theorem 1 first. According to the definitions of $\hat{\boldsymbol{\beta}}_R^*$ and $\hat{\boldsymbol{\beta}}_{R2}^{(k)}$, it follows from inequality (A.6) that

$$\hat{\boldsymbol{\beta}}_{R2}^* \equiv \lim_{k\to\infty} \hat{\boldsymbol{\beta}}_2^{(k)} = 0 \tag{A.11}$$

holds with the probability tending to 1.

Next, we will show that $P(\hat{\boldsymbol{\beta}}_{R1}^* = \hat{\boldsymbol{\alpha}}^*) \to 1$. Consider Equation (A.2), we define $\boldsymbol{\gamma}^* = 0$ if $\boldsymbol{\beta}_2 = 0$. Note that for any fixed large n, from Equation (A.2), we have

$$\lim_{\boldsymbol{\beta}_2\to 0} \boldsymbol{\gamma}^*(\boldsymbol{\beta}) = 0.$$

Furthermore, multiply $(\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta}))$ on both sides of Equation (A.1), then we can get

$$\lim_{\boldsymbol{\beta}_2\to 0} \boldsymbol{\alpha}^*(\boldsymbol{\beta}) = (\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1))^{-1} \boldsymbol{P}_n^{(1)} = f(\boldsymbol{\beta}_1). \tag{A.12}$$

Combining Equation (A.11) and (A.12), we have

$$\sup_{\boldsymbol{\beta}_1 \in [1/K_0, K_0]^q} \|f(\boldsymbol{\beta}_1) - \boldsymbol{\alpha}^*(\boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_{R2}^{(k)})\| \to 0, \text{ as } k \to \infty. \tag{A.13}$$

Since $f(\cdot)$ is a contract mapping, then it follows from Equation (A.10) that

$$\|f(\hat{\boldsymbol{\beta}}_{R1}^{(k)}) - \hat{\boldsymbol{\alpha}}^*\| = \|f(\hat{\boldsymbol{\beta}}_{R1}^{(k)}) - f(\hat{\boldsymbol{\alpha}}^*)\| \le \frac{1}{c}\|\hat{\boldsymbol{\beta}}_{R1}^{(k)} - \hat{\boldsymbol{\alpha}}^*\|. \quad (c > 1) \tag{A.14}$$

Let $h_k = \|\hat{\boldsymbol{\beta}}_{R1}^{(k)} - \hat{\boldsymbol{\alpha}}^*\|$, then from (A.13) and (A.14) we get

$$\begin{aligned} h_{k+1} = \|\boldsymbol{\alpha}^*(\hat{\boldsymbol{\beta}}_R^{(k)}) - \hat{\boldsymbol{\alpha}}^*\| &\leq \|\boldsymbol{\alpha}^*(\hat{\boldsymbol{\beta}}_R^{(k)}) - f(\hat{\boldsymbol{\beta}}_{R1}^{(k)})\| + \|f(\hat{\boldsymbol{\beta}}_{R1}^{(k)}) - \hat{\boldsymbol{\alpha}}^*\| \\ &\leq \eta_k + \frac{1}{c}\, h_k, \text{ for some small } \eta_k > 0. \end{aligned}$$

From (A.13), for any $\epsilon \geq 0$, there exists $N > 0$, such that $|\eta_k| < \epsilon$. Employing some recursive calculation, we have $h_k \to 0$ as $k \to \infty$. Hence, with probability tending to one,

$$\|\hat{\boldsymbol{\beta}}_{R1}^{(k)} - \hat{\boldsymbol{\alpha}}^*\| \to 0 \text{ as } k \to \infty.$$

Since $\hat{\boldsymbol{\beta}}_{R1}^* \equiv \lim_{k\to\infty} \hat{\boldsymbol{\beta}}_{R1}^{(k)}$ and from the uniqueness of fixed-point, we have $P(\hat{\boldsymbol{\beta}}_{R1}^* = \hat{\boldsymbol{\alpha}}^*) \to 1$. That is, condition (1) holds.

Finally, based on Equation (A.2) and (C6) and note that $\lambda_n/n = o_p(n^{-1/2})$, we get $\sqrt{n}(\hat{\boldsymbol{\beta}}_{R1}^* - \boldsymbol{\beta}_{01}) \approx \sqrt{n}(\hat{\boldsymbol{b}}_1 - \boldsymbol{\beta}_{01})$, where $\hat{\boldsymbol{b}}_1$ is the first q elements of $\hat{\boldsymbol{b}}$. Then conclusion (3) follows from the asymptotic normality of $\hat{\boldsymbol{b}}$ Schaubel et al. (2006).

A.1.3 Proof of Theorem 2

Let

$$Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) \equiv \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p} \frac{\beta_j^2}{\tilde{\beta}_j^2},$$

and $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})$.

On the one hand, from $Q(\hat{\boldsymbol{\beta}}|\tilde{\boldsymbol{\beta}}) \leq Q(0|\tilde{\boldsymbol{\beta}})$, we have

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 + \lambda_n \sum_{i=1}^{p} \frac{\hat{\beta}_m^2}{\tilde{\beta}_m^2} \leq \|\boldsymbol{y}\|^2.$$

Therefore,

$$\|\hat{\boldsymbol{\varepsilon}}\| \leq \|\boldsymbol{y}\|.$$

On the other hand, when $\hat{\beta}_m \neq 0$, considering the derivative

$$\frac{\partial Q(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}})}{\partial \beta_m}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = -2\boldsymbol{x}'_m(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2\lambda_n \cdot \frac{\hat{\beta}_m}{\tilde{\beta}_m^2} = 0,$$

where $m \in \{1, \ldots, p\}$, we have

$$\hat{\beta}_m = \frac{\tilde{\beta}_m^2}{\lambda_n} \cdot \boldsymbol{x}'_m \hat{\boldsymbol{\varepsilon}}. \tag{A.15}$$

Taking limitation on both sides of Equation (A.15), that is, $\lim_{k\to\infty} \hat{\beta}_m = \lim_{k\to\infty} \tilde{\beta}_m = \hat{\beta}^*_{Rm}$ hold with probability tending to 1, then we can obtain that

$$\frac{1}{\hat{\beta}^*_{Ri}} = \frac{1}{\lambda_n}\boldsymbol{x}'_i\hat{\boldsymbol{\varepsilon}} \quad \text{and} \quad \frac{1}{\hat{\beta}^*_{Rj}} = \frac{1}{\lambda_n}\boldsymbol{x}'_j\hat{\boldsymbol{\varepsilon}}.$$

Therefore,

$$\left|\frac{1}{\hat{\beta}^*_{Ri}} - \frac{1}{\hat{\beta}^*_{Rj}}\right| \leq \frac{1}{\lambda_n}\|\hat{\boldsymbol{\varepsilon}}\| \cdot \|\boldsymbol{x}'_i - \boldsymbol{x}'_j\| \leq \frac{1}{\lambda_n}\|\boldsymbol{y}\|\sqrt{2(1-\rho_{ij})}.$$

This completes the proof.

A.2 Proofs of Theorem 3 and 4

In this section, we will sketch the proofs of the oracle and grouping effect properties of $\widehat{\boldsymbol{\beta}}^*_P$ described in Theorems 3 and 4.

A.2.1 Preliminaries and Lemmas

For this, define

$$\begin{pmatrix} \boldsymbol{\alpha}^*(\boldsymbol{\beta}) \\ \boldsymbol{\gamma}^*(\boldsymbol{\beta}) \end{pmatrix} = g(\boldsymbol{\beta}) = \{\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta})\}^{-1} \mathbf{v}_n, \tag{A.16}$$

and the partition of the matrix $(\boldsymbol{\Omega}_n/n)^{-1}$ as

$$(\boldsymbol{\Omega}_n/n)^{-1} = \begin{pmatrix} A & B \\ B^\top & G \end{pmatrix},$$

where A is a $q_n \times q_n$ matrix. Note that since $\boldsymbol{\Omega}_n$ is nonsingular, it follows by multiplying $\boldsymbol{\Omega}_n^{-1}\{\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta})\}$ and subtracting $\boldsymbol{\beta}_0$ on both sides of Eq. (A.16) that we have

$$\begin{pmatrix} \boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} \\ \boldsymbol{\gamma}^* \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} A\boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^* + B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^* \\ B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^* + G\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^* \end{pmatrix} = \hat{\boldsymbol{b}} - \boldsymbol{\beta}_0, \tag{A.17}$$

where $\hat{\boldsymbol{b}} = \boldsymbol{\Omega}_n^{-1}\mathbf{v}_n$, $\boldsymbol{D}_1(\boldsymbol{\beta}_1) = \mathrm{diag}(\beta_1^{-2}, \ldots, \beta_{q_n}^{-2})$ and $\boldsymbol{D}_2(\boldsymbol{\beta}_2) = \mathrm{diag}(\beta_{q_n+1}^{-2}, \ldots, \beta_{p_n}^{-2})$.

To prove Theorem 3, we also need the following two lemmas.

Lemma 3. *Let δ_n be a sequence of positive real numbers satisfying $\delta_n \to \infty$ and $\delta_n^2 p_n/\lambda_n \to 0$. Define $H_n = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top : \boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q_n}, \|\boldsymbol{\beta}_2\| \le \delta_n\sqrt{p_n/n}\}$, where $K_0 > 1$ is a constant such that $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^{q_n}$. Then under regularity conditions* (D1)–(D8) *and with probability tending to* 1*, we have*

(i) $\sup_{\boldsymbol{\beta}\in H_n} \|\boldsymbol{\gamma}^*(\boldsymbol{\beta})\|/\|\boldsymbol{\beta}_2\| < 1/c_0$ *for some constant* $c_0 > 1$.

(ii) g is a mapping from H_n to itself.

Proof. First according to Li et al. (2010), we have that $\mathrm{E}(\|\hat{\boldsymbol{b}} - \boldsymbol{\beta}_0\|^2) = \mathrm{tr}(\boldsymbol{\Omega}_n^{-1}\Xi_n\boldsymbol{\Omega}_n^{-1})/n =$

$O(p_n/n)$, and hence $\|\hat{\boldsymbol{b}} - \boldsymbol{\beta}_0\|^2 = O_p(p_n/n)$. It then follows from Eq. (A.17) that

$$\sup_{\boldsymbol{\beta} \in H_n} \|\boldsymbol{\gamma}^* + \lambda_n B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^*/n + \lambda_n G \boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\| = O_p(\sqrt{p_n/n}).$$

By Condition (D5) and the fact that

$$\|BB^\top\| - \|A^2\| \leq \|BB^\top + A^2\| \leq \|(\boldsymbol{\Omega}_n/n)^{-2}\| < c^2,$$

we can deduce that $\|B\| \leq \sqrt{2}c$. Furthermore, based on Conditions (D5) and (D6), and given that $\boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q_n}$ and $\|\boldsymbol{\alpha}^*\| \leq \|g(\boldsymbol{\beta})\| \leq \|\hat{\boldsymbol{b}}\| = O_p(\sqrt{p_n})$, we have

$$\sup_{\boldsymbol{\beta} \in H_n} \|\lambda_n B^\top \boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^*/n\| = o_p(\sqrt{p_n/n}). \tag{A.18}$$

Since $\lambda_{\min}(\boldsymbol{G}) > 1/c$, it follows from (A.17) that, with probability tending to 1,

$$\|\lambda_n \boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\|/c - \|\boldsymbol{\gamma}^*\| \leq \sup_{\boldsymbol{\beta} \in H_n} \|\boldsymbol{\gamma}^* + \lambda_n G \boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\| \leq \delta_n \sqrt{p_n/n}. \tag{A.19}$$

Let $m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2} = (\gamma_1^*/\beta_{q_n+1}, \gamma_2^*/\beta_{q_n+2}, \ldots, \gamma_{p_n-q_n}^*/\beta_{p_n})^\top$. It then follows from the Cauchy–Schwarz inequality and the assumption $\|\boldsymbol{\beta}_2\| \leq \delta_n\sqrt{p_n/n}$ that

$$\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \|\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*\|\delta_n\sqrt{p_n/n},$$

and

$$\|\boldsymbol{\gamma}^*\| = \|\{\boldsymbol{D}_2(\boldsymbol{\beta}_2)\}^{-1/2} m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \times \|\boldsymbol{\beta}_2\| \leq \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\|\delta_n\sqrt{p_n/n}, \tag{A.20}$$

for all large n. Thus from Eqs. (A.19)–(A.20), we have the inequality

$$\frac{\lambda_n}{nC}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| - \|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\|\delta_n\sqrt{p_n/n} \leq \delta_n\sqrt{p_n/n}.$$

From $p_n\delta_n^2/\lambda_n \to 0$, we immediately have, for $c_0 > 1$,

$$\|m_{\boldsymbol{\gamma}^*/\boldsymbol{\beta}_2}\| \leq \frac{1}{\lambda_n/(p_n\delta_n^2 c) - 1} < 1/c_0, \tag{A.21}$$

with probability tending to 1. Hence it follows from Eqs. (A.20)–(A.21) that, as $n \to \infty$,

$$\|\boldsymbol{\gamma}^*\| < \|\boldsymbol{\beta}_2\| \leq \delta_n\sqrt{p_n/n} \to 0, \tag{A.22}$$

which implies that conclusion (i) holds.

To prove (ii), we only need to verify that $\boldsymbol{\alpha}^* \in [1/K_0, K_0]^{q_n}$ with probability tending to 1 since (A.22) guarantees that $\|\boldsymbol{\gamma}^*\| \leq \delta_n\sqrt{p_n/n}$ with probability tending to 1. Analogously, given Condition (D5), $\boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q_n}$ and $\|\boldsymbol{\alpha}^*\| < O_p(\sqrt{p_n})$, we have

$$\sup_{\beta \in H_n} \|\lambda_n A \boldsymbol{D}_1(\boldsymbol{\beta}_1)\boldsymbol{\alpha}^*/n\| = o_p(\sqrt{p_n/n}).$$

Then from Eq. (A.17), we have

$$\sup_{\boldsymbol{\beta} \in H_n} \|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} + \lambda_n B \boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\| = O_p(\sqrt{p_n/n}) \leq \delta_n\sqrt{p_n/n}, \tag{A.23}$$

and according to Eqs. (A.19) and (A.22), we have $\|\lambda_n \boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\| \leq 2c\delta_n\sqrt{p_n/n}$. Then based on Condition (D5), we know that as $n \to \infty$ and with probability tending

to 1,

$$\sup_{\beta\in H_n} \|\lambda_n B\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*/n\| \le \lambda_n\|B\| \sup_{\boldsymbol{\beta}\in H_n} \|\boldsymbol{D}_2(\boldsymbol{\beta}_2)\boldsymbol{\gamma}^*\|/n \le 2\sqrt{2}c^2\delta_n\sqrt{p_n/n}. \quad \text{(A.24)}$$

Therefore from Eqs. (A.23)–(A.24), we can get, as $n \to \infty$,

$$\sup_{\boldsymbol{\beta}\in H_n} \|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01}\| \le (2\sqrt{2}c^2 + 1)\delta_n\sqrt{p_n/n} \to 0$$

with probability tending to 1, which implies that for any $\epsilon > 0$, $\Pr(\|\boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01}\| \le \epsilon) \to 1$. Since $\boldsymbol{\beta}_{01} \in [1/K_0, K_0]^{q_n}$, thus $\boldsymbol{\alpha}^* \in [1/K_0, K_0]^{q_n}$ holds for large n. Thus (ii) is true. This concludes the proof of Lemma 3.

□

Lemma 4. *Under regularity conditions* (D1)–(D8) *and with probability tending to* 1, *the equation* $\boldsymbol{\alpha} = \{\boldsymbol{\Omega}_n^{(1)} + \lambda_n\boldsymbol{D}_1(\boldsymbol{\alpha})\}^{-1}\mathbf{v}_n^{(1)}$ *has a unique fixed-point* $\hat{\boldsymbol{\alpha}}^*$ *in the domain* $[1/K_0, K_0]^{q_n}$.

Proof. Define

$$f(\boldsymbol{\alpha}) = (f_1(\boldsymbol{\alpha}), \ldots, f_{q_n}(\boldsymbol{\alpha}))^\top = \{\boldsymbol{\Omega}_n^{(1)} + \lambda_n\boldsymbol{D}_1(\alpha)\}^{-1}\mathbf{v}_n^{(1)}, \quad \text{(A.25)}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{q_n})^\top$. By multiplying $(\boldsymbol{\Omega}_n^{(1)})^{-1}\{\boldsymbol{\Omega}_n^{(1)} + \lambda_n\boldsymbol{D}_1(\boldsymbol{\alpha})\}$ and then subtracting $\boldsymbol{\beta}_{01}$ on both sides of (A.25), we have

$$f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01} + \lambda_n(\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\alpha)f(\boldsymbol{\alpha}) = (\boldsymbol{\Omega}_n^{(1)})^{-1}\mathbf{v}_n^{(1)} - \boldsymbol{\beta}_{01} = \hat{\boldsymbol{\beta}}_1(\text{OLS}) - \boldsymbol{\beta}_{01},$$

where $\hat{\boldsymbol{\beta}}_1(\text{OLS})$ is the first q_n-dimensional subvector of $\hat{\boldsymbol{b}}$. Therefore,

$$\sup_{\alpha\in[1/K_0,K_0]^{q_n}} \|f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01} + \lambda_n(\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\alpha)f(\boldsymbol{\alpha})\| = O_p(\sqrt{q_n/n}).$$

Similar to Eq. (A.18), it can be shown that

$$\sup_{\alpha\in[1/K_0,K_0]^{q_n}} \|\lambda_n(\boldsymbol{\Omega}_n^{(1)}/n)^{-1}\boldsymbol{D}_1(\alpha)f(\boldsymbol{\alpha})/n\| = o_p(\sqrt{q_n/n}).$$

Thus, as $n \to \infty$,

$$\sup_{\alpha\in[1/K_0,K_0]^{q_n}} \|f(\boldsymbol{\alpha}) - \boldsymbol{\beta}_{01}\| \le \delta_n\sqrt{q_n/n} \to 0,$$

which implies that $f(\boldsymbol{\alpha}) \in [1/K_0, K_0]^{q_n}$ with probability tending to 1. That is, $f(\boldsymbol{\alpha})$ is a mapping from $[1/K_0, K_0]^{q_n}$ to itself.

Also by multiplying $\boldsymbol{\Omega}_n^{(1)} + \lambda_n\boldsymbol{D}_1(\boldsymbol{\alpha})$ and taking derivative with respect to $\boldsymbol{\alpha}$ on both sides of A.25, we have

$$\{\boldsymbol{\Omega}_n^{(1)}/n + \lambda_n\boldsymbol{D}_1(\boldsymbol{\alpha})/n\}\dot{f}(\boldsymbol{\alpha}) + \lambda_n\text{diag}\{-2f_1(\alpha)/\alpha_1^3, \ldots, -2f_{q_n}(\alpha)/\alpha_{q_n}^3\}/n = 0,$$

where $\dot{f}(\boldsymbol{\alpha}) = \partial f(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^\top$. Then

$$\sup_{\alpha\in[1/K_0,K_0]^{q_n}} \|\{\boldsymbol{\Omega}_n^{(1)}/n + \lambda_n\boldsymbol{D}_1(\boldsymbol{\alpha})/n\}\dot{f}(\boldsymbol{\alpha})\| =$$

$$\sup_{\alpha\in[1/K_0,K_0]^{q_n}} 2\lambda_n\|\text{diag}\{f_1(\boldsymbol{\alpha})/\alpha_1^3, \ldots, f_q(\boldsymbol{\alpha})/\alpha_{q_n}^3\}\|/n = o_p(1).$$

According to Condition (D6) and the fact that $\boldsymbol{\alpha} \in [1/K_0, K_0]^{q_n}$, we can derive

$$\begin{aligned}\|\{\boldsymbol{\Omega}_n^{(1)}/n + \lambda_n \boldsymbol{D}_1(\boldsymbol{\alpha})/n\}\dot{f}(\boldsymbol{\alpha})\| &\geq \|\boldsymbol{\Omega}_n^{(1)}\dot{f}(\boldsymbol{\alpha})/n\| - \|\lambda_n \boldsymbol{D}_1(\boldsymbol{\alpha})\dot{f}(\boldsymbol{\alpha})/n\| \\ &\geq (1/c - \lambda_n K_0^2/n)\|\dot{f}(\boldsymbol{\alpha})\|.\end{aligned}$$

Thus we have that $\sup_{\boldsymbol{\alpha}\in[1/K_0,K_0]^{q_n}} \|\dot{f}(\boldsymbol{\alpha})\| \to 0$, which implies that f is a contraction mapping from $[1/K_0, K_0]^{q_n}$ to itself with probability tending to 1. Hence according to the Contraction Mapping Theorem, there exists a unique fixed-point $\hat{\boldsymbol{\alpha}}^* \in [1/K_0, K_0]^{q_n}$ such that

$$\hat{\boldsymbol{\alpha}}^* = \{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\}^{-1}\mathbf{v}_n^{(1)}. \tag{A.26}$$

This concludes the proof of Lemma 4. □

A.2.2 Proof of Theorem 3

First we will prove conclusion (ii). For this, according to the definitions of $\hat{\boldsymbol{\beta}}_{P2}^*$ and $\hat{\boldsymbol{\beta}}_{P2}^{(k)}$, it follows from (A.22) that

$$\hat{\boldsymbol{\beta}}_{P2}^* = \lim_{k\to\infty} \hat{\boldsymbol{\beta}}_{P2}^{(k)} = 0 \tag{A.27}$$

holds with probability tending to 1.

Next we will show that $\Pr(\hat{\boldsymbol{\beta}}_{P1}^* = \hat{\boldsymbol{\alpha}}^*) \to 1$. For this, consider (A.17) and define $\boldsymbol{\gamma}^* = 0$ if $\boldsymbol{\beta}_2 = 0$. Note that for any fixed large n, from Eq. (A.17), we have

$$\lim_{\boldsymbol{\beta}_2\to 0} \boldsymbol{\gamma}^*(\boldsymbol{\beta}) = 0.$$

Furthermore, multiplying by $\{\boldsymbol{\Omega}_n + \lambda_n \boldsymbol{D}(\boldsymbol{\beta})\}$ on both sides of Eq. (A.16), we can get

$$\lim_{\boldsymbol{\beta}_2 \to 0} \boldsymbol{\alpha}^*(\boldsymbol{\beta}) = \{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\boldsymbol{\beta}_1)\}^{-1} \mathbf{v}_n^{(1)} = f(\boldsymbol{\beta}_1). \tag{A.28}$$

By combining (A.27) and (A.28), it follows that, as $k \to \infty$,

$$\eta_k = \sup_{\boldsymbol{\beta}_1 \in [1/K_0, K_0]^{q_n}} \| f(\boldsymbol{\beta}_1) - \boldsymbol{\alpha}^*(\boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_{P2}^{(k)}) \| \to 0. \tag{A.29}$$

Since f is a contract mapping, (A.26) yields

$$\| f(\hat{\boldsymbol{\beta}}_{P1}^{(k)}) - \hat{\boldsymbol{\alpha}}^* \| = \| f(\hat{\boldsymbol{\beta}}_{P1}^{(k)}) - f(\hat{\boldsymbol{\alpha}}^*) \| \leq \frac{1}{c} \| \hat{\boldsymbol{\beta}}_{P1}^{(k)} - \hat{\boldsymbol{\alpha}}^* \| \tag{A.30}$$

with $c > 1$. Let $h_k = \| \hat{\boldsymbol{\beta}}_{P1}^{(k)} - \hat{\boldsymbol{\alpha}}^* \|$. It then follows from (A.29) and (A.30) that

$$h_{k+1} = \| \boldsymbol{\alpha}^*(\hat{\boldsymbol{\beta}}^{(k)}) - \hat{\boldsymbol{\alpha}}^* \| \leq \| \boldsymbol{\alpha}^*(\hat{\boldsymbol{\beta}}^{(k)}) - f(\hat{\boldsymbol{\beta}}_{P1}^{(k)}) \| + \| f(\hat{\boldsymbol{\beta}}_{P1}^{(k)}) - \hat{\boldsymbol{\alpha}}^* \| \leq \eta_k + h_k / c.$$

From (A.29), for any $\epsilon \geq 0$, there exists $N > 0$ such that when $k > N$, $|\eta_k| < \epsilon$. Employing some recursive calculation, we have $h_k \to 0$ as $k \to \infty$. Hence, with probability tending to 1, we have, as $k \to \infty$,

$$\| \hat{\boldsymbol{\beta}}_{P1}^{(k)} - \hat{\boldsymbol{\alpha}}^* \| \to 0.$$

It follows that $\Pr(\hat{\boldsymbol{\beta}}_{P1}^* = \hat{\boldsymbol{\alpha}}^*) \to 1$ holds since $\hat{\boldsymbol{\beta}}_{P1}^{(k)} \to \hat{\boldsymbol{\beta}}_{P1}^*$ as $k \to \infty$ and by the uniqueness of the fixed-point.

Finally, based on (A.26), we have $\sqrt{n}\,(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\beta}_{01}) = \Pi_1 + \Pi_2$, where

$$\Pi_1 = \sqrt{n}\,[\{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\}^{-1}\boldsymbol{\Omega}_n^{(1)} - I_{q_n}]\boldsymbol{\beta}_{01}$$

and

$$\Pi_2 = \sqrt{n}\,\{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\}^{-1}(\mathbf{v}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{01}).$$

It follows from the first order resolvent expansion formula that

$$\{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\}^{-1} = (\boldsymbol{\Omega}_n^{(1)})^{-1} - \lambda_n(\boldsymbol{\Omega}_n^{(1)})^{-1}\boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\{\boldsymbol{\Omega}_n^{(1)} + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\}^{-1}. \quad \text{(A.31)}$$

This yields

$$\Pi_1 = -\frac{\lambda_n}{\sqrt{n}}\,(\boldsymbol{\Omega}_n^{(1)}/n)^{-1}\ \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)\{\boldsymbol{\Omega}_n^{(1)}/n + \lambda_n \boldsymbol{D}_1(\hat{\boldsymbol{\alpha}}^*)/n\}^{-1}\,\boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{01}/n.$$

From Conditions (D5) and (D6), we have, as $n \to \infty$,

$$\|\Pi_1\| = O_p(\lambda_n\sqrt{q_n/n}) \to 0. \quad \text{(A.32)}$$

Furthermore, it follows from (A.31) that

$$\Pi_2 = \sqrt{n}\,\{(\boldsymbol{\Omega}_n^{(1)}/n)^{-1} - o_p(n^{-1/2})\}(\mathbf{v}_n^{(1)}/n - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{01}/n) \quad \text{(A.33)}$$

$$= (\boldsymbol{\Omega}_n^{(1)}/n)^{-1}(\mathbf{v}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{01})/\sqrt{n} + o_p(1), \quad \text{(A.34)}$$

where

$$\sqrt{n}\,(\mathbf{v}_n^{(1)} - \boldsymbol{\Omega}_n^{(1)}\boldsymbol{\beta}_{01}) = \sqrt{n}\sum_{i=1}^{n} U_i^{(1)} \to \mathcal{N}[0, \mathrm{E}\{U_1^{(1)}U_1^{(1)\top}\}]$$

with $U_i^{(1)}$ consisting of the first q_n components of U_i. The covariance $\mathrm{E}\{U_1^{(1)}U_1^{(1)\top}\}$ can be estimated by $\hat{\Xi}_n^{(1)}$, where $\boldsymbol{Z}_i^{*(1)}(t)$ and $\bar{\boldsymbol{Z}}^{*(1)}(t)$ consist of the first q_n components of $\boldsymbol{Z}_i^*(t)$ and $\bar{\boldsymbol{Z}}^*(t)$, respectively. Thus Π_2 converges in distribution to a mean-zero normal distribution whose covariance can be consistently estimated by $\{n^{-1}\boldsymbol{\Omega}_n^{(1)}\}^{-1}\ \hat{\Xi}_n^{(1)}\ \{n^{-1}\boldsymbol{\Omega}_n^{(1)}\}^{-1}$, and the conclusion (iii) of Theorem 3 holds by combining (A.32) and (A.34).

Thus the proof of Theorem 3 is complete. □

A.2.3 Proof of Theorem 4

Let

$$Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_P^{(k)}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{\ell=1}^{p_n} \beta_\ell^2/(\hat{\boldsymbol{\beta}}_{P\ell}^{(k)})^2,$$

and $\hat{\boldsymbol{\varepsilon}}^{(k+1)} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_P^{(k+1)}$, where $\hat{\boldsymbol{\beta}}_P^{(k+1)} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_P^{(k)})$. On the one hand, from $Q(\hat{\boldsymbol{\beta}}_P^{(k+1)}|\hat{\boldsymbol{\beta}}_P^{(k)}) \le Q(0|\hat{\boldsymbol{\beta}}_P^{(k)})$, we have

$$\|\hat{\boldsymbol{\varepsilon}}^{(k+1)}\|^2 + \lambda_n \sum_{\ell=1}^{p} (\hat{\beta}_{P\ell}^{(k+1)})^2/(\hat{\beta}_{P\ell}^{(k)})^2 \le \|\boldsymbol{y}\|^2.$$

Therefore, $\|\hat{\boldsymbol{\varepsilon}}^{(k+1)}\| \le \|\boldsymbol{y}\|$. On the other hand, when $\hat{\beta}_{P\ell} \neq 0$, note that

$$\frac{\partial}{\partial \beta_\ell} Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(k)})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k+1)}} = -2\boldsymbol{x}_\ell^\top \hat{\boldsymbol{\varepsilon}}^{(k+1)} + 2\lambda_n \times \hat{\beta}_{P\ell}^{(k+1)}/(\hat{\beta}_{P\ell}^{(k)})^2 = 0,$$

where $\ell \in \{1, \ldots, p_n\}$. It then follows that

$$\hat{\beta}_{P\ell}^{(k+1)} = (\hat{\beta}_{P\ell}^{(k)})^2 \times \boldsymbol{x}_\ell^\top \hat{\boldsymbol{\varepsilon}}^{(k+1)}/\lambda_n. \tag{A.35}$$

Since $\lim_{k\to\infty}\hat{\beta}_{P\ell}^{(k+1)} = \lim_{k\to\infty}\hat{\beta}_{P\ell}^{(k)} = \hat{\beta}_{P\ell}^{*}$ and by taking the limitation on both sides of (A.35), we have that

$$1/\hat{\beta}_{Pi}^{*} = \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\varepsilon}}^{*}/\lambda_n, \quad 1/\hat{\beta}_{Pj}^{*} = \boldsymbol{x}_j^{\top}\hat{\boldsymbol{\varepsilon}}^{*}/\lambda_n$$

hold with probability tending to 1 for any $i,\ j \in \{1,\ldots,p_n\}$ and $\hat{\beta}_{Pi}^{*} \times \hat{\beta}_{Pj}^{*} \neq 0$, where $\hat{\boldsymbol{\varepsilon}}^{*} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}^{*}$. Therefore

$$|1/\hat{\beta}_{Pi}^{*} - 1/\hat{\beta}_{Pj}^{*}| \leq \|\hat{\boldsymbol{\varepsilon}}^{*}\| \times \|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\lambda_n \leq \|\boldsymbol{y}\|\sqrt{2(1-\rho_{ij})}\lambda_n.$$

This completes the proof of Theorem 4. □

A.3 Proof of Theorem 5

This section consists of three parts. Section A.3.1 gives two lemmas as well as their proofs that will be needed for the proof of Theorem 5. Section A.3.2 sketches the proof of Theorem 5 and Section A.3.3 gives the closed forms of $\hat{\Sigma}_{\theta}\left(\hat{\theta}\right)$, $\hat{B}\left(\hat{\theta},\hat{\gamma}\right)$ and $\hat{D}\left(\hat{\theta},\hat{\gamma}\right)$ associated with the special link functions $g(x) = \log(x)$ and x, respectively.

A.3.1 Two Lemmas to Prove Theorem 5

Lemma 5. *Assume that the regularity conditions (E1)-(E7) hold. Then we have*

$$\begin{aligned}\left.\frac{\partial U_{\theta}\left(\theta;\gamma_0,\hat{\mu}\left(t;\theta,\gamma_0\right)\right)}{\partial\theta}\right|_{\theta=\theta_0} &\to -E\int_0^{\tau}\Delta(t)\frac{V\left(t,\theta_0,\gamma_0\right)}{s^{(0)}\left(t;\gamma_0\right)}\mu_0\left(t\right)\exp\left(\gamma_0' Z\right)d\Lambda_0\left(t\right)\\ &= -B\left(\theta_0,\gamma_0\right)\end{aligned}$$

in probability, where

$$V(t,\theta,\gamma) = E\left[\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta' X(t,t))\}X(t,t)\right.$$
$$\left.\times\left\{X(t,t)-\tilde{x}(t;\theta,\gamma)\right\}'\exp(\theta' X(t,t)+\gamma' Z(t))\right]\nu_0(t).$$

Proof. Let

$$\hat{B}(\theta,\gamma) = -\left.\frac{\partial U_\theta(\theta;\gamma,\hat{\mu}(t;\theta,\gamma))}{\partial\theta}\right|_{\theta=\theta}$$

and

$$\hat{B}(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,dO_i(r)\,\Delta_i(t)\frac{1}{nS_n^{(0)}(t;\gamma)}$$
$$\left[\sum_{l=1}^{n}\int_0^\tau K_h(t-s)\,\Delta_l(t)X_l(t,s)\,\dot{g}\{\hat{\mu}(t;\theta,\gamma)\exp(\theta' X_l(t,s))\}\right.$$
$$\left.\times\left\{\left.\frac{\partial\hat{\mu}(t;\theta,\gamma)}{\partial\theta}\right|_{\theta=\theta}+\hat{\mu}(t;\theta,\gamma)X_l(t,s)\right\}'\exp(\theta' X_l(t,s)+\gamma' Z_l(s))\,dO_l(s)\right]dN_i(t) \tag{A.36}$$

Taking differentiation of $U_\mu(\hat{\mu}(t;\theta,\gamma);\gamma)=0$ on both sides with respect to θ yields

$$\left.\frac{\partial\hat{\mu}(t;\theta,\gamma)}{\partial\theta}\right|_{\theta=\theta} = -\tilde{X}(t;\theta,\gamma)\,\hat{\mu}(t;\theta,\gamma), \tag{A.37}$$

where

$$\tilde{X}(t;\theta,\gamma) = \frac{Q_n^{(1)}(t;\theta,\gamma)}{Q_n^{(0)}(t;\theta,\gamma)}.$$

Combining (A.36) and (A.37), we obtain

$$\begin{aligned}\left.\frac{\partial U_\theta\left(\theta;\gamma,\hat{\mu}\left(t;\theta,\gamma\right)\right)}{\partial\theta}\right|_{\theta=\theta} &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h\left(t-r\right)dO_i\left(r\right)\Delta_i(t)\frac{1}{nS_n^{(0)}\left(t;\gamma\right)} \\ &\times\left[\sum_{l=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_l(t)\dot{g}\{\hat{\mu}\left(t;\theta,\gamma\right)\exp\left(\theta' X_l(t,s)\right)\}X_l\left(t,s\right)\right. \\ &\left.\times\left\{X_l(t,s)-\tilde{X}\left(t;\theta,\gamma\right)\right\}'\exp\left(\theta' X_l(t,s)+\gamma' Z_l(s)\right)dO_l\left(s\right)\right]\hat{\mu}\left(t;\theta,\gamma\right)dN_i\left(t\right).\end{aligned}$$

From Proposition 1 in the supplementary material of Cao et al. (2015a), in a neighborhood of γ_0, $S_n^{(0)}\left(t;\gamma\right)$ converge to $s^{(0)}\left(t;\gamma\right)$ a.s. and uniformly in γ. By a similar argument, in a neighborhood of γ_0 and θ_0, it is not hard to show that $Q_n^{(k)}\left(t;\theta,\gamma\right)$ converges to some nonrandom function $q^{(k)}\left(t;\theta,\gamma\right)$ a.s. and uniformly in θ and γ. Hence, $\tilde{X}\left(t;\theta,\gamma\right)$ also converges to $\tilde{x}\left(t;\theta,\gamma\right)=q^{(1)}\left(t;\theta,\gamma\right)/q^{(0)}\left(t;\theta,\gamma\right)$ a.s. by the continuous mapping theorem. Let

$$\begin{aligned}\hat{V}\left(t,\theta,\gamma\right) &= \frac{1}{n}\sum_{l=1}^{n}\int_0^\tau K_h\left(t-r\right)\Delta_l(t)\dot{g}\{\hat{\mu}\left(t;\theta,\gamma\right)\exp\left(\theta' X_l(t,r)\right)\}X_l\left(t,r\right) \\ &\times\left\{X_l(t,r)-\tilde{X}\left(t;\theta,\gamma\right)\right\}'\exp\left(\theta' X_l(t,r)+\gamma' Z_l(r)\right)dO_l\left(r\right).\end{aligned}$$

By the uniform law of large numbers and uniform convergence of $\hat{\mu}\left(t;\theta_0,\gamma_0\right)$ to

$\mu(t;\theta_0,\gamma_0) \equiv \mu_0(t)$, $\hat{V}(t,\theta_0,\gamma_0)$ converges to

$$\begin{aligned}
V(t,\theta_0,\gamma_0) =& E\left[\int_0^\tau K_h(t-s)\,\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta_0' X(t,s))\}X(t,s)\right. \\
&\left.\times\left\{X(t,s)-\tilde{X}(t;\theta_0,\gamma_0)\right\}'\exp(\theta_0' X(t,s)+\gamma_0' Z(s))\,dO(s)\right] \\
=& E\left[\int_0^\tau K_h(t-s)\,\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta_0' X(t,s))\}X(t,s)\right. \\
&\left.\times\{X(t,s)-\tilde{x}(t;\theta_0,\gamma_0)\}'\exp(\theta_0' X(t,s)+\gamma_0' Z(s))\,\nu_0(s)\right]ds+o_p(1) \\
=& E\left[\int_z K(z)\,\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta_0' X(t,t-hz))\}X(t,t-hz)\,\nu_0(t-hz)\right. \\
&\left.\times\{X(t,t-hz)-\tilde{x}(t;\theta,\gamma)\}'\exp(\theta_0' X(t,t-hz)+\gamma' Z(t-hz))\right]dz+o_p(1).
\end{aligned}$$

By Taylor expansion with respect to h, under conditions (E3) and (E5), we have

$$\begin{aligned}
V(t,\theta_0,\gamma_0) = E\left[\Delta(t)\dot{g}\{\mu_0(t)\exp(\theta_0' X(t,t))\}X(t,t)\right. \\
\left.\{X(t,t)-\tilde{x}(t;\theta_0,\gamma_0)\}'\exp(\theta_0' X(t,t)+\gamma_0' Z(t))\right]\nu_0(t)+O\left(h^2\right).
\end{aligned}$$

Then, by the law of large numbers and conditions (E5) and (E7), $\hat{B}(\theta_0,\gamma_0)$ converges to

$$\begin{aligned}
B(\theta_0,\gamma_0) &= E\left[\sum_{i=1}^n\int_0^\tau\int_0^\infty K_h(t-r)\,dO_i(r)\,\Delta_i(t)\frac{V(t,\theta_0,\gamma_0)}{s^{(0)}(t;\gamma)}\mu_0(t)\,dN_i(t)\right]+O\left(h^2\right) \\
&= E\left[\int_0^\tau \Delta(t)\frac{V(t,\theta_0,\gamma_0)}{s^{(0)}(t;\gamma_0)}\mu_0(t)\exp(\gamma_0' Z(t))\,d\Lambda_0(t)\right]+O\left(h^2\right) \\
&= E\left[\int_0^\tau \Delta(t)\frac{V(t,\theta_0,\gamma_0)}{s^{(0)}(t;\gamma_0)}\mu_0(t)\exp(\gamma_0' Z(t))\,d\Lambda_0(t)\right]+o(1),
\end{aligned}$$

as $n\to\infty$ and $h\to 0$. □

Lemma 6. *Assume that the regularity conditions (E1)-(E7) hold. Then*

$$\left.\frac{\partial U_\theta\left(\theta_0;\gamma,\hat{\mu}\left(t;\theta_0,\gamma\right)\right)}{\partial\gamma}\right|_{\gamma=\gamma_0}$$

converges to

$$D\left(\theta_0,\gamma_0\right)=-E\int_0^\tau\Delta(t)\frac{\nu_0\left(t\right)}{s_n^{(0)}\left(t;\gamma\right)}\left\{X\left(t,t\right)-\frac{q^{(1)}\left(t,\theta_0,\gamma_0\right)}{q^{(0)}\left(t,\theta_0,\gamma_0\right)}\right\}g\{\mu_0\left(t\right)\exp\left(\theta_0'X(t,t)\right)\}$$
$$\times\left\{Z(s)-\frac{s^{(1)}\left(t;\gamma_0\right)}{s^{(0)}\left(t;\gamma_0\right)}\right\}\exp\left(2\gamma_0'Z(t)\right)d\Lambda_0\left(t\right)$$

in probability.

Proof. The derivative of $U_\theta\left(\theta;\gamma,\hat{\mu}\left(t;\theta,\gamma\right)\right)$ with respect to γ is

$$\left.\frac{\partial U_\theta\left(\theta;\gamma,\hat{\mu}\left(t;\theta,\gamma\right)\right)}{\partial\gamma}\right|_{\gamma=\gamma}=\frac{1}{n}\sum_{i=1}^n\int_0^\tau\sum_{k=1}^{K_i}K\left(t-R_{ik}\right)\Delta_i(t)\left[-\left.\frac{\partial\hat{\mu}\left(t;\theta,\gamma\right)}{\partial\gamma}\right|_{\gamma=\gamma}\frac{Q_n^{(1)}\left(t,\theta,\gamma\right)}{nS_n^{(0)}\left(t;\gamma\right)}\right.$$
$$-\frac{1}{nS_n^{(0)}\left(t;\gamma\right)}\left[\sum_{l=1}^n\int_0^\tau K\left(t-s\right)\Delta_l(t)X_l\left(t,s\right)g\{\hat{\mu}\left(t;\theta,\lambda\right)\right.$$
$$\left.\left.\times\exp\left(\theta'X_l(t,s)\right)\}\left\{Z_l(s)-\frac{S_n^{(1)}\left(t;\gamma\right)}{S_n^{(0)}\left(t;\gamma\right)}\right\}\exp\left(\gamma'Z_l(s)\right)dO_l\left(s\right)\right]dN_i\left(t\right).\qquad\text{(A.38)}$$

Similarly to the proof of Lemma 5, take differentiation of $U_\mu\left(\hat{\mu}\left(t;\theta,\gamma\right);\gamma\right)$ with respect to γ, we obtain

$$\left.\frac{\partial\hat{\mu}\left(t;\theta,\gamma\right)}{\partial\gamma}\right|_{\gamma=\gamma}=-\frac{1}{nQ_n^{(0)}\left(t,\theta,\gamma\right)}\left[\sum_{l=1}^n\int_0^\tau K\left(t-s\right)\Delta_l(t)g\{\hat{\mu}\left(t;\theta,\gamma\right)\exp\left(\theta'X_l(t,s)\right)\}\right.$$
$$\left.\times\left\{Z_l(s)-\frac{S_n^{(1)}\left(t;\gamma\right)}{S_n^{(0)}\left(t;\gamma\right)}\right\}\exp\left(\gamma'Z_l(s)\right)dO_l\left(s\right)\right].\qquad\text{(A.39)}$$

Plugging in (A.39) into (A.38), we have

$$\left.\frac{\partial U_\theta\left(\theta;\gamma,\hat{\mu}\left(t;\theta,\gamma\right)\right)}{\partial\gamma}\right|_{\gamma=\gamma} = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \sum_{k=1}^{K_i} K\left(t-R_{ik}\right)\Delta_i(t)\frac{1}{nS_n^{(0)}\left(t;\gamma\right)}$$
$$\times\left[\sum_{l=1}^{n}\int_0^\tau K\left(t-s\right)\Delta_l(t)\left\{X_l\left(t,s\right)-\frac{Q_n^{(1)}\left(t,\theta,\gamma\right)}{Q_n^{(0)}\left(t,\theta,\gamma\right)}\right\}g\{\hat{\mu}\left(t;\theta,\gamma\right)\exp\left(\theta' X_l(t,s)\right)\}\right.$$
$$\left.\times\left\{Z_l(s)-\frac{S_n^{(1)}\left(t;\gamma\right)}{S_n^{(0)}\left(t;\gamma\right)}\right\}\exp\left(\gamma' Z_l(s)\right)dO_l\left(s\right)\right]dN_i\left(t\right).$$

By the convergence of $\hat{\mu}\left(t;\theta_0,\gamma_0\right)$, since $h\to 0$, by a similar argument in the proof of Lemma 5, it is not hard to show that

$$\frac{1}{n}\sum_{l=1}^{n}\int_0^\tau K\left(t-s\right)\Delta_l(t)\left\{X_l\left(t,s\right)-\frac{Q_n^{(1)}\left(t,\theta_0,\gamma_0\right)}{Q_n^{(0)}\left(t,\theta_0,\gamma_0\right)}\right\}g\{\hat{\mu}\left(t;\theta_0,\gamma_0\right)\exp\left(\theta_0' X_l(t,s)\right)\}$$
$$\times\left\{Z_l(s)-\frac{S_n^{(1)}\left(t;\gamma_0\right)}{S_n^{(0)}\left(t;\gamma_0\right)}\right\}\exp\left(\gamma_0' Z_l(s)\right)dO_l\left(s\right)$$

converges to

$$E\left[\Delta(t)\left\{X\left(t,t\right)-\frac{q^{(1)}\left(t,\theta_0,\gamma_0\right)}{q^{(0)}\left(t,\theta_0,\gamma_0\right)}\right\}g\{\mu_0\left(t\right)\exp\left(\theta_0' X(t,t)\right)\}\right.$$
$$\left.\times\left\{Z(t)-\frac{s^{(1)}\left(t;\gamma_0\right)}{s^{(0)}\left(t;\gamma_0\right)}\right\}\exp\left(\gamma_0' Z(t)\right)\right]\nu_0\left(t\right),$$

in probability. Therefore, by the law of large numbers, we can show

$$\begin{aligned}
\left.\frac{\partial U_\theta\left(\theta_0;\gamma,\hat{\mu}\left(t;\theta_0,\gamma\right)\right)}{\partial\gamma}\right|_{\gamma=\gamma_0} =& -E\int_0^\tau\int_0^\infty K\left(t-r\right)dO\left(r\right)\Delta(t)\frac{1}{s_n^{(0)}\left(t;\gamma_0\right)} \\
&\times\left[\left\{X\left(t,t\right)-\frac{q^{(1)}\left(t,\theta_0,\gamma_0\right)}{q^{(0)}\left(t,\theta_0,\gamma_0\right)}\right\}g\{\mu_0\left(t\right)\exp\left(\theta_0'X(t,t)\right)\}\right. \\
&\left.\times\left\{Z(t)-\frac{s^{(1)}\left(t;\gamma_0\right)}{s^{(0)}\left(t;\gamma_0\right)}\right\}\exp\left(\gamma_0'Z(t)\right)\right]dN\left(t\right)+o_p\left(1\right) \\
=& -E\int_0^\tau\Delta(t)\frac{\nu_0\left(t\right)}{s_n^{(0)}\left(t;\gamma\right)}\left\{X\left(t,t\right)-\frac{q^{(1)}\left(t,\theta_0,\gamma_0\right)}{q^{(0)}\left(t,\theta_0,\gamma_0\right)}\right\} \\
&\times g\{\mu_0\left(t\right)\exp\left(\theta_0'X(t,t)\right)\}\left\{Z(s)-\frac{s^{(1)}\left(t;\gamma_0\right)}{s^{(0)}\left(t;\gamma_0\right)}\right\} \\
&\times\exp\left(2\gamma_0'Z(t)\right)d\Lambda_0\left(t\right)+O\left(h^2\right)+o_p\left(1\right).
\end{aligned}$$

This lemma then follows. □

A.3.2 Detailed Proof of Theorem 5

By Taylor expansion with respect to $\hat{\theta}$, we obtain

$$\begin{aligned}
\left(nh\right)^{1/2}U_\theta\left(\hat{\theta};\hat{\gamma},\hat{\mu}\left(t;\hat{\theta},\hat{\gamma}\right)\right) =& \left(nh\right)^{1/2}U_\theta\left(\theta_0;\hat{\gamma},\hat{\mu}\left(t;\theta_0,\hat{\gamma}\right)\right) \\
&+\left(nh\right)^{1/2}\left.\frac{\partial U_\theta\left(\theta;\hat{\gamma},\hat{\mu}\left(t;\theta,\hat{\gamma}\right)\right)}{\partial\theta}\right|_{\theta=\theta^*}\left(\hat{\theta}-\theta\right). \qquad \text{(A.40)}
\end{aligned}$$

where $\theta^*$is a line segment between θ_0 and $\hat{\theta}$. By Lemma 5, the continuous mapping theorem and the consistency of $\hat{\theta}$ and $\hat{\gamma}$,

$$\left.\frac{\partial U_\theta\left(\theta;\hat{\gamma},\hat{\mu}\left(t;\hat{\theta},\hat{\gamma}\right)\right)}{\partial\theta}\right|_{\theta=\theta^*}\rightarrow B\left(\theta_0,\gamma_0\right). \qquad \text{(A.41)}$$

Then, it is necessary to establish the asymptotic distribution of $(nh)^{1/2} U_\theta\left(\theta_0; \hat{\gamma}, \hat{\mu}\left(t; \theta_0, \hat{\gamma}\right)\right)$. By Taylor expansion with respect to $\hat{\gamma}$,

$$
\begin{aligned}
& (nh)^{1/2} U_\theta\left(\theta_0; \hat{\gamma}, \hat{\mu}\left(t; \theta_0, \hat{\gamma}\right)\right) \\
= & (nh)^{1/2} U_\theta\left(\theta_0; \gamma_0, \hat{\mu}\left(t; \theta_0, \gamma_0\right)\right) + (nh)^{1/2} \left.\frac{\partial U_\theta\left(\theta_0; \gamma, \hat{\mu}\left(t; \theta_0, \gamma\right)\right)}{\partial \gamma}\right|_{\gamma=\gamma^*} \left(\hat{\gamma} - \gamma_0\right) \\
= & (nh)^{1/2} U_\theta\left(\theta_0; \gamma_0, \hat{\mu}\left(t; \theta_0, \gamma_0\right)\right) \\
& + (nh)^{1/2} \left.\frac{\partial U_\theta\left(\theta_0; \gamma_0, \hat{\mu}\left(t; \theta_0, \gamma\right)\right)}{\partial \gamma}\right|_{\gamma=\gamma^*} \hat{A}^{-1}\left(\gamma_0\right) \\
& \times \left[\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \int_0^\infty K_h(t-r)\{Z_i(r) - \frac{S_n^{(1)}\left(t, \gamma_0\right)}{S_n^{(0)}\left(t, \gamma_0\right)}\} dO_i\left(r\right) dN_i\left(t\right)\right]
\end{aligned} \tag{A.42}
$$

where γ^* is a line segment between γ and $\hat{\gamma}$. The last equationcomes from Theorem 2 of Cao et al. (2015a).

By Lemma 6 and the consistency of $\hat{\gamma}$ to γ_0,

$$
\left.\frac{\partial U_\theta\left(\theta_0; \gamma, \hat{\mu}\left(t; \theta_0, \gamma\right)\right)}{\partial \gamma}\right|_{\gamma=\gamma^*} \rightarrow D\left(\theta_0, \gamma_0\right) \tag{A.43}
$$

in probability.

It is then necessary to deal with $U_\theta\left(\theta_0; \gamma_0, \hat{\mu}\left(t; \theta_0, \gamma_0\right)\right)$. Take the linear expansion of $g\{\hat{\mu}\left(t; \theta_0, \gamma_0\right) \exp\left(\theta_0' X_l(t, s)\right)\}$ at $\mu_0\left(t\right)$ in $U_\theta\left(\theta_0; \gamma_0, \hat{\mu}\left(t; \theta_0, \gamma_0\right)\right)$, under condition

(E5), we have

$$
\begin{aligned}
& U_\theta\left(\theta_0;\gamma_0,\hat{\mu}\left(t;\theta_0,\gamma_0\right)\right) \\
=&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\int_0^\tau K_h\left(t-R_{ik}\right)\Delta_i(t)\left[X_i\left(t,R_{ik}\right)Y_i\left(t\right)-\frac{Q_n^{(1)}\left(t;\theta_0,\gamma_0\right)}{S_n^{(0)}\left(t;\gamma_0\right)}\right]dN_i\left(t\right) \\
=&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\int_0^\tau K_h\left(t-R_{ik}\right)\Delta_i(t)\left[X_i\left(t,R_{ik}\right)Y_i\left(t\right)\right. \\
&-\frac{\sum_{l=1}^{n}\int_0^\tau K\left(t-s\right)\Delta_l(t)X_l\left(t,s\right)g\{\mu_0\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\gamma_0'Z_l(s)\right)dO_l\left(s\right)}{S_n^{(0)}\left(t;\gamma_0\right)} \\
&-\frac{\hat{\mu}\left(t;\theta_0,\gamma_0\right)-\mu_0\left(t\right)}{S_n^{(0)}\left(t;\gamma_0\right)}\left[\sum_{l=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_l(t)X_l\left(t,s\right)\right. \\
&\left.\left.\times\ \dot{g}\{\mu^*\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\theta_0'X_l(t,s)+\gamma_0'Z_l(s)\right)dO_l\left(s\right)\right]dN_i\left(t\right), \qquad \text{(A.44)}
\end{aligned}
$$

where $\mu^*\left(t\right)$ lies between $\hat{\mu}\left(t;\theta_0,\gamma_0\right)$ and $\mu_0\left(t\right)$.

Similarly, take the linear expansion of $g\{\hat{\mu}\left(t;\theta_0,\gamma_0\right)\exp\left(\theta_0'X_l(t,s)\right)\}$ at $\mu_0\left(t\right)$ in $U_\mu\left(\hat{\mu}\left(t;\theta_0,\gamma_0\right);\gamma_0\right)$,

$$
\begin{aligned}
U_\mu\left(\hat{\mu}\left(t;\theta_0,\gamma_0\right);\gamma_0\right)=&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_i}K_h\left(t-R_{ik}\right)\Delta_i(t)\left[Y_i\left(t\right)\right. \\
&-\frac{\sum_{l=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_l(t)g\{\mu_0\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\gamma_0'Z_l(s)\right)dO_l\left(s\right)}{S_n^{(0)}\left(t;\gamma_0\right)} \\
&-\frac{\hat{\mu}\left(t;\theta_0,\gamma_0\right)-\mu_0\left(t\right)}{S_n^{(0)}\left(t;\gamma_0\right)} \\
&\left.\times\sum_{l=1}^{n}\int_0^\tau K_h\left(t-s\right)\Delta_l(t)\dot{g}\{\mu^{**}\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\theta_0'X_l(t,s)+\gamma_0'Z_l(s)\right)dO_l\left(s\right)\right] \\
&\times dN_i\left(t\right). \qquad \text{(A.45)}
\end{aligned}
$$

From (A.45) and some algebra, we have

$$\hat{\mu}\left(t;\theta_0,\gamma_0\right)-\mu_0\left(t\right)=$$
$$\left\{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)\left[Y_l\left(t\right)-g\{\mu_0\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\right]\exp\left(\gamma_0'Z_l(s)\right)dO_l\left(s\right)\right\}$$
$$/\left\{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)\dot{g}\{\mu^{**}\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\theta_0'X_l(t,s)+\gamma_0'Z_l(s)\right)dO_l\left(s\right)\right\} \tag{A.46}$$

where $\mu^{**}\left(t\right)$ lies between $\hat{\mu}\left(t;\theta_0,\gamma_0\right)$ and $\mu_0\left(t\right)$.

Then, combining (A.44) and (A.46), we obtain

$$\begin{aligned}
&\left(nh\right)^{1/2}U_\theta\left(\theta_0;\gamma_0,\hat{\mu}\left(t;\theta_0,\gamma_0\right)\right)\\
=&\left(nh\right)^{1/2}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\int_0^{\tau}K\left(t-R_{ik}\right)\Delta_i(t)\left[X_i\left(t,R_{ik}\right)Y_i\left(t\right)\right.\\
&-\frac{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)X_l\left(t,s\right)g\{\mu_0\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\gamma_0'Z_l(s)\right)dO_l\left(s\right)}{\sum_{l=1}^{n}\int_0^{\tau}K_h(t-s)\Delta_l(t)\exp\{\gamma_0'Z_l(s)\}dO_l\left(s\right)}\\
&-\left\{\frac{\sum_{l=1}^{n}\sum_{k=1}^{K_l}K_h\left(t-R_{lk}\right)\Delta_l(t)Y_l\left(t\right)dN_l\left(t\right)}{\sum_{l=1}^{n}\sum_{k=1}^{K_l}K_h\left(t-R_{lk}\right)\Delta_l(t)dN_l\left(t\right)}\right.\\
&\left.-\frac{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)g\{\mu_0\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\gamma_0'Z_l(s)\right)dO_l\left(s\right)}{\sum_{l=1}^{n}\int_0^{\tau}K_h(t-s)\Delta_l(t)\exp\{\gamma_0'Z_l(s)\}dO_l\left(s\right)}\right\}\\
&\left.\times\frac{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)X_l\left(t,s\right)\dot{g}\{\mu^{*}\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\theta_0'X_l(t,s)+\gamma_0'Z_l(s)\right)dO_l\left(s\right)}{\sum_{l=1}^{n}\int_0^{\tau}K_h\left(t-s\right)\Delta_l(t)\dot{g}\{\mu^{**}\left(t\right)\exp\left(\theta_0'X_l(t,s)\right)\}\exp\left(\theta_0'X_l(t,s)+\gamma_0'Z_l(s)\right)dO_l\left(s\right)}\right]\\
&\times dN_i\left(t\right)\\
&+o_p\left(\left(nh\right)^{1/2}\right)\\
=&\left(nh\right)^{1/2}\tilde{U}_\theta\left(\theta_0;\gamma_0\right)+o_p\left(\left(nh\right)^{1/2}\right),
\end{aligned} \tag{A.47}$$

where

$$\tilde{U}_\theta(\theta_0;\gamma_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)\,[\{X_i(t,r)-\tilde{x}(t)\}\,Y_i(t)$$
$$-\frac{\sum_{l=1}^{n}\int_0^\tau K_h(t-s)\,\Delta_l(t)\,\{X_l(t,s)-\tilde{x}(t)\}\,g\{\mu_0(t)\exp(\theta_0' X_l(t,s))+\gamma_0' Z_l(s)\}dO_l(s)}{\sum_{l=1}^{n}\int_0^\tau K_h(t-s)\Delta_l(t)\exp\{\gamma_0' Z_l(s)\}dO_l(s)}\Bigg]$$
$$\times dO_i(r)\,dN_i(t).$$

The last equation comes from the uniform convergence of $\hat{\mu}(t;\theta_0,\gamma_0)$ to $\mu_0(t)$ and

$$(nh)^{1/2}\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)\left[X_i(t,r)\,Y_i(t)-\frac{Q_n^{(1)}(t;\theta_0,\gamma_0)}{S_n^{(1)}(t;\gamma_0)}\right]dO_i(r)\,dN_i(t)$$

is $o_p\left((nh)^{1/2}\right)$ under condition (E5).

Plugging (A.43) and (A.47) into (A.42), we have,

$$\begin{aligned}
&(nh)^{1/2}\,U_\theta(\theta_0;\hat{\gamma},\hat{\mu}(t;\theta_0,\hat{\gamma}))\\
&=(nh)^{1/2}\,\tilde{U}_\theta(\theta_0;\gamma_0)+(nh)^{1/2}\,D(\theta_0,\gamma_0)\,\hat{A}^{-1}(\gamma_0)\\
&\quad\times\left(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\Delta_i(t)\left\{Z_i(r)-\frac{S_n^{(1)}(t,\gamma_0)}{S_n^{(0)}(t,\gamma_0)}\right\}dO_i(r)\,dN_i(t)\right)+o_p\left((nh)^{1/2}\right)\\
&=(nh)^{1/2}\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)\left[X_i(t,r)\,Y_i(t)-\frac{Q_n^{(1)}(t)}{S_n^{(0)}(t,\gamma_0)}\right.\\
&\quad\left.+\,D(\theta_0,\gamma_0)\,\hat{A}^{-1}(\gamma_0)\left\{Z_i(r)-\frac{S_n^{(1)}(t,\gamma_0)}{S_n^{(0)}(t,\gamma_0)}\right\}\right]dO_i(r)\,dN_i(t)+o_p\left((nh)^{1/2}\right)
\end{aligned}$$

Next we need to derive the variance of $(nh)^{1/2}\,U_\theta(\theta_0;\hat{\gamma},\hat{\mu}(t;\theta_0,\hat{\gamma}))$. It is calculated

as

$$\begin{aligned}\Sigma_\theta(\theta_0) =& var\left[(nh)^{1/2} U_\theta(\theta_0; \hat{\gamma}, \hat{\mu}(t; \theta_0, \hat{\gamma}))\right] \\ =& nhE\left[U_\theta(\theta_0; \hat{\gamma}, \hat{\mu}(t; \theta_0, \hat{\gamma}))^{\otimes 2}\right] - nhE\left[U_\theta(\theta_0; \hat{\gamma}, \hat{\mu}(t; \theta_0, \hat{\gamma}))\right]^{\otimes 2}. \end{aligned} \tag{A.48}$$

For the first term on the right hand side of (A.48), it can be shown that, by the similar arguments in Cao et al. (2015a) and Cao et al. (2015b) and the convergence of $\hat{\gamma}$, under (E3) and (E5)

$$\begin{aligned}& nhE\left[U_\theta(\theta_0; \hat{\gamma}, \hat{\mu}(t; \theta_0, \hat{\gamma}))^{\otimes 2}\right] \\ =& hE\left[\left\{\int_0^\tau \int_0^\infty K_h(t-r)\Delta(t)\left[X(t,r)Y(t) - \frac{Q_n^{(1)}(t; \theta_0, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right.\right.\right. \\ & \left.\left.\left. + D(\theta_0, \gamma_0)\hat{A}^{-1}(\gamma_0)\left\{Z(r) - \frac{S_n^{(1)}(t, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right\}\right] dO(r) dN(t)\right\}^{\otimes 2}\right] + O(h) + o_p(1) \\ =& hE\left[\int_0^\tau \int_0^\infty K_h(t-r)\Delta(t)\left\{X(t,r)Y(t) - \frac{Q_n^{(1)}(t; \theta_0, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right.\right. \\ & \left.\left. + D(\theta_0, \gamma_0)\hat{A}^{-1}(\gamma_0)\left\{Z(r) - \frac{S_n^{(1)}(t, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right\}\right\}^{\otimes 2} \exp(\gamma_0' Z(t))\right] \nu_0(r) dr d\Lambda_0(t) \\ & + O(h) + o_p(1) \\ =& \int_0^\tau \int_z K^2(z) dz E\left[\int_0^\tau \Delta(t)\left\{X(t, t-hz)Y(t) - \frac{Q_n^{(1)}(t; \theta_0, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right.\right. \\ & \left.\left. + D(\theta_0, \gamma_0)\hat{A}^{-1}(\gamma_0)\left\{Z(t-hz) - \frac{S_n^{(1)}(t, \gamma_0)}{S_n^{(0)}(t, \gamma_0)}\right\}\right\}^{\otimes 2} \exp(\gamma_0' Z(t))\right] \\ & \times \nu_0(t-hz) d\Lambda_0(t) + O(h) + o_p(1) \\ =& \int_0^\tau \int_z K^2(z) dz\end{aligned}$$

$$\times E\left[\left\{X\left(t,t\right)Y\left(t\right)-\frac{Q_{n}^{(1)}\left(t;\theta_{0},\gamma_{0}\right)}{S_{n}^{(0)}\left(t,\gamma_{0}\right)}+D\left(\theta_{0},\gamma_{0}\right)\hat{A}^{-1}\left(\gamma_{0}\right)\left\{Z(t)-\frac{S_{n}^{(1)}\left(t,\gamma_{0}\right)}{S_{n}^{(0)}\left(t,\gamma_{0}\right)}\right\}\right\}^{\otimes2}\right.$$

$$\left.\times\,\Delta(t)\exp\left(\gamma_{0}'Z(t)\right)\right]\nu_{0}\left(t\right)d\Lambda_{0}\left(t\right)+O\left(h\right)+o_{p}\left(1\right)$$

$$=\int_{0}^{\tau}\int_{z}K^{2}\left(z\right)dzE\left[\Delta(t)\exp\left(\gamma_{0}'Z(t)\right)\left\{X\left(t,t\right)Y\left(t\right)-\frac{q^{(1)}\left(t;\theta_{0},\gamma_{0}\right)}{s^{(0)}\left(t,\gamma_{0}\right)}\right.\right.$$

$$\left.\left.+\;D\left(\theta_{0},\gamma_{0}\right)A^{-1}\left(\gamma_{0}\right)\left\{Z(t)-\frac{s^{(1)}\left(t,\gamma_{0}\right)}{s^{(0)}\left(t,\gamma_{0}\right)}\right\}\right\}^{\otimes2}\right]\nu_{0}\left(t\right)d\Lambda_{0}\left(t\right)+O\left(h\right)+o_{p}\left(1\right)$$

For the second term on the right hand side of (A.48),

$$nhE\left[U_{\theta}\left(\theta_{0};\hat{\gamma},\hat{\mu}\left(t;\theta_{0},\hat{\gamma}\right)\right)\right]^{\otimes2}=o_{p}\left(1\right),$$

because $U_{\theta}\left(\theta_{0};\hat{\gamma},\hat{\mu}\left(t;\theta_{0},\hat{\gamma}\right)\right)=o_{p}\left(\left(nh\right)^{-1/2}\right)$. Since $h\to0$ as $n\to\infty$ and $\int_{z}K^{2}\left(z\right)dz<\infty$,

$$\Sigma_{\theta}\left(\theta_{0}\right)=\int_{z}K^{2}\left(z\right)dz\int_{0}^{\tau}E\left[\Delta(t)\exp\left(\gamma_{0}'Z(t)\right)\right.$$

$$\left.\times\left\{X\left(t,t\right)Y\left(t\right)-\frac{q^{(1)}\left(t;\theta_{0},\gamma_{0}\right)}{s^{(0)}\left(t,\gamma_{0}\right)}+D\left(\theta_{0},\gamma_{0}\right)A^{-1}\left(\gamma_{0}\right)\left\{Z(t)-\frac{s^{(1)}\left(t,\gamma_{0}\right)}{s^{(0)}\left(t,\gamma_{0}\right)}\right\}\right\}^{\otimes2}\right]$$

$$\times\nu_{0}\left(t\right)d\Lambda_{0}\left(t\right)+o_{p}\left(1\right).\tag{A.49}$$

The last step is to verify that Lyapunov condition holds. Define

$$H_{i}=\left(nh\right)^{1/2}\frac{1}{n}\int_{0}^{\tau}\int_{0}^{\infty}K_{h}\left(t-r\right)\Delta(t)\left[X_{i}\left(t,r\right)Y_{i}\left(t\right)-\frac{Q_{n}^{(1)}\left(t;\theta_{0},\gamma_{0}\right)}{S_{n}^{(0)}\left(t,\gamma_{0}\right)}\right.$$

$$\left.+\;D\left(\theta_{0},\gamma_{0}\right)\hat{A}^{-1}\left(\gamma_{0}\right)\left\{Z_{i}(r)-\frac{S_{n}^{(1)}\left(t,\gamma_{0}\right)}{S_{n}^{(0)}\left(t,\gamma_{0}\right)}\right\}\right]dO_{i}\left(r\right)dN_{i}\left(t\right)\Bigg].$$

It is not hard to show that $E\left[H_{i}\right]=o\left(\left(h/n\right)^{1/2}\right)$. From (A.49), $E\left[H_{1}^{2}\right]=O\left(\left(nh\right)^{-1}\right)$.

To verify Lyapunov condition, it is seen that

$$\begin{aligned}\sum_{i=1}^{n} E\left[\left(H_i - E\left[H_i\right]\right)^3\right] &= n\left\{E\left[H_1^3\right] - 3E\left[H_1^2\right]E\left[H_1\right] + 4E\left[H_1\right]^3\right\} \\ &= n\left\{E\left[H_1^3\right] - 3O\left((nh)^{-1}\right)o\left((h/n)^{1/2}\right) + 4o\left((h/n)^{3/2}\right)\right\} \\ &= nE\left[H_1^3\right] + o\left(1\right). \end{aligned} \tag{A.50}$$

Similar to calculating $\Sigma_\theta\left(\theta_0\right)$, due to the fact that $\int_z K^3\left(z\right)dz = 0$ induced by the symmetry of $K(\cdot)$, we have

$$E\left[H_1^3\right] = O\left((nh)^{3/2}n^{-3}h^{-3}h\right) = O\left(n^{-3/2}h^{-1/2}\right). \tag{A.51}$$

Combining (A.50) and (A.51), we obtain

$$\sum_{i=1}^{n} E\left[\left(H_i - E\left[H_i\right]\right)^3\right] = O\left((nh)^{-1/2}\right) \to 0$$

as $n \to \infty$.

Therefore,

$$\left(nh\right)^{1/2} U_\theta\left(\theta_0; \hat{\gamma}, \hat{\mu}\left(t; \theta_0, \hat{\gamma}\right)\right) \to N\left(0, \Sigma_\theta\left(\theta_0\right)\right). \tag{A.52}$$

Theorem 5 then follows by combining (A.40), (A.41) and (A.52).

A.3.3 Special Forms of $\hat{\Sigma}_\theta\left(\hat{\theta}\right)$, $\hat{B}\left(\hat{\theta},\hat{\gamma}\right)$ and $\hat{D}\left(\hat{\theta},\hat{\gamma}\right)$

For the linear case where $g\left(x\right)=\log\left(x\right)$,

$$\hat{\Sigma}_\theta\left(\hat{\theta}\right)=\frac{1}{n^2}\sum_{i=1}^{n}\left[\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})\left[Y_i\left(T_{ij}\right)\left\{X_i\left(T_{ij},R_{ik}\right)-\bar{X}\left(T_{ij}\right)\right\}\right.\right.$$
$$\left.-\left\{\overline{X\left(T_{ij}\right)^{\otimes 2}}-\bar{X}\left(T_{ij}\right)^{\otimes 2}\right\}\hat{\theta}\right]$$
$$\left.-\hat{D}\left(\hat{\theta},\hat{\gamma}\right)\hat{A}^{-1}\int_0^\tau\int_0^\tau K_h(u-s)\{Z_i(s)-\bar{Z}(u;\gamma)\}dO_i\left(s\right)dN_i(u)\right]^{\otimes 2},$$

$$\hat{B}\left(\hat{\theta},\hat{\gamma}\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})Y_i\left(T_{ij}\right)\left\{\overline{X\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}}-\bar{X}\left(T_{ij};\hat{\gamma}\right)^{\otimes 2}\right\},$$

and

$$\hat{D}\left(\hat{\theta},\hat{\gamma}\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\Delta_i(T_{ij})Y_i\left(T_{ij}\right)$$
$$\times\left[-\frac{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)X_l\left(T_{ij},R_{lu}\right)Z_l\left(R_{lu}\right)'\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}\right.$$
$$\left.+\bar{X}\left(T_{ij}\right)\bar{Z}'\left(T_{ij}\right)\right]-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)$$
$$\times\left\{\frac{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)X_l\left(T_{ij},R_{lu}\right)^{\otimes 2}\theta Z_l'\left(R_{lu}\right)\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}\right.$$
$$\left.-\frac{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)X_l\left(T_{ij},R_{lu}\right)^{\otimes 2}\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}{\sum_{l=1}^{n}\sum_{u=1}^{K_l}K_h\left(T_{ij}-R_{lu}\right)I\left(C_l\geq T_{ij}\right)\exp\left(\hat{\gamma}'Z_l\left(R_{lu}\right)\right)}\theta\bar{Z}'\left(T_{ij}\right)\right\}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K_i}K_h\left(T_{ij}-R_{ik}\right)\times\left\{-2\theta'\bar{X}\left(T_{ij}\right)\frac{\partial\bar{X}'\left(T_{ij}\right)}{\partial\gamma}\right\}.$$

For the proportional case where $g(x) = x$,

$$\hat{\Sigma}_\theta\left(\hat{\theta}\right) = n^{-1}\sum_{i=1}^{n}\left[\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)Y_i(t)\left[X_i(t,r)\right.\right.$$
$$\left.-\frac{\sum_{l=1}^n\int_0^\tau K_h(t-s)\,\Delta_l(t)X_l(t,s)\exp\left(\theta' X_l(t,s)+\gamma' Z_l(s)\right)dO_l(s)}{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)\exp\left(\theta' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}\right]dO_i(r)\,dN_i(t)$$
$$\left.+\hat{D}\left(\hat{\theta},\hat{\gamma}\right)\hat{A}^{-1}(\hat{\gamma})\int_0^\tau\int_0^\tau K_h(t-r)\{Z_i(r)-\bar{Z}(t;\hat{\gamma})\}dO_i(r)\,dN_i(t)\right]^{\otimes 2},$$

$$\hat{B}\left(\hat{\theta},\hat{\gamma}\right) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)Y_i(t)$$
$$\times\left[\frac{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)X_j^{\otimes 2}(t,s)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}\right.$$
$$\left.-\left\{\frac{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)X_j(t,s)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}\right\}^{\otimes 2}\right]$$
$$\times dO_i(r)\,dN_i(t)$$

and

$$\hat{D}\left(\hat{\theta},\hat{\gamma}\right) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\infty K_h(t-r)\,\Delta_i(t)Y_i(t)$$
$$\times\left[\frac{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)X_j(t,s)\,Z_j'(s)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}\right.$$
$$\left.-\frac{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)X_j(t,s)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}{\sum_{j=1}^n\int_0^\tau K_h(t-s)\,\Delta_j(t)\exp\left(\hat{\theta}' X_j(t,s)+\hat{\gamma}' Z_j(s)\right)dO_j(s)}\right.$$

$$\times \left\{ \frac{\sum_{j=1}^{n} \int_0^\tau K_h\left(t-s\right) \Delta_j(t) Z_j\left(s\right) \exp\left(\hat{\theta}' X_j\left(t,s\right) + \hat{\gamma}' Z_j\left(s\right)\right) dO_j\left(s\right)}{\sum_{j=1}^{n} \int_0^\tau K_h\left(t-s\right) \Delta_j(t) \exp\left(\hat{\theta}' X_j\left(t,s\right) + \hat{\gamma}' Z_j\left(s\right)\right) dO_j\left(s\right)} \right\}' \Bigg]$$

$$\times dO_i\left(r\right) dN_i\left(t\right).$$

Bibliography

Andersen, P. K. (1997). *Statistical Models Based on Counting Processes*. Springer.

Balakrishnan, N., & Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *The Annals of Statistics*, *37*(3), 1112–1149.

Cai, J., & Schaubel, D. E. (2004). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis*, *10*(2), 121–138.

Cao, H., Churpek, M. M., Zeng, D., & Fine, J. P. (2015a). Analysis of the proportional hazards model with sparse longitudinal covariates. *Journal of the American Statistical Association*, *110*(511), 1187–1196.

Cao, H., Zeng, D., & Fine, J. P. (2015b). Regression analysis of sparse asynchronous longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(4), 755–776.

Cao, H., Li, J., & Fine, J. P. (2016). On last observation carried forward and asynchronous longitudinal regression analysis. *Electronic Journal of Statistics*, *10*(1), 1155–1180.

Chen, L., & Cao, H. (2017). Analysis of asynchronous longitudinal data with partially linear models. *Electronic Journal of Statistics*, *11*(1), 1549–1569.

Chen, X., & Wang, Q. (2013). Variable selection in the additive rate model for recurrent event data. *Computational Statistics & Data Analysis*, *57*(1), 491–503.

Cook, R. J., & Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer-Verlag New York.

Dai, L., Chen, K., Sun, Z., Liu, Z., & Li, G. (2018). Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis*, *168*, 334–351.

Dicker, L., Huang, B., & Lin, X. (2013). Variable selection and estimation with the seamless-L_0 penalty models. *Statistica Sinica*, *23*, 929–962.

Diggle, P., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74–99.

Fleming, T. R., & Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Frommlet, F., & Nuel, G. (2016). An adaptive ridge procedure for L_0 regularization. *PLOS ONE*, *11*(2), e0148620.

Gorodnitsky, I., & Rao, B. (1997). Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, *45*(3), 600–616.

Han, M., Song, X., & Sun, L. (2014). Joint modeling of longitudinal data with informative observation times and dropouts. *Statistica Sinica*, *24*(4), 1487–1504.

Hand, D. J., & Crowder, M. J. (1996). *Practical Longitudinal Data Analysis*. London New York: Chapman & Hall/CRC.

He, X., Tong, X., & Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis*, *15*(2), 177–196.

Hu, X. J., Lagakos, S. W., & Lockhart, R. A. (2009). Marginal analysis of panel counts through estimating functions. *Biometrika*, *96*(2), 445–456.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*.

Lawless, J. F., & Nadeau, C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events. *Technometrics*, *37*(2), 158–168.

Lawson, C. (1961). *Contributions to the theory of linear least maximum approximation* (Doctoral dissertation, University of California, Los Angeles).

Li, N., Sun, L., & Sun, J. (2010). Semiparametric transformation models for panel count data with dependent observation process. *Statistics in Biosciences*, *2*(2), 191–210.

Li, N., Zhao, H., & Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Statistics in Medicine*, *32*(17), 3039–3054.

Lin, D. Y., & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, *81*(1), 61–71.

Lin, D. Y., Wei, L. J., Yang, I., & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 711–730.

Lin, D. Y., Wei, L. J., & Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association*, *96*(454), 620–628.

Lin, H., Scharfstein, D. O., & Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 791–813.

Liu, Z., & Li, G. (2016). Efficient regularized regression with L_0 penalty for variable selection and network construction. *Computational and Mathematical Methods in Medicine*, *2016*, Article ID 3456153.

Mazumder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*(495), 1125–1138.

Qu, L., Sun, L., & Song, X. (2018). A joint modeling approach for longitudinal data with informative observation times and a terminal event. *Statistics in Biosciences*, *10*(3), 609–633.

Schaubel, D. E., Zeng, D., & Cai, J. (2006). A semiparametric additive rates model for recurrent event data. *Lifetime Data Analysis*, *12*(4), 389–406.

Song, X., Mu, X., & Sun, L. (2012). Regression analysis of longitudinal data with time-dependent covariates and informative observation times. *Scandinavian Journal of Statistics*, *39*(2), 248–258.

Sun, J., & Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. Springer-Verlag New York.

Sun, J., Park, D.-H., Sun, L., & Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, *100*(471), 882–889.

Sun, J., Sun, L., & Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association*, *102*(480), 1397–1406.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.

Tong, X., He, X., Sun, L., & Sun, J. (2009a). Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics*, *36*(4), 620–635.

Tong, X., Zhu, L., & Sun, J. (2009b). Variable selection for recurrent event data via nonconcave penalized estimating function. *Lifetime Data Analysis*, *15*(2), 197–215.

Wohl, D. A., Zeng, D., Stewart, P., Glomb, N., Alcorn, T., Jones, S., ... van der Horst, C. (2005). Cytomegalovirus viremia, mortality, and end-organ disease among patients with AIDS receiving potent antiretroviral therapies. *Journal of Acquired Immune Deficiency Syndromes*, *38*(5), 538–544.

Zhang, H., Sun, J., & Wang, D. (2013). Variable selection and estimation for multivariate panel count data via the seamless-L_0 penalty. *Canadian Journal of Statistics*, *41*(2), 368–385.

Zhao, H., Li, Y., & Sun, J. (2013). Semiparametric analysis of multivariate panel count data with dependent observation processes and a terminal event. *Journal of Nonparametric Statistics*, *25*(2), 379–394.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

VITA

Dayu Sun received his Bachelor of Science in 2013 and Master of Philosophy in 2016 from The Hong Kong Polytechnic University. Then he joined the Ph.D. program in the Department of Statistics at the University of Missouri-Columbia in August 2016.