Survival Analysis of Complex Featured Data with Measurement Error

by

Li-Pang Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Gang Li Professor, Department of Biostatistics, School of Public Health University of California Los Angeles
Supervisor(s):	Grace Yi Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Richard Cook Professor, Department of Statistics and Actuarial Science, University of Waterloo
	Mary Thompson Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Shai Ben-David Professor, Department of Computer Science, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Survival analysis plays an important role in many fields, such as cancer research, clinical trials, epidemiological studies, actuarial science, and so on. A large body of methods on analyzing survival data have been developed. However, many important problems have still not been fully explored. In this thesis, we focus on the analysis of survival data with complex features.

In Chapter 1, we review relevant topics including survival analysis, the measurement error model, the graphical model, and variable selection.

Graphical models are useful in characterizing the dependence structure of variables. They have been commonly used for analysis of high-dimensional data, including genetic data and data with network structures. Many estimation procedures have been developed under various graphical models with a stringent assumption that the associated variables must be measured precisely. In applications, this assumption, however, is often unrealistic and mismeasurement in variables is usually presented in data. In Chapter 2, we investigate the high-dimensional graphical model with error-prone variables. We propose valid estimation procedures to account for measurement error effects. Theoretical results are established for the proposed methods and numerical studies are reported to assess the performance of our proposed methods.

In Chapter 3, we consider survival analysis with network structures and measurement error in covariates. In survival data analysis, the Cox proportional hazards (PH) model is perhaps the most widely used model to feature the dependence of survival times on covariates. While many inference methods have been developed under such a model or its variants, those models are not adequate for handling data with complex structured covariates. High-dimensional survival data often entail several features: (1) many covariates are inactive in explaining the survival information, (2) active covariates are associated in a network structure, and (3) some covariates are error-contaminated. To hand such kinds of survival data, we propose graphical proportional hazards measurement error models, and develop inferential procedures for the parameters of interest. Our proposed models significantly enlarge the scope of the usual Cox PH model and have great flexibility in characterizing survival data. Theoretical results are established to justify the proposed methods. Numerical studies are conducted to assess the performance of the proposed methods.

In Chapter 4, we focus on sufficient dimension reduction for high-dimensional survival data with covariate measurement error. Sufficient dimension reduction (SDR) is an important tool in regression analysis which reduces the dimension of covariates without losing

predictive information. Several methods have been proposed to handle data with either censoring in the response or measurement error in covariates. However, little research is available to deal with data having these two features simultaneously. Moreover, the analysis becomes more challenging when data contain ultrahigh-dimensional covariates. In Chapter 4, we examine this problem. We start with considering the cumulative distribution function in regular settings and propose a valid SDR method to incorporate the effects of both censored data and covariates measurement error. Next, we extend the proposed method to handle ultrahigh-dimensional data. Theoretical results of the proposed methods are established. Numerical studies are reported to assess the performance of the proposed methods.

In Chapter 5, we slightly switch our attention to examine sampling issues concerning survival data. Specifically, we discuss survival analysis for left-truncated and right-censored data with covariate measurement error. Many methods have been developed for analyzing survival data which commonly involve right-censoring. These methods, however, are challenged by complex features pertinent to the data collection as well as the nature of data themselves. Typically, biased samples caused by left-truncation or length-biased sampling and measurement error are often accompanying with survival analysis. While such data frequently arise in practice, little work has been available in the literature. In Chapter 5, we study this important problem and explore valid inference methods for handling left-truncated and right-censored survival data with measurement error under the widely used Cox model. We exploit a flexible estimator for the survival model parameters which does not require specification of the baseline hazard function. To improve the efficiency, we further develop an augmented non-parametric maximum likelihood estimator. We establish asymptotic results for the proposed estimators and examine the efficiency and robustness issues of the proposed estimators. The proposed methods enjoy appealing features that the distributions of the covariates and of the truncation times are left unspecified. Numerical studies are reported to assess the performance of the proposed methods.

In Chapter 6, we study outstanding issues on model selection and model averaging for survival data with measurement error. Model selection plays a critical role in statistical inference and a vast literature has been devoted to this topic. Despite extensive research attention on model selection, research gaps still remain. An important but unexplored problem concerns model selection for truncated and censored data with measurement error. Although analysis of left-truncated and right-censored (LTRC) data has received extensive interests in survival analysis, there has been no research on model selection for LTRC data, let alone LTRC data involving with measurement error. In Chapter 6, we take up this important problem and develop inferential procedures to handle model selection for LTRC data with measurement error in covariates. Our development employs the local model

misspecification framework and emphasizes the use of the focus information criterion (FIC). We develop valid estimators using the model averaging scheme and establish theoretical results to justify the validity of our methods. Numerical studies are conducted to assess the performance of the proposed methods.

Finally, Chapter 7 summarizes the thesis with discussions.

Acknowledgements

Foremost, I would like to express my deep and sincere gratitude to my supervisor Dr. Grace Y. Yi for her guidance and continuous support of my Ph.D. study and research. She has a strong insight of statistics and she always stimulates me to explore new statistical topics during academic discussions. In addition, she always encourages me and teaches me how to become a good researcher. Her encouragements and supervisions make me become stronger. I am very lucky to have Dr. Yi as my supervisor.

I thank Dr. Richard Cook, Dr. Mary Thompson, Dr. Gang Li (University of California Los Angeles), and Dr. Shai Ben-David (Department of Computer Science at the University of Waterloo) for serving as my committee members and providing thoughtful and invaluable comments.

I took six courses during my Ph.D. study. Here I would like to express my gratitude to those instructors, Dr. Richard Cook, Dr. Joel Dubin, Dr. Peisong Han, Dr. Pengfei Li, Dr. Grace Yi, and Dr. Yeying Zhu, for their generous help, valuable advices and guidances when I took courses.

Many thanks go to the department staffs, especially Mary Lou Dufton and Greg Preston, for their excellent administrative supports. In addition, I am very lucky to have Junhan, Haoxin, and Qihuang as my academic friends in my Ph.D. study. I would like to thank them for providing support, encouragement, fun, and generous help.

Finally, I would like to thank my family for their love and support. I also thank my wife, Lingyu, for her support and some academic discussions. She is not only my lover, but also my academic partner.

Dedication

To my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we review relevant topics for the thesis, including survival analysis, measurement error models, graphical models, and variable selection.

1.1 Survival Data Analysis

1.1.1 Cox Model with Right-Censoring

For $i = 1, \cdots, n$, let T_i be the failure time and C_i be the censoring time. Let $f(t)$ and $S(t)$ denote the density function and the survivor function of T_i, respectively. Let $f_C(t)$ and $S_C(t)$ denote the density function and the survivor function of C_i, respectively. Let X_i denote the covariate vector of dimension p for a subject i. We assume that the $\{T_i, C_i, X_i\}$ are independent for $i = 1, \cdots, n$. Let $Y_i = \min\{T_i, C_i\}$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. The observed data consist of $\{(y_i, \delta_i, x_i) : i = 1, \cdots, n\}$, where the (y_i, δ_i, x_i) are realization values for (Y_i, Δ_i, X_i).

Let $N_i(t) = I(Y_i \leq t, \delta_i = 1)$ denote the number of the observed failures for the ith subject up to and including time t, and let $R_i(t) = I(Y_i \geq t)$ indicate whether or not the ith subject is at risk of failure at time t.

In survival analysis, suitable assumptions are often imposed to describe various censoring mechanisms. These assumptions include independent random censoring, Type I censoring, Type II censoring, and non-informative censoring (Lawless 2003, Section 2.2.1). In this thesis, we mainly consider the independent random censoring which means that T_i and C_i are independent, given the covariate X_i.

The Cox proportional hazards (PH) model is widely used to study the relationship between the failure time and the covariates. This model is formulated as

$$\lambda(t|X_i = x_i) = \lambda_0(t) \exp\left(x_i^\top \beta\right), \tag{1.1}$$

where $\lambda_0(\cdot)$ is the unspecified baseline hazard function, and β is the p-dimensional unknown parameter.

Based on model (1.1), the *partial likelihood* (Cox 1972; Lawless 2003, Section 7.1.1) is given by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(x_i^\top \beta\right)}{\sum_{j=1}^{n} \exp\left(x_j^\top \beta\right) I\left(y_i < y_j\right)} \right\}^{\delta_i},$$

and an estimator of β is obtained as $\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\, L(\beta)$, or equivalently, by solving $U(\beta) = 0$ for β, where

$$U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \int_0^\tau \left\{ x_i - \frac{\sum_{j=1}^{n} x_j \exp\left(x_j^\top \beta\right) R_j(u)}{\sum_{j=1}^{n} \exp\left(x_j^\top \beta\right) R_j(u)} \right\} dN_i(u) \tag{1.2}$$

is the *partial likelihood score function*.

1.1.2 Left-Truncation

The time from the onset of an initiating event to the disease event (or failure) is usually of interest in epidemiological and biomedical research. In the prevalent cohort sampling design, individuals can be recruited in the study if the failure time is larger than the time of recruitment. In other words, individuals might not be observed because they experience the failure event before the time of recruitment. Such a phenomenon caused by delayed entry is called the *left-truncation* and tends to produce a biased sample. Meanwhile, right-censoring may appear to those individuals who are recruited in the study, as described in Section 1.1.1.

To be more specific, let u_i and r_i denote, respectively, the calendar time of the initiating event and the failure event of a subject i with $u_i < r_i$. Let ξ be the calendar time of the

recruitment with $u_i < \xi < r_i$. Let $\widetilde{T}_i = r_i - u_i$ be the failure time and let $\widetilde{A}_i = \xi - u_i$ denote the truncation time. Let $f(t)$ and $S(t)$ be the density function and the survivor function of $\widetilde{T}_i$, respectively. A subject i can be recruited to the study only when $\widetilde{T}_i \geq \widetilde{A}_i$. Based on such selection criterion, we denote T_i and A_i, respectively, the *observed* failure time and truncation time of a subject i who is recruited in the study. It is known that for $i = 1, \cdots, n$, the joint distribution (T_i, A_i) has the same distribution as $(\widetilde{T}_i, \widetilde{A}_i)$ given $\widetilde{T}_i \geq \widetilde{A}_i$, i.e., $(T_i, A_i) \stackrel{d}{=} (\widetilde{T}_i, \widetilde{A}_i) | \widetilde{T}_i \geq \widetilde{A}_i$ (Huang et al. 2012; Huang and Qin 2013).

For these recruited subjects, either the failure event or the censoring happens. For $i = 1, \cdots, n$, let C_i denote the censoring time after the recruitment and therefore $C_i + A_i$ is the total censoring time for a subject i. Similar to the notation in Section 1.1.1, let $Y_i = \min\{T_i, A_i + C_i\}$ and $\Delta_i = I(T_i \leq A_i + C_i)$. Figure 1.1 gives an illustration of the relationship among those defined variables. We assume that $\{Y_i, A_i, \Delta_i\}$ are independent and identically distributed for $i = 1, \cdots, n$. Therefore, we have data $\{(y_i, a_i, \delta_i) : i = 1, \cdots, n\}$, where (y_i, a_i, δ_i) are realization values for (Y_i, A_i, Δ_i).

For the development of the estimation, Lawless (2003, Section 2.4.1) considered the likelihood function

$$L = \prod_{i=1}^{n} \frac{\{f(y_i)\}^{\delta_i} \{S(y_i)\}^{1-\delta_i}}{S(a_i)}.$$

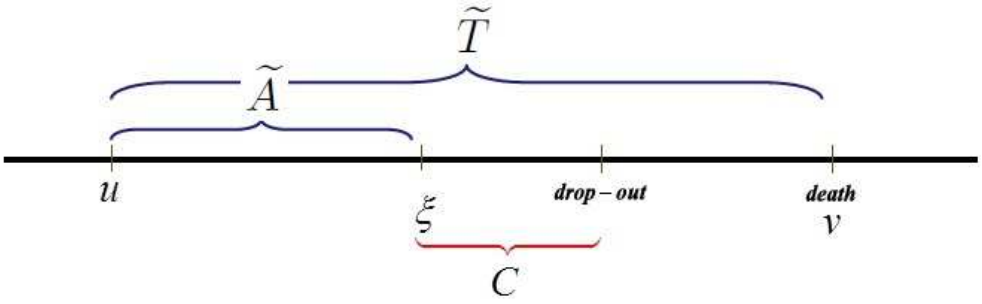


Figure 1.1: Schematic depiction of LTRC

1.2 Measurement Error Model

In many applications, we may not always have accurate measurements. Instead, the variables are usually collected with error. In this section, we introduce some measurement error models for continuous covariates and misclassification models for discrete covariates.

1.2.1 Modelling Measurement Error with Continuous Variables

In practice, measurement error in covariates usually arises due to various reasons. For $i = 1, \cdots, n$, let X_i be the p-dimensional true covariate with mean μ_X and covariance matrix Σ_X, and let X_i^* denote the surrogate, or observed covariate, of X_i with mean μ_{X^*} and covariance matrix Σ_{X^*}. If X_i and X_i^* are continuous, then the relationship between X_i and X_i^* may be described by the following measurement error models (Yi 2017, Section 2.6):

- Classical Additive Model

$$X_i^* = X_i + \epsilon_i, \tag{1.3}$$

  where ϵ_i is independent of X_i, and the ϵ_i are independent and identically distributed with mean zero and covariance matrix Σ_ϵ.

- Berkson Model

$$X_i = X_i^* + \epsilon_i,$$

  where ϵ_i is independent of X_i, and the ϵ_i are independent and identically distributed with mean zero and covariance matrix Σ_ϵ.

- Multiplicative Model

$$X_i^* = X_i \epsilon_i,$$

  where ϵ_i is independent of X_i, and the ϵ_i are independent and identically distributed with mean one and covariance matrix Σ_ϵ.

In some applications, Σ_ϵ is assumed to be known. However, it is usually unknown in practice. To give a valid estimate of Σ_ϵ, we usually need additional information to feature the measurement error process. In the following, we introduce two scenarios to estimate Σ_ϵ.

Scenario I: Repeated measurements

Since we have repeated measurements, then the measurement error model (1.3) becomes

$$X_{ij}^* = X_i + \epsilon_{ij}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, n_i$, where the X_{ij}^* represent the jth repeated measurement of X_i, the $\epsilon_{ij} \sim N(0, \Sigma_\epsilon)$ and are independent of X_i. It is easily seen that Σ_ϵ can be estimated by

$$\widehat{\Sigma}_\epsilon = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(X_{ij}^* - \bar{X}_i^*\right) \left(X_{ij}^* - \bar{X}_i^*\right)^\top}{\sum_{i=1}^{n} (n_i - 1)},$$

where $\bar{X}_i^* = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}^*$.

Scenario II: Validation data

Suppose that $\mathcal{M}$ is the subject set for the main study containing n subjects and $\mathcal{V}$ is the subject set for the external validation study containing m subjects. Assume that $\mathcal{M}$ and $\mathcal{V}$ do not overlap. Therefore, the available data contain measurements $\{(t_i, c_i, \delta_i, x_i^*) : i \in \mathcal{M}\}$ from the main study and $\{(x_i, x_i^*) : i \in \mathcal{V}\}$ from the validation sample. Hence, for the measurement error model, we have

$$X_i^* = X_i + \epsilon_i$$

for $i \in \mathcal{M} \cup \mathcal{V}$, where the $\epsilon_i \sim N(0, \Sigma_\epsilon)$ and are independent of X_i.

Since

$$\begin{aligned} \text{var}(X_i^*) &= E\left\{\text{var}(X_i^*|X_i)\right\} + \text{var}\left\{E(X_i^*|X_i)\right\} \\ &= E\left(\Sigma_\epsilon\right) + \text{var}(X_i) \\ &= \Sigma_\epsilon + \Sigma_X, \end{aligned}$$

therefore, Σ_ϵ can be estimated by

$$\widehat{\Sigma}_\epsilon = \widehat{\Sigma}_{X^*} - \widehat{\Sigma}_X,$$

where $\widehat{\Sigma}_{X^*} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left(X_i^* - \bar{X}_i^*\right) \left(X_i^* - \bar{X}_i^*\right)^\top$ and $\widehat{\Sigma}_X = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left(X_i - \bar{X}_i\right) \left(X_i - \bar{X}_i\right)^\top$, and $\bar{X}_i^* = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} X_i^*$.

1.2.2 Modelling Misclassification with Discrete Variables

When both X_i and X_i^* are discrete, (mis)classification probabilities are frequently used to characterize the relationship of X_i and X_i^*, given by

$$P\left(X_i^* = x^* | X_i = x\right) \tag{1.4}$$

for x^* and x which represent all possible values of X_i^* and X_i, respectively. In situations where $X_i, X_i^* \in \{0,1\}$, (1.4) gives $p_{i00} = P(X_i^* = 0|X_i = 0)$ and $p_{i11} = P(X_i^* = 1|X_i = 1)$, which are often called *specificity* and *sensitivity*, respectively, as well as $p_{i01} = P(X_i^* = 0|X_i = 1)$ and $p_{i10} = P(X_i^* = 1|X_i = 0)$.

To see the relationship between X_i and X_i^* more closely, we express p_{i00}, p_{i01}, p_{i10}, and p_{i11} in the matrix form.

First, we note that

$$\begin{aligned} P(X_i^* = 0) &= \sum_{j=0}^{1} P(X_i^* = 0, X_i = j) \\ &= \sum_{j=0}^{1} P(X_i^* = 0|X_i = j)P(X_i = j) \\ &= \sum_{j=0}^{1} p_{i0j} P(X_i = j). \end{aligned}$$

Similarly, we have

$$P(X_i^* = 1) = \sum_{j=0}^{1} p_{i1j} P(X_i = j).$$

Therefore, we have

$$\begin{pmatrix} P(X_i^* = 0) \\ P(X_i^* = 1) \end{pmatrix} = \mathbf{P}_i \begin{pmatrix} P(X_i = 0) \\ P(X_i = 1) \end{pmatrix}, \tag{1.5}$$

where $\mathbf{P}_i = \begin{pmatrix} p_{i00} & p_{i01} \\ p_{i10} & p_{i11} \end{pmatrix}$.

By (1.5), we determine $X_i^* = 0$ or 1 by the probability $P(X_i^* = j)$ with $j = 0, 1$. To ease notation, we let $MC[\mathbf{P}](X_i)$ denote the misclassification operator indicated by (1.5), i.e., (1.5) is notationally written as

$$X_i^* = MC[\mathbf{P}](X_i).$$

Such a misclassification operator was used by Carroll et al. (2006, p.125) and Küchenhoff et al. (2006) for a misclassified binary variable.

Meanwhile, consistent with Carroll et al. (2006, p.125), we assume that $\mathbf{P}_i$ has the spectral decomposition. That is, there exists a diagonal matrix $\mathbf{D}_i$ and the corresponding matrix of eigenvectors Ω_i, such that $\mathbf{P}_i = \Omega_i \mathbf{D}_i \Omega_i^{-1}$.

1.3 Graphical Model

In this section, we review some basics for graphical models. We first introduce the concept of graph and then describe some graphical models. Finally, we outline several commonly used methods concerning the graph theory.

1.3.1 Basic Concepts

Let V be the set of vertices and let $E \subset V \times V$ denote the set of edges. A *graph* is usually expressed as $G = (V, E)$. There are two types of graph, including an *undirected graph* and a *directed graph*. Figure 1.2 illustrates the concept of a graph which is constructed by nodes and edges. The main difference between an undirected graph and a directed graph is that the undirected graph only considers the relationship/pairwise dependence between any two nodes by using edges, while in the directed graph, an arrow is added to edges. The directed graph emphasizes that the ordering of the variables is taken into account and the relationship between any two nodes is not *reversible* (e.g., $i \to j$ cannot imply $j \to i$).

The left panel of Figure 1.2 is an undirected graph with $V = \{1, 2, 3, 4, 5\}$ and $E = \{(1, 2), (1, 5), (2, 3), (2, 4), (3, 4), (4, 5)\}$. On the other hand, the directed graph (shown by the right panel of Figure 1.2) shows not only the network structure of nodes but also the direction between two different nodes. For the applications, undirected graphs are usually applied to the study of network structures in biological data or the social network studies, while directed graphs are frequently applied for causal inference (Edwards 2000, Chapter 8). In this thesis, our discussion focuses on undirected graphs; the detailed descriptions of directed graphs can be found in Edwards (2000, Chapter 7).

1.3.2 Model Formulation

In this subsection, we introduce the model with the graphical structure incorporated. Let $X = (X_1, \cdots, X_p)^\top$ where p is a positive integer. Suppose that X_s is a binary random variable for $s \in V$, then we have

$$\mathbb{P}_\theta(x_1, \cdots, x_p) = \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \mathbb{A}(\Theta)\right), \tag{1.6}$$

where $\Theta = [\theta_{st}]$ denotes a $p \times p$ symmetric matrix, θ_{st} is the parameter associated with the pairwise dependence between X_s and X_t, and $\mathbb{A}(\Theta)$ is the normalizing constant. Model

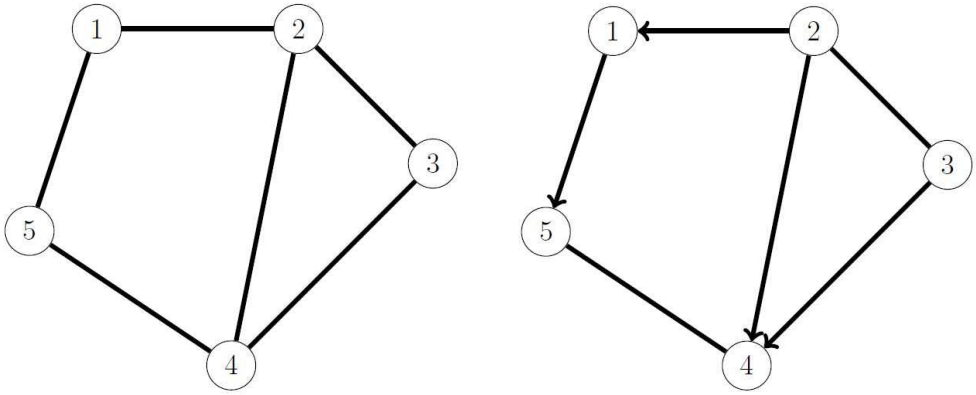


Figure 1.2: Graphical structures. The left graph is undirected; the right graph is directed.

(1.6) is called the *Ising* model (Ravikumar et al. 2010). In model (1.6), the parameter θ_s for $s \in V$ describes the main effects, while θ_{st} for $(s,t) \in E$ conveys the information of the pairwise dependence between variables X_s and X_t, and X_s and X_t are said to be dependent if $\theta_{st} \neq 0$. We assume that $\theta_{st} = \theta_{ts}$ for $s \neq t$ in the undirected graphical structure.

On the other hand, suppose that X follows the Gaussian distribution with mean μ and covariance Σ. The distribution of X is written as

$$\mathbb{P}_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}. \tag{1.7}$$

In order to express (1.7) by the graphical model as shown in (1.6), we do the re-parametrization by letting $\Sigma = \Theta^{-1}$ and $\mu = -\Theta^{-1}\gamma$, yielding

$$\mathbb{P}_{\gamma,\Theta}(x) = \exp\left\{\sum_{s\in V} \gamma_s x_s - \frac{1}{2}\sum_{(s,t)\in E} \theta_{st} x_s x_t - \mathbb{A}(\Theta)\right\}, \tag{1.8}$$

where $\gamma \in \mathbb{R}^p$, $\Theta \in \mathbb{R}^{p\times p}$, and $\mathbb{A}(\Theta) = -\frac{1}{2}\log\det\left(\frac{\Theta}{2\pi}\right)$, so that $\int \mathbb{P}_{\gamma,\Theta}(x)dx = 1$. The model (1.8) is called the *Gaussian graphical model* (Hastie et al. 2015, p.245).

In addition to the well-known Ising model and the Gaussian graphical model, some extensions are available. The first extended model is the mixed graphical model (Lee and Hastie 2015; Hastie et al. 2015, p.259). Let X_{C} denote the p-dimensional continuous random vector and let X_{D} be the q-dimensional discrete random vector. The mixed graphical

model is formulated by

$$\begin{aligned}\mathbb{P}_{\Theta}(x_{\mathrm{C}}, x_{\mathrm{D}}) &= \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p}\beta_{st}x_{\mathrm{C},s}x_{\mathrm{C},t} + \sum_{t=1}^{p}\alpha_t x_{\mathrm{C},t}\right.\\ &\quad \left. + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(x_{\mathrm{D},j})x_{\mathrm{C},s} + \sum_{j=1}^{q}\sum_{r=1}^{q}\psi_{rj}(x_{\mathrm{D},r}, x_{\mathrm{D},j})\right), \end{aligned} \tag{1.9}$$

where Θ represents a set of parameters $\{\{\beta_{st}\}, \{\alpha_s\}, \{\rho_{sj}\}, \{\psi_{rj}\}\}$. Different from the graphical model which contains either all continuous variables (e.g., the Gaussian graphical model) or all discrete variables (e.g., the Ising model), the model (1.9) extends the pairwise dependence for the continuous and discrete variables.

The second extension is the the graphical model via the exponential family distribution (Yang et al. 2015). The graphical model is formulated by

$$\mathbb{P}_{\beta,\Theta}(x) = \exp\left\{\sum_{r\in V}\beta_r\mathbb{B}(x_r) + \sum_{(s,t)\in E}\theta_{st}\mathbb{B}(x_s)\mathbb{B}(x_t) + \sum_{r\in V}\mathbb{C}(x_r) - \mathbb{A}(\beta,\Theta)\right\}, \tag{1.10}$$

where $\beta = (\beta_1, \cdots \beta_p)^{\top}$ is the p-dimensional parameter vector, $\Theta = [\theta_{st}]$ is a non-diagonal $p \times p$ symmetric matrix, and $\mathbb{B}(\cdot)$ and $\mathbb{C}(\cdot)$ are given functions. The function $\mathbb{A}(\beta, \Theta)$ is the normalizing constant which makes (1.10) integrated as 1; it is also called the *log-partition function*, given by

$$\mathbb{A}(\beta,\Theta) = \log\int \exp\left\{\sum_{r\in V}\beta_r\mathbb{B}(x_r) + \sum_{(s,t)\in E}\theta_{st}\mathbb{B}(x_s)\mathbb{B}(x_t) + \sum_{r\in V}\mathbb{C}(x_r)\right\}dx.$$

The graphical model (1.10) gives a broad class of models which essentially cover basic exponential family distributions. For example, if $\mathbb{B}(X) = \frac{X}{\sigma}$ and $\mathbb{C}(X) = -\frac{X^2}{2\sigma^2}$ where σ is a positive constant, then (1.10) reduces to (1.8). If $\mathbb{B}(X) = X$ and $\mathbb{C}(X) = 0$ with $X \in \{0, 1\}$, then (1.10) reduces to (1.6). Furthermore, taking $\mathbb{B}(X) = -X$ and $\mathbb{C}(X) = 0$ with $X \in [0, \infty)$ yields the exponential graphical model distribution

$$\mathbb{P}(x) = \exp\left(-\sum_{s=1}^{p}\theta_s x_s + \sum_{s=1}^{p}\sum_{t=1}^{p}\theta_{st}x_s x_t + \mathbb{A}(\Theta)\right),$$

provided that $\theta_s > 0$ and $\theta_{st} \geq 0$. In addition, replacing $\mathbb{B}(X)$ and $\mathbb{C}(X)$ in (1.10), respectively, by X and $-\log(X!)$ gives the Poisson graphical model

$$\mathbb{P}(x) = \exp\left[\sum_{s=1}^{p}\{\theta_s x_s - \log(x_s!)\} + \sum_{s=1}^{p}\sum_{t=1}^{p}\theta_{st}x_s x_t + \mathbb{A}(\Theta)\right],$$

provided that $\theta_{st} \leq 0$.

1.3.3 Existing Methods

As described in Section 1.3.2, the main interest of the graphical model is to estimate Θ. However, the main challenge is that the underlying graph structure is unknown, and the parameter Θ is sparse and contains zero entries.

A useful method of estimating Θ is the *graphical lasso* (Friedman et al. 2008). The key idea of the graphical lasso is to construct the penalized likelihood function based on (1.8) and the LASSO penalty function. The detailed algorithm can be found in Friedman et al. (2008) and Hastie et al. (2015, Section 9.3). However, the graphical lasso method mainly focus on the Gaussian graphical model, and the method based on the other models is not fully explored.

Alternatively, the *conditional inference*, or the *neighbourhood-based likelihood*, is a more flexible method which can be used for any distribution function. It was first proposed by Meinshausen and Bühlmann (2006). The application of the conditional inference on the Ising model was presented by Ravikumar et al. (2010). In the following presentation, we take the Gaussian graphical model (1.8) as an example. The detailed inference procedure is also available in Hastie et al. (2015, Section 9.4). The basic idea is shown in Figure 1.3.

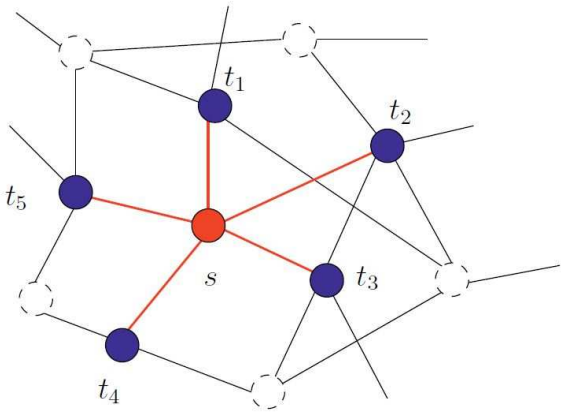


Figure 1.3: Visualization of the conditional inference

Without loss of generality, as shown in Figure 1.3, we fix a node s, and define the

neighbourhood set

$$\mathcal{N}(s) = \{t \in V : (s,t) \in E\}. \tag{1.11}$$

To estimate the neighbourhood set of s, it suffices to study the inference of $X_s|X_{V\setminus\{s\}}$. Since the random vector X follows a multivariate Gaussian and is generated by (1.8), then $X_s|X_{V\setminus\{s\}}$ is also Gaussian (Hastie et al. 2015, p.255). By some algebra, we have

$$\mathbb{P}\left(X_s|X_{V\setminus\{s\}};\theta_s^*\right) \propto \exp\left\{-\left(X_s - \sum_{t\in V\setminus\{s\}} \theta_{st}^* X_t\right)^2\right\}. \tag{1.12}$$

Let $\theta_s^* = \left(\theta_{s1}^*, \cdots, \theta_{s(s-1)}^*, \theta_{s(s+1)}^*, \cdots, \theta_{sp}^*\right)$. Then the estimator $\widehat{\theta}_s^*$ is given by

$$\widehat{\theta}_s^* = \underset{\theta_s^*}{\operatorname{argmin}} \left\{\sum_{i=1}^{n}\left(X_s^{(i)} - \sum_{t\in V\setminus\{s\}} \theta_{st}^* X_t^{(i)}\right)^2 + \lambda \left\|\theta_s^*\right\|_1\right\}, \tag{1.13}$$

where $\left\|\theta_s^*\right\|_1 = \sum_{t\in V\setminus\{s\}} |\theta_{st}^*|$ and λ is a tuning parameter.

Finally, the estimated neighbourhood set is given by

$$\widehat{\mathcal{N}}(s) = \left\{t \in V : \widehat{\theta}_{st}^* \neq 0\right\}. \tag{1.14}$$

By repeating the procedure (1.13) for all $s \in V$, we have $\widehat{\theta}_s^*$ and $\widehat{\mathcal{N}}(s)$ for $s \in V$.

In practice, based on the penalized likelihood approach (1.13), $\widehat{\theta}_{st}$ is not usually equal to $\widehat{\theta}_{ts}$. Meinshausen and Bühlmann (2006) and Hastie et al. (2015, p.255) presented the AND/OR rule to determine the estimated edge set $\widehat{E}$. For any two different nodes s and t, the AND rule allows $(s,t) \in \widehat{E}$ if both $s \in \widehat{\mathcal{N}}(t)$ and $t \in \widehat{\mathcal{N}}(s)$ hold, while the OR rule declares $(s,t) \in \widehat{E}$ if either $s \in \widehat{\mathcal{N}}(t)$ OR $t \in \widehat{\mathcal{N}}(s)$ is true.

1.4 Variable Selection and Dimension Reduction

In this section, we discuss several methods for variable selection and dimension reduction. For variable selection, we first review some classical methods for regression analysis, and then introduce the focus information criterion proposed by Claeskens and Hjort (2003) and Hjort and Claeskens (2003). After that, we discuss the penalized regression where different types of penalty functions were considered. Moreover, analysis of ultrahigh-dimensional data will also be introduced. Finally, we describe a basic idea of dimension reduction.

1.4.1 Classical Criteria

For the problems of selecting variables, there are several useful methods and are fully described in the classical linear regression analysis, including the Akaike information criterion (AIC) (Akaike 1973), the Bayesian information criterion (BIC) (Schwarz 1978), Mallow's C_p (Mallow 1973), and so on. In this section, we mainly review AIC and BIC.

Suppose that $\ell(\beta)$ is the log-likelihood function of β, and let $X_i = (X_{i1}, \cdots, X_{ip})$ denote the p-dimensional vector of covariates for $i = 1, \cdots, n$. Let $\mathcal{S}$ be the set of all possible combinations of the components of X with $|\mathcal{S}| = 2^p$. We call $S \in \mathcal{S}$ a *candidate model*. Based on a candidate model S, we have the corresponding log-likelihood function $\ell_S(\beta)$, and the estimator of β based on the candidate model S is given by $\widehat{\beta}_S = \underset{\beta}{\operatorname{argmax}}\, \ell_S(\beta)$. The AIC is defined as

$$\mathrm{AIC}(S) = 2\ell_S(\widehat{\beta}_S) - 2\dim(S),$$

where $\dim(S)$ is the number of elements in the set S. AIC can be viewed as a penalized log-likelihood criterion to balance the goodness of fit and the number of estimated parameters. The optimal candidate model is determined by the highest AIC score.

An alternative approach is BIC, which is formulated by

$$\mathrm{BIC}(S) = 2\ell_S(\widehat{\beta}_S) - \log(n)\dim(S).$$

The best candidate model is determined by the highest BIC score. Different from the AIC method, the penalty term in BIC involves the sample size, and it conveys a stronger penalty for complexity, especially when $n \geq 8$ (e.g., Claeskens and Hjort 2008, p.70).

1.4.2 Focus Information Criterion

In survival analysis, sometimes we are interested in certain specific quantities such as the hazards ratio or the survivor function. Here we call such 'specific quantity' the *focus parameter*, and denote it by μ.

It is expected that the focus parameter μ is usually the function of the initial parameter β. A crucial issue is to determine the variables which are informative for estimating μ. As introduced in Section 1.4.1, some methods, such as AIC or BIC, can be used to select variables. However, the best candidate model determined by AIC or BIC may not be the best model for a given focus parameter μ.

To overcome this, Claeskens and Hjort (2003) proposed a new variable selection method, called the *focus information criterion* (FIC). The key idea of FIC is to determine the best candidate model by selecting the minimizer of the mean squared error (MSE) of the estimator $\widehat{\mu}$ of μ. The following steps summarize the procedure based on the FIC. The detailed description can be found in Claeskens and Hjort (2003) and Claeskens and Hjort (2008, Chapter 6).

Step 1 : Let $\ell_S(\beta)$ be the log-likelihood function based on a candidate model $S \in \mathcal{S}$, and let $\widehat{\beta}_S$ denote the maximum likelihood estimator (MLE) of β based on a candidate model.

Step 2 : Based on the likelihood theory, one can develop the asymptotic distribution of the estimator $\widehat{\beta}_S$.

Step 3 : Since the focus parameter μ is the function of parameter β, then by the invariance property of MLE, the estimator of μ based on a candidate model S is given by $\widehat{\mu}_S = \mu(\widehat{\beta}_S)$. Furthermore, by Step 2, we also develop the asymptotic distribution of the estimator $\widehat{\mu}_S$ for all $S \in \mathcal{S}$. Let $\text{bias}(\widehat{\mu}_S)$ and $\text{var}(\widehat{\mu}_S)$ denote the bias and the variance of $\widehat{\mu}_S$, respectively.

Step 4 : Based on a candidate model S, the MSE of $\widehat{\mu}_S$ is given by $\text{MSE}(\widehat{\mu}_S) = \{\text{bias}(\widehat{\mu}_S)\}^2 + \text{var}(\widehat{\mu}_S)$.

Step 5 : The candidate model S^* which minimizes $\text{MSE}(\widehat{\mu}_S)$ is the best candidate model for the focus parameter.

1.4.3 Penalized Regression

Different from AIC or BIC, we introduce the penalized likelihood method where the penalty term is a function of the parameter. Let $\rho(\beta)$ denote the penalty function, then the general form of the penalized likelihood function is given by

$$\ell(\beta) + \lambda \sum_{i=1}^{p} \rho(\beta_i),$$

where λ is called the *tuning parameter*. Useful penalty functions are as follows.

- Adaptive LASSO (Zou 2006): $\rho(\beta_i) = w_i \, |\beta_i|$, where w_i is a weight.

- LASSO (Tibshirani 1996): $\rho(\beta_i) = |\beta_i|$.
- SCAD (Fan and Li 2001): $\rho'(t) = I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t \geq \lambda)$ for $t > 0$, where $(x)_+ = \max\{x, 0\}$ and $a > 2$ is a fixed parameter.
- MCP (Zhang 2010): $\rho'(t) = \frac{(a\lambda - t)_+}{a\lambda}$ for $t > 0$, where $a > 1$ is a fixed parameter.
- SICA (Lv and Fan 2009): $\rho(t) = \frac{(a+1)t}{a+t}$ for $t > 0$, where $a > 0$ is a fixed parameter.

The LASSO method with the ℓ_1 penalty is the earliest approach among penalty functions. As pointed by Tibshirani (1996), the LASSO method achieves to shrink non-informative parameters to zero and to obtain the estimators for those informative parameters simultaneously. However, as discussed in Zou (2006), the LASSO shrinkage may produce biased estimates, and some necessary conditions should be imposed. To overcome this problems, Zou (2006) proposed the adaptive LASSO by adding weights in the ℓ_1 penalty. Different from the LASSO method, the SCAD, MCP, and SICA methods are based on the non-convex penalty functions, and they still achieve the oracle properties.

1.4.4 Ultrahigh-Dimensional Statistical Analysis and Feature Screening

In high-dimensional statistical analysis, since not all variables are informative, then it is necessary to select important variables. However, even though a number of variable selection methods have been discussed in Section 1.4.3, those methods are restricted to the case where the dimension of variables p is smaller than the sample size n, i.e., $p < n$. In applications, such as gene expression data, proteomics studies, and biomedical imaging, we usually encounter the *ultrahigh-dimensional data* in the sense that the dimension p is greater than the sample size n, i.e., $p \gg n$. It is difficult to apply the methods in Section 1.4.3 to analyze the ultrahigh-dimensional data due to the inaccuracy of estimation and the highly computational cost.

To overcome this problem, Fan and Lv (2008) considered linear regression models and proposed a feature screening method that is based on correlation learning, called sure independence screening, to reduce dimensionality from high to a moderate scale that is below the sample size. Similar idea was extended to different models. For example, Fan and Song (2010) developed sure independent screening method in generalized linear models. In survival analysis, Fan et al. (2010) developed feature screening method based on the Cox model. Song et al. (2014) proposed rank based independent screening for high-dimensional

survival data. Yan et al. (2017) studied the Spearman rank correlation screening for censored data.

1.4.5 Sufficient Dimension Reduction

Different from selecting informative variables in previous subsections, in this subsection, we aim to introduce *sufficient dimension reduction* (SDR) which is useful in reducing the dimension of covariates but not losing predictive information of covariates when developing non-parametric models.

Let $T \in \mathbb{R}$ be the univariate response and let X denote the p-dimensional vector of covariates. The spirit of SDR is to find a $p \times d$ basis matrix B such that

$$T \perp\!\!\!\perp X | B^\top X,$$

where "$\perp\!\!\!\perp$" stands for the statistical independence and d is usually called the structure dimension which is unknown. To estimate B and d, a number of methods has been proposed, including the inverse regression (Li 1991; Li and Wang 2007; Zhu et al. 2010), the minimum average variance estimation (Zhu and Zeng 2006; Xia 2007; Wang and Xia 2008; Yin and Li 2011), and the semiparametric framework (Ma and Zhu 2012, 2013). When B is obtained, then the subsequent analysis can be based on the lower dimensional variables $\{T, B^\top X\}$ without losing information.

1.5 Thesis Topics and Outline of the Thesis

Although a large number of methods have been available for survival analysis (e.g., Kalfleisch and Prentice 2002; Lawless 2003; Cook and Lawless 2018), research gaps still remain. Many problems in survival analysis with complex features in data have not been explored. We now list important problems of our interest and give detailed descriptions in the following subsections:

- Graphical models with error-prone variables: bias analysis and valid inference methods.
- Analysis of noisy survival data under graphical proportional hazards measurement error models.

- Sufficient dimension reduction for analysis of high-dimensional survival data with error-prone variables.
- Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error.
- Model selection and model averaging for analysis of truncated and censored data with measurement error.

This thesis consists of seven chapters. The remainder of the thesis is organized as follows. In Chapter 2, we focus on the high-dimensional analysis of (mixture) graphical model with mismeasurement. We propose the simulation-based conditional inference to derive the estimators and determine the graphical structure. We also point out that the estimated graphical structure based on the error-prone variables would not recover the underlying true graphical structure if we do not correct the error effect appropriately.

In Chapter 3, we study high-dimensional statistical inference for survival analysis. Different from the usual variable selection problem in survival analysis, we consider joint modeling of the survival response and the graphically structured covariates, which allows us to explore the informative main effects and the pairwise dependence among the covariates. Meanwhile, we also consider the measurement error and misclassification in covariates. We propose the simulation-based three-stage procedure to correct error effects, determine the informative variables, obtain the pairwise graphical structure, and derive the estimators simultaneously.

In Chapter 4, we investigate the sufficient dimension reduction problem with survival data and covariate measurement error as well as ultrahigh-dimension. We develop the semiparametric estimation procedure which does not require usual assumptions imposed in the past literature. We also propose feature screening method with the effects of censored responses and covariate measurement error taken into account and extend the semiparametric estimation method to deal with ultrahigh-dimensional sufficient dimension reduction.

In Chapter 5, we consider the survival analysis with left-truncated and right-censored data subject to covariate measurement error. We first discuss the corrected conditional likelihood approach which was outlined by Yi and Lawless (2007). To improve the efficiency of the estimator, we propose the augmented pseudo-likelihood method. For the measurement error model, we consider two different scenario where the parameters in the measurement error model can be known or unknown.

In Chapter 6, we are interested in the focus parameter based on left-truncated and right-censored survival data with covariate measurement error. Instead of directly using

AIC or BIC to determine the candidate model for the focus parameter, we provide the valid inferential procedure, which allows us to determine the suitable and best candidate model for any focus parameter.

Finally, the thesis is concluded with a discussion in Chapter 7.

1.6 Background and Literature Review for Each Topic

1.6.1 Graphical Models with Error-Prone Variables

In the era of Big Data, high-dimensional data become more accessible than before and frequently arise from many areas, including genomic studies, cancer research, and medical health record frameworks. Understanding the association structure, or the *network structure* of variables, is often of prime interest. To characterize such dependence structures of variables, *graphical models* are commonly used. For example, the *Ising model* and the *Gaussian graphical model* are two popular models to describe association structures for binary and continuous variables, respectively. Many inference methods have been proposed for those models. To name a few, Ravikumar et al. (2010) proposed an inferential procedure to estimate the graph for the Ising model. Yuan and Lin (2007) considered the Gaussian graphical model and adopted an interior point optimization method to obtain the network structure. Friedman et al. (2008) proposed the graphical lasso to select the variables and estimate the model parameters. The Gaussian graphical model with complex features, such as latent variables, was also explored by Zhou et al. (2009), Ravikumar et al. (2011), Sun and Li (2012), Tan et al. (2016), Dalal and Rajarantam (2017), and Fan et al. (2017), among many others.

Extensions of the Gaussian graphical model and the Ising model have also been explored in the literature. The exponential family graphical model, which treats the Ising and Gaussian graphical models as special cases, was developed by Yang et al. (2015). *Mixed graphical models* were proposed to handle the settings where the variables contain both continuous and discrete variables. For example, Lee and Hastie (2015) discussed the pseudo-likelihood method to deal with the mixture of the Gaussian graphical model and the Ising model. Cheng et al. (2017) proposed the group lasso to conduct inferences under the mixed graphical models. Chen et al. (2015) and Zhang et al. (2017) studied mixed graphical models via the exponential family distribution.

Even though analysis of graphical models has been widely explored, research gaps still remain. A typical feature that is left unattended to is about *measurement error in*

variables, which usually appears in applications. For example, in biological studies, protein signaling networks play a central role in the etiology of many diseases. A major challenge, commented by Bandara (2009), is due to measurement noise which primarily attributes large uncertainty of parameter estimation. It was observed that the misleading results may be produced if measurement error effects are ignored.

Concerning regression analysis, research on measurement error has attracted extensive attention. It has been well understood that ignoring measurement error effects often yields seriously biased and misleading results (e.g., Carroll et al. 2006; Yi 2017). A large body of research papers have been available in the literature to address measurement error effects for different settings (e.g., Carroll 1989; Carroll et al. 1996; Carroll et al. 2004; Carroll et al. 2007; Carroll et al. 2009; Yi et al. 2015; Yi and He 2017; Yi et al. 2019). For detailed discussions, see monographs Biemer et al. (1991), Buonaccorsi (2010), Fuller (1987), Gustafson (2004), Carroll et al. (2006), and Yi (2017).

While there has been great attention on measurement error on regression analysis, there has been little work investigating measurement error effects on graphical model analysis, an area that has proven useful for featuring complex association structures among the variables. Driven by this, we consider this important problem and explore *Graphical models with Error-prone Measurement (GEM)* in Chapter 2. We consider undirected graphical models which are described by the exponential family distribution. We investigate the asymptotic biases of the naive method which disregards measurement error effects. We examine all the three scenarios of mismeasurement: (1) all the error-prone variables are continuous, (2) all the error-prone variables are discrete, and (3) error-prone variables include both continuous and discrete variables. Furthermore, we develop valid inference procedures to address mismeasurement effects. We establish theoretical results for the proposed methods. To the best of our knowledge, there has been no research to explore this problem.

1.6.2 Analysis of Noisy Survival Data with Graphical Proportional Hazards Measurement Error Models

Survival analysis has been proven useful in many areas including cancer research, clinical trials, epidemiological studies, actuarial science, and so on. A large body of methods have been developed for various survival models. Among them, methods concerning the Cox proportional hazards (PH) model have attracted the most research attention. Comprehensive discussion on those methods can be found in Kalbfleisch and Prentice (2002), Lawless (2003), and the references therein.

While the Cox proportional hazards model has been widely used, this model and its extensions are inadequate for handling data with complex features. In the era of Big Data, high-dimensional survival data become available and such data entail new features that traditional survival data do not possess: (1) many covariates are inactive in explaining the survival information, (2) active covariates are associated in a network structure, and (3) some covariates are error-contaminated.

To handle the first feature of survival data with high-dimensional covariates, several methods have also been developed based on the Cox PH model. For example, Fan and Li (2002) used the SCAD penalty function to select important variables for the Cox PH model. Cai et al. (2005) considered variable selection with multivariate failure time data. Zhang et al. (2007) developed the adaptive lasso method for the Cox PH model. Yan and Huang (2012) explored the adaptive group lasso for the Cox regression model with time-varying coefficients. Huang et al. (2013) studied the penalized partial likelihood with the L_1-penalty for the Cox model. Li and Ma (2013) discussed analysis of survival data with high-dimensional genetic covariates.

Regarding survival data with error-prone covariates, many inference methods have been developed for the Cox PH models with error-contaminated covariates since the seminal paper by Prentice (1982). For instance, Nakamura (1992) developed an approximate corrected partial likelihood method. Huang and Wang (2000) proposed a nonparametric approach for settings with repeated measurements for mismeasured covariates. Xie et al. (2001) explored a least squares method to calibrate the induced hazard function. Song and Huang (2005) presented a conditional score approach for estimation of the model parameters. Yi and Lawless (2007) developed a weakly parametric approach to correct for measurement error effects based on the likelihood formulation. Wang et al. (1997), Wang (1999), Shaw and Prentice (2012) and Zhao and Prentice (2014) considered the regression calibration method to address measurement error for the Cox PH model. Detailed discussion on this topic can be found in Yi (2017, Chapter 3).

While available methods address certain features of high-dimensional survival data, none of them incorporate network structured covariates into modeling and analyzing survival data. Furthermore, there have no methods dealing with all the three features altogether, even though data with such features arise commonly in applications. To fill this gap, we propose graphical proportional hazards measurement error models in Chapter 3. Our models extend the scope of the conventional Cox PH model and accommodate network structures of covariates. We utilize the formulation of graphical models (Ravikumar et al. 2010; Lee and Hastie 2015) and consider a broad variety of covariates which follow the exponential family distribution (Yang et al. 2015), extending the commonly used models including the Gaussian graphical model (e.g., Yuan and Lin 2007; Friedman et al. 2008;

Sun and Li 2012; Wang 2015; Tan et al. 2016; Dalal and Rajarantam 2017) and the Ising model (e.g., Ravikumar et al. 2010). Regarding covariate mismeasurement, we consider general settings where both continuous covariates and discrete covariates may be subject to measurement error. We establish the theoretical results to justify the validity of the proposed method.

1.6.3 Sufficient Dimension Reduction for High-Dimensional Survival Data with Error-Prone Variables

Survival analysis has been proven useful in many areas including cancer research, clinical trials, epidemiological studies, actuarial science, and so on. A primary interest in survival analysis is to study the association between survival times and covariates of interests. Many parametric or semiparametric survival models are proposed for survival analysis, including the Cox proportional hazards model (Cox 1972), the proportional odds model (Bennett 1983), the additive hazards model (Lin and Ying 1994), and the accelerated failure-time model (Cox and Oakes 1984). Although those models are useful in applications, they may still be inadequate to handle real problems due to the lack of the knowledge of the suitability of a particular model. Motivated by this, non-parametric regression models are employed in applications. Non-parametric models offer the flexibility of modeling and protect us against the risk of model misspecification. However, such models are hampered by the high dimension of covariates. To offer a flexible yet parsimonious model formulation, *sufficient dimension reduction* (SDR) become useful in reducing the dimension of covariates but not losing predictive information of covariates.

For uncensored data, various methods have been proposed to reduce the dimension of covariates, including the inverse regression (Li 1991; Li and Wang 2007; Zhu et al. 2010), the minimum average variance estimation (Zhu and Zeng 2006; Xia 2007; Wang and Xia 2008; Yin and Li 2011), and the semiparametric framework (Ma and Zhu 2012, 2013); some details can be found in Cook (1998) and Li (2018).

For right-censored survival data, a number of methods have also been developed for dimension reduction. To name a few, Li et al. (1999) examined the sliced inverse regression method to estimate the central space (CS) of dimension reduction directions. Xia et al. (2010) considered semiparametric models and proposed the minimum average variance estimation using the inverse censoring weighting scheme. Lue et al. (2011) explored the spline method and principal Hessian directions (PHD) approach. Lu and Li (2011) discussed the sliced inverse regression with inverse probability weights and implemented the variable selection approach for sparse data with a large dimension. Nadkarni et al.

(2011) developed a minimum discrepancy approach using the inverse censoring weighting method to build a inverse regression estimator and applied the bootstrapping method to estimate the structural dimension. Zhao and Zhou (2014) examined sufficient dimension reduction using marginal regression models to analyze recurrent event data. Zhao et al. (2017) developed the multi-index model using the martingale approach.

While those methods are useful for different settings, they are inapplicable for error-contaminated data, an ubiquitous feature in applications. As noted by Carroll and Li (1992), when covariates are subject to measurement error, misleading results are often yielded if measurement error effects are ignored when performing sufficient dimension reduction. To address measurement error effects in the SDR framework, Carroll and Li (1992) proposed the "corrected" covariates for the implementation of sliced inverse regression. Li and Yin (2007) established the invariance law for correcting measurement error effects. Zhang et al. (2014) developed the cumulative slicing estimation method using the "corrected" covariates, which extended the development of Zhu et al. (2010). In the presence of both censored data and measurement error in covariates, however, there has been no available work, to the best of our knowledge.

Another limitation of most available methods in the SDR framework lies in the assumption that the dimension (p) of covariates is smaller than the sample size (n), i.e., $p < n$. In practice, however, high-dimensional data have become more accessible than ever, and the $p \gg n$ problem is an important yet challenging topic that deserves careful research. It is not trivial to directly apply conventional SDR methods to estimate the CS when the dimension p is higher than the sample size n, partly because the covariance matrix of the covariates X, say Σ_X, is usually singular due to $p \gg n$.

To analyze data with $p \gg n$, *feature screening* is typically applied before performing standard analysis. Different feature screening methods have been proposed for varying settings. For example, Zhu et al. (2011) proposed model-free feature screening for ultrahigh-dimensional data. Li et al. (2012) developed the distance correlation approach for feature screening. Yu et al. (2013) proposed the Dantzig selection approach with the sliced inverse regression. Yin and Hilafu (2015) and Hilafu and Yin (2017) studied the dimension reduction framework for $p \gg n$ through the sequential method. However, these approaches apply only to the data with complete responses and precisely measured covariates. Even though limited research has been directed to perform feature screening for censored responses (e.g., Song et al. 2014; Yan et al. 2017; Chen et al. 2019), little work has been available to deal with sufficient dimension reduction in the concurrent presence of censored responses, covariate measurement error, and ultrahigh-dimensional covariates.

Driven by the lack of methods for handling such data, we develop methods for han-

dling dimension reduction for censored data with covariate measurement error as well as ultrahigh-dimension in Chapter 4. We consider the single-index conditional distribution model which covers many useful survival models, and based on them, we develop the semiparametric estimation procedure. Our method does not require usual assumptions such as linearity and constant variance conditions that are imposed by other authors for a similar problem (e.g., Li 1991). Our method employs the "corrected" covariates to correct for measurement error effects and applies the conditional expectation scheme to remove the bias caused by censoring. To handle the ultrahigh-dimensional SDR problem, we propose a two-stage procedure, where in the first stage, we develop model-free feature screening method with the effects of censored responses and covariate measurement error taken into account, and in the second stage, we extend the semiparametric estimation method to build the single-index conditional distribution model. Theoretical results of the proposed methods are established accordingly.

1.6.4 Left-Truncated and Right-Censored Survival Data with Covariate Measurement Error

Survival analysis has been proven useful in many areas including cancer research, clinical trials, epidemiological studies, actuarial science, and so on. A large body of methods have been developed under various survival models. Among them, methods on the Cox proportional hazard model have attracted the most research attention. Comprehensive discussion on those methods can be found in Kalbfleisch and Prentice (2002), Lawless (2003), and the references therein.

Those methods, however, break down when data have complex features pertinent to the data collection and the natures of variables. Typically, simultaneous presence of biased samples caused by left-truncation or length-biased sampling and measurement error in covariates pose considerable challenges in survival analysis. Focusing on measurement error only, a large number of research papers have emerged since Prentice (1982). To name a few, Nakamura (1992) developed an approximate corrected partial likelihood method which was extended by Buzas (1998) and Hu and Lin (2002). Huang and Wang (2000) proposed a nonparametric approach for settings with repeated measurements for mismeasured covariates. Xie et al. (2001) explored a least squares method to calibrate the induced hazard function. Song and Huang (2005) presented a conditional score approach for estimation of the model parameters. Other approaches include Augustin (2004), Greene and Cai (2004), Li and Ryan (2006), Küchenoff, Bender and Langner (2007), and the references therein. A review on this topic was given by Yi (2017, Chapter 3).

On the other hand, left-truncation is a common characteristic of survival studies which arises when study subjects do not enter the study at the same time. In the presence of left-truncation, individuals with shorter survival times are less likely to be recruited in the study, thus resulting in a biased sample. Sizable methods have been available for analyzing such data. For instance, Qin and Shen (2010) proposed the weighted estimating equation approach. Qin et al. (2011) described an EM algorithm for estimation involving infinite dimensional parameters. Huang et al. (2012) examined a profile likelihood method for parameter estimation for which the distribution of the truncation time was restricted as a uniform distribution. Wu et al. (2018) proposed a pairwise likelihood method of handling left-truncated data. With joint modeling of longitudinal covariates and survival outcomes, Su and Wang (2012) proposed a semiparametric method to handle the feature of left-truncation.

While there have been methods of dealing with survival data with different features, to the best of our knowledge, no systematic methods have been available for handling the those features simultaneously (Yi and Lawless 2007). In Chapter 5, we consider this important problem and develop inference methods for analysis of left-truncated right-censored survival data with measurement error. To delineate the survival process, we employ the most widely used framework - the Cox proportional hazards model; to postulate the measurement error process, we extend the classical additive model, the model most popularly considered in the literature of measurement error models, to facilitate measurement error that is induced from both a systematic way and a random manner. We exploit a flexible estimator for the survival model parameters which does not require specification of the baseline hazard function. To improve the efficiency, we further develop an augmented non-parametric maximum likelihood estimator. We establish asymptotic results for the proposed estimators and examine the issues of efficiency and model misspecifications. While the proposed methods generalize the scope of existing work on survival data, the extensions turn out neither trivial nor straightforward. The proposed methods enjoy appealing features that the distributions of the true covariates and of the truncation times are left unspecified, and they are easy to implement.

Our work is partially motivated by the Worcester Heart Attack Study (WHAS500) data (Hosmer et al. 2008) which involve left-truncation and right-censoring data. Data were collected over thirteen 1-year periods beginning in 1975 and extending in 2001 on all patients with acute myocardial infarction (MI) admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. Basically, three types of time are recorded: time of the hospital admission, time of the hospital discharge, and time of the last follow-up (which is either death or censoring time). The total follow-up length is defined as the time gap between the hospital admission and the last follow-up, and the hospital

stay time is defined as the time length between the hospital admission and the hospital discharge. Data can only be collected for those individuals whose total follow-up length is larger than the hospital stay time, creating left-truncation (e.g., Kalbfleisch and Prentice 2002, Section 1.3; Lawless 2003, Section 2.4). It is interesting to study how the risk factors are associated with the survival times after the patients are discharged from the hospital. To conduct sensible analyses, it is imperative to account for possible measurement error effects that are induced from error-prone covariates.

1.6.5 Model Selection and Model Averaging for Analysis of Truncated and Censored Data with Measurement Error

Model selection plays an important role in statistical inference, and various model selection criteria have been proposed, including the Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978), Cross Validation, and Mallow's C_p (Mallow, 1973). To incorporate the feature of the quantity of interest, Claeskens and Hjort (2003) proposed the focus information criterion (FIC) for model selection. Several extensions of the FIC method have been developed for distinct settings. For example, Claeskens and Carroll (2007) studied the FIC method for the partially linear model. Zhang and Liang (2011) implemented the FIC method to generalized additive partial linear models. Xu et al. (2014) discussed the FIC method based on weighted composite quantile regression.

While those extensions branch out the scope of the FIC method, they are not applicable to handle many practical problem. Typically, those methods fails to handle truncated and censored data with measurement error, which arise ubiquitously from many areas including clinical trials, epidemiological studies, actuarial science, and so on.

In analysis of censored data or survival data, some methods have been proposed for variable selection for different survival models. For example, Liang and Zou (2008) studied the AIC strategy on the accelerated lifetime model. Fan and Li (2002) proposed the penalized log-partial likelihood function for the Cox model. Hjort and Claeskens (2006) considered the Cox model with right censored data. Wang et al. (2015) examined the FIC method for panel count data. Du et al. (2017) discussed the development of the quantile regression with right-censoring.

Model selection with survival data is significantly challenged by other features which are commonly possessed by practical data. Typically, survival data with both left-truncation and covariate measurement error are quite common in applications. However, little work has been available to address these two features simultaneously as noted by Yi and Lawless

(2007) and Yi (2017), while addressing either measurement error or biased sampling has attracted extensive attention in survival analysis. Regarding survival data with covariate measurement error, there have been many methods in the literature (e.g., Prentice 1982, Nakamura 1992, Yi 2017, Chapter 3). Concerning biased sampling, left-truncation is a common source which usually arises when a subject dies before the recruitment of study subjects. Different methods have been proposed to account for biased sampling effects; see, for example, Qin and Shen (2010), Huang et al. (2012), Wu et al. (2017), among others.

Although many methods have been available to address some specific features of data (such as measurement error and/or left-truncation) and/or model building, there has been, to the best of our knowledge, no research on dealing with all these issues simultaneously. In Chapter 6, we investigate this important topic and develop valid inference methods which simultaneously accommodate measurement error effects and sampling issues as well as model building for censored survival data. Our model selection development takes the local model misspecification framework (Claeskens and Hjort 2003; Hjort and Claeskens 2003) and specifically focuses on the FIC criterion.

We further explore estimation of the model parameters for conducting post-selection inference. Traditional statistical analysis often first builds the model by selecting important variables and then, based on the model, carries out statistical inferences. This procedure, however, as pointed out by Clyde and George (2004) and Wang et al. (2009), among others, ignores the uncertainty induced from the variable selection process, thus producing estimators with invalid characterization of the associated variability. To circumvent this issue, we take one step back by not producing an estimator from a singlely selected model from a class of candidates, but instead, we average a set of candidate models with suitable weights attached, and then produce an estimator of the model parameter accordingly. We establish the asymptotic properties for the proposed estimator. Our development extends the scope of the usual *model averaging* strategy that has been used frequently under different selection criteria. For example, Hansen (2008) considered least square estimation with the Mallow criterion. Model averaging based on the jackknife (or cross-validation) criterion was discussed by Hansen and Racine (2012). From the Bayesian perspective, Raftery et al. (1997) proposed Bayesian model averaging for linear regression models. Hoeting et al. (1999) summarized techniques of Bayesian model averaging (BMA) and their applications to settings with generalized linear models or survival models. With the FIC selection criterion, many authors explored the model averaging techniques; see Claeskens and Hjort (2003), Claeskens and Carroll (2007), Hjort and Claeskens (2003), Hjort and Claeskens (2006), Wang et al. (2012), Wang et al. (2016), and Zhang and Liang (2011).

Chapter 2

Graphical Models with Error-Prone Variables: Bias Analysis and Valid Inference Methods

2.1 Notation and Models

2.1.1 The Graphical Model

Let $X = (X_1, \cdots, X_p)^\top$ be a p-dimensional random vector. We use a *graph*, denoted as $G = (V, E)$, to describe the relationship among the components of X, where $V = \{1, \cdots, p\}$ includes all the indices of random variables and $V \times V$ contains all the pairs of indices in V. A random variable X_r is called a *vertex* of the graph G if $r \in V$; a pair of random variables $\{X_r, X_s\}$ is called an *edge* of the graph G if $(r, s) \in E \subset V \times V$.

To characterize the distribution of a random vector X, we consider the graphical model with the exponential family distribution,

$$P(X; \boldsymbol{\theta}, \Theta) = \exp\left\{\sum_{r \in V} \theta_r \mathbb{B}(X_r) + \sum_{(s,t) \in E} \theta_{st} \mathbb{B}(X_s)\mathbb{B}(X_t) + \sum_{r \in V} \mathbb{C}(X_r) - \mathbb{A}(\boldsymbol{\theta}, \Theta)\right\}, \quad (2.1)$$

where $\boldsymbol{\theta} = (\theta_1, \cdots \theta_p)^\top$ is a p-dimensional vector of parameters, $\Theta = [\theta_{st}]$ is a $p \times p$ symmetric matrix with zero diagonal elements, and $\mathbb{B}(\cdot)$ and $\mathbb{C}(\cdot)$ are given functions. The

function $\mathbb{A}(\boldsymbol{\theta}, \Theta)$ is the normalizing constant, also called the *log-partition function*, which makes (2.1) be integrated as 1:

$$\mathbb{A}(\boldsymbol{\theta}, \Theta) = \log \int \exp \left\{ \sum_{r \in V} \theta_r \mathbb{B}(X_r) + \sum_{(s,t) \in E} \theta_{st} \mathbb{B}(X_s) \mathbb{B}(X_t) + \sum_{r \in V} \mathbb{C}(X_r) \right\} dX.$$

Formulation (2.1) gives a broad class of models which essentially can cover any distributions. For example, if for $r \in V$, we set $\mathbb{B}(X_r) = \frac{X_r}{\sigma_r}$ and $\mathbb{C}(X_r) = -\frac{X_r^2}{2\sigma_r^2}$ with σ_r being a positive constant, then (2.1) is proportional to

$$\exp \left(\sum_{r \in V} \frac{1}{\sigma_r} \theta_r X_r + \sum_{(s,t) \in E} \frac{1}{\sigma_r \sigma_t} \theta_{st} X_s X_t - \sum_{r \in V} \frac{X_r^2}{2\sigma_r^2} \right), \tag{2.2}$$

yielding the well-known *Gaussian graphical model* (Friedman et al. 2008, Hastie et al. 2015, Lee and Hastie 2015). If we constraint θ_r to be 0 for all $r \in V$ and let $\mathbb{B}(X) = X$ and $\mathbb{C}(X) = 0$ with $X \in \{0, 1\}$, then (2.1) reduces to

$$\exp \left\{ \sum_{(s,t) \in E} \theta_{st} X_s X_t - \mathbb{A}(\Theta) \right\}, \tag{2.3}$$

which is the *Ising model* without the singleton for the simplicity (Ravikumar et al. 2010). The structure (2.1) was discussed by Yang et al. (2015) and Chen et al. (2015) in detail.

For every $r \in V$, let $X_{V \setminus \{r\}}$ denote the $(p-1)$-dimensional subvector of X with its rth component deleted, i.e., $X_{V \setminus \{r\}} = (X_1, \cdots X_{r-1}, X_{r+1}, \cdots, X_p)^\top$. Define the *neighbourhood* of r:

$$\mathcal{N}(r) = \{t \in V \setminus \{r\} : (r, t) \in E\}, \tag{2.4}$$

which is the set containing all the indices of random variables X_t that are dependent on X_r. By Proposition 1 in Yang et al. (2015), the conditional distribution of X_r given $X_{V \setminus \{r\}}$ can be expressed as

$$P\left(X_r | X_{V \setminus \{r\}}\right) = \exp \left\{ \theta_r X_r + X_r \sum_{t \in \mathcal{N}(r)} \theta_{rt} X_t + \mathbb{C}(X_r) - D\left(\theta_r + \sum_{t \in \mathcal{N}(r)} \theta_{rt} X_t \right) \right\}, \tag{2.5}$$

where $D(\cdot)$ is the normalizing constant ensuring the integration of the right-hand side (2.5) equal to one.

2.1.2 Measurement Error and Misclassification

In many applications, random variables are often subject to mismeasurement. Let X^* denote the observed or surrogate version of X. To emphasize the type of the random vector, we use X^{C} (or X^{D}) to replace X if all components of X are continuous (or discrete); when X contains both discrete and continuous components, we write $X = \left(X^{\mathrm{C}\top}, X^{\mathrm{D}\top}\right)^\top$ where X^{C} and X^{D} are the subvectors of continuous and discrete components, respectively. Similarly, we use $X^{*\mathrm{C}}$ and $X^{*\mathrm{D}}$ to express the surrogate vector X^*. In the following, we describe different ways of modeling the relationship between $X^{*\mathrm{C}}$ and X^{C} as well as the relationship between $X^{*\mathrm{D}}$ and X^{D}.

We start with the case where X contains both continuous and discrete subvectors, i.e., $X = \left(X^{\mathrm{C}\top}, X^{\mathrm{D}\top}\right)^\top$. The classical additive measurement error model (Carroll et al. 2006, Chapter 1; Yi 2017, Chapter 2) is assumed to describe the relationship between $X^{*\mathrm{C}}$ and X^{C}:

$$X^{*\mathrm{C}} = X^{\mathrm{C}} + \epsilon, \tag{2.6}$$

where ϵ is independent of X as well as $X^{*\mathrm{D}}$, and $\epsilon \sim N(0, \Sigma_\epsilon)$ with covariance matrix Σ_ϵ. To highlight the idea, Σ_ϵ is assumed known for now.

To feature the relationship between $X^{*\mathrm{D}}$ and X^{D}, we first write the vectors of all possible values of X^{D} as $x_{(1)}, x_{(2)}, \cdots, x_{(m)}$, where m is a positive integer. Assume that

$$P\left(X^{*\mathrm{D}} = x_{(k)} \middle| X^{\mathrm{D}} = x_{(l)}, X^{\mathrm{C}}\right) = P\left(X^{*\mathrm{D}} = x_{(k)} \middle| X^{\mathrm{D}} = x_{(l)}\right). \tag{2.7}$$

Let

$$p_{kl} = P\left(X^{*\mathrm{D}} = x_{(k)} | X^{\mathrm{D}} = x_{(l)}\right) \tag{2.8}$$

be the (mis)classification probability for $k, l = 1, \cdots, m$, and define the $m \times m$ (mis) classification matrix

$$\mathbf{P} = [p_{kl}]_{m \times m} \tag{2.9}$$

with element (k, l) given by (2.8) for $k, l = 1, \cdots, m$.

Noting that

$$\begin{aligned} P\left(X^{*\mathrm{D}} = x_{(k)}\right) &= \sum_{l=1}^{m} P\left(X^{*\mathrm{D}} = x_{(k)}, X^{\mathrm{D}} = x_{(l)}\right) \\ &= \sum_{l=1}^{m} P\left(X^{*\mathrm{D}} = x_{(k)} | X^{\mathrm{D}} = x_{(l)}\right) P\left(X^{\mathrm{D}} = x_{(l)}\right) \\ &= \sum_{l=1}^{m} p_{kl} P\left(X^{\mathrm{D}} = x_{(l)}\right) \end{aligned}$$

for all $k = 1, \cdots, m$, or equivalently, the matrix expression

$$\begin{pmatrix} P\left(X^{*\mathrm{D}} = x_{(1)}\right) \\ \vdots \\ P\left(X^{*\mathrm{D}} = x_{(m)}\right) \end{pmatrix} = \mathbf{P} \begin{pmatrix} P\left(X^{\mathrm{D}} = x_{(1)}\right) \\ \vdots \\ P\left(X^{\mathrm{D}} = x_{(m)}\right) \end{pmatrix}, \tag{2.10}$$

we obtain the constraints for the surrogate $X^{*\mathrm{D}}$ and the true vector X^{D}. To ease notation, we let $MC\left[\mathbf{P}\right](X^{\mathrm{D}})$ denote the misclassification operator and write (2.10) as $X^{*\mathrm{D}} = MC\left[\mathbf{P}\right](X^{\mathrm{D}})$. This expression extends the misclassification operator used by Carroll et al. (2006, p.125) and Küchenhoff et al. (2006) who considered a misclassified binary random variable.

To highlight the idea, in the following development we assume that $\mathbf{P}$ is known; in the last section we discuss the setting where $\mathbf{P}$ is unknown. Consistent with Carroll et al. (2006, p.125), suppose that $\mathbf{P}$ has the spectral decomposition $\mathbf{P} = \Omega \mathbf{D} \Omega^{-1}$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues of $\mathbf{P}$ and Ω is the corresponding matrix of eigenvectors.

In the case where X contains only continuous variables, then only model (2.6) is imposed to describe the relationship between X^* and X where the independence requirement is altered to be the independence between ϵ and X. When X includes only discrete components, then only model (2.8) is needed and the assumption (2.7) disappears.

2.2 Impact of Naive Analysis

In the presence of measurement error or misclassification, it is important to study the impact of ignoring such a feature. That is, if we carry out by replacing the true measurement for X with its surrogate value X^*, then how does the resulting estimator behavior? To

answer this question, we consider a naive analysis which disregards the difference between X^* and X.

For $r \in V$, let $\theta(r) = \left(\theta_r, \theta_{\backslash r}^\top\right)^\top$ with $\theta_{\backslash r} = \left(\theta_{r1}, \cdots, \theta_{r(r-1)}, \theta_{r(r+1)}, \cdots, \theta_{rp}\right)^\top$. For $i = 1, \cdots, n$, the *naive log likelihood function* is determined by

$$\ell_{nv}(\theta(r)) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \log\left\{P\left(X_r^{*(i)} | X^{*(i)}_{V\backslash\{r\}}\right)\right\}, \tag{2.11}$$

where $P\left(X_r^{*(i)} | X^{*(i)}_{V\backslash\{r\}}\right)$ is determined by (2.5) with X replaced by X^*, i.e.,

$$P\left(X_r^* | X^*_{V\backslash\{r\}}\right) = \exp\left\{\theta_r X_r^* + X_r^* \sum_{t\in\mathcal{N}(r)} \theta_{rt} X_t^* + \mathbb{C}(X_r^*) - D\left(\theta_r + \sum_{t\in\mathcal{N}(r)} \theta_{rt} X_t^*\right)\right\}.$$

To carry out variable selection for the variables associated with the parameter vector $\theta_{\backslash r}$, we implement the lasso penalty function for $\theta_{\backslash r}$ (Tibshirani 1996) and obtain the *naive estimator* of $\theta(r)$:

$$\widehat{\theta}_{nv}(r) = \underset{\theta(r)}{\operatorname{argmin}} \left\{\ell_{nv}(\theta(r)) + \lambda_n \left\|\theta_{\backslash r}\right\|_1\right\}, \tag{2.12}$$

where λ_n is the tuning parameter and

$$\left\|\theta_{\backslash r}\right\|_1 = \sum_{j\neq r} |\theta_{rj}|. \tag{2.13}$$

Now we discuss the asymptotic bias of the naive estimator $\widehat{\theta}_{nv}(r)$ for $r \in V$. For a vector a with elements a_i's, let $\|a\|_\infty$ denote the infinity norm defined as $\max_i |a_i|$. For $r \in V$, define $\mathcal{D}_r = E\left\{D''\left(\theta_{0;r} + X^{(i)\top}_{V\backslash\{r\}}\theta_{0;\backslash r}\right)\right\}$ and $\mathcal{Q}_r = E\left\{\left(X^{(i)\top}_{V\backslash\{r\}} X^{(i)\top}_{V\backslash\{r\}}\right) D''\left(\theta_{0;r} + X^{(i)\top}_{V\backslash\{r\}}\theta_{0;\backslash r}\right)\right\}$, and let $\Sigma_{\epsilon;\backslash r}$ be the covariance matrix Σ_ϵ with the rth row and the rth column deleted.

Theorem 2.2.1 *Assume that regularity conditions in Section 3.1 of Yang et al. (2015) hold. Then for any $r \in V$, there exist constants $\widetilde{\alpha} \in (0,1)$ and $\widetilde{\rho} > 0$ such that*

$$\begin{aligned}\left\|\widehat{\theta}_{nv}(r) - \theta_0(r)\right\|_\infty \;\geq\; & \left\{\|\mathcal{Q}_r\|_\infty + \left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_\infty\right\}^{-1}\left(\left\|\frac{\partial \ell_{nv}(\theta_0)}{\partial\theta}\right\|_\infty - \frac{\lambda_n\widetilde{\alpha}}{4(2-\widetilde{\alpha})} - 2\lambda_n\right) \\ & - \left\{1 + \|\mathcal{Q}_r\|_\infty^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_\infty\right\}^{-1} 5\widetilde{\rho}\lambda_n.\end{aligned}$$

Theorem 2.2.1 indicates that the naive estimator $\widehat{\theta}_{nv}(r)$ is not close to the true parameter $\theta_0(r)$ in the infinity norm since a lower bound of $\left\|\widehat{\theta}_{nv}(r) - \theta_0(r)\right\|_\infty$ is positive, provided that the tuning parameter λ_n is smaller than $\left\{\frac{16-7\widetilde{\alpha}}{4(2-\widetilde{\alpha})} + 5\widetilde{\rho}\,\|\mathcal{Q}_r\|_\infty\right\}^{-1} \left\|\frac{\partial \ell_{nv}(\theta_0)}{\partial \theta}\right\|_\infty$, as shown in Appendix A.3. This suggests that the estimated graph generally differs from the true graph. In Section 2.5 we also carry out simulation studies which confirm the theoretical result in Theorem 2.2.1. Consequently, it is imperative to correct for the measurement error effects in order to obtain valid results. In the following two sections we develop inferential procedures which account for measurement error effects.

2.3 Correction Method with Either Continuous or Discrete Variables but not Both

In this section, we consider cases where X contains either all continuous or all discrete random variables, i.e., $X = X^{\mathrm{C}}$ or $X = X^{\mathrm{D}}$. To address mismeasurement effects, we develop a simulation-based three-stage neighbourhood-set likelihood method. The basic idea is to first depict how the bias induced from mismeasurement in the variables is related to the degree of mismeasurement, and then use this relation to extrapolate it to the case without mismeasurement. Such an idea has the similarity to the simulation-extrapolation (SIMEX) approach proposed by Cook and Stefanski (1994) and the misclassification SIMEX (MC-SIMEX) method considered by Küchenhoff et al. (2006). However, the development here is more complex in technical details and the establishment of theoretical results is a lot more challenging.

2.3.1 Inferential Procedures

Stage 1 : Simulation

Let B be a given positive integer and let $\mathcal{Z} = \{\zeta_0, \zeta_1, \cdots, \zeta_M\}$ be a sequence of prespecified values with $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_M$, where M is a positive integer, and ζ_M is a prespecified positive number such as 1.

For a given subject i with $i = 1, \cdots, n$ and $b = 1, \cdots, B$, if the random vector X is continuous with $X = X^{\mathrm{C}}$, then we generate $U_b^{(i)}$ from $N(0, \Sigma_\epsilon)$, and define $W_b^{(i)}(\zeta)$ as

$$W_b^{(i)}(\zeta) = X^{*\mathrm{C}(i)} + \sqrt{\zeta} U_b^{(i)} \tag{2.14}$$

for every $\zeta \in \mathcal{Z}$. If the random vector X is discrete with $X = X^{\mathrm{D}}$, then we generate $W_b^{(i)}(\zeta)$ by

$$W_b^{(i)}(\zeta) = MC[\mathbf{P}^{\zeta}]\left(X^{*\mathrm{D}(i)}\right) \tag{2.15}$$

for every $\zeta \in \mathcal{Z}$, where $\mathbf{P}^{\zeta} = \Omega \mathbf{D}^{\zeta} \Omega^{-1}$.

Stage 2 : Selection

For $r \in V$, replacing X in (2.5) by $W_b^{(i)}(\zeta)$ gives

$$\begin{aligned} & P\left(W_{b,r}^{(i)}(\zeta) \middle| W_{b,V\setminus\{r\}}^{(i)}(\zeta)\right) \\ = \; & \exp\left\{\theta_r W_{b,r}^{(i)}(\zeta) + W_{b,r}^{(i)}(\zeta) \sum_{t\in\mathcal{N}(r)} \theta_{rt} W_{b,t}^{(i)}(\zeta)\right. \\ + \; & \left.\mathbb{C}\left(W_{b,r}^{(i)}(\zeta)\right) - D\left(\theta_r + \sum_{t\in\mathcal{N}(r)} \theta_{rt} W_{b,r}^{(i)}(\zeta)\right)\right\}, \end{aligned} \tag{2.16}$$

and hence, the log-likelihood based on (2.16) is given by

$$\ell_{b,\zeta}(\theta(r)) = -\frac{1}{n}\sum_{i=1}^{n}\left\{P\left(W_{b,r}^{(i)}(\zeta)\middle| W_{b,V\setminus\{r\}}^{(i)}(\zeta)\right)\right\}. \tag{2.17}$$

Then for the given b and ζ, we calculate

$$\widehat{\theta}(r;\zeta,b) = \underset{\theta(r)}{\operatorname{argmin}}\left\{\ell_{b,\zeta}(\theta(r)) + \lambda_n \left\|\theta_{\setminus r}\right\|_1\right\} \tag{2.18}$$

and hence, we define

$$\widehat{\theta}(r;\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}(r;\zeta,b). \tag{2.19}$$

Stage 3 : Extrapolation

Grouping the estimators obtained from (2.19), we obtain the sequence

$$\mathbf{S}_r = \left\{\left(\zeta, \widehat{\theta}(r;\zeta)\right) : \zeta \in \mathcal{Z}\right\}$$

for each $r \in V$. Then we regress $\widehat{\theta}(r;\zeta)$ over ζ by fitting a model

$$\widehat{\theta}\,(r;\zeta) = \mathcal{G}\,(\zeta,\Gamma) + \delta \tag{2.20}$$

to the sequence $\mathbf{S}_r$, where $\mathcal{G}(\cdot,\cdot)$ is a regression function, Γ is the vector of associated parameters, and δ is the noise term. Parameter Γ can be estimated by applying the least squares method to the sequence $\mathbf{S}_r$; and we let $\widehat{\Gamma}$ denote the resulting estimates of Γ.

Finally, we extrapolate model (2.20) by letting $\zeta = -1$ and calculate the predicted vector

$$\widehat{\theta}(r) = \mathcal{G}(-1,\widehat{\Gamma}). \tag{2.21}$$

To obtain the estimator of Θ, one may consider to repeat the same procedures for $r \in V$. However, this procedure contains a flaw since Θ is a symmetric matrix with $\theta_{rt} = \theta_{tr}$ for $t \neq r$, but $\widehat{\theta}_{rt}$ is not necessarily equal to $\widehat{\theta}_{tr}$ for $t \neq r$. To obtain a reasonable estimate of Θ, one may apply the AND or OR rule proposed by Meinshausen and Bühlmann (2006). Corresponding to the elements of $\theta(r)$ defined before (2.11), we write $\widehat{\theta}(r)$ as $\left(\widehat{\theta}_r, \widehat{\theta}_{\setminus r}^{\top}\right)^{\top}$ with $\widehat{\theta}_{\setminus r} = \left(\widehat{\theta}_{r1}, \cdots, \widehat{\theta}_{r(r-1)}, \widehat{\theta}_{r(r+1)}, \cdots, \widehat{\theta}_{rp}\right)^{\top}$. Then the estimated neighbourhood of r is given by $\widehat{\mathcal{N}}(r) = \left\{t \in V \setminus \{r\} : \widehat{\theta}_{rt} \neq 0\right\}$. For any two different nodes r and t, the AND rule declares that (r,t) belongs to the estimated edge set $\widehat{E}$ if both $r \in \widehat{\mathcal{N}}(t)$ and $t \in \widehat{\mathcal{N}}(r)$ hold, while the OR rule allows $(r,t) \in \widehat{E}$ if either $r \in \widehat{\mathcal{N}}(t)$ or $t \in \widehat{\mathcal{N}}(r)$. In this chapter, we use the AND rule.

In implementing the proposed method, choosing sensible tuning parameters is critical. Suggested by Wang et al. (2007), BIC tends to perform well in general, especially in the setting with a penalized likelihood function. Here we employ the BIC approach to select the tuning parameter λ_n. Specifically, we let $\widehat{\theta}(r;\zeta,b,\lambda_n)$ denote the estimator obtained from (2.18) by spelling out the dependence on the tuning parameter. For the given b and ζ, define

$$BIC\,(\lambda_n) = 2n\ell_{b,\zeta}\left(\widehat{\theta}(r;\zeta,b,\lambda_n)\right) + \log(n) \times \mathrm{df}\left(\widehat{\theta}(r;\zeta,b,\lambda_n)\right), \tag{2.22}$$

where $\mathrm{df}\left(\widehat{\theta}(r;\zeta,b,\lambda_n)\right)$ represents the number of non-zero elements in $\widehat{\theta}(r;\zeta,b,\lambda_n)$. The optimal tuning parameter λ_n, denoted by $\widehat{\lambda}_n$, is determined by minimizing (2.22) within suitable ranges of λ_n. As a result, the estimator of $\theta(r)$ is determined by $\widehat{\theta}(r;\zeta,b) = \widehat{\theta}(r;\zeta,b,\widehat{\lambda}_n)$.

2.3.2 Theoretical Results

Here we establish theoretical results to justify the validity of the proposed method in Section 2.3.1.

For any $r \in V$, similar to the notation $\theta(r)$ in Section 2.2, let $\theta_0(r) = \left(\theta_{0;r}, \theta_{0;\backslash r}^\top\right)^\top$ denote the true value of the parameter $\theta(r)$, where

$$\theta_{0;\backslash r} = \left(\theta_{0;r1}, \cdots, \theta_{0;r(r-1)}, \theta_{0;r(r+1)}, \cdots, \theta_{0;rp}\right)^\top$$

is the true value of $\theta_{\backslash r}$. Let $\mathcal{S}_r = \{t \in V \setminus \{r\} : \theta_{rt} \neq 0\}$ denote the set indexing nonzero elements of $\theta_{\backslash r}$ and let $\mathcal{S}_r^c$ be its complement. Write $d_r = |\mathcal{S}_r|$. Let $\theta_{0;\mathcal{S}_r} = (\theta_{0;rt} : t \in \mathcal{S}_r)$ denote the subvector of $\theta_{0;\backslash r}$ containing nonzero elements. Then we write

$$\theta_0(r) = \left(\theta_{0;r}, \theta_{0;\mathcal{S}_r}^\top, \theta_{0;\mathcal{S}_r^c}^\top\right)^\top.$$

Similarly, let $\widehat{\theta}_{\mathcal{S}_r} = \left(\widehat{\theta}_{rt} : t \in \mathcal{S}_r\right)$ denote the subvector of $\widehat{\theta}_{\backslash r}$ containing nonzero estimates, and we write estimator (2.21) as $\widehat{\theta}(r) = \left(\widehat{\theta}_r, \widehat{\theta}_{\mathcal{S}_r}^\top, \widehat{\theta}_{\mathcal{S}_r^c}^\top\right)^\top$.

Let $\nabla_\alpha f(\alpha) = \frac{\partial f(\alpha)}{\partial \alpha}$ and $\nabla_\alpha^2 f(\alpha) = \frac{\partial^2 f(\alpha)}{\partial \alpha \partial \alpha^\top}$ denote the operators of differentiating the function $f(\alpha)$ with respect to α. For the log likelihood function (2.17), let $Q = \nabla^2_{\theta(r)} \ell_{b,\zeta}(\theta_0(r))$. Write

$$Q = \begin{pmatrix} Q_{\mathcal{S}_r\mathcal{S}_r} & Q_{\mathcal{S}_r\mathcal{S}_r^c} \\ Q_{\mathcal{S}_r^c\mathcal{S}_r} & Q_{\mathcal{S}_r^c\mathcal{S}_r^c} \end{pmatrix}$$

as the block matrix of Q with $Q_{\mathcal{S}_r\mathcal{S}_r} = \nabla^2_{\theta_{\mathcal{S}_r}(r)} \ell_{b,\zeta}(\theta_0(r))$, $Q_{\mathcal{S}_r^c\mathcal{S}_r^c} = \nabla^2_{\theta_{\mathcal{S}_r^c}(r)} \ell_{b,\zeta}(\theta_0(r))$, $Q_{\mathcal{S}_r\mathcal{S}_r^c} = \nabla_{\theta_{\mathcal{S}_r^c}(r)} \left\{\nabla_{\theta_{\mathcal{S}_r}(r)} \ell_{b,\zeta}(\theta_0(r))\right\}$, and $Q_{\mathcal{S}_r^c\mathcal{S}_r} = Q_{\mathcal{S}_r\mathcal{S}_r^c}^\top$.

Let $\mathcal{G}'(\zeta, \Gamma) = \frac{\partial \mathcal{G}(\zeta,\Gamma)}{\partial \Gamma}$ and $\mathcal{G}_\Gamma = (\mathcal{G}'(\zeta, \Gamma) : \zeta \in \mathcal{Z})$. For a given constant a, let $sign(a)$ be the sign function which takes value $+1$ if $a > 0$, value -1 if $a < 0$, and 0 otherwise. For a vector (or a matrix) A, $sign(A)$ is defined to be the vector (or the matrix) whose element corresponding to the element a of A is $sign(a)$.

Theorem 2.3.1 *Under regularity conditions (A1)-(A5) in Appendix A.1, we have the following results:*

(a) *Sparsity recovery:*
For every node $r \in V$, the estimated neighbourhood set is equal to the true neighbourhood, i.e.,

$$\widehat{\mathcal{N}}(r) = \mathcal{N}(r),$$

with a large probability.

(b) *Boundness of the estimator:*
For $r \in V$, let $\theta_{0;\mathcal{S}_r}(r) = \left(\theta_{0;r}, \theta_{0;\mathcal{S}_r}^\top\right)^\top$ denote the true value for the subvector of nonzero parameters associated with r, and let $\widehat{\theta}_{\mathcal{S}_r}(r) = \left(\widehat{\theta}_r, \widehat{\theta}_{\mathcal{S}_r}^\top\right)^\top$ denote its estimator obtained based on (2.21). For the infinity norm, we have

$$\left\|\widehat{\theta}_{\mathcal{S}_r}(r) - \theta_{0;\mathcal{S}_r}(r)\right\|_\infty \leq \mathcal{A}\frac{6d_r^{\frac{3}{2}}\lambda_n}{\rho_1}$$

with a large probability, where $\mathcal{A} = \left\|\mathcal{G}'(-1,\Gamma)\left(\mathcal{G}_\Gamma^\top\mathcal{G}_\Gamma\right)^{-1}\mathcal{G}_\Gamma^\top\right\|_1$.

(c) *Sign recovery:*

$$sign\left(\widehat{\theta}_{\mathcal{S}_r}(r)\right) = sign\left(\theta_{0;\mathcal{S}_r}(r)\right)$$

with a large probability.

Theorem 2.3.1 (a) shows that the estimated neighbourhood set is equal to the true neighbourhood set with a large probability for any node $r \in V$, hence suggesting that the estimated graph is equal to the true graphical structure with a large probability. Specifically, based on the definition of the neighbourhood set, the edge set is given by

$$E = \{(s,r) : s,t \in V\} = \bigcup_{r\in V}\{(s,r) : s \in \mathcal{N}(r)\}.$$

In addition, the estimated edge set is determined by

$$\widehat{E} = \bigcup_{r\in V}\{(s,r) : s \in \widehat{\mathcal{N}}(r)\}.$$

Theorem 2.3.1 (a) indicates that there is a constant α_r close to one such that

$$P\{\widehat{\mathcal{N}}(r) = \mathcal{N}(r)\} > \alpha_r.$$

Then we have

$$\begin{aligned} P(\widehat{E} = E) &= P\left(\bigcup_{r\in V}\{(s,r) : s \in \widehat{\mathcal{N}}(r)\} = \bigcup_{r\in V}\{(s,r) : s \in \mathcal{N}(r)\}\right) \\ &\geq \max_{r\in V} P\left(\{(s,r) : s \in \widehat{\mathcal{N}}(r)\} = \{(s,r) : s \in \mathcal{N}(r)\}\right) \\ &> \max_{r\in V} \alpha_r, \end{aligned}$$

which is lower bounded by a constant close to one due to the assumption that the number of nodes of the considered graphical models is fixed. Regarding the subvector of nonzero parameters, Theorem 2.3.1 (b) offers an upper bound for the difference between its estimator and the true value, and Theorem 2.3.1 (c) says that with a high probability, the sign of the estimator is the same as the sign of the true parameter value.

2.4 Inference of Mixed Graphical Model with Both Measurement Error and Misclassification

In this section, we consider a general case with $X = \left(X^{\mathrm{C}\top}, X^{\mathrm{D}\top}\right)^\top$ subject to mismeasurement, where X^{C} is a p_{C}-dimensional continuous random vector and X^{D} is a p_{D}-dimensional discrete random vector. The observed surrogate vector $X^* = \left(X^{*\mathrm{C}\top}, X^{*\mathrm{D}\top}\right)^\top$ is described by the models (2.6) and (2.10).

2.4.1 Model and Method

To show the nature of the variables in X, let V_{C} and V_{D} denote the sets containing all the indices of continuous and discrete random variables, respectively. Let E_{C} and E_{D} represent the sets of edges restricted to the pairs of the indices in V_{C} and V_{D}, respectively, and let E_{CD} denote the set of *heterogeneous* edges for the pairs of the indices in V_{C} and V_{D}, i.e., $E_{\mathrm{CD}} = \left\{(r,t') : r \in V_{\mathrm{C}}, t' \in V_{\mathrm{D}}, \text{ and } X_r^{\mathrm{C}} \text{ and } X_{t'}^{\mathrm{D}} \text{ are dependent}\right\}$. For $r \in V_{\mathrm{C}}$, let

$$\mathcal{N}_{\mathrm{C}}(r) = \{t \in V_{\mathrm{C}} : (r,t) \in E_{\mathrm{C}}\}$$

be the *homogeneous* neighbourhood of r containing all the indices of continuous random variables X_t^{C} that are dependent on X_r^{C}, and let

$$\mathcal{N}_{\mathrm{CD}}(r) = \{t' \in V_{\mathrm{D}} : (r,t') \in E_{\mathrm{CD}}\}$$

be the *heterogeneous* neighbourhood of r containing all the indices of discrete random variables $X_{t'}^{\mathrm{D}}$ that are dependent on X_r^{C}.

For $r' \in V_{\mathrm{D}}$, define

$$\mathcal{N}_{\mathrm{D}}(r') = \{t' \in V_{\mathrm{D}} : (r', t') \in E_{\mathrm{D}}\}$$

and

$$\mathcal{N}_{\mathrm{DC}}(r') = \{t \in V_{\mathrm{C}} : (t, r') \in E_{\mathrm{CD}}\}.$$

Then the mixed graphical model, derived from (2.1), is formed as

$$\begin{aligned} P\left(X^{\mathrm{C}}, X^{\mathrm{D}}\right) &= \exp\left\{\sum_{r\in V_{\mathrm{C}}} \theta_r^{\mathrm{C}} X_r^{\mathrm{C}} + \sum_{(r,t)\in E_{\mathrm{C}}} \theta_{rt}^{\mathrm{C}} X_r^{\mathrm{C}} X_t^{\mathrm{C}} \right. \\ &\quad + \sum_{(r,t')\in E_{\mathrm{CD}}} \theta_{rt'}^{\mathrm{CD}} X_r^{\mathrm{C}} X_{t'}^{\mathrm{D}} + \sum_{r'\in V_{\mathrm{D}}} \theta_{r'}^{\mathrm{D}} X_{r'}^{\mathrm{D}} + \sum_{(r',t')\in E_{\mathrm{D}}} \theta_{r't'}^{\mathrm{D}} X_{r'}^{\mathrm{D}} X_{t'}^{\mathrm{D}} \\ &\quad \left. + \sum_{r\in V_{\mathrm{C}}} \mathbb{C}(X_r^{\mathrm{C}}) + \sum_{r'\in V_{\mathrm{D}}} \mathbb{C}(X_{r'}^{\mathrm{D}}) - \mathbb{A}_{\mathrm{mix}}(\boldsymbol{\theta}, \Theta)\right\}, \end{aligned} \tag{2.23}$$

where $\mathbb{A}_{\mathrm{mix}}(\boldsymbol{\theta}, \Theta)$ is the normalizing constant of (2.23); θ_r^{C} and $\theta_{r'}^{\mathrm{D}}$ are the parameters corresponding to X_r^{C} and $X_{r'}^{\mathrm{D}}$ for $r \in V_{\mathrm{C}}$ and $r' \in V_{\mathrm{D}}$, respectively; θ_{rt}^{C} and $\theta_{r't'}^{\mathrm{D}}$ are the parameters indicating the pairwise dependence of the variables in E_{C} and E_{D}; and $\theta_{rt'}^{\mathrm{CD}}$ is the parameter showing the pairwise dependence of $(X_r^{\mathrm{C}}, X_{t'}^{\mathrm{D}})$ for $r \in V_{\mathrm{C}}$ and $t' \in V_{\mathrm{D}}$. Different from the setting in Chen et al. (2015), our setting can clearly detect the homogeneous edges and heterogeneous edges.

Generalizing the methods of Section 2.3, we propose the simulation-based three-stage neighbourhood likelihood method to correct for mismeasurement effects on estimation of the mixed graphical model.

Stage 1 : Simulation

Given B, $\mathcal{Z} = \{\zeta_0, \cdots, \zeta_M\}$, and M as defined in Stage 1 of Section 2.3.1, we generate the working data $W_b^{\mathrm{C}(i)}(\zeta)$ and $W_b^{\mathrm{D}(i)}(\zeta)$ by (2.14) and (2.15), respectively.

Stage 2 : Estimation

For any $r \in V_{\mathrm{C}}$ and $r' \in V_{\mathrm{D}}$, we define

$$\begin{aligned} \theta_{\setminus r}^{\mathrm{C}} &= \left(\theta_{r1}^{\mathrm{C}}, \cdots, \theta_{r(r-1)}^{\mathrm{C}}, \theta_{r(r+1)}^{\mathrm{C}}, \cdots, \theta_{rp_{\mathrm{C}}}^{\mathrm{C}}\right)^\top, \\ \theta_{\setminus r'}^{\mathrm{D}} &= \left(\theta_{r'1}^{\mathrm{D}}, \cdots, \theta_{r'(r'-1)}^{\mathrm{D}}, \theta_{r'(r'+1)}^{\mathrm{D}}, \cdots, \theta_{r'p_{\mathrm{D}}}^{\mathrm{D}}\right)^\top, \\ \theta_r^{\mathrm{CD}} &= \left(\theta_{r1}^{\mathrm{CD}}, \cdots, \theta_{rp_{\mathrm{D}}}^{\mathrm{CD}}\right)^\top, \quad \text{and} \quad \theta_{r'}^{\mathrm{DC}} = \left(\theta_{1r'}^{\mathrm{CD}}, \cdots, \theta_{p_{\mathrm{C}}r'}^{\mathrm{CD}}\right)^\top. \end{aligned} \tag{2.24}$$

Now we describe estimators separately according to the nature of the variables. Similar to (2.5), for subject $i = 1, \cdots, n$, we first calculate

$$P\left(W_{b,r}^{\mathrm{C(i)}}(\zeta)\middle| W_{b,V_{\mathrm{C}}\setminus\{r\}}^{\mathrm{C}(i)}(\zeta), W_{b}^{\mathrm{D}(i)}(\zeta)\right) = \exp\left\{W_{b,r}^{\mathrm{C}(i)}(\zeta)\,\eta^{\mathrm{C}(i)} - D(\eta^{\mathrm{C}(i)})\right\} \quad (2.25)$$

with $\eta^{\mathrm{C}(i)} = \theta_r^{\mathrm{C}} + \sum_{t\in\mathcal{N}_{\mathrm{C}}(r)} \theta_{rt}^{\mathrm{C}} W_{b,t}^{\mathrm{C}(i)}(\zeta) + \sum_{t'\in V_{\mathrm{D}}} \theta_{rt'}^{\mathrm{CD}} W_{b,t'}^{\mathrm{D}(i)}(\zeta)$. Then we find the estimators, based on (2.25),

$$\begin{aligned}
&\left(\widehat{\theta}_r^{\mathrm{C}}(\zeta,b), \widehat{\theta}_{\setminus r}^{\mathrm{C}}(\zeta,b), \widehat{\theta}_r^{\mathrm{CD}}(\zeta,b)\right) \\
= &\underset{\left(\theta_r^{\mathrm{C}}, \theta_{\setminus r}^{\mathrm{C}}, \theta_r^{\mathrm{CD}}\right)}{\operatorname{argmin}} \left[\frac{-1}{n}\sum_{i=1}^{n} \log\left\{P\left(W_{b,r}^{\mathrm{C}(i)}(\zeta)\middle| W_{b,V_{\mathrm{C}}\setminus\{r\}}^{\mathrm{C}(i)}(\zeta), W_{b}^{\mathrm{D}(i)}(\zeta)\right)\right\}\right. \\
&\left. + \lambda_{n1}\left\|\theta_{\setminus r}^{\mathrm{C}}\right\|_1 + \lambda_{n2}\left\|\theta_r^{\mathrm{CD}}\right\|_1\right], \qquad (2.26)
\end{aligned}$$

where λ_{n1} and λ_{n2} are tuning parameters.

Next, by analogy, for subject $i = 1, \cdots, n$, we calculate

$$P\left(W_{b,r'}^{\mathrm{D(i)}}(\zeta)\middle| W_{b,V_{\mathrm{D}}\setminus\{r'\}}^{\mathrm{D}(i)}(\zeta), W_{b}^{\mathrm{C}(i)}(\zeta)\right) = \exp\left\{W_{b,r'}^{\mathrm{D}(i)}(\zeta)\,\eta^{\mathrm{D}(i)} - D(\eta^{\mathrm{D}(i)})\right\} \quad (2.27)$$

with $\eta^{\mathrm{D}(i)} = \theta_{r'}^{\mathrm{D}} + \sum_{t'\in\mathcal{N}_{\mathrm{D}}(r')} \theta_{r't'}^{\mathrm{D}} W_{b,t'}^{\mathrm{D}(i)}(\zeta) + \sum_{t\in V_{\mathrm{C}}} \theta_{r't}^{\mathrm{DC}} W_{b,t}^{\mathrm{C}(i)}(\zeta)$. Then we find the estimators, using (2.27),

$$\begin{aligned}
&\left(\widehat{\theta}_{r'}^{\mathrm{D}}(\zeta,b), \widehat{\theta}_{\setminus r'}^{\mathrm{D}}(\zeta,b), \widehat{\theta}_{r'}^{\mathrm{DC}}(\zeta,b)\right) \\
= &\underset{\left(\theta_{r'}^{\mathrm{D}}, \theta_{\setminus r'}^{\mathrm{D}}, \theta_{r'}^{\mathrm{DC}}\right)}{\operatorname{argmin}} \left[\frac{-1}{n}\sum_{i=1}^{n} \log\left\{P\left(W_{b,r'}^{\mathrm{D}(i)}(\zeta)\middle| W_{b,V_{\mathrm{D}}\setminus\{r'\}}^{\mathrm{D}(i)}(\zeta), W_{b}^{\mathrm{C}(i)}(\zeta)\right)\right\}\right. \\
&\left. + \lambda_{n3}\left\|\theta_{\setminus r'}^{\mathrm{D}}\right\|_1 + \lambda_{n2}\left\|\theta_{r'}^{\mathrm{DC}}\right\|_1\right], \qquad (2.28)
\end{aligned}$$

where λ_{n2} and λ_{n3} are tuning parameters.

Finally, we calculate

$$\widehat{\theta}_r^{\mathrm{C}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_r^{\mathrm{C}}(\zeta,b), \tag{2.29a}$$

$$\widehat{\theta}_{r'}^{\mathrm{D}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{r'}^{\mathrm{D}}(\zeta,b), \tag{2.29b}$$

$$\widehat{\theta}_{\setminus r}^{\mathrm{C}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{\setminus r}^{\mathrm{C}}(\zeta,b), \tag{2.29c}$$

$$\widehat{\theta}_{\setminus r'}^{\mathrm{D}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{\setminus r'}^{\mathrm{D}}(\zeta,b), \tag{2.29d}$$

$$\widehat{\theta}_r^{\mathrm{CD}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_r^{\mathrm{CD}}(\zeta,b), \quad \text{and} \tag{2.29e}$$

$$\widehat{\theta}_{r'}^{\mathrm{DC}}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{r'}^{\mathrm{DC}}(\zeta,b). \tag{2.29f}$$

Stage 3 : Extrapolation

Similar to Stage 3 in Section 2.3.1, we fit a regression model each of the six sequences obtained from (2.29a) − (2.29f), and extrapolate each model to $\zeta = -1$ and then obtain the estimators $\widehat{\theta}_r^{\mathrm{C}}$, $\widehat{\theta}_{r'}^{\mathrm{D}}$, $\widehat{\theta}_{\setminus r}^{\mathrm{C}}$, $\widehat{\theta}_{\setminus r'}^{\mathrm{D}}$, $\widehat{\theta}_r^{\mathrm{CD}}$ and $\widehat{\theta}_{r'}^{\mathrm{DC}}$.

Similar to the discussion in Section 2.3.1, we note that the vectors θ_r^{CD} and $\theta_{r'}^{\mathrm{DC}}$ in (2.26) and (2.28) share the same parameter $\theta_{rr'}^{\mathrm{CD}}$ with $r \neq r'$, but the preceding procedure does not necessarily yield identical estimator $\widehat{\theta}_{rr'}^{\mathrm{CD}}$ in $\widehat{\theta}_r^{\mathrm{CD}}$ and $\widehat{\theta}_{r'}^{\mathrm{DC}}$. To deal with this discrepancy, we can apply the AND or OR rule to determine the estimators and the estimated graph. Furthermore, for $(r,t) \in E_{\mathrm{C}}$, $(r',t') \in E_{\mathrm{D}}$, and $(r,t') \in E_{\mathrm{CD}}$, the estimated neighbourhood sets for $\mathcal{N}_{\mathrm{C}}(r)$, $\mathcal{N}_{\mathrm{D}}(r')$, $\mathcal{N}_{\mathrm{CD}}(r)$ and $\mathcal{N}_{\mathrm{DC}}(r')$ are given by $\widehat{\mathcal{N}}_{\mathrm{C}}(r) = \left\{t \in V_{\mathrm{C}} : \widehat{\theta}_{rt}^{\mathrm{C}} \neq 0\right\}$, $\widehat{\mathcal{N}}_{\mathrm{D}}(r') = \left\{t' \in V_{\mathrm{D}} : \widehat{\theta}_{r't'}^{\mathrm{D}} \neq 0\right\}$, $\widehat{\mathcal{N}}_{\mathrm{CD}}(r) = \left\{t' \in V_{\mathrm{D}} : \widehat{\theta}_{rt'}^{\mathrm{CD}} \neq 0\right\}$, and $\widehat{\mathcal{N}}_{\mathrm{DC}}(r') = \left\{t \in V_{\mathrm{C}} : \widehat{\theta}_{tr'}^{\mathrm{CD}} \neq 0\right\}$, respectively.

2.4.2 Theoretical Results

In this subsection, we establish theoretical results for the estimators proposed in Section 2.4.1. For $r \in V_{\mathrm{C}}$ and $r' \in V_{\mathrm{D}}$, we write $\theta_{\mathrm{C}}(r) = \left(\theta^{\mathrm{C}}_{r}, \theta^{\mathrm{C}\top}_{\setminus r}, \theta^{\mathrm{CD}\top}_{r}\right)^{\top}$ and $\theta_{\mathrm{D}}(r') = \left(\theta^{\mathrm{D}}_{r'}, \theta^{\mathrm{D}\top}_{\setminus r'}, \theta^{\mathrm{DC}\top}_{r'}\right)^{\top}$ for the parameters defined in (2.24), and we write their true values analogously as $\theta_{0;\mathrm{C}}(r) = \left(\theta^{\mathrm{C}}_{0;r}, \theta^{\mathrm{C}\top}_{0;\setminus r}, \theta^{\mathrm{CD}\top}_{0;r}\right)^{\top}$ and $\theta_{0;\mathrm{D}}(r') = \left(\theta^{\mathrm{D}}_{0;r'}, \theta^{\mathrm{D}\top}_{0;\setminus r'}, \theta^{\mathrm{DC}\top}_{0;r'}\right)^{\top}$. We also write their estimators constructed in Section 2.4.1 analogously as $\widehat{\theta}_{\mathrm{C}}(r) = \left(\widehat{\theta}^{\mathrm{C}}_{r}, \widehat{\theta}^{\mathrm{C}\top}_{\setminus r}, \widehat{\theta}^{\mathrm{CD}\top}_{r}\right)^{\top}$ and $\widehat{\theta}_{\mathrm{D}}(r') = \left(\widehat{\theta}^{\mathrm{D}}_{r'}, \widehat{\theta}^{\mathrm{D}\top}_{\setminus r'}, \widehat{\theta}^{\mathrm{DC}\top}_{r'}\right)^{\top}$.

For $r \in V_{\mathrm{C}}$, let $\mathcal{S}_{\mathrm{C},r}(V_{\mathrm{C}}) = \left\{t \in V_{\mathrm{C}} : \theta^{\mathrm{C}}_{rt} \neq 0\right\}$ and $\mathcal{S}_{\mathrm{C},r}(V_{\mathrm{D}}) = \left\{t' \in V_{\mathrm{D}} : \theta^{\mathrm{CD}}_{rt'} \neq 0\right\}$. Similarly, for $r' \in V_{\mathrm{D}}$, define $\mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{D}}) = \{t' \in V_{\mathrm{D}} : \theta^{\mathrm{D}}_{r't'} \neq 0\}$ and $\mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{C}}) = \{t \in V_{\mathrm{C}} : \theta^{\mathrm{DC}}_{tr'} \neq 0\}$. We further define $\mathcal{S}_{\mathrm{C},r} = \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{C}}) \cup \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{D}})$ and $\mathcal{S}_{\mathrm{D},r'} = \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{D}}) \cup \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{C}})$, and write $d_{\mathrm{C},r} = |\mathcal{S}_{\mathrm{C},r}|$ and $d_{\mathrm{D},r'} = |\mathcal{S}_{\mathrm{D},r'}|$. For $r \in V_{\mathrm{C}}$, let $\theta_{0;\mathcal{S}_{\mathrm{C},r}} = (\theta^{\mathrm{C}}_{0;rt}, \theta^{\mathrm{CD}}_{0;rt'} : t \in \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{C}})$ and $t' \in \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{D}}))^{\top}$ denote the column subvector of $\left(\theta^{\mathrm{C}\top}_{0;\setminus r}, \theta^{\mathrm{CD}\top}_{0;r}\right)^{\top}$ containing nonzero elements, and hence $\theta_{0;\mathrm{C}}(r)$ can be also written as $\left(\theta^{\mathrm{C}}_{0;r}, \theta^{\top}_{0;\mathcal{S}_{\mathrm{C},r}}, \theta^{\top}_{0;\mathcal{S}^c_{\mathrm{C},r}}\right)^{\top}$. By analogy, for $r' \in V_{\mathrm{D}}$, let $\theta_{0;\mathcal{S}_{\mathrm{D},r'}} = \left(\theta^{\mathrm{D}}_{0;r't'}, \theta^{\mathrm{CD}}_{0;tr'} : t \in \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{C}})\right.$ and $\left.t' \in \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{D}})\right)^{\top}$ be the column subvector of $\left(\theta^{\mathrm{D}\top}_{0;\setminus r'}, \theta^{\mathrm{DC}\top}_{0;r'}\right)^{\top}$ containing nonzero elements, and $\theta_{0;\mathrm{D}}(r')$ can then be re-written as $\theta_{0;\mathrm{D}}(r') = \left(\theta^{\mathrm{D}}_{0;r'}, \theta^{\top}_{0;\mathcal{S}_{\mathrm{D},r'}}, \theta^{\top}_{0;\mathcal{S}^c_{\mathrm{D},r'}}\right)^{\top}$.

Similarly, for $r \in V_{\mathrm{C}}$, let $\widehat{\theta}_{\mathcal{S}_{\mathrm{C},r}} = \left(\widehat{\theta}^{\mathrm{C}}_{rt}, \widehat{\theta}^{\mathrm{CD}}_{rt'} : t \in \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{C}})\right.$ and $\left.t' \in \mathcal{S}_{\mathrm{C},r}(V_{\mathrm{D}})\right)^{\top}$ denote the column subvector of $(\widehat{\theta}^{\mathrm{C}\top}_{\setminus r}, \widehat{\theta}^{\mathrm{CD}\top}_{r})^{\top}$ containing nonzero estimates; and for $r' \in V_{\mathrm{D}}$, let $\widehat{\theta}_{\mathcal{S}_{\mathrm{D},r'}} = \left(\widehat{\theta}^{\mathrm{D}}_{r't'}, \widehat{\theta}^{\mathrm{DC}}_{tr'} : t \in \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{C}})\right.$ and $\left.t' \in \mathcal{S}_{\mathrm{D},r'}(V_{\mathrm{D}})\right)^{\top}$ be the column subvector of $(\widehat{\theta}^{\mathrm{D}\top}_{\setminus r'}, \widehat{\theta}^{\mathrm{DC}\top}_{r'})^{\top}$ containing nonzero elements. Therefore, the two estimators $\widehat{\theta}_{\mathrm{C}}(r)$ and $\widehat{\theta}_{\mathrm{D}}(r')$ can also be written as $\widehat{\theta}_{\mathrm{C}}(r) = \left(\widehat{\theta}^{\mathrm{C}}_{r}, \widehat{\theta}^{\top}_{\mathcal{S}_{\mathrm{C},r}}, \widehat{\theta}^{\top}_{\mathcal{S}^c_{\mathrm{C},r}}\right)^{\top}$ and $\widehat{\theta}_{\mathrm{D}}(r') = \left(\widehat{\theta}^{\mathrm{D}}_{r'}, \widehat{\theta}^{\top}_{\mathcal{S}_{\mathrm{D},r'}}, \widehat{\theta}^{\top}_{\mathcal{S}^c_{\mathrm{D},r'}}\right)^{\top}$, respectively.

Theorem 2.4.1 *Under regularity conditions (A1) − (A6) in Appendix A.1 and given $r \in V_C$ and $r' \in V_D$, the following properties hold:*

(a) Sparsity recovery: for the homogeneous-neighbourhood,

$$\widehat{\mathcal{N}}_C(r) = \mathcal{N}_C(r) \quad \textit{and} \quad \widehat{\mathcal{N}}_D(r') = \mathcal{N}_D(r')$$

with a large probability; and for the heterogeneous-neighbourhood,

$$\widehat{\mathcal{N}}_{CD}(r) = \mathcal{N}_{CD}(r) \quad \textit{and} \quad \widehat{\mathcal{N}}_{DC}(r') = \mathcal{N}_{DC}(r')$$

with a large probability;

(b) *Boundness of the estimators:*
For $r \in V_C$, let $\theta_{0;C,\mathcal{S}_{C,r}}(r) = \left(\theta^C_{0;r}, \theta^\top_{0;\mathcal{S}_{C,r}}\right)^\top$ denote the true value for the subvector of nonzero parameters associated with r, and let $\widehat{\theta}_{C;\mathcal{S}_{C,r}}(r) = \left(\widehat{\theta}^C_r, \widehat{\theta}^\top_{\mathcal{S}_{C,r}}\right)^\top$ denote its estimator. For $r' \in V_D$, let $\theta_{0;D,\mathcal{S}_{D,r'}}(r') = \left(\theta^D_{0;r'}, \theta^\top_{0;\mathcal{S}_{D,r'}}\right)^\top$ denote the true value for the subvector of nonzero parameters associated with r', and let $\widehat{\theta}_{D;\mathcal{S}_{D,r'}}(r') = \left(\widehat{\theta}^D_{r'}, \widehat{\theta}^\top_{\mathcal{S}_{D,r'}}\right)^\top$ denote its estimator. Then for the infinity norm, we have

$$\left\|\widehat{\theta}_{C;\mathcal{S}_{C,r}}(r) - \theta_{0;C,\mathcal{S}_{C,r}}(r)\right\|_\infty \leq \frac{6\sqrt{d_{C,r}}\lambda_n}{\rho_1}$$

and

$$\left\|\widehat{\theta}_{D;\mathcal{S}_{D,r'}}(r') - \theta_{0;D,\mathcal{S}_{D,r'}}(r')\right\|_\infty \leq \frac{6\sqrt{d_{D,r'}}\lambda_n}{\rho_1}$$

with large probabilities.

(c) *Sign recovery:*

$$sign\left(\widehat{\theta}_{C;\mathcal{S}_{C,r}}(r)\right) = sign\left(\theta_{0;C,\mathcal{S}_{C,r}}(r)\right) \quad \textit{and} \quad sign\left(\widehat{\theta}_{D;\mathcal{S}_{D,r'}}(r')\right) = sign\left(\theta_{0;D,\mathcal{S}_{D,r'}}(r')\right)$$

with large probabilities.

The proof of the theorem is given in Appendix A.5.

2.5 Numerical Studies

In this section, we conduct numerical studies to assess the performance of the proposed estimators for a variety of settings. We first design the simulation settings and then present the simulation results. Finally, the proposed method is implemented to analyze a real dataset.

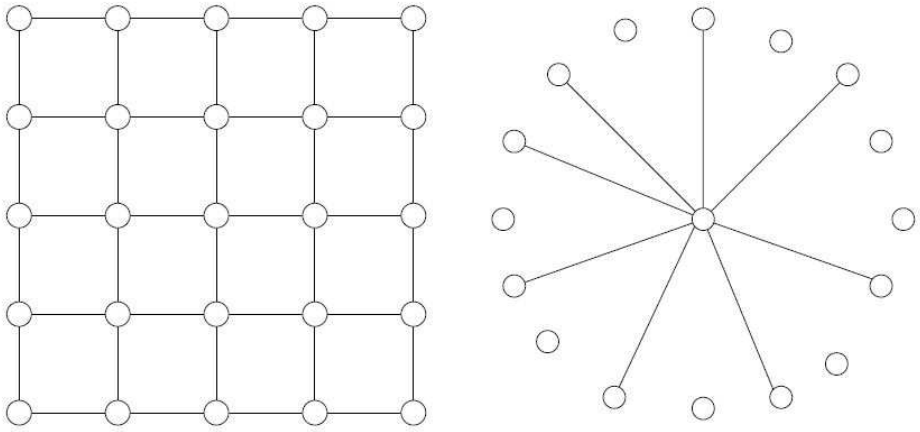


Figure 2.1: The left-hand-side structure is a *Lattice* and the right-hand-side structure is a *Hub*.

2.5.1 Model Settings

Let Θ_0 be the $p \times p$ matrix which is specified to have the network structure, a lattice or a hub structure, as shown in Figure 2.1. Let X denote the p-dimensional random vector which follows the exponential family distribution (2.1), where the continuous random vector X^{C} assumes the structure (2.2), and the discrete random vector X^{D} assumes the form (2.3). Moreover, let p_{C} be the dimension of X^{C} and let p_{D} be the dimension of X^{D}.

For the measurement error process, we consider the following three scenarios:

Scenario 1: *Only continuous variables are subject to measurement error*
In this scenario, all error-prone random variables are continuous, i.e., $X = X^{\mathrm{C}}$, and they assume the classical additive measurement error model

$$X^{*\mathrm{C}} = X^{\mathrm{C}} + \epsilon, \tag{2.30}$$

where ϵ is independent of X^{C}, $\epsilon \sim N(0, \Sigma_\epsilon)$, and Σ_ϵ is a $p \times p$ diagonal matrix with entries σ_ϵ^2, with σ_ϵ^2 set as $0.15^2, 0.5^2$ or 0.75^2 to reflect increasing degrees of measurement error.

Scenario 2: *Only binary variables are subject to misclassification*
In this scenario, all the error-contaminated random variables are considered to be

binary, taking value 1 or -1, i.e., $X = X^{\mathrm{D}}$. In contrast to the misclassification probabilities defined by (2.8), we consider that

$$p_{ll} = P\left(X^{*\mathrm{D}} = x_{(l)} | X^{\mathrm{D}} = x_{(l)}\right) \tag{2.31}$$

assumes a common value, say π, for $l = 1, \cdots, m$, where $m = 2^p$, representing the cardinality of the set $\{-1, 1\}^p$. Thus, the misclassification matrix (2.9) is the $m \times m$ matrix

$$\mathbf{P} = \begin{bmatrix} \pi & 1-\pi & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{2}(1-\pi) & \pi & \frac{1}{2}(1-\pi) & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{2}(1-\pi) & \pi & \frac{1}{2}(1-\pi) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-\pi & \pi \end{bmatrix},$$

where we set $\pi = 0.7, 0.8$, or 0.9 to reflect different degrees of misclassification.

Scenario 3: *Both measurement error and misclassification exist*

In this scenario, we examine the case where both continuous and discrete random variables are subject to mismeasurement by combining Scenarios 1 and 2 with additional assumptions that ϵ in (2.30) is independent of $X^{*\mathrm{D}}$ and (2.7) holds. Consistent with the notation in Section 2.4.1, let $X = \left(X^{\mathrm{C}\top}, X^{\mathrm{D}\top}\right)^\top$ be the vector of the true random vectors and let $X^* = \left(X^{*\mathrm{C}\top}, X^{*\mathrm{D}\top}\right)^\top$ denote the surrogate random vectors with dimension $p = p_{\mathrm{C}} + p_{\mathrm{D}}$.

In implementing the proposed methods, we set $B = 500$ and partition the interval $[0, 2]$ into subintervals with the equal width 0.25 so the resulting cutpoints are set as the values of ζ. We take the regression functions $\mathcal{G}(\cdot, \cdot)$ in (2.20) to be the quadratic function, as suggested by Carroll et al. (2006, p.126). In each setting, we consider different combinations of the sample size n and the dimension of X. In Scenario 1, we set $(n, p_{\mathrm{C}}) = (400, 20)$, $(400, 100)$, or $(200, 400)$; in Scenario 2 we examine $(n, p_{\mathrm{D}}) = (400, 20)$, $(400, 15)$, or $(15, 20)$; and in Scenario 3 we set $(n, p_{\mathrm{C}}, p_{\mathrm{D}}) = (400, 10, 10)$, $(400, 90, 10)$, or $(200, 280, 20)$. We perform 500 simulations for each setting.

2.5.2 Simulation Results

We examine the accuracy of the estimator of Θ by employing the L_1-norm and the Frobenius norm, respectively, given by

$$\|\Delta_\Theta\|_1 = \max_j \sum_i \left|\widehat{\Theta}_{ij} - \Theta_{0,ij}\right|$$

and

$$\|\Delta_\Theta\|_F = \sqrt{\sum_i \sum_j \left|\widehat{\Theta}_{ij} - \Theta_{0,ij}\right|^2},$$

where $\Delta_\Theta = \widehat{\Theta} - \Theta_0$.

To examine the accuracy of variable selection for the graphical structure, we examine the *specificity* (Spe) and the *sensitivity* (Sen) for the estimator $\widehat{\Theta}$. The specificity is defined as the proportion of zero coefficients that are correctly estimated to be zero, and the sensitivity is defined as the proportion of non-zero coefficients that are correctly estimated to be non-zero. The simulation results of both the naive and the proposed methods are reported in Tables 2.1-2.3. As a reference for comparisons, we also use the true values of X for the estimation, and denote this method as "true".

It is apparent that the naive method yields seriously biased results. The values of the L_1-norm and the Frobenius norm are noticeably large whereas the specificities are small for various settings. Although the sensitivities are all good for Scenario 1, they tend to be far off value 1 in Scenarios 2 and 3. As the degree of mismeasurement increases, the bias incurred in the naive method becomes more substantial.

On the contrary, the proposed method obviously outperforms the naive method. The values of the L_1-norm and the Frobenius norm are fairly small, and the specificities and the sensitivities are high for all the three scenarios. As expected, the good performance of the proposed method deteriorates as mismeasurement becomes more severe.

2.5.3 Analysis of Cell-Signalling Data

We implement the proposed method to analyze the cell-signalling data which were discussed by Sachs et al. (2005). This dataset contains $p = 11$ proteins and $n = 7466$ cells. For a given cell, 11 proteins are dyed by different colors using phosphorylation. The amount of dyed proteins can be measured by flow cytometry (a technique that measures

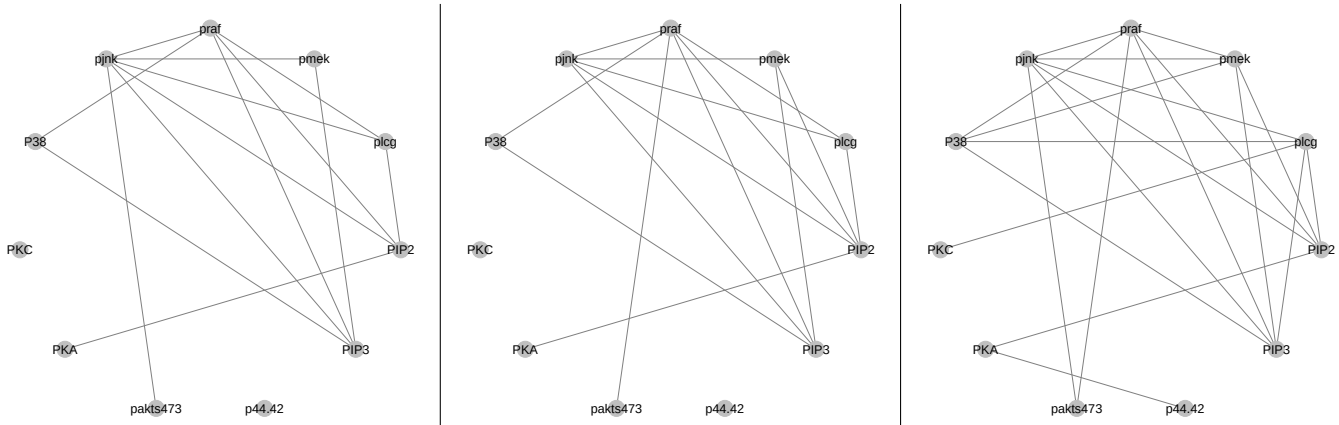
the amount of proteins in a population of cells), so here X represents the amount of a specific protein in one cell. According to Sachs et al. (2005), the cell signaling is a communication process that controls cell activities. When an external signal (e.g., growth factor) binds to its specific cell surface receptor, the activated receptor will interact with signaling proteins inside cell, which triggers a cascade of information flow or signalling pathway. The signaling pathway involves chemical, physical or locational modifications of protein-protein interaction, which leads to a specific cell response such as inducing the transcription and translation to produce certain proteins. It is important to understand the relationship among various signaling proteins/molecules by investigating signaling pathways and the dependence structure of proteins.

To this end, several authors analyzed the data with different approaches. Sachs et al. (2003) fitted a directed acyclic graph (DAG) to the data, and Friedman et al. (2008) implemented the graphical lasso method (GLASSO) to estimate the network structure of the proteins. However, those methods do not address the effects due to mismeasurement, a common phenomenon that is common with the measurement of cell signaling, as pointed out by Bandara et al. (2009) and Yörük et al. (2011).

In our analysis here, we address the feature of mismeasurement and apply the proposed method to analyze this dataset containing error-prone continuous variables. Since the dataset has no additional information such as repeated measurements or validation data for quantifying the degree of measurement error, we conduct sensitivity analyses to investigate how the analysis results are affected by different magnitudes of measurement error. To be precise, let Σ be the sample covariance matrix and we consider $\Sigma + \Sigma_e$ to be the covariance matrix Σ_ϵ for the measurement error model (2.6), where Σ_e is the diagonal matrix with diagonal elements being a common value σ_e^2. We specifically consider $\sigma_e^2 = 0.15^2, 0.5^2$ and 0.75^2, representing an increasing degree of measurement error. The estimated networks are displayed in Figure 2.2. In comparison, we also examine the naive analysis discussed in Section 2.2, and the result is displayed in Figure 2.3.

Figure 2.2 demonstrates that the estimation of the network structure is clearly influenced by the degree of measurement error. Although only one edge is differently identified by incorporating $\sigma_e^2 = 0.15^2$ or 0.50^2 (i.e., `pakts473` is connected with `pjnk` or `praf`), the differences between the settings with $\sigma_e^2 = 0.50^2$ and 0.75^2 are more noticeable. Three extra edges (i.e., `P38` and `plcg`; `PKC` and `plcg`; `PKA` and `p44.42`) are identified with the measurement error degree increased from $\sigma_e^2 = 0.50^2$ to 0.75^2, which also include the edges identified for the setting with $\sigma_e^2 = 0.15^2$. On the other hand, the naive method produces a more complex network structure and the result is clearly different from the proposed method which accounts for the measurement error effects. The naive method indicates more connected variables than the method which corrects for different magnitudes of mea-

surement error. These studies demonstrate that in the presence of measurement error in the variables, ignoring such a feature may produce spurious correlation structures among the variables.



$\sigma_e^2 = 0.15^2$ $\sigma_e^2 = 0.50^2$ $\sigma_e^2 = 0.75^2$

Figure 2.2: Graphical structures of 11 proteins with different degrees of mismeasurement in cell-signalling data.

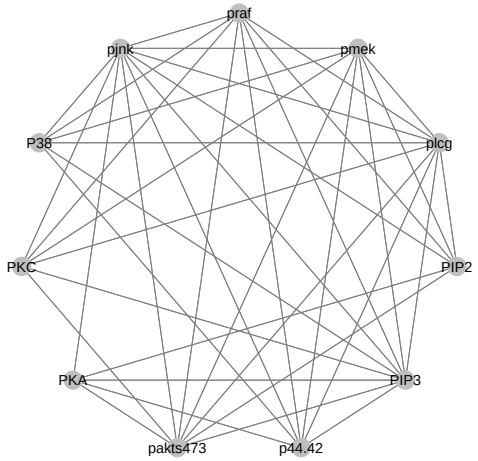


Figure 2.3: Graphical structure of 11 proteins with ignorance of mismeasurement in cell-signalling data.

Table 2.1: Simulation results for the estimators of Θ based on Scenario 1

Model	(n, p_C)	σ_ϵ	Method	Estimator of Θ_0			
				$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 20)$	0.15	naive	2.773	17.986	0.533	1.000
			corrected	1.691	2.021	0.988	1.000
		0.50	naive	2.711	15.833	0.207	1.000
			corrected	1.741	3.768	1.000	1.000
		0.75	naive	3.111	15.653	0.071	1.000
			corrected	1.956	3.318	0.988	1.000
			true	1.316	1.954	1.000	1.000
	$(400, 100)$	0.15	naive	2.555	86.625	0.749	1.000
			corrected	0.822	6.061	1.000	1.000
		0.50	naive	3.434	90.056	0.237	1.000
			corrected	1.064	7.320	0.995	1.000
		0.75	naive	3.911	96.357	0.094	1.000
			corrected	1.549	14.247	0.995	1.000
			true	0.664	4.924	1.000	1.000
	$(200, 400)$	0.15	naive	4.895	106.474	0.298	1.000
			corrected	1.234	34.828	1.000	0.996
		0.50	naive	8.243	521.504	0.082	0.903
			corrected	2.072	56.988	0.999	0.993
		0.75	naive	11.209	621.102	0.030	1.000
			corrected	1.908	106.474	0.998	0.956
			true	0.586	1.954	1.000	1.000
Hub	$(400, 20)$	0.15	naive	4.857	10.375	0.637	1.000
			corrected	1.933	2.448	1.000	0.944
		0.50	naive	4.678	9.716	0.357	1.000
			corrected	1.354	0.799	1.000	1.000
		0.75	naive	4.574	10.173	0.110	1.000
			corrected	1.209	0.674	1.000	1.000
			true	1.651	1.207	1.000	1.000
	$(400, 100)$	0.15	naive	9.108	43.036	0.735	1.000
			corrected	2.259	3.183	1.000	1.000
		0.50	naive	9.671	47.410	0.432	1.000
			corrected	2.366	3.477	1.000	1.000
		0.75	naive	9.676	49.876	0.163	1.000
			corrected	2.659	3.146	0.998	1.000
			true	1.677	1.364	1.000	1.000
	$(200, 400)$	0.15	naive	10.179	206.529	0.433	1.000
			corrected	2.735	15.989	0.999	0.995
		0.50	naive	12.633	261.196	0.165	1.000
			corrected	3.280	27.334	0.998	0.989
		0.75	naive	15.659	332.615	0.041	1.000
			corrected	3.581	31.680	0.997	0.955
			true	0.206	7.686	1.000	1.000

Table 2.2: Simulation results for the estimators of Θ based on Scenario 2

Model	(n, p_{D})	π	Method	Estimator of Θ_0			
				$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 20)$	0.70	naive	2.856	27.741	0.876	0.871
			corrected	1.586	6.607	0.988	0.974
		0.80	naive	2.571	23.328	0.840	0.868
			corrected	1.606	5.473	1.000	0.974
		0.90	naive	2.447	21.853	0.858	0.889
			corrected	1.564	5.129	0.994	0.972
			true	0.879	3.335	1.000	0.996
	$(400, 15)$	0.70	naive	2.670	19.047	0.779	0.681
			corrected	1.981	12.111	1.000	0.912
		0.80	naive	2.843	17.959	0.823	0.727
			corrected	2.338	10.231	1.000	0954
		0.90	naive	2.469	15.573	0.856	0.743
			corrected	1.590	7.038	1.000	0.909
			true	0.885	2.105	1.000	0.964
	$(15, 20)$	0.70	naive	7.769	64.859	0.083	1.000
			corrected	6.034	49.371	0.922	0.952
		0.80	naive	5.313	55.756	0.143	1.000
			corrected	5.130	25.334	0.944	0.903
		0.90	naive	5.017	54.200	0.159	1.000
			corrected	4.130	21.670	0.968	0.968
			true	1.760	4.999	0.986	0.973
Hub	$(400, 20)$	0.70	naive	6.781	15.941	0.676	0.844
			corrected	3.648	4.281	0.961	0.911
		0.80	naive	7.013	15.918	0.692	0.833
			corrected	2.971	4.205	0.967	0.956
		0.90	naive	5.059	10.779	0.720	0.904
			corrected	1.615	1.074	1.000	1.000
			true	1.510	0.977	1.000	1.000
	$(400, 15)$	0.70	naive	4.511	11.085	0.628	0.846
			corrected	2.100	4.8000	0.920	0.906
		0.80	naive	4.572	11.818	0.668	0.615
			corrected	1.903	3.955	0.930	0.923
		0.90	naive	3.879	8.063	0.658	1.000
			corrected	1.611	1.833	1.000	1.000
			true	0.996	0.661	1.000	1.000
	$(15, 20)$	0.70	naive	13.580	71.308	0.050	1.000
			corrected	7.356	36.161	0.918	0.944
		0.80	naive	10.354	43.816	0.054	1.000
			corrected	7.653	26.891	0.938	0.933
		0.90	naive	9.406	36.304	0.055	1.000
			corrected	4.973	18.947	0.929	0.954
			true	2.931	4.650	0.996	0.961

Table 2.3: Simulation results for the estimators of Θ based on Scenario 3

Model	$(n, p_{\mathrm{C}}, p_{\mathrm{D}})$	(σ_ϵ, π)	Method	Estimator of Θ_0			
				$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 10, 10)$	$(0.15, 0.9)$	naive	1.878	10.671	0.811	0.839
			corrected	1.111	3.787	0.982	0.939
		$(0.50, 0.8)$	naive	1.921	11.277	0.746	0.806
			corrected	1.630	5.264	0.952	0.942
		$(0.75, 0.7)$	naive	2.305	11.869	0.686	0.839
			corrected	1.692	6.218	0.953	0.967
			true	0.853	2.210	1.000	0.977
	$(400, 90, 10)$	$(0.15, 0.9)$	naive	1.893	53.422	0.718	0.744
			corrected	1.685	18.655	0.998	0.928
		$(0.50, 0.8)$	naive	2.110	64.900	0.646	0.744
			corrected	1.800	18.708	0.993	0.928
		$(0.75, 0.7)$	naive	2.266	58.259	0.581	0.739
			corrected	1.812	23.935	0.987	0.933
			true	1.200	9.068	1.000	0.965
	$(200, 280, 20)$	$(0.15, 0.9)$	naive	2.538	131.564	0.581	0.866
			corrected	1.914	50.116	0.994	0.958
		$(0.50, 0.8)$	naive	5.223	174.617	0.278	0.872
			corrected	3.767	110.748	0.967	0.958
		$(0.75, 0.7)$	naive	5.991	169.521	0.260	0.879
			corrected	4.136	118.195	0.959	0.956
			true	0.971	16.218	0.999	0.958
Hub	$(400, 10, 10)$	$(0.15, 0.9)$	naive	3.397	5.098	0.599	1.000
			corrected	1.142	1.200	0.945	1.000
		$(0.50, 0.8)$	naive	5.190	8.414	0.500	0.667
			corrected	2.301	4.042	0.941	0.956
		$(0.75, 0.7)$	naive	5.351	11.984	0.439	1.000
			corrected	2.266	7.578	0.959	0.944
			true	0.468	0.755	1.000	0.984
	$(400, 90, 10)$	$(0.15, 0.9)$	naive	9.670	27.398	0.769	0.705
			corrected	5.447	14.104	0.992	0.947
		$(0.50, 0.8)$	naive	10.235	35.899	0.628	0.716
			corrected	7.811	24.237	0.958	0.947
		$(0.75, 0.7)$	naive	12.733	60.029	0.673	0.715
			corrected	7.274	29.614	0.921	0.953
			true	3.337	11.003	0.998	0.973
	$(200, 280, 20)$	$(0.15, 0.9)$	naive	9.401	83.764	0.401	1.000
			corrected	4.585	32.320	0.987	1.000
		$(0.50, 0.8)$	naive	13.650	132.395	0.206	0.875
			corrected	11.254	117.863	0.957	0.977
		$(0.75, 0.7)$	naive	11.708	159.760	0.176	1.000
			corrected	6.257	120.503	0.951	0.996
			true	1.101	1.694	0.999	1.000

Chapter 3

Analysis of Noisy Survival Data under Graphical Proportional Hazards Measurement Error Models

3.1 Notation and Model Setup

3.1.1 The Graphical Model

Let $X = (X_1, \cdots, X_p)^\top$ be a p-dimensional covariates. We use the *graph* to describe the relationship among the components of X. We call each component of X as a vertex and use a line segment, called an edge, to connect two associated components. Specifically, the graph is defined as $G = (V, E)$, where V is the set of vertices with $|V| = p$ and $E \subset V \times V$ is the set of edges. We now use the exponential family distribution to describe the graphical structure of X. The graphical model is formed as

$$P(X; \beta, \Theta) = \exp\left\{\sum_{r\in V} \beta_r \mathbb{B}(X_r) + \sum_{(s,\nu)\in E} \theta_{s\nu}\mathbb{B}(X_s)\mathbb{B}(X_\nu) + \sum_{r\in V} \mathbb{C}(X_r) - \mathbb{A}(\beta, \Theta)\right\}, \quad (3.1)$$

where $\beta = (\beta_1, \cdots \beta_p)^\top$ is the p-dimensional parameter vector, $\Theta = [\theta_{s\nu}]$ is a non-diagonal $p \times p$ symmetric matrix, and $\mathbb{B}(\cdot)$ and $\mathbb{C}(\cdot)$ are given functions. The function $\mathbb{A}(\beta, \Theta)$ is normalizing constant which makes (3.1) be integrated as 1.

Formulation (3.1) gives the broad class of models which essentially can cover any distributions. For example, if $\mathbb{B}(X) = \frac{X}{\sigma}$ and $\mathbb{C}(X) = -\frac{X^2}{2\sigma^2}$ where σ is a positive constant, then

(3.1) yields the well-known *Gaussian graphical model* (Friedman et al. 2008; Hastie et al. 2015). If $\mathbb{B}(X) = X$ and $\mathbb{C}(X) = 0$ with $X \in \{0, 1\}$, then (3.1) reduces to the *Ising model* without the singleton for the simplicity (Ravikumar et al. 2010). The structure (3.1) was discussed by Yang et al. (2015) in detail.

3.1.2 The Cox Model

For an individual, let $\widetilde{T}$ and $\widetilde{C}$ be the failure time and the censoring time, respectively, and let $\delta = I\left(\widetilde{T} \leqslant \widetilde{C}\right)$ be the censoring indicator. Let $T = \min\left\{\widetilde{T}, \widetilde{C}\right\}$ and let X be a p-dimensional random vector of covariates. In standard survival analysis, the Cox proportional hazard (PH) model (Cox, 1972) is often employed with the hazard function specified as

$$\lambda(t|X) = \lambda_0(t) \exp\left\{g\left(X; \alpha\right)\right\}, \tag{3.2}$$

where $\lambda_0(\cdot)$ is the unspecified baseline hazard function, and $g\left(X; \alpha\right)$ is the link function of the linear predictor with the covariate vector X and the unknown parameter α. For instance, $g\left(X; \alpha\right) = \alpha^\top X$ is a common choice.

Model (3.2) is perhaps the most widely used model for handling survival data. However, there is a limitation. The covariates appear equally in the model formulation, and the possible dependence structures of the covariates are not incorporated. The covariate vector X may possess complex association or network structures, and their effects on the survival process cannot be appropriately described if such structures are not accommodated in modeling and/or estimation procedures. To deal with such settings, we describe X using the graphical model and assume that X follows a distribution specified by (3.1); to link the survival time with the covariates, we extend (3.2) with the structure of (3.1) accommodated.

One approach is to set $\exp\left\{g(X; \alpha)\right\}$ in (3.2) to include the terms of X in $P\left(X; \beta, \Theta\right)$ determined by (3.1), where α is the parameter vector consisting of the elements of β and Θ. This immediately yields a generalized Cox PH model

$$\lambda(t|X) \;=\; \lambda_0^*(t) \exp\left\{\sum_{r\in V} \beta_r \mathbb{B}(X_r) + \sum_{(s,\nu)\in E} \theta_{s\nu} \mathbb{B}(X_s)\mathbb{B}(X_\nu) + \sum_{r\in V} \mathbb{C}(X_r)\right\}, \tag{3.3}$$

where $\lambda_0^*(t)$ is the baseline hazard function; and $\mathbb{B}(\cdot)$, $\mathbb{C}(\cdot)$, the β_r, and the $\theta_{s\nu}$ are defined as in (3.1). One may equivalently re-write (3.3) as

$$\lambda(t|X) = \lambda_0(t) \exp\left\{\sum_{r\in V} \beta_r \mathbb{B}(X_r) + \sum_{(s,\nu)\in E} \theta_{s\nu} \mathbb{B}(X_s)\mathbb{B}(X_\nu) + \sum_{r\in V} \mathbb{C}(X_r) - \mathbb{A}\left(\beta, \Theta\right)\right\}, \tag{3.4}$$

where $\lambda_0(t)$ is treated as a baseline hazard function that differs from $\lambda_0^*(t)$ by a constant factor $\exp\{\mathbb{A}(\beta, \Theta)\}$, and the function $\mathbb{A}(\cdot)$ is used to make the exponential function in (3.4) be identical to $P(X; \beta, \Theta)$ in (3.1).

While (3.3) and (3.4) can both equivalently describe a generalized Cox PH model with graphic structures of covariates X reflected, using (3.4) may be more convenient due to the probability nature of (3.1). As a result, we use (3.4) for the following development.

Without loss of general interest, we take $\mathbb{B}(X_r)$ as the linear function $\mathbb{B}(X_r) = X_r$ for $r \in V$ in (3.4) although other forms can be specified for the function $\mathbb{B}(\cdot)$. Since $\mathbb{C}(X_r)$ does not contain information of parameters β and Θ, sometimes, it is more convenient to express (3.4) as

$$\lambda(t|X) \propto \lambda_0(t) \exp\left\{\sum_{r\in V} \beta_r X_r + \sum_{(s,\nu)\in E} \theta_{s\nu} X_s X_\nu - \mathbb{A}(\beta, \Theta)\right\}, \tag{3.5}$$

where parameter β_r reflects marginal effects of the covariate X_r, and $\theta_{s\nu}$ is the parameter to determine the dependence structure of two covariates X_s and X_ν. Model (3.5) can be viewed as an extension of the usual Cox PH model by adding all the pairwise interaction terms of the covariate variables.

Inference about the parameters β and Θ may proceed with the partial likelihood method. To see this, consider a random sample of n subjects and we use the same symbols as before with subscript i added to the corresponding quantities for subject i. Let $N_i(t) = I(T_i < t, \delta_i = 1)$ and $Y_i(t) = I(T_i \geqslant t)$. Given the model (3.5) with right-censoring, we construct the log partial likelihood

$$\begin{aligned} \ell(\beta, \Theta) \;=\; & \sum_{i=1}^{n} \int \left[\left(\sum_{r\in V} X_r^{(i)} \beta_r + \sum_{(s,\nu)\in E} X_s^{(i)} X_\nu^{(i)} \theta_{s\nu} \right) \right. \\ & \left. - \log\left\{ \sum_{j=1}^{n} \exp\left(\sum_{r\in V} X_r^{(i)} \beta_r + \sum_{(s,\nu)\in E} X_s^{(i)} X_\nu^{(i)} \theta_{s\nu} \right) Y_j(t) \right\} \right] dN_i(t). \end{aligned} \tag{3.6}$$

3.1.3 Measurement Error and Misclassification

In practice, covariates, either continuous or discrete, are often subject to mismeasurement. That is, we may encounter *measurement error* in continuous covariates or *misclassification*

in discrete covariates. Suppose that X is written as $X = \left(X_{\mathrm{C}}^{\top}, X_{\mathrm{D}}^{\top}\right)^{\top}$ where X_{C} is the subvector consisting of continuous components and X_{D} is the subvector consisting of discrete components. Let p_{C} and p_{D} denote the dimension of X_{C} and X_{D}, respectively. Let X^* denote the observed or surrogate version of X, and we write $X^* = \left({X_{\mathrm{C}}^*}^{\top}, {X_{\mathrm{D}}^*}^{\top}\right)^{\top}$ where ${X_{\mathrm{C}}^*}^{\top}$ and ${X_{\mathrm{D}}^*}^{\top}$ are the observed versions of X_{C} and X_{D}, respectively, and $p = p_{\mathrm{C}} + p_{\mathrm{D}}$. The true covariates X and its surrogate X^* are linked by the following models.

Conditional on $\{X_{\mathrm{C}}, X_{\mathrm{D}}, X_{\mathrm{D}}^*\}$, X_{C}^* follows the classical additive measurement error model (Carroll et al. 2006; Yi 2017)

$$X_{\mathrm{C}}^* = X_{\mathrm{C}} + \epsilon, \tag{3.7}$$

where ϵ is independent of $\left\{X, X_{\mathrm{D}}^*, \widetilde{T}, \widetilde{C}\right\}$, and $\epsilon \sim N(0, \Sigma_\epsilon)$ with covariance matrix Σ_ϵ.

In terms of the subvector X_{D} of discrete components, we let $x_{(1)}, x_{(2)}, \cdots,$ and $x_{(m)}$ denote all the possible values of X_{D}. We assume that $P\left(X_{\mathrm{D}}^* = x_{(k)} | X_{\mathrm{D}} = x_{(l)}, X_{\mathrm{C}}\right) = P\left(X_{\mathrm{D}}^* = x_{(k)} | X_{\mathrm{D}} = x_{(l)}\right)$ for $k, l = 1, \cdots, m$, and let $p_{kl} = P\left(X_{\mathrm{D}}^* = x_{(k)} | X_{\mathrm{D}} = x_{(l)}\right)$ be the (mis)classification probability for $k, l = 1, \cdots, m$ (Yi 2017, p.71, Chapter 2). For ease of exposition, we define the $m \times m$ (mis)classification matrix $\mathbf{P} = [p_{kl}]$ whose element (k, l) is given by p_{kl} for $l, k = 1, \cdots, m$. In addition, we have $P\left(X_{\mathrm{D}}^* = x_{(k)}\right) = \sum_{l=1}^{m} p_{kl} P\left(X_{\mathrm{D}} = x_{(l)}\right)$ for all $k = 1, \cdots, m$, leading to the matrix expression

$$\begin{pmatrix} P\left(X_{\mathrm{D}}^* = x_{(1)}\right) \\ \vdots \\ P\left(X_{\mathrm{D}}^* = x_{(m)}\right) \end{pmatrix} = \mathbf{P} \begin{pmatrix} P\left(X_{\mathrm{D}} = x_{(1)}\right) \\ \vdots \\ P\left(X_{\mathrm{D}} = x_{(m)}\right) \end{pmatrix}. \tag{3.8}$$

Therefore, the surrogate vector X_{D}^* can be generated from the true covariate vector X_{D} through (3.8). To ease notation, we let $MC\left[\mathbf{P}\right](X_{\mathrm{D}})$ denote the misclassification operator indicated by (3.8) and notationally write (3.8) as $X_{\mathrm{D}}^* = MC\left[\mathbf{P}\right](X_{\mathrm{D}})$. Such a misclassification operator was used by Carroll et al. (2006, p.125) and Küchenhoff et al. (2006) for a misclassified binary variable. To highlight the key idea, we assume that Σ_ϵ and $\mathbf{P}$ are known for now. In addition, as discussed in Carroll et al. (2006, p.125), we assume that $\mathbf{P}$ has the spectral decomposition $\mathbf{P} = \Omega \mathbf{D} \Omega^{-1}$, where $\mathbf{D}$ is the diagonal matrix with diagonal elements being the eigenvalues of $\mathbf{P}$, and Ω is the corresponding matrix of eigenvectors.

3.2 The Methodology

We consider the case where X is subject to mismeasurement, as described in Section 3.1.3, and some components of X are unimportant in the model (3.5). To conduct valid inference,

we need to not only correct for the mismeasurement effects, but also select important covariate variables.

3.2.1 Inferential Procedures

To address mismeasurement effects and select active covariate variables simultaneously, we develop a simulation-based three-stage procedure.

Stage 1 : Simulation

Let B be a given positive integer and let $\mathcal{Z} = \{\zeta_0, \zeta_1, \cdots, \zeta_M\}$ be a sequence of prespecified values with $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_M$, where M is a positive integer, and ζ_M is a prespecified positive number such as $\zeta_M = 1$.

For a given subject i with $i = 1, \cdots, n$ and $b = 1, \cdots, B$, we generate $U_b^{(i)}$ from $N(0, \Sigma_\epsilon)$. Then for vector $X_{\mathrm{C}}^{*(i)}$ and we define $W_{\mathrm{C},b}^{(i)}(\zeta)$ as

$$W_{\mathrm{C},b}^{(i)}(\zeta) = X_{\mathrm{C}}^{*(i)} + \sqrt{\zeta} U_b^{(i)} \tag{3.9}$$

for every $\zeta \in \mathcal{Z}$. For the discrete vector $X_{\mathrm{D}}^{(i)}$, we generate $W_{\mathrm{D},b}^{(i)}(\zeta)$ by the operator

$$W_{\mathrm{D},b}^{(i)}(\zeta) = MC\left[\mathbf{P}^{\zeta}\right] X_{\mathrm{D}}^{*(i)}, \tag{3.10}$$

where $\mathbf{P}^{\zeta} = \Omega \mathbf{D}^{\zeta} \Omega^{-1}$. Let $W_b^{(i)}(\zeta) = \left(W_{\mathrm{C},b}^{(i)\top}(\zeta), W_{\mathrm{D},b}^{(i)\top}(\zeta)\right)^{\top}$ and we call $W_b^{(i)}(\zeta)$ the *working data* for any $b = 1, \cdots, B$, $\zeta \in \mathcal{Z}$ and $i = 1, \cdots, n$.

Stage 2 : Selection

Let $\ell_{b,\zeta}(\beta, \Theta)$ denote the partial likelihood function (3.6) with $X^{(i)}$ replaced by $W_b^{(i)}(\zeta)$, which is given by

$$\begin{aligned} & \ell_{b,\zeta}(\beta, \Theta) \\ = & -\sum_{i=1}^{n} \int \left[\left(\sum_{r \in V} W_{b,r}^{(i)}(\zeta) \beta_r + \sum_{(s,\nu) \in E} W_{b,s}^{(i)}(\zeta) W_{b,\nu}^{(i)}(\zeta) \theta_{s\nu} \right) \right. \\ & \left. - \log \left\{ \sum_{j=1}^{n} \exp \left(\sum_{r \in V} W_{b,r}^{(i)}(\zeta) \beta_r + \sum_{(s,\nu) \in E} W_{b,s}^{(i)}(\zeta) W_{b,\nu}^{(i)}(\zeta) \theta_{s\nu} \right) Y_j(t) \right\} \right] dN_i(t). \end{aligned} \tag{3.11}$$

To do the variable selection, we propose to use different penalty functions for β and Θ and implement the adaptive lasso procedure. The penalty function for β is given by

$$\rho_1(\beta) = \sum_{r \in V} w_r |\beta_r|, \tag{3.12}$$

where $w = (w_1, \cdots, w_p)$ is the vector of weights. As suggested by Zou (2006), the weight can be set as $w_r = |\beta_r|^{-\gamma_1}$ for any $\gamma_1 > 0$ and $r \in V$. For the parameter Θ associated with graphical structure, the penalty function is given by

$$\rho_2(\Theta) = \sum_{\nu \neq s} v_{s\nu} |\theta_{s\nu}|, \tag{3.13}$$

where the weight $v_{s\nu}$ can be set as $v_{s\nu} = |\theta_{s\nu}|^{-\gamma_2}$ for some $\gamma_2 > 0$. To find a value of $v_{s\nu}$, we may first obtain a consistent estimate of Θ and then take the weight as $\widehat{v}_{s\nu} = \left|\widehat{\theta}_{s\nu}\right|^{-\gamma_2}$.

As a result, for the given b and ζ, the proposed estimator is given by

$$\left(\widehat{\beta}_b(\zeta), \widehat{\Theta}_b(\zeta)\right) = \underset{\beta,\Theta}{\operatorname{argmin}} \left\{\ell_{b,\zeta}(\beta, \Theta) + \lambda_{n1}\rho_1(\beta) + \lambda_{n2}\rho_2(\Theta)\right\}, \tag{3.14}$$

where λ_{n1} and λ_{n2} are the tuning parameters associated with (3.12) and (3.13), respectively. Moreover, we define

$$\widehat{\beta}(\zeta) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\beta}_b(\zeta) \quad \text{and} \quad \widehat{\Theta}(\zeta) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\Theta}_b(\zeta). \tag{3.15}$$

Stage 3 : Extrapolation

For the two sequences $\left\{\left(\zeta, \widehat{\beta}(\zeta)\right) : \zeta \in \mathcal{Z}\right\}$ and $\left\{\left(\zeta, \widehat{\Theta}(\zeta)\right) : \zeta \in \mathcal{Z}\right\}$ obtained from (3.15), we fit a regression model to each of the two sequences

$$\widehat{\beta}(\zeta) = \varphi_1(\zeta; \Gamma_1) + \epsilon_1 \quad \text{and} \quad \widehat{\Theta}(\zeta) = \varphi_2(\zeta; \Gamma_2) + \epsilon_2, \tag{3.16}$$

where $\varphi_1(\cdot;\cdot)$ and $\varphi_2(\cdot;\cdot)$ are the user-specific regression functions, Γ_1 and Γ_2 are the associated parameters, and ϵ_1 and ϵ_2 are the noise terms. Parameters Γ_1 and Γ_2 can be estimated by the least square method; and we let $\widehat{\Gamma}_1$ and $\widehat{\Gamma}_2$ denote the resulting estimates of Γ_1 and Γ_2, respectively.

Finally, we calculate the predicted values

$$\widehat{\beta} = \varphi_1\left(-1; \widehat{\Gamma}_1\right) \quad \text{and} \quad \widehat{\Theta} = \varphi_2\left(-1; \widehat{\Gamma}_2\right) \tag{3.17}$$

and take $\widehat{\beta}$ as the estimator of β. Note that Θ is a symmetric matrix, i.e., $\theta_{s\nu} = \theta_{\nu s}$ for $s \neq \nu$, but $\widehat{\theta}_{s\nu}$ is usually not equal to $\widehat{\theta}_{\nu s}$ for $s \neq \nu$. Hence, to obtain a reasonable estimate of Θ, we apply the AND rule proposed by Meinshausen and Bühlmann (2006) so that $\widehat{\theta}_{s\nu} = \widehat{\theta}_{\nu s} = \max\{\widehat{\theta}_{s\nu}, \widehat{\theta}_{\nu s}\}$.

The key idea of the proposed three-stage procedure is to use simulated surrogate measurements to delineate the patterns of different degrees of measurement error on inference results. The first and third stages generalize the simulation-extrapolation (SIMEX) method (Cook and Stefanski 1994) and the MC-SIMEX method (Küchenhoff et al. 2006) which are respectively applicable to error-contaminated continuous and discrete covariates. Our steps embrace a more general setting where error-prone covariates are a mixture of continuous and discrete covariates. Different from the conventional graphical model framework which focuses on either continuous random variables or discrete random variables, we allow both continuous and discrete random variables to be accommodated and they may be subject to mismeasurement.

The second stage of the proposed method undertakes the selection of important variables for settings with different magnitudes of mismeasurement. Although we adopt the adaptive lasso (Zou 2006) which was developed for variable selection in the absence of measurement error, it is imperative to address the impact of measurement error on variable selection in this step.

3.2.2 Implementation Algorithm

To implement the three-stage procedure described in Section 3.2.1, we apply the *coordinate-descent* approach (e.g., Ravikumar et al. 2010; Yang et al. 2015). For the given $b = 1, \cdots, B$ and $\zeta \in \mathcal{Z}$, we carry out the following steps:

Step 1 : Choose an initial value of Θ, and denote it as $\widehat{\Theta}^{(0)}$.

Step 2 : Given $\widehat{\Theta}^{(k-1)}$ with $k = 1, 2, \cdots$, update β by finding

$$\widehat{\beta}^{(k)} = \underset{\beta}{\operatorname{argmin}} \left\{ \ell_{b,\zeta}\left(\beta \,\middle|\, \widehat{\Theta}^{(k-1)}\right) + \lambda_{n1}\rho_1(\beta) \right\}. \tag{3.18}$$

Step 3 : Given $\widehat{\beta}^{(k)}$ with $k = 1, 2, \cdots$, update $\widehat{\Theta}$ by finding

$$\widehat{\Theta}^{(k)} = \underset{\Theta}{\operatorname{argmin}} \left\{ \ell_{b,\zeta} \left(\Theta \left| \widehat{\beta}^{(k)} \right. \right) + \lambda_{n2} \rho_2(\Theta) \right\}. \tag{3.19}$$

Step 4 : Repeat Steps 2 and 3 until convergence, and let $\widehat{\beta}_b(\zeta)$ and $\widehat{\Theta}_b(\zeta)$, respectively, denote the limit of $\widehat{\beta}^{(k)}$ and $\widehat{\Theta}^{(k)}$ as $k \to \infty$.

The implementation of the proposed procedure requires starts with an initial value of Θ. Although our numerical experience does not suggest sensitivity of the results to a specific choice of an initial value, a reasonably chosen initial value of Θ is often helpful for the implementation. Since Θ is mainly used to indicate the pairwise relationship among the covariate components, so intuitively, an initial value of Θ can be set as the covariance matrix of covariate with the diagonal elements replaced by zeros, i.e., setting. $\widehat{\Theta}^{(0)} = \operatorname{cov}(W_b(\zeta)) - \operatorname{diag}\{\operatorname{cov}(W_b(\zeta))\}$, where $\operatorname{diag}(A)$ represents the diagonal matrix of A for any square matrix A.

In implementing the proposed method, choosing sensible tuning parameters is critical. Suggested by Wang et al. (2007), BIC tends to outperform among those procedures, especially in the setting with a penalized likelihood function. Consequently, we employ the BIC approach to select the tuning parameters λ_{n1} and λ_{n2}.

3.2.3 Estimation of the Cumulative Baseline Hazards Function

Estimation of the estimator of $\Lambda_0(\cdot)$ can be carried out after the model parameters β and Θ are obtained as described in Sections 3.2.1 and 3.2.2. For $j \in V$ and $(s, \nu) \in E$, let $\widehat{\beta}_j$ and $\widehat{\theta}_{s\nu}$ denote the corresponding elements of $\widehat{\beta}$ and $\widehat{\Theta}$ which are determined by (3.17). Define $\widehat{\mathcal{S}}_1 = \left\{ j \in V : \ \widehat{\beta}_j \neq 0 \right\}$, $\widehat{\mathcal{S}}_2 = \left\{ (s, \nu) \in E : \ \widehat{\theta}_{s\nu} \neq 0 \right\}$, and $\widehat{\mathcal{N}} = \widehat{\mathcal{S}}_1 \cup \widehat{\mathcal{S}}_2$. Let $\widehat{\beta}_{\mathcal{S}_1}$ and $\widehat{\Theta}_{\mathcal{S}_2}$ denote the subvector of $\widehat{\beta}$ and submatrix of $\widehat{\Theta}$ which contain informative variables and dependent pairs in $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively.

For the data generated at Stage 1 in Section 3.2.1, for $b = 1, \cdots, B$ and $\zeta \in \mathcal{Z}$, we calculate

$$\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t; b, \zeta) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\sum_{i=1}^{n} g\left(W_b^{(i)}(\zeta); \widehat{\beta}_{\mathcal{S}_1}, \widehat{\Theta}_{\mathcal{S}_2} \right) Y_i(u)} \tag{3.20}$$

for a given time t, where

$$g\left(W_b^{(i)}(\zeta);\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right)=\exp\left\{\sum_{r\in V\cap\widehat{\mathcal{S}}_1}W_{b,r}^{(i)}(\zeta)\widehat{\beta}_r+\sum_{(s,\nu)\in E\cap\widehat{\mathcal{S}}_2}W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\widehat{\theta}_{s\nu}\right\}.$$

Taking averaging on (3.20) with respect to b gives

$$\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;\zeta)=\frac{1}{B}\sum_{b=1}^{B}\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;b,\zeta)\ \text{ for }\ \zeta\in\mathcal{Z},\tag{3.21}$$

where t is a given time. Then the estimate of $\Lambda_0(t)$ at a given time point $t\in[0,\tau]$ can be obtained by adapting Stage 3 in Section 3.2.1.

3.3 Theoretical Results

Let $\nabla_\alpha f(\alpha)=\frac{\partial f(\alpha)}{\partial\alpha}$ and $\nabla_\alpha^2 f(\alpha)=\frac{\partial^2 f(\alpha)}{\partial\alpha\partial\alpha^\top}$ denote the operators of differentiating the function $f(\alpha)$ with respect to α. Define

$$U_{\beta;b,\zeta}(\beta,\Theta)=\nabla_\beta\ell_{b,\zeta}(\beta,\Theta)\ \text{ and }\ U_{\Theta;b,\zeta}(\beta,\Theta)=\nabla_\Theta\ell_{b,\zeta}(\beta,\Theta).\tag{3.22}$$

By the arguments of Carroll et al. (2006, p.126), we have that

$$W_{\mathrm{C},b}^{(i)}(\zeta)\sim N\left(X_{\mathrm{C}}^{(i)},(1+\zeta)\Sigma_\epsilon\right)\ \text{ and }\ W_{\mathrm{D},b}^{(i)}(\zeta)=MC[\mathbf{P}^{1+\zeta}]\left(X_{\mathrm{D}}^{(i)}\right).$$

When $\zeta\to-1$, $W_{\mathrm{C},b}^{(i)}(\zeta)$ and $W_{\mathrm{D},b}^{(i)}(\zeta)$ are respectively close to $X_{\mathrm{C}}^{(i)}$ and $X_{\mathrm{D}}^{(i)}$ in the sense that $(1+\zeta)\Sigma_\epsilon$ is close to zero and $\mathbf{P}^{1+\zeta}$ is close to the identity matrix. Then similar to the derivations of Lawless (2003, p.351), we can show that

$$E\left\{U_{\beta;b,\zeta}(\beta,\Theta)\right\}=0\text{ and }E\left\{U_{\Theta;b,\zeta}(\beta,\Theta)\right\}=0.\tag{3.23}$$

We now further define

$$I_{\beta;b,\zeta}(\beta,\Theta)=\nabla_\beta^2\ell_{b,\zeta}(\beta,\Theta)\ \text{ and }\ I_{\Theta;b,\zeta}(\beta,\Theta)=\nabla_\Theta^2\ell_{b,\zeta}(\beta,\Theta).\tag{3.24}$$

Let

$$G_{b,\zeta}(u;\beta,\Theta)=\sum_{i=1}^{n}\exp\left(\sum_{r\in V}W_{b,r}^{(i)}(\zeta)\beta_r+\sum_{(s,\nu)\in E}W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\theta_{s\nu}\right)Y_i(u),\tag{3.25}$$

and write $G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta) = \nabla_\beta G_{b,\zeta}(u;\beta,\Theta)$ and $G^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta) = \nabla_\Theta G_{b,\zeta}(u;\beta,\Theta)$.

Let β_0 and Θ_0 denote true values of β and Θ. We first present the consistency of the proposed estimator $\left(\widehat{\beta}^\top, \mathrm{vec}(\widehat{\Theta})^\top\right)^\top$, where $\mathrm{vec}(\cdot)$ stands for the column vectorization of a matrix. For a vector a, we write $\|a\|_2^2 = a^\top a$.

Theorem 3.3.1 *Under regularity conditions (C1) − (C9) in Appendix B.1, we have that*

$$\left\| \left(\widehat{\beta}^\top, vec(\widehat{\Theta})^\top\right)^\top - \left(\beta_0^\top, vec\left(\Theta_0\right)^\top\right)^\top \right\|_2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Next, we discuss the property of recovery which demonstrate how the selected variables can reflect the underlying true structure. Let

$$\mathcal{S}_1 = \{r \in V :\ \beta_r \neq 0\}, \quad \mathcal{S}_2 = \{(s,\nu) \in E :\ \theta_{s\nu} \neq 0\},$$

$d_\beta = |\mathcal{S}_1|$, and $d_\Theta = |\mathcal{S}_2|$. Let $\mathcal{N} = \mathcal{S}_1 \cup \mathcal{S}_2$ denote the set containing the truly informative variables and the dependent pairs. For a given constant a, let $sign(a)$ be the sign function which takes value $+1$ if $a > 0$, value -1 if $a < 0$, and 0 otherwise. For a vector (or a matrix) A, $sign(A)$ is defined to be the vector (or the matrix) whose element corresponding to the element a of A is $sign(a)$. In Appendix B.4, we prove the following results.

Theorem 3.3.2 *Under regularity conditions in Appendix B.1, the following properties hold:*

(a) (Sparsity recovery): $P\left(\widehat{\mathcal{N}} = \mathcal{N}\right) \to 1$ *as* $n \to \infty$.

(b) (Sign recovery): $sign(\widehat{\beta}) = sign(\beta_0)$ *and* $sign(\widehat{\Theta}) = sign(\Theta_0)$ *with a large probability.*

Theorem 3.3.2 (a) basically says that those informative variables and dependent pairs of the covariate components can be selected consistently. Theorem 3.3.2 (b) shows that the sign of the estimators is always identical to the sign of the true parameters. We call it 'sign edge recovery' if both (a) and (b) hold (Ravikumar et al. 2010).

Next, we establish the asymptotic distribution for the corresponding estimators. Let $\beta_{0;\mathcal{S}_1}$ and $\Theta_{0;\mathcal{S}_2}$ denote the subvector of β_0 and submatrix of Θ_0, respectively. For (3.16), we write $\varphi_{\Gamma_j,j} = \frac{\partial \varphi_j(\mathcal{Z};\Gamma_j)}{\partial \Gamma_j}$ for $j = 1, 2$. Furthermore, we write $\varphi'_\Gamma = \begin{pmatrix} \varphi_{\Gamma_1,1} & 0 \\ 0 & \varphi_{\Gamma_2,2} \end{pmatrix}$ and

$\varphi'(-1;\Gamma) = \begin{pmatrix} \frac{\partial\varphi_1(-1;\Gamma_1)}{\partial\Gamma_1} & 0 \\ 0 & \frac{\partial\varphi_2(-1;\Gamma_2)}{\partial\Gamma_2} \end{pmatrix}$. Define

$$\begin{aligned}\Phi_0(c,d) &= \left(c^\top, d^\top\right)\begin{pmatrix}\mathcal{U}\\ \mathcal{V}\end{pmatrix} \\ &\quad + \frac{1}{2}\left(c^\top, d^\top\right)\begin{pmatrix}\mathcal{I}_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) & 0\\ 0 & \mathcal{I}_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}\begin{pmatrix}c\\ d\end{pmatrix},\end{aligned}$$

where $\mathcal{U}$ and $\mathcal{V}$ are random variables having normal distributions $N\left(0,\Sigma_\beta\right)$ and $N\left(0,\Sigma_\Theta\right)$, respectively. Let

$$\mathbf{MVN} = \underset{c,d}{\operatorname{argmin}}\, \varphi'\left(-1;\Gamma\right)\left(\varphi_\Gamma'^\top\varphi_\Gamma'\right)^{-1}\varphi_\Gamma'^\top\Phi_0(c,d). \tag{3.26}$$

Theorem 3.3.3 *(Asymptotic Normality) Suppose that $\lambda_{n1}n^{-1/2}\to 0$ and $\lambda_{n2}n^{-1/2}\to 0$ as $n\to\infty$. Then under regularity conditions in Appendix B.1, we have that as $n\to\infty$,*

$$\sqrt{n}\left(\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}, vec\left(\widehat{\Theta}_{\mathcal{S}_2}\right) - vec\left(\Theta_{0;\mathcal{S}_2}\right)\right) \xrightarrow{d} \mathbf{MVN}.$$

Finally, we discuss the asymptotic property for the estimator of the cumulative baseline hazard function $\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t)$. Define

$$\mathcal{G}_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) = E\left\{Y_i(u)\exp\left(\sum_{r\in V\cap\mathcal{S}_1} W_{b,r}^{(i)}(\zeta)\beta_{0r} + \sum_{(s,\nu)\in E\cap\mathcal{S}_2} W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\theta_{0s\nu}\right)\right\},$$

$$\begin{aligned}&\mathcal{G}_{\beta;b,\zeta}^{(1)}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\\ &= E\left[Y_i(u)\left\{W_{b;\mathcal{S}_1}^{(i)}(\zeta)\right\}\exp\left\{\sum_{r\in V\cap\mathcal{S}_1} W_{b,r}^{(i)}(\zeta)\beta_{0r} + \sum_{(s,\nu)\in E\cap\mathcal{S}_2} W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\theta_{0s\nu}\right\}\right],\end{aligned}$$

and

$$\begin{aligned}\mathcal{G}_{\Theta;b,\zeta}^{(1)}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) &= E\left[Y_i(u)\left\{\left(W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\right)_{s\neq\nu}\right\}\right.\\ &\quad\left.\times\exp\left\{\sum_{r\in V\cap\mathcal{S}_1} W_{b,r}^{(i)}(\zeta)\beta_{0r} + \sum_{(s,\nu)\in E\cap\mathcal{S}_2} W_{b,s}^{(i)}(\zeta)W_{b,\nu}^{(i)}(\zeta)\theta_{0s\nu}\right\}\right].\end{aligned}$$

Let $\varphi'_\Lambda(\zeta;\Gamma_\Lambda) = \frac{\partial \varphi_\Lambda(\zeta;\Gamma_\Lambda)}{\partial \Gamma_\Lambda}$ and define $\varphi'_{\Gamma,\Lambda} = (\varphi'_\Lambda(\zeta;\Gamma_\Lambda) : \zeta \in \mathcal{Z})$. Let

$$\mathbf{W}_i(t;b,\zeta) = \int_0^t \left[\frac{\sum_{i=1}^n \left\{ dN_i(u) - g\left(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) Y_i(t) d\Lambda_0(u) \right\}}{\mathcal{G}_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)} \right],$$

then $\mathbf{W}_i(t;\zeta) = \frac{1}{B}\sum_{b=1}^B \mathbf{W}_i(t;b,\zeta)$ and $\mathbf{W}_i(t;\mathcal{Z}) = (\mathbf{W}_i(t;\zeta) : \zeta \in \mathcal{Z})$.

Define

$$\mathbf{W}_i(t) = \varphi'_\Lambda(-1,\Gamma_\Lambda)\left(\varphi'^\top_{\Gamma,\Lambda}\varphi'_{\Gamma,\Lambda}\right)^{-1}\varphi'^\top_{\Gamma,\Lambda}\mathbf{W}_i(t;\mathcal{Z}).$$

Let $\mathcal{W}(t)$ be the Gaussian process with mean zero and covariance $E\{\mathbf{W}_i(t)\mathbf{W}_i(s)\}$.

Define

$$\begin{aligned} F_{\beta;b,\zeta}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) &= \int_0^t \frac{E\left\{dN_i(u)\right\}\mathcal{G}^{(1)}_{\beta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)}{\left\{\mathcal{G}_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\right\}^2} \quad \text{and} \\ F_{\Theta;b,\zeta}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) &= \int_0^t \frac{E\left\{dN_i(u)\right\}\mathcal{G}^{(1)}_{\Theta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)}{\left\{\mathcal{G}_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\right\}^2}. \end{aligned}$$

Theorem 3.3.4 *Under regularity conditions in Appendix B.1, we have that as $n \to \infty$,*

$$\begin{aligned} &\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t) - \Lambda_0(t)\right\} \\ \xrightarrow{d}\ &\mathcal{W}(t) + \varphi'_{\Gamma,\Lambda}(-1;\Gamma_\Lambda(t))\left(\varphi'^\top_{\Gamma,\Lambda}\varphi'_{\Gamma,\Lambda}\right)^{-1}\varphi'^\top_{\Gamma,\Lambda}\begin{pmatrix} F_{\beta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}^\top \times \mathbf{MVN}. \end{aligned}$$

3.4 Numerical Studies

In this section, we conduct numerical studies to assess the performance of the proposed estimators for a variety of settings, and also implement the methods to analyze a real dataset.

3.4.1 Model Settings

We use model (3.1) to generate the p-dimensional true covariate X where the p-dimensional parameter β_0 is given by $\beta_0 = \left(\underbrace{1, \cdots, 1}_{[\frac{p}{4}]}, \underbrace{-1, \cdots, -1}_{[\frac{p}{4}]}, \underbrace{0, \cdots, 0}_{1-2[\frac{p}{4}]} \right)$. The (p^2-p)-dimensional parameter Θ_0 is specified to have the network structure, the lattice or hub structure, as shown in Figure 3.1.

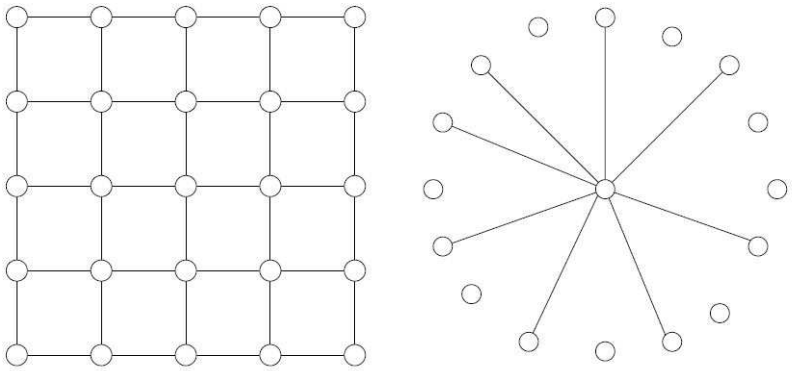


Figure 3.1: The left-hand-side structure is a *Lattice* with $p = 25$ and the right-hand-side structure is a *Hub* with $p = 17$.

Once X is generated, we use (3.5) with the baseline hazard function $\lambda_0(t) = 2t$ to generate the failure time T by letting

$$T = \sqrt{-\exp\left\{\sum_{r\in V}\beta_{0r}X_r + \sum_{(s,\nu)\in E}\theta_{0s\nu}X_sX_\nu + \sum_{r\in V}\mathbb{C}(X_r) - A\left(\beta_0, \Theta_0\right)\right\}\log\left(1-U\right)},$$

where U is simulated from the uniform distribution $U(0,1)$. Let C be the censoring time generated from the uniform distribution $U(0,c)$, where c is a constant that is chosen to yield about 50% censoring rate.

For surrogate measurements, we consider the following three scenarios.

Scenario I: *Only continuous covariates are subject to measurement error*

In this scenario, all error-prone covariates are continuous with $X = X_{\mathrm{C}}$. We consider

the measurement error model (3.7) where $\epsilon \sim N(0, \Sigma_\epsilon)$, Σ_ϵ is a $p \times p$ diagonal matrix with entries σ_ϵ^2. Here we let $\sigma_\epsilon^2 = 0.15^2, 0.5^2$ and 0.75^2 to reflect increasing degrees of measurement error in X_{C}.

Scenario II: *Only binary covariates are subject to misclassification*
In this scenario, all the error-contaminated covariates are considered to be binary take value 1 or -1. That is, using the notation in Section 3.1.3, we have $X = X_{\mathrm{D}}$ and $p = p_{\mathrm{D}}$. We consider that $p_{ll} = P\left(X_{\mathrm{D}}^* = x_{(l)} | X_{\mathrm{D}} = x_{(l)}\right)$ assumes a common value, say π, $l = 1, \cdots, m$ where $m = 2^p$ representing the cardinality of the set $\{-1, 1\}^p$, and we set $\pi = 0.2, 0.5$, or 0.8 to reflect different degrees of misclassification.

Scenario III: *Both measurement error and misclassification in covariates*
In this scenario, we examine the case where both continuous and discrete covariates are subject to mismeasurement by combining Scenarios I and II. Consistent with the notation in Section 3.1.3, $X = \left(X_{\mathrm{C}}^\top, X_{\mathrm{D}}^\top\right)^\top$ is the vector of the true covariates and $X^* = \left(X_{\mathrm{C}}^{*\top}, X_{\mathrm{D}}^{*\top}\right)^\top$ is the vector of surrogate covariates, we have X_{C}^* and X_{D}^* are independently generated by Scenarios I and II, respectively, and the dimension of X_{C} and X_{C} is p_{C} and p_{D}, respectively.

In implementing the proposed method, we set $B = 500$ and partition the interval $[0, 2]$ into subintervals with the equal width 0.25 with the resulting cutpoints set as the values of ζ. We take the regression functions $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ in (3.16) to be the quadratic polynomial functions, as suggested in Carroll et al. (2006, p.126). In each setting, we set the sample size $n = 400$ and examine different dimensions of X. In Scenario I, we set $p = p_{\mathrm{C}} = 10$ or 50; in Scenario II we examine $p = p_{\mathrm{D}} = 10$ or 15; and in Scenario III we set $p = p_{\mathrm{C}} + p_{\mathrm{D}} = 10$ or 50 with $(p_{\mathrm{C}}, p_{\mathrm{D}}) = (5, 5)$ and $(40, 10)$. We perform 500 simulations for each setting.

3.4.2 Simulation Results

To assess the performance of the estimator of β, we report several measures, the L_1-norm

$$\|\Delta_\beta\|_1 = \sum_i \left|\widehat{\beta}_i - \beta_{0,i}\right|$$

the L_2-norm

$$\|\Delta_\beta\|_2 = \sqrt{\sum_i \left(\widehat{\beta}_i - \beta_{0,i}\right)^2},$$

where $\Delta_\beta = \widehat{\beta} - \beta_0$. In addition, we calculate the number of the correctly selected variables (#CS) and the number of the falsely excluded variables (#FE).

To examine the accuracy of the estimator of Θ, we employ the L_1-norm and the Frobenius norm, respectively, given by

$$\|\Delta_\Theta\|_1 = \max_j \sum_i \left|\widehat{\Theta}_{ij} - \Theta_{0,ij}\right| \quad \text{and} \quad \|\Delta_\Theta\|_F = \sqrt{\sum_i \sum_j \left|\widehat{\Theta}_{ij} - \Theta_{0,ij}\right|^2},$$

where $\Delta_\Theta = \widehat{\Theta} - \Theta_0$.

We also examine the *specificity* (Spe) and the *sensitivity* (Sen), where the specificity is defined as the proportion of zero coefficients that were correctly estimated to be zero, and the sensitivity is defined as the proportion of non-zero coefficients that were correctly estimated to be non-zero.

For Scenarios I, II, and III, we compare the performance of the estimators obtained from applying the proposed method to the surrogate covariates as opposed to the estimators obtained from fitting the data with the true covariate measurements. We use the adaptive lasso with the penalty functions (3.12) and (3.13) as well as the lasso method. In comparison, we also examine the *naive estimators* of β and Θ which are derived by directly implementing the observed covariates X_i^* in (3.6).

In Tables 3.1- 3.3, we report the numerical results of our proposed method and the naive approach as well as those obtained from the true covariate measurements. It is clear and expected that the results obtained from using the true covariate measurements are the best with the smallest norms under all settings. Regarding the performance on the true measurements of the lasso and the adaptive lasso, the adaptive lasso tends to slightly outperform the lasso in terms of the specificity and the finite sample biases, indicated by the norms. In terms of correctly selecting variables, the lasso method performs better than the adaptive lasso. Both methods perform equally well in terms of falsely excluding variables and sensitivity, producing nearly perfect results.

The same patterns are observed for data with different degrees of measurement error and/or different network structures in covariates. The simulation results also demonstrate the impact of measurement error on inferential procedures. The performance of the proposed method would deteriorate as measurement error becomes more substantial. Furthermore, it is revealed that the naive method performs unsatisfactorily, with considerable finite sample biases produced and unreliable variable selection and exclusion results.

3.4.3 Analysis of NKI Breast Cancer Data

In this section, we implement our proposed method to analyze the breast cancer data collected by the Netherlands Cancer Institute (NKI) (van de Vijver et al. 2002). Tumors from 295 women with breast cancer were collected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. Tumors of those patients were primarily invasive breast cancer carcinoma that were about 5 cm in diameter. Patients at diagnosis were 52 years or younger and the diagnosis was done from 1984 to 1995. Of all those patients, 79 patients died before the study ended, yielding approximately the 73.2% censoring rate.

With those patients, about 25000 gene expressions were also collected. Presented by van de Vijver et al. (2002, p. 2002), among all the gene expressions, 70 genes with previously determined average profiles are useful for tumor diagnosis. Therefore, in our analysis here, we focus on those 70 genes with good prognosis and study their relationship with survival times.

Our goal is to select and estimate those gene expressions which are associated with the tumor development, where incorporating the network structure of those gene expressions is of particular interest. Consistent with He and Yi (2009), we treat log intensity as the covariates and implement the joint model (3.5) to analyze data. Since this dataset contains no information to characterize the degree of measurement error that is accompanying with the gene expressions, here we conduct sensitivity analyses to investigate the measurement error effects on analysis results. Specifically, let Σ be the covariance matrix of the gene expressions. For sensitivity analyses, we consider $\Sigma+\Sigma_e$ to be the covariance matrix for the measurement error model (3.7), where Σ_e is the diagonal matrix with diagonal elements being a common value σ_e^2, which is specified as $\sigma_e^2 = 0.15^2, 0.50^2$, or 0.75^2 to feature a setting with minor, moderate or severe measurement error. In addition, for the penalty function, we examine both lasso and adaptive lasso. The analysis results are summarized in Table 3.4. It is observed that the lasso method select more variables than the adaptive lasso method for each setting with a given degree of measurement error. The variables selected by the adaptive lasso method are a subset of those selected by the lasso method for each scenario. The selection results obtained from the lasso method tend to vary more noticeably than those produced from the adaptive lasso. The variables selected by the adaptive lasso method seem to be fairly insensitive to the change of the measurement error degrees we consider. Furthermore, the results produced from the naive method with different penalty functions differ from those obtained from the proposed method with measurement error effects accounted for. Regarding the estimation results for Θ, we display the gene network results in Figures 3.2 and 3.3, where we observe that the lasso method gives more complex association network than the adaptive lasso method.

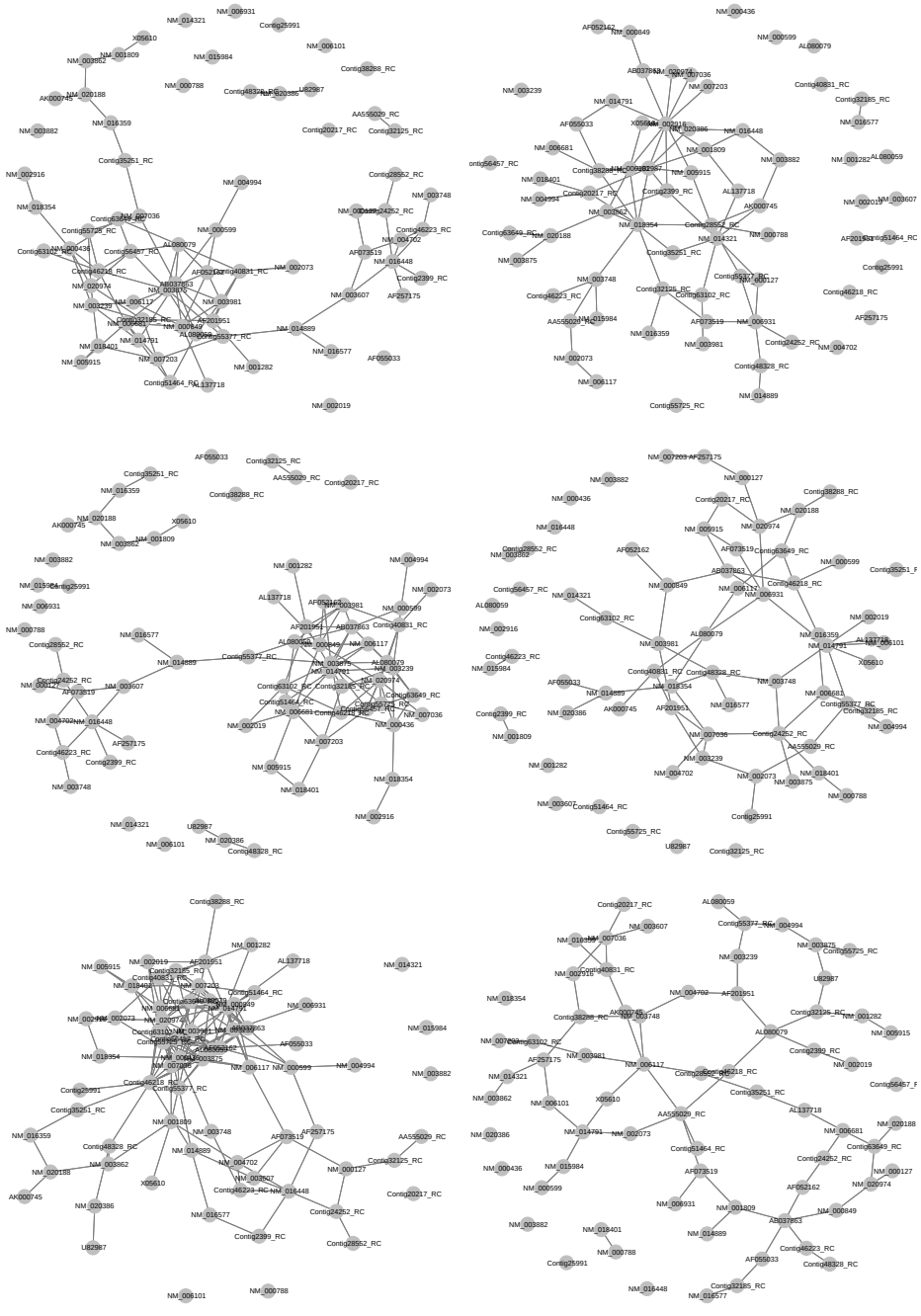


Figure 3.2: Graphical structures of 70 good prognosis genes in NKI Breast Cancer Data obtained from different measurement error degrees imposed: from top to bottom corresponds to $\sigma_e^2 = 0.15^2, 0.5^2$ or 0.75^2. The left and right columns are, respectively, obtained from the lasso and adaptive lasso methods.

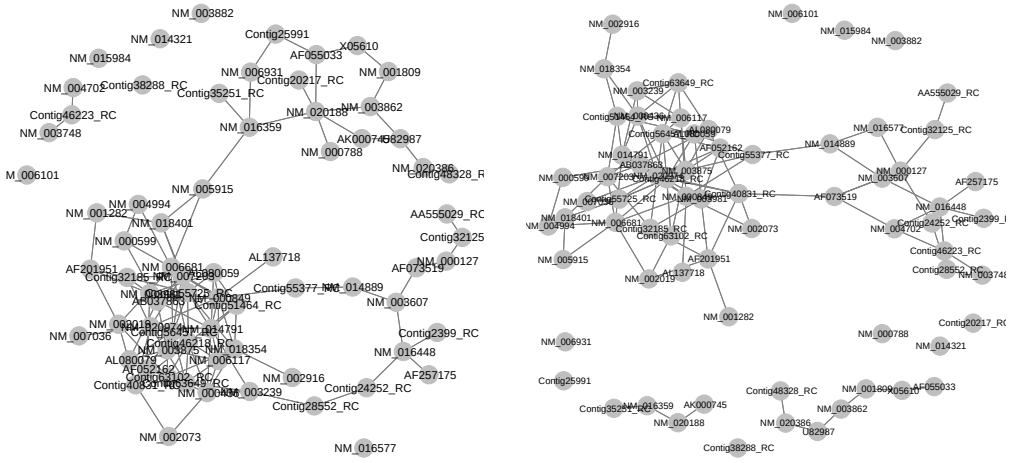


Figure 3.3: Graphical structures of 70 good prognosis genes in NKI Breast Cancer Data obtained from the naive method which ignores mismeasurement in covariates. The left and right figures are, respectively, obtained from the lasso and adaptive lasso methods.

Table 3.1: Simulation results for the proposed estimators of (β, Θ) based on Scenario I

Network	(n, p, d_β)	σ_ϵ	Method	Estimator of β_0				Estimator of Θ_0			
				$\|\Delta_\beta\|_1$	$\|\Delta_\beta\|_2$	#CS	#FE	$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 10, 6)$	0.15	LASSO-Naive	4.595	3.478	8.940	0.240	2.230	10.538	0.838	0.923
			ALASSO-Naive	4.229	2.860	8.200	0.100	2.222	6.795	0.865	1.000
			LASSO	0.808	0.094	8.980	0.040	1.958	2.558	0.946	1.000
			ALASSO	0.751	0.080	7.400	0.000	1.788	2.035	0.972	0.969
		0.50	LASSO-Naive	5.388	4.758	8.460	0.280	2.691	18.638	0.811	0.923
			ALASSO-Naive	5.336	4.545	8.600	0.200	2.571	16.465	0.846	1.000
			LASSO	1.118	0.179	8.140	0.080	2.584	5.619	0.946	1.000
			ALASSO	1.042	0.152	6.980	0.020	2.031	4.416	0.973	0.923
		0.75	LASSO-Naive	5.568	5.104	7.280	0.440	3.798	21.066	0.811	0.906
			ALASSO-Naive	5.489	4.894	7.893	0.300	2.738	19.105	0.846	1.000
			LASSO	0.670	0.070	7.500	0.020	2.994	7.313	0.946	1.000
			ALASSO	0.453	0.038	7.500	0.000	2.548	5.311	0.973	0.923
		$\times$	LASSO	0.642	0.057	8.480	0.000	0.454	1.909	0.945	1.000
			ALASSO	0.379	0.021	7.500	0.000	0.314	1.701	0.973	0.982
	$(400, 50, 26)$	0.15	LASSO-Naive	25.993	25.903	30.460	0.120	3.686	36.871	1.000	0.518
			ALASSO-Naive	22.391	17.529	29.400	0.060	3.473	31.076	1.000	0.763
			LASSO	3.062	0.497	37.260	0.020	1.555	9.218	0.999	0.988
			ALASSO	2.166	0.184	27.700	0.000	1.123	4.797	0.988	0.979
		0.50	LASSO-Naive	26.006	25.947	30.220	0.124	4.445	57.927	1.000	0.517
			ALASSO-Naive	24.392	19.609	31.500	0.008	3.724	48.327	0.998	0.762
			LASSO	5.408	0.845	39.680	0.000	1.306	8.021	0.996	0.941
			ALASSO	4.557	0.709	26.740	0.000	1.161	4.739	0.979	0.947
		0.75	LASSO-Naive	26.006	25.961	30.980	0.115	4.522	66.971	1.000	0.517
			ALASSO-Naive	24.777	20.731	30.200	0.007	3.817	53.975	0.941	0.770
			LASSO	6.030	1.138	40.440	0.000	1.603	7.824	0.985	1.000
			ALASSO	4.680	0.670	26.080	0.000	1.009	5.686	0.973	0.957
		$\times$	LASSO	1.182	0.073	37.320	0.000	1.112	4.302	1.000	0.988
			ALASSO	0.608	0.016	26.600	0.000	0.806	2.198	1.000	0.953

Hub	$(400, 10, 6)$	0.15	LASSO-Naive	4.325	2.991	8.880	0.240	3.098	5.374	0.776	1.000
			ALASSO-Naive	3.789	2.226	8.900	0.100	3.094	4.168	0.810	1.000
			LASSO	0.685	0.079	8.900	0.020	1.299	2.353	0.905	1.000
			ALASSO	0.605	0.061	7.980	0.000	0.952	1.564	0.976	1.000
		0.50	LASSO-Naive	5.132	4.283	8.720	0.240	3.491	9.656	0.846	1.000
			ALASSO-Naive	4.993	4.089	8.600	0.200	2.936	8.998	0.857	1.000
			LASSO	0.736	0.094	8.520	0.000	2.496	6.087	0.905	1.000
			ALASSO	0.643	0.072	7.480	0.000	2.349	5.740	0.976	1.000
		0.75	LASSO-Naive	5.441	4.822	8.825	0.420	3.948	11.937	0.905	1.000
			ALASSO-Naive	5.416	4.322	8.800	0.320	3.361	11.272	0.905	1.000
			LASSO	1.161	0.217	8.220	0.000	3.049	8.960	0.956	1.000
			ALASSO	0.897	0.124	6.680	0.000	2.717	8.076	0.976	1.000
		$\times$	LASSO	0.787	0.089	8.500	0.000	0.982	1.402	0.977	1.000
			ALASSO	0.306	0.013	7.480	0.000	0.600	1.118	1.000	1.000
	$(400, 50, 26)$	0.15	LASSO-Naive	24.003	23.953	32.280	0.110	18.663	90.601	1.000	0.638
			ALASSO-Naive	20.701	14.575	32.340	0.070	15.092	42.119	0.834	1.000
			LASSO	4.714	0.827	35.740	0.010	11.746	29.543	0.987	0.978
			ALASSO	2.824	0.380	27.700	0.000	11.470	27.318	0.989	0.979
		0.50	LASSO-Naive	25.991	25.934	33.210	0.098	18.764	91.017	1.000	0.638
			ALASSO-Naive	20.667	14.438	32.600	0.080	12.523	44.159	0.805	1.000
			LASSO	5.919	1.210	39.360	0.000	11.975	30.613	0.986	1.000
			ALASSO	4.269	0.622	26.280	0.000	11.088	29.881	0.988	0.957
		0.75	LASSO-Naive	25.982	25.936	33.102	0.096	19.854	92.361	1.000	0.638
			ALASSO-Naive	23.080	17.534	31.160	0.080	14.474	54.346	0.842	1.000
			LASSO	5.832	1.211	40.640	0.020	12.863	31.404	0.987	1.000
			ALASSO	4.564	0.828	26.620	0.000	11.868	29.461	0.988	0.979
		$\times$	LASSO	3.379	0.730	34.600	0.000	9.384	24.398	0.987	1.000
			ALASSO	1.549	0.102	26.320	0.000	9.170	13.210	0.989	0.979

Table 3.2: Simulation results for the proposed estimators of (β, Θ) based on Scenario II

Network	(n, p, d_β)	π	Method	Estimator of β_0				Estimator of Θ_0			
				$\|\Delta_\beta\|_1$	$\|\Delta_\beta\|_2$	#CS	#FE	$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 10, 6)$	0.20	LASSO-Naive	3.726	3.330	8.240	0.280	3.649	10.847	0.757	0.846
			ALASSO-Naive	3.743	3.321	8.220	0.220	3.624	10.397	0.676	0.846
			LASSO	0.844	0.180	8.220	0.000	2.942	8.678	0.973	0.946
			ALASSO	0.741	0.114	7.780	0.000	2.884	8.099	0.973	0.966
		0.50	LASSO-Naive	3.504	2.763	8.480	0.210	2.489	9.178	0.676	1.000
			ALASSO-Naive	3.477	2.760	8.940	0.340	2.464	6.432	0.757	1.000
			LASSO	0.811	0.108	8.460	0.000	2.385	5.423	0.946	0.923
			ALASSO	0.709	0.092	7.020	0.000	1.652	2.614	0.973	1.000
		0.80	LASSO-Naive	2.805	1.898	7.940	0.080	2.124	5.878	0.621	1.000
			ALASSO-Naive	2.787	1.710	7.420	0.100	1.842	5.738	0.838	0.923
			LASSO	0.830	0.107	8.680	0.000	1.609	3.783	0.946	1.000
			ALASSO	0.633	0.068	6.010	0.000	1.101	2.018	0.973	1.000
		$\times$	LASSO	0.488	0.044	9.480	0.000	1.021	5.182	0.966	1.000
			ALASSO	0.308	0.047	7.050	0.000	0.739	4.002	0.967	1.000
	$(400, 15, 8)$	0.20	LASSO-Naive	6.786	5.699	13.350	0.040	2.852	20.328	0.801	0.954
			ALASSO-Naive	6.744	5.655	12.340	0.040	2.705	17.496	0.790	0.955
			LASSO	0.936	0.140	13.620	0.000	2.697	12.352	0.978	0.954
			ALASSO	0.919	0.133	10.480	0.000	1.948	10.584	1.000	0.955
		0.50	LASSO-Naive	6.145	4.559	12.480	0.260	2.609	13.488	0.790	0.954
			ALASSO-Naive	6.104	4.492	12.160	0.240	2.642	10.939	0.779	0.955
			LASSO	0.866	0.124	12.56	0.000	2.388	10.635	0.977	0.954
			ALASSO	0.853	0.122	9.220	0.000	1.798	7.317	0.989	1.000
		0.80	LASSO-Naive	4.343	2.204	13.040	0.140	2.825	9.325	0.757	1.000
			ALASSO-Naive	4.244	2.122	12.720	0.160	2.048	6.290	0.856	1.000
			LASSO	0.894	0.085	11.840	0.000	1.821	4.824	0.978	1.000
			ALASSO	0.861	0.084	8.620	0.000	1.629	4.596	0.978	1.000
		$\times$	LASSO	0.868	0.075	12.560	0.000	1.541	4.144	0.965	1.000
			ALASSO	0.485	0.027	9.250	0.000	1.366	3.983	0.988	1.000

Hub	$(400, 10, 6)$	0.20	LASSO-Naive	3.574	3.197	7.680	0.060	4.187	11.730	0.762	0.625
			ALASSO-Naive	3.555	3.175	7.380	0.040	4.118	11.904	0.738	0.750
			LASSO	1.472	0.558	8.140	0.000	3.370	9.349	0.952	0.950
			ALASSO	1.369	0.439	7.160	0.000	1.944	4.941	0.986	1.000
		0.50	LASSO-Naive	3.408	2.609	7.580	0.260	3.983	11.349	0.762	0.750
			ALASSO-Naive	3.312	2.563	7.220	0.400	3.898	10.095	0.762	0.625
			LASSO	1.007	0.211	8.700	0.000	2.642	7.856	0.933	1.000
			ALASSO	0.906	0.128	7.180	0.000	2.407	5.759	0.929	1.000
		0.80	LASSO-Naive	3.308	2.609	7.580	0.260	3.813	11.349	0.762	0.750
			ALASSO-Naive	2.702	1.614	7.600	0.160	3.226	8.419	0.738	1.000
			LASSO	0.675	0.065	8.760	0.000	2.449	5.418	0.928	0.975
			ALASSO	0.641	0.058	6.530	0.000	1.564	3.180	0.976	0.975
		$\times$	LASSO	0.494	0.049	9.480	0.000	1.527	5.360	0.955	1.000
			ALASSO	0.438	0.043	7.500	0.000	0.536	1.098	0.976	1.000
	$(400, 15, 8)$	0.20	LASSO-Naive	7.622	6.429	13.700	0.040	6.222	19.981	0.609	0.923
			ALASSO-Naive	7.245	6.144	12.240	0.000	6.193	19.845	0.698	0.923
			LASSO	2.827	1.525	13.30	0.000	3.612	10.899	0.920	1.000
			ALASSO	2.054	0.980	10.200	0.000	3.422	6.389	0.970	1.000
		0.50	LASSO-Naive	6.614	4.855	13.480	0.260	6.609	18.939	0.790	0.955
			ALASSO-Naive	6.421	4.640	12.720	0.200	5.654	18.285	0.719	0.769
			LASSO	2.791	0.815	12.500	0.000	3.004	6.409	0.970	1.000
			ALASSO	2.424	0.641	9.520	0.000	2.543	5.130	0.970	1.000
		0.80	LASSO-Naive	5.343	3.204	13.040	0.140	5.821	14.824	0.757	1.000
			ALASSO-Naive	4.547	2.372	12.620	0.160	5.069	12.069	0.729	1.000
			LASSO	2.539	0.785	11.560	0.000	2.445	5.322	0.980	1.000
			ALASSO	2.414	0.617	9.380	0.000	1.766	3.207	0.990	1.000
		$\times$	LASSO	1.277	0.108	12.560	0.000	1.721	5.157	0.953	1.000
			ALASSO	0.698	0.082	9.380	0.000	0.923	2.889	0.989	1.000

Table 3.3: Simulation results for the proposed estimators of (β, Θ) based on Scenario III

Network	(n, p, d_β)	(σ_ϵ, π)	Method	Estimator of β_0				Estimator of Θ_0			
				$\|\Delta_\beta\|_1$	$\|\Delta_\beta\|_2$	#CS	#FE	$\|\Delta_\Theta\|_1$	$\|\Delta_\Theta\|_F$	Spe	Sen
Lattice	$(400, 10, 6)$	(0.75,0.20)	LASSO-Naive	4.446	3.760	8.470	0.230	2.962	12.535	0.743	0.962
			ALASSO-Naive	4.165	3.440	8.530	0.231	2.590	10.693	0.796	0.884
			LASSO	0.827	0.126	8.520	0.000	2.490	5.620	0.942	0.923
			ALASSO	0.731	0.093	7.590	0.000	2.231	4.067	0.973	0.938
		(0.50,0.50)	LASSO-Naive	4.187	3.501	7.610	0.260	2.820	13.472	0.716	0.953
			ALASSO-Naive	4.138	3.304	7.657	0.210	2.423	12.412	0.842	0.961
			LASSO	0.965	0.145	8.380	0.000	2.485	5.521	0.946	0.915
			ALASSO	0.775	0.112	7.100	0.000	1.842	3.516	0.973	1.000
		(0.15,0.80)	LASSO-Naive	4.407	3.652	8.750	0.270	2.518	11.822	0.805	1.000
			ALASSO-Naive	3.986	3.091	8.200	0.150	2.174	7.596	0.771	0.923
			LASSO	0.781	0.085	8.100	0.000	2.550	5.584	0.946	1.000
			ALASSO	0.563	0.054	6.751	0.000	1.824	2.664	0.973	1.000
		$\times$	LASSO	0.584	0.051	7.480	0.000	1.087	3.545	0.946	1.000
			ALASSO	0.348	0.034	6.275	0.000	0.526	1.854	0.967	1.000
	$(400, 50, 26)$	(0.75,0.20)	LASSO-Naive	28.397	26.830	30.165	0.078	4.687	63.649	0.901	0.736
			ALASSO-Naive	25.761	21.193	30.270	0.024	4.261	47.355	0.866	0.863
			LASSO	3.462	0.612	37.662	0.000	2.126	10.740	0.957	0.951
			ALASSO	2.705	0.377	29.478	0.000	1.537	7.581	1.000	0.975
		(0.50,0.50)	LASSO-Naive	26.076	25.269	30.350	0.122	4.527	54.433	0.895	0.736
			ALASSO-Naive	24.248	20.591	29.830	0.024	3.813	40.908	0.895	0.859
			LASSO	3.137	0.486	37.614	0.000	1.847	9.329	0.986	0.948
			ALASSO	2.615	0.316	28.748	0.000	1.459	6.018	0.989	1.000
		(0.15,0.80)	LASSO-Naive	25.168	24.054	30.750	0.120	3.754	35.871	0.876	0.759
			ALASSO-Naive	23.318	19.651	29.050	0.100	3.276	31.683	0.928	0.882
			LASSO	2.001	0.319	36.441	0.000	2.214	8.547	0.978	1.000
			ALASSO	1.541	0.153	28.594	0.000	1.319	4.141	0.978	1.000
		$\times$	LASSO	1.025	0.085	32.168	0.000	1.825	4.723	0.914	1.000
			ALASSO	0.547	0.021	26.030	0.000	1.088	3.091	0.985	1.000

Hub	$(400, 10, 6)$	(0.75,0.20)	LASSO-Naive	4.508	4.010	8.253	0.240	4.668	11.834	0.859	0.813
			ALASSO-Naive	4.486	3.749	8.090	0.180	3.740	11.588	0.822	0.875
			LASSO	1.317	0.387	8.184	0.000	3.710	10.654	0.955	0.983
			ALASSO	1.113	0.283	6.914	0.000	2.331	6.510	0.965	1.000
		(0.50,0.50)	LASSO-Naive	4.270	3.446	8.150	0.200	3.737	10.503	0.804	0.813
			ALASSO-Naive	4.153	3.326	7.910	0.163	3.417	9.547	0.810	0.875
			LASSO	0.895	0.153	8.673	0.000	3.374	7.316	0.938	1.000
			ALASSO	0.775	0.108	7.344	0.000	2.378	5.750	0.954	1.000
		(0.15,0.80)	LASSO-Naive	3.817	2.800	8.230	0.150	3.456	7.759	0.769	0.875
			ALASSO-Naive	3.245	1.920	8.250	0.130	3.160	6.890	0.774	1.000
			LASSO	0.680	0.075	8.863	0.000	2.596	7.385	0.928	0.937
			ALASSO	0.523	0.059	7.256	0.000	1.258	2.372	0.976	0.958
		×	LASSO	0.541	0.059	7.990	0.000	1.255	5.736	0.905	1.000
			ALASSO	0.372	0.028	7.490	0.000	0.468	1.250	0.976	1.000
	$(400, 50, 26)$	(0.75,0.20)	LASSO-Naive	25.802	24.183	32.401	0.068	19.031	96.171	0.805	0.781
			ALASSO-Naive	25.523	13.837	31.700	0.040	14.335	57.096	0.817	0.861
			LASSO	4.355	1.365	36.173	0.000	7.738	20.156	0.958	1.000
			ALASSO	3.309	0.904	28.944	0.000	6.145	17.926	0.970	1.000
		(0.50,0.50)	LASSO-Naive	25.603	24.395	33.345	0.047	18.687	94.978	0.895	0.797
			ALASSO-Naive	23.544	13.539	31.660	0.044	13.089	41.222	0.862	0.880
			LASSO	4.105	1.013	36.080	0.000	7.495	18.510	0.970	1.000
			ALASSO	3.147	0.632	27.639	0.000	6.816	16.055	0.970	1.000
		(0.15,0.80)	LASSO-Naive	24.673	24.079	32.660	0.013	18.242	92.713	0.879	0.819
			ALASSO-Naive	22.624	13.474	31.480	0.011	12.270	32.094	0.878	1.000
			LASSO	3.027	0.806	35.544	0.000	6.955	15.435	0.938	1.000
			ALASSO	2.619	0.499	27.619	0.000	5.423	12.263	0.984	1.000
		×	LASSO	2.328	0.319	33.489	0.000	4.055	14.778	0.953	1.000
			ALASSO	1.124	0.092	26.944	0.000	2.747	8.049	0.989	1.000

Table 3.4: Sensitivity analyses for NKI Breast Cancer Data: estimators of selected variables

Gene ID	$\sigma_e^2 = 0.15^2$		$\sigma_e^2 = 0.50^2$		$\sigma_e^2 = 0.75^2$		Naive	
	LASSO	ALASSO	LASSO	ALASSO	LASSO	ALASSO	LASSO	ALASSO
Contig036649_RC	-0.342	–	–	–	–	–	-0.241	-0.159
Contig46218_RC	-0.520	–	-0.486	–	-0.505	-0.641	0.329	–
NM_016359	1.548	1.782	2.463	2.029	3.066	2.128	-0.193	-0.251
AA555029_RC	-1.941	-1.970	-2.473	-2.186	-2.854	-2.178	1.048	0.883
NM_003748	-2.780	-3.031	-3.546	-3.323	-3.843	-3.282	-1.272	-0.885
Contig38288_RC	1.606	1.230	1.339	1.155	1.237	1.190	-1.639	-1.311
NM_003862	1.560	2.321	3.266	2.291	3.104	2.133	-0.239	0.507
Contig28552_RC	-4.245	-3.700	-4.981	-3.699	-4.728	-3.754	2.691	0.969
Contig32125_RC	-3.511	-3.673	-4.368	-3.834	-4.719	-3.513	-1.615	-1.583
AB037863	-0.607	–	-0.505	–	1.051	–	-1.167	-1.546
NM_020188	-0.756	-0.532	–	–	-0.650	–	1.158	–
Contig55377_RC	0.580	–	–	–	-0.329	–	-0.165	–
Contig25991	-0.433	–	0.696	–	–	–	-0.114	–
NM_003875	0.621	–	–	–	–	–	1.766	–
NM_006101	2.049	1.650	2.531	1.462	2.723	1.663	–	–
NM_003882	-0.662	–	–	–	–	–	-0.678	–
NM_003607	-3.045	-2.388	-3.377	-1.925	-3.167	-2.147	-0.203	–
NM_000849	–	–	–	–	0.393	–	-0.321	–
NM_016577	0.521	–	–	–	–	–	–	–
Contig48328_RC	–	–	0.543	–	–	–	0.277	–
Contig46223_RC	-1.484	-1.569	-2.197	-1.522	-1.661	-1.443		
NM_006117	-0.596	–	–	-0.538	-0.850	-0.638	-0.613	–
AK000745	-0.424	–	–	–	–	–	0.259	–
NM_003239	–	–	–	–	-1.605	–	-0.145	–

Gene ID	$\sigma_e^2 = 0.15^2$		$\sigma_e^2 = 0.50^2$		$\sigma_e^2 = 0.75^2$		Naive	
	LASSO	ALASSO	LASSO	ALASSO	LASSO	ALASSO	LASSO	ALASSO
NM_014791	-0.393	–	–	–	-1.150	-0.634	-0.141	–
X05610	3.465	3.566	4.265	3.653	5.298	3.744	-0.159	0.641
NM_018401	0.366	–	–	–	-0.346	–	-3.309	–
AL080079	–	–	-0.540	–	-0.439	–	–	-0.837
NM_006931	-2.470	-1.760	-2.146	-1.840	-1.751	-1.950	-0.977	–
AF257175	-4.332	-3.966	-4.494	-4.055	-5.149	-3.891	–	–
NM_614321	6.619	4.680	5.765	4.245	5.594	4.396	–	–
Contig55725_RC	-0.301	–	-0.345	–	-0.492	–	0.149	0.130
Contig24252_RC	-0.995	–	–	–	–	–	–	–
AF201905	0.773	–	0.532	–	–	–	–	–
NM_005915	–	–	-0.343	–	–	–	0.583	–
NM_001282	–	–	-0.431	–	–	–	-1.006	-0.634
NM_000599	–	–	–	–	-0.336	–	–	–
NM_020386	0.942	0.838	1.211	1.060	1.828	1.071	-0.257	-0.322
NM_014889	–	–	-0.525	–	–	–	3.116	–
AF055033	-0.702	–	–	–	–	–	-0.343	–
Contig20217_RC	–	–	0.668	–	–	–	–	–
NM_001809	-0.447	–	–	–	–	–	-0.401	-0.278
Contig2399_RC	-0.371	–	-0.351	–	–	–	1.102	1.566
NM_007036	0.355	–	–	–	0.601	–	-2.789	–
NM_018354	–	–	–	–	-0.359	–	–	–
#S	36	16	27	16	29	18	34	16

Chapter 4

Sufficient Dimension Reduction for Analysis of High-Dimensional Survival Data with Error-Prone Variables

4.1 Preliminaries

In this section, we introduce the preliminaries of sufficient dimension reduction (SDR), survival analysis, and measurement error models.

4.1.1 SDR and Conditional Distribution

Let $T \in \mathbb{R}$ be the univariate response, and let X be the p-dimensional vector of covariates, where p is often a large positive integer. The spirit of sufficient dimension reduction (SDR) is to find a $p \times d$ matrix $B = (\beta_1, \cdots, \beta_d)$ such that

$$T \perp\!\!\!\perp X | B^\top X, \tag{4.1}$$

where "$\perp\!\!\!\perp$" stands for the statistical independence, and β_j is a p-dimensional vector for $j = 1, \cdots, d$. Here d can be viewed as the dimension of the reduced covariates and is smaller than p, and B is often called a basis.

Let $\mathcal{S}(B)$ represent the SDR subspace which is spanned by the column vectors of B. Cook (1994) showed that the intersection of all such $\mathcal{S}(B)$ exists. Consequently, such an intersection is called the *central subspace* (CS) for the regression of T on X. Let $\mathcal{S}_{T|X}$ denote the CS with the structural dimension $d = \dim(\mathcal{S}_{T|X})$ which is usually unknown. If B is obtained, then the subsequent analysis can be based on the lower dimensional variables $\{T, B^\top X\}$ without losing information.

We now consider the cumulative distribution function of T given $X = x$:

$$F_{T|X}(t|x) = F(t, B^\top x), \tag{4.2}$$

where $F_{T|X}(t|x) \triangleq P(T \leq t|X = x)$ and $F(\cdot, \cdot)$ is an unknown nonnegative function. Let $\mathcal{F}_{T|X}(t|x) = 1 - F_{T|X}(t|x)$ denote the survivor function of T given X. Then the (conditional) hazards function of T, given $X = x$, is given by

$$\lambda(t|x) = \frac{\frac{d}{dx} F_{T|X}(t|x)}{\mathcal{F}_{T|X}(t|x)}, \tag{4.3}$$

which is uniquely determined by (4.2). This suggests that (4.2) can be broadly used to describe any survival models. The following examples give some commonly used survival models.

Example 1 If $F(t, B^\top x) = 1 - \exp\left\{-t^2 \exp\left(B^\top x\right)\right\}$, then $\lambda(t|x) = 2t \exp\left(B^\top x\right)$ is the Cox proportional hazard model with the baseline hazards function $\lambda_0(t) = 2t$ (Cox 1972).

Example 2 If $F(t, B^\top x) = \frac{\exp\left(t - B^\top x\right)}{1 + \exp\left(t - B^\top x\right)}$, then (4.2) gives the proportional odds model (Bennett 1983).

Example 3 If $F(t, B^\top x) = 1 - \exp\left\{-\left(t^2 + tB^\top x\right)\right\}$, then $\mathcal{F}_{T|X}(t|x) = \exp\left\{-\left(t^2 + tB^\top x\right)\right\}$ and, equivalently, $\lambda(t|x) = 2t + B^\top x$, which is the additive hazards model with the baseline hazards function $\lambda_0(t) = 2t$ (Lin and Ying 1994).

4.1.2 Survival Data with Measurement Error

Let the response T represent the survival time for a subject. We consider the setting where T is associated with a p-dimensional covariate X where p is large. We are interested in finding the CS, $\mathcal{S}_{T|X}$, to study the relationship between the survival time T and covariates

X. In survival analysis, T is usually incomplete due to the presence of the censoring time for a subject, denoted as C. Let $Y = \min\{T, C\}$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Directly implementing the SDR methods on the observed variable Y and X is equivalent to studying $\mathcal{S}_{Y|X}$, which is generally not equal to $\mathcal{S}_{T|X}$. With the non-informative censoring time, Xia et al. (2010) pointed out that the CS for the regression of Y on X is the direct sum of the CS for the regressions of T on X and that of C on X, which means that $\mathcal{S}_{Y|X} = \mathcal{S}_{T|X} + \mathcal{S}_{C|X} = \left\{\mathbf{v}_1 + \mathbf{v}_2 : \mathbf{v}_1 \in \mathcal{S}_{T|X}, \mathbf{v}_2 \in \mathcal{S}_{C|X}\right\}$ and $\mathcal{S}_{T|X} \cap \mathcal{S}_{C|X} = \phi$. This shows that $\mathcal{S}_{Y|X}$ is not equal to $\mathcal{S}_{T|X}$ in general, suggesting that using the existing dimension reduction methods to the response Y only yields the estimator of $\mathcal{S}_{Y|X}$ instead of $\mathcal{S}_{T|X}$, the quality of primary interest.

On top of the issue of censoring, another challenge is posed by that covariates X are commonly error-contaminated. Ignoring measurement error in inferential procedures can yield seriously biased results. To feature this, let X^* denote the surrogate, or observed covariate, of X. Let Σ_{X^*} and Σ_X be the covariance matrices of X^* and X, respectively. We consider the measurement error model (Carroll and Li 1992; Li and Yin 2007; Zhang et al. 2014)

$$X^* = \gamma + \Gamma X + \epsilon, \tag{4.4}$$

where ϵ is independent of $\{X, T, C\}$, $\epsilon \sim N(0, \Sigma_\epsilon)$, γ is an s-dimensional vector of parameters, and Γ is an $s \times p$ matrix of parameters which may be known or unknown. As discussed in Carroll and Li (1992), Li and Yin (2007) and Zhang et al. (2014), we may consider

$$U = LX^* \tag{4.5}$$

as the "corrected" covariates in terms of X^*, where

$$L = \text{cov}(X, X^*)\Sigma_{X^*}^{-1} = \Sigma_X \Gamma^\top \Sigma_{X^*}^{-1}. \tag{4.6}$$

This "corrected" covariate U can be used to study the CS $\mathcal{S}_{T|X}$, as shown by the following proposition given by Li and Yin (2007).

Proposition 4.1.1 *(Li and Yin 2007)*
Suppose that X follows a normal distribution with mean μ_X and covariance matrix Σ_X, and model (4.4) is assumed. Then $\mathcal{S}_{T|X} = \mathcal{S}_{T|U}$ with $U = LX^*$.

Proposition 4.1.1 is a general invariance law in the sense that replacing X by U still preserves the CS $\mathcal{S}_{T|X}$; it applies to the setting in Section 4.1.1 as well. Moreover, this proposition basically shows that for a matrix B,

$$T \perp\!\!\!\perp X | B^\top X \Longleftrightarrow T \perp\!\!\!\perp U | B^\top U, \tag{4.7}$$

where "$\Longleftrightarrow$" means the equivalence between the two statements.

4.1.3 Determination of "Corrected" Covariates

Suppose that we have a sample of n subjects and that for $i = 1, \cdots, n$, $\{Y_i, \Delta_i, X_i\}$ has the same distribution as $\{Y, \Delta, X\}$ and $\{y_i, \delta_i, x_i\}$ represents realizations of $\{Y_i, \Delta_i, X_i\}$. Let τ denote a finite value which is no smaller than the maximum survival times in the sample. Suppose that for $i = 1, \cdots, n$, (X_i^*, X_i) has the same distribution as (X^*, X) and let (x_i^*, x_i) denote the realizations of (X_i^*, X_i).

Note that model (4.4) yields $\Sigma_{X^*} = \Gamma\Sigma_X\Gamma^\top + \Sigma_\epsilon$, and that Σ_{X^*} can be estimated by its empirical estimator based on the available measurements of X^*, given by $\widehat{\Sigma}_{X^*} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^* - \bar{X}_i^*\right)\left(X_i^* - \bar{X}_i^*\right)^\top$ with $\bar{X}_i^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$. To estimate L, we need only to handle Σ_ϵ and Γ. Consequently, we consider the following three scenarios.

Scenario I : *Both Σ_ϵ and Γ are known.*

In this scenario, L is determined by (4.6), which allows us to directly calculate the "corrected" covariates U using (4.5).

Scenario II : *Γ is known, Σ_ϵ is unknown, and repeated measurements of X are available.*

Suppose that two repeated measurements of X, $\{X_{ir}^* : r = 1, 2; i \in \mathcal{R}\}$, are collected for additional m subjects, where $\mathcal{R}$ denotes the index set for those subjects. Consistent with Carroll and Li (1992), we take Γ to be $I_{p\times p}$ for ease of discussions. Then the measurement error model based on repeated measurements is given by

$$X_{ir}^* = \gamma + X_i + \epsilon_{ir} \tag{4.8}$$

for $i \in \mathcal{R}$ and $r = 1, 2$, where $\epsilon_{ir} \sim N(0, \Sigma_\epsilon)$ and ϵ_{ir} is independent of $\{X_i, T_i, C_i\}$.

In this case,

$$\begin{aligned}\Sigma_{X^*} &= \mathrm{var}(X_{ir}^*) \\ &= \mathrm{var}\left\{E(X_{ir}^*|X_i)\right\} + E\left\{\mathrm{var}(X_{ir}^*|X_i)\right\} \\ &= \Sigma_X + \Sigma_\epsilon. \end{aligned} \tag{4.9}$$

Noting that for every $i \in \mathcal{R}$, Σ_ϵ and Σ_X can be, respectively, expressed by

$$\Sigma_\epsilon = \frac{1}{2}\mathrm{var}\left(X_{i1}^* - X_{i2}^*\right) \tag{4.10}$$

and

$$\begin{aligned}\Sigma_X &= \frac{1}{4}\mathrm{var}\left(X_{i1}^* + X_{i2}^*\right) - \frac{1}{2}\Sigma_\epsilon \\ &= \frac{1}{4}\left\{\mathrm{var}\left(X_{i1}^* + X_{i2}^*\right) - \mathrm{var}\left(X_{i1}^* - X_{i2}^*\right)\right\}.\end{aligned}$$

Therefore, by (4.9), we have

$$\Sigma_{X^*} = \frac{1}{4}\left\{\mathrm{var}\left(X_{i1}^* + X_{i2}^*\right) - \mathrm{var}\left(X_{i1}^* - X_{i2}^*\right)\right\}, \tag{4.11}$$

and by (4.6), we have

$$L = \left(\Sigma_{X^*} - \Sigma_\epsilon\right)\Sigma_{X^*}^{-1}$$

which can be estimated by

$$\widehat{L} = \left(\widehat{\Sigma}_{X^*} - \widehat{\Sigma}_\epsilon\right)\widehat{\Sigma}_{X^*}^{-1} \tag{4.12}$$

with $\widehat{\Sigma}_\epsilon$ and $\widehat{\Sigma}_{X^*}$ being empirical estimators of (4.10) and (4.11), respectively.

Scenario III : *Both Σ_ϵ and Γ are unknown and validation data are available.*

Suppose that $\mathcal{M}$ is the set of n subjects for the main study and $\mathcal{V}$ is the set of m subjects for the external validation study. That is, $\mathcal{M}$ and $\mathcal{V}$ do not overlap, the available data for the main study and the validation sample are $\{(t_i, c_i, \delta_i, x_i^*) : i \in \mathcal{M}\}$ and $\{(x_i, x_i^*) : i \in \mathcal{V}\}$, respectively. Hence, the measurement error model (4.4) gives that

$$X_i^* = \gamma + \Gamma X_i + \epsilon_i$$

for $i \in \mathcal{M} \cup \mathcal{V}$, where $\epsilon_i \sim N(0, \Sigma_\epsilon)$ and ϵ_i is independent of $\{(X_i, T_i, C_i)\}$ for $i \in \mathcal{M} \cup \mathcal{V}$.

Let $\mu_X = E(X_i)$ and $\mu_{X^*} = E(X_i^*)$. Then using the validation data $\{(x_i, x_i^*) : i \in \mathcal{V}\}$, we estimate μ_X and μ_{X^*} by $\widehat{\mu}_X = \frac{1}{m}\sum_{i\in\mathcal{V}} x_i$ and $\widehat{\mu}_{X^*} = \frac{1}{m}\sum_{i\in\mathcal{V}} x_i^*$, respectively. Then a consistent estimate of $\mathrm{cov}(X_i, X_i^*)$ is given by

$$\widehat{\mathrm{cov}(X_i, X_i^*)} = \frac{1}{m}\sum_{i\in\mathcal{V}}\left(x_i - \widehat{\mu}_X\right)\left(x_i^* - \widehat{\mu}_{X^*}\right)^\top,$$

and hence, by (4.6), L can be estimated by

$$\widehat{L} = \widehat{\mathrm{cov}(X_i, X_i^*)}\widehat{\Sigma}_{X^*}^{-1}.$$

Once the estimator $\widehat{L}$ is obtained by either repeated measurements or validation data in Scenarios II or III, we adjust the surrogate covariate X_i^* by $\widehat{U}_i = \widehat{L}X_i^*$ and let $\widehat{u}_i$ denote a realization of $\widehat{U}_i$.

4.2 Methodology

In this section, we consider the model (4.2) and propose an estimation method to handle data with both right-censoring and mismeasurement. To be more specific, we first apply (4.5) to correct the measurement error effects and also adjust for the censoring effects due to censored responses; next, we propose valid inferential procedures to estimate B and d without imposing additional conditions, such as the linearity condition (e.g., Li 1991) which is commonly used in the conventional SDR methods.

Before we present the details of our proposed method, we make a few comments here.

- Although Li and Yin (2007) provided a method to correct for measurement error effects in dimension reduction, their approach mainly focused on establishing the validity of $U = LX^*$ (given by (4.5)) instead of focusing on the estimation of the parameter B. Furthermore, their approach was directed to handling complete responses but not censored responses caused by right-censoring.

- The idea of using $U = LX^*$ in (4.5) to correct for measurement error effects was also considered in the dimension reduction frameworks by other authors such as Carroll and Li (1992) and Zhang et al. (2014). However, their settings were still targeted to complete responses instead of censored responses caused by right-censoring. Moreover, their approaches employed "sliced inverse regression" or "cumulative slicing estimation", which typically requires the so-called *linearity condition*.

- While our development here has relevance to existing work, the problem we consider has an additional feature that the response is subject to censoring, which significantly complicates the development of estimation procedures as well as the establishment of theoretical results. Moreover, we develop a semiparametric inference approach which requires minimal model assumptions (e.g., the "linearity condition" imposed by Li (1991) is not needed in our procedures).

4.2.1 Method Setup and Correction of Measurement Error

For $t > 0$, let $N_T(t) = I(T \leq t)$ be the indicator function of T. Then (4.2) is equivalently expressed as

$$E\{N_T(t)|X\} = F\left(t, B^\top x\right). \tag{4.13}$$

If T is fully observed and there is no censoring, then the following proposition, stated by Wang and Xia (2008) and proved by Zeng and Zhu (2010), can be used to connect (4.1) and (4.2).

Proposition 4.2.1 *For any matrix B, "$T \perp\!\!\!\perp X|B^\top X$" is equivalent to*

$$\text{"}P(T \leq t|\, X = x) = P\left(T \leq t|\, B^\top X = B^\top x\right) \text{ for any } t \in \mathbb{R}^1 \text{ and } X \in \mathbb{R}^p\text{."}$$

Proposition 4.2.1 shows that the central space of T is closely related to the central mean space of $I(T \leq t)$. Combining Propositions 4.1.1 and 4.2.1 yields

$$T \perp\!\!\!\perp U|B^\top U \Longleftrightarrow E\{N_T(t)|U\} = E\left\{N_T(t)|B^\top U\right\} \tag{4.14}$$

for any $t > 0$. Equation (4.14) shows that the measurement error effects can be corrected for by using (4.5); then the usual dimension reduction techniques may be employed to derive estimators of B.

In survival analysis, however, T is usually incomplete due to censoring; we have only (Y, Δ) as described in Section 4.1.1. Conventional methods for sufficient dimension reduction are not valid any more in this case (e.g., Xia et al. 2010; Lu and Li 2011). To accommodate censored responses, one may proceed with the inverse weighted scheme. For any given $y > 0$, let $N_Y(y) = I(Y \leq y)$ be the random indicator variable. For given $y > 0$ and $U = u$,

$$\begin{aligned}
E\left\{\left.\frac{\Delta I(Y \leq y)}{P(C \geq Y|U = u)}\right| U = u\right\} &= E\left\{\left.\frac{I(T \leq C)\, I(Y \leq y)}{P(C \geq Y|U = u)}\right| U = u\right\} \\
&= E\left[E\left\{\left.\frac{I(T \leq C)\, I(T \leq y)}{P(C \geq T|U = u)}\right| \Delta = 1, U = u\right\}\right| U = u\Bigg] \\
&= E\{I(T \leq y)|\, U = u\} \\
&= E\{N_T(y)|\, U = u\}. \qquad (4.15)
\end{aligned}$$

The identity (4.15) allows us to study the expectation $E\{N_T(y)|\, U\}$ for the survival time T by using the observed time Y, where the inverse weight $P(C \geq Y|U = u)$ is imposed

to correct for the censoring effect. A similar strategy of (4.15) was used by Xia et al. (2010) and Lu and Li (2011). However, the main drawback of (4.15) lies in the requirement of $P(C \geq Y|U = u)$ which is basically unknown; $P(C \geq Y|U = u)$ involves U, which in turn involves B, the target of our primary interest. Although in principle, it is possible to employ non-parametric methods, such as the local linear estimation (Xia et al. 2010) or the Kaplan-Meier estimator (Lu and Li 2011) to estimate $P(C \geq Y|U = u)$, the resulting inferential procedures are complex and the results are often not efficient.

Driven by these concerns, we explore a different inference method. First, we present the following proposition whose proof is given in Appendix C.3.1.

Proposition 4.2.2

$$\begin{aligned} E\{N_T(y)|U\} &= E\{N_Y(y)|\Delta = 1, U\} \\ &= E\left\{N_Y(y)|\Delta = 1, B^\top U\right\} \quad \textit{for} \;\; y > 0. \end{aligned}$$

Proposition 4.2.2 shows that the observed measurement Y can be used to develop the regression model by conditioning on $\Delta = 1$. This property gives us a simple and more straightforward basis for the following development, which is based on the regression model

$$F(y, u) = E\{N_Y(y)|\,\Delta = 1, U = u\} \tag{4.16}$$

for given $y > 0$ and u.

4.2.2 Estimation Procedures

As noted by Ma and Zhu (2013), the basis matrix B is not unique even though $\mathcal{S}_{T|U}$ is; for any full rank $d \times d$ matrix $\mathbf{A}$, $B\mathbf{A}$ may generate the same column space as B does. In order to uniquely map CS to a basis matrix, Ma and Zhu (2013) suggested to consider the decomposition $B = (B_u^\top, B_l^\top)^\top$ for any $p \times d$ matrix B with rank d, where B_u is a $d \times d$ matrix whose inverse B_u^{-1} exists, and B_l is a $(p-d) \times d$ matrix. With this decomposition for B, setting $\mathbf{A} = B_u^{-1}$ gives that

$$BB_u^{-1} = (I_{d\times d}, B_l^\top B_u^{-1})^\top. \tag{4.17}$$

This suggests that for any $p \times d$ matrix B of rank d, by (4.17), it suffices to consider $B_l^\top B_u^{-1}$, a $(p-d) \times d$ matrix which can be of any form. As a result, by (4.17), we consider the set of all $p \times d$ matrices of the form

$$B = \left(I_{d\times d}, C^\top\right)^\top \tag{4.18}$$

with C being a $(p-d)\times d$ matrix with $(p-d)d$ unknown parameters. Estimation of B is equivalent to estimation of C.

To estimate (4.16), we implement the kernel estimation, yielding the estimator

$$\widehat{F}\left(y, B^{\top}u\right) = \frac{\sum_{i=1}^{n} \Delta_i I\left(Y_i \leq y\right) \mathcal{K}_h\left(B^{\top}U_i - B^{\top}u\right)}{\sum_{i=1}^{n} \Delta_i \mathcal{K}_h\left(B^{\top}U_i - B^{\top}u\right)}, \tag{4.19}$$

where $\mathcal{K}_h(u) = \prod_{j=1}^{d} \frac{1}{h} K\left(\frac{u_j}{h}\right)$, h is a positive bandwidth, and $K(v)$ is a qth-order kernel function with $\int K(v)dv = 1$, $\int v^k K(v)dv = 0$ for $k = 1, \cdots, q-1$, $\int v^q K(v)dv < \infty$, and q is a positive constant.

With the estimator of $F(\cdot,\cdot)$ by (4.19), we use the cross-validation (CV) criterion to construct the CV value

$$CV\left(B, d, h\right) = \frac{1}{n}\sum_{i=1}^{n} \int_0^{\tau} \left\{ I\left(Y_i \leq y\right) - \widehat{F}^{(-i)}(y, B^{\top}U_i) \right\}^2 d\widehat{F}_Y(y), \tag{4.20}$$

where $\widehat{F}_Y(\cdot)$ is the empirical distribution function of Y_i and $\widehat{F}^{(-i)}(y, B^{\top}u)$ is the estimator of (4.19) with the ith subject being deleted. The estimator of (B, d, h) can be derived by minimizing (4.20), i.e,

$$\left(\widehat{B}, \widehat{d}, \widehat{h}\right) = \underset{B,d,h}{\operatorname{argmin}}\, CV\left(B, d, h\right). \tag{4.21}$$

4.2.3 Computational Algorithm

The implementation of the minimization problem (4.21) can be realized by the following computational algorithm.

Step 1: For $d = 0$, calculate

$$CV[0] = \frac{1}{n}\sum_{i=1}^{n} \int_0^{\tau} \left\{ I\left(Y_i \leq y\right) - \frac{\sum_{i=1}^{n} \Delta_i I(Y_i \leq y)}{\sum_{i=1}^{n} \delta_i} \right\}^2 d\widehat{F}_Y(y).$$

Step 2: For any given $d \geq 1$, let $\left(\widehat{B}_d, \widehat{h}_d\right)$ denote the estimators which are obtained by minimizing (4.20):

$$\left(\widehat{B}_d, \widehat{h}_d\right) = \underset{B,h}{\operatorname{argmin}} CV\left(B, d, h\right).$$

Let $CV[d] = CV\left(\widehat{B}_d, d, \widehat{h}_d\right)$.

Step 3: Continue Step 2 until $d = \widehat{d}$ with $CV[\widehat{d}+1] > CV[\widehat{d}]$. As a result, the final estimators are $\left(\widehat{B}, \widehat{d}, \widehat{h}\right) = \left(\widehat{B}_{\widehat{d}}, \widehat{d}, \widehat{h}_{\widehat{d}}\right)$.

4.3 Theoretical Results

Let $\text{vec}(\cdot)$ denote the vectorization operation that stacks the columns of a matrix, and let $\|\cdot\|$ represent the Frobenius norm of a matrix. Define $a^{\otimes 2} = aa^\top$ for any vector a. To emphasize the involvement of the estimator $\widehat{L}$, we write $\widehat{U}_i = \widehat{L}X_i^*$ as defined in Section 4.1.3. For any function $f(\alpha)$, let $\nabla_\alpha^j f(\alpha)$ denote the jth order derivative of the function $f(\cdot)$ with respect to α. Let B_0 and d_0 denote the true values of the parameter and its structural dimension, respectively. Let h_0 be the optimal bandwidth. We first present the consistency of the estimators $(\widehat{B}, \widehat{d}, \widehat{h})$ whose proof is placed in Appendix C.4.1.

Theorem 4.3.1 *Under regularity conditions in Appendix C.1, for any $\eta > 0$, as $n \to \infty$,*

$$\widehat{B} \xrightarrow{p} B_0 \text{ and } P\left(\widehat{d} = d_0, \left|\frac{\widehat{h}}{h_0} - 1\right| < \eta\right) \to 1.$$

For $l = 0, 1$ and $j = 0, 1, 2$, let

$$\widehat{\mathbb{F}}_{l,B,L}^{(j)}(y,u) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \left\{I(Y_i \leq y)\right\}^l \nabla_{\text{vec}(B)}^j \mathcal{K}\left(B^\top U_i - u\right) \tag{4.22}$$

and

$$\widehat{\mathbb{F}}_{l,B,\widehat{L}}^{(j)}(y,u) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \left\{I(Y_i \leq y)\right\}^l \nabla_{\text{vec}(B)}^j \mathcal{K}\left(B^\top \widehat{U}_i - u\right). \tag{4.23}$$

Furthermore, let $f_{B^\top U}(B^\top u)$ denote the density function of $B^\top U$, and define

$$\mathbb{F}^{(j)}_{l,B,L}(y,u) = \nabla^j_{\mathrm{vec}(B)} \left\{F(y, B^\top u)\right\}^l E(\Delta_i) E\left\{ (U_i - u)^{\otimes j} \middle| B^\top U_i = B^\top u \right\} f_{B^\top U}(B^\top u),$$

and

$$\widetilde{\mathbb{F}}^{(j)}_{i,l,B,L}(y,u) = \Delta_i \left\{I\left(Y_i \le y\right)\right\}^l \nabla^j_{\mathrm{vec}(B)} \mathcal{K}\left(B^\top U_i - u\right) - \mathbb{F}^{(j)}_{l,B,L}(y,u)$$

for $j = 0, 1, 2$ and $l = 0, 1$. Define

$$\zeta^{(0)}_{i,B_0}(y,u) = \sum_{l=0}^{1} \widetilde{\mathbb{F}}^{(0)}_{i,l,B,L}(y,u) \frac{\left\{-F\left(y, B_0^\top u\right)\right\}^{1-l}}{\mathbb{F}_{0,B,L}\left(y, B_0^\top u\right)}. \tag{4.24}$$

Theorem 4.3.2 *Under regularity conditions in Appendix C.1, then*

$$\sup_{y,u} \left| \widehat{F}\left(y, \widehat{B}^\top u\right) - F(y, B_0^\top u) - \frac{1}{n} \sum_{i=1}^{n} \zeta^{(0)}_{i,B_0}(y,u) \right| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

The proof of Theorem 4.3.2 is deferred to Appendix C.4.2.

Finally, we present the asymptotic distribution of the estimator $\widehat{B}$. Define

$$F^{(0)}(y,u) = \frac{\mathbb{F}^{(1)}_{1,B,L}(y,u)}{\mathbb{F}^{(0)}_{0,B,L}(y,u)}, \tag{4.25}$$

$$F^{(1)}(y,u) = \sum_{l=0}^{1} \frac{\left\{-F(y,u)\right\}^l \mathbb{F}^{(1)}_{1-l,B,L}(y,u)}{\mathbb{F}^{(0)}_{0,B,L}(y,u)}, \tag{4.26}$$

and

$$F^{(2)}(y,u) = \sum_{l_1=0}^{1} \sum_{l_2=0}^{1} 2^{l_1} \left\{ \frac{-\mathbb{F}^{(2-l_1)}_{0,B,L}(y,u)}{\mathbb{F}^{(0)}_{0,B,L}(y,u)} \right\}^{l_1+l_2} \left(\frac{\mathbb{F}^{(\{2-l_1\}\{1-l_2\})}_{1,B,L}(y,u)}{\mathbb{F}^{(0)}_{0,B,L}(y,u)} \right).$$

Let

$$U(B_0) = \int_0^\tau \left\{ I\left(Y_i \le y\right) - F^{(0)}(y, B_0^\top U_i) \right\} F^{(1)}(y, B_0^\top U_i) dF_Y(y)$$

and

$$\mathcal{A} = 2E\left(\int_0^\tau \left[\left\{F^{(1)}\left(y, B_0^\top U_i\right)\right\}^{\otimes 2} - F^{(2)}\left(y, B_0^\top U_i\right)\left\{I\left(Y_i \le y\right) - F^{(0)}\left(y, B_0^\top U_i\right)\right\}\right] dF_Y(y)\right).$$

We define

$$\mathcal{T}(B_0) = E\left[\int_0^\tau \left\{F^{(1)}(y, B_0^\top U_i)\right\}^{\otimes 2} dF_Y(y)\right]\Sigma_{X^*}.$$

Theorem 4.3.3 *Suppose that regularity conditions in Appendix C.1 holds.*

(a) Assume that L is known, then as $n \to \infty$,

$$\sqrt{n}\left\{vec(\widehat{B}) - vec(B_0)\right\} \xrightarrow{d} N\left(0, \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right),$$

where $\mathcal{B} = E\left\{U^{\otimes 2}(B_0)\right\}$.

(b) Assume that L is unknown and estimated based on either repeated measurements or validation data. Let Φ_i be $\left\{\left(X_{i1}^* - X_{i2}^*\right)\left(X_{i1}^* - X_{i2}^*\right)^\top - 2\Sigma_\epsilon\right\}$ if L is estimated based on repeated measurements, and let Φ_i be $\left\{\left(X_i - \mu_X\right)\left(X_i^* - \mu_{X^*}\right)^\top - \Sigma_{XX^*}\right\}$ if L is estimated from validation data. Then as $n \to \infty$,

$$\sqrt{n}\left\{vec(\widehat{B}) - vec(B_0)\right\} \xrightarrow{d} N\left(0, \mathcal{A}_L^{-1}\mathcal{B}_L\mathcal{A}_L^{-1}\right),$$

where $\mathcal{A}_L = \mathcal{A}$ and $\mathcal{B}_L = E\left[\left\{U(B_0) + \mathcal{T}(B_0)\Phi_i\right\}^{\otimes 2}\right]$.

4.4 SDR with Ultrahigh-Dimensional Covariates

In this section, we explore the SDR with ultrahigh-dimensional covariates. We first propose a valid feature screening method to deal with ultrahigh-dimensional censored data with measurement error. In the second step, we develop estimation procedures for parameter B and the structural dimension d based on the selected covariates.

4.4.1 Review of the Distance Correlation Method

To start, we briefly review the distance correlation (DC) method proposed by Székely et al. (2007).

Suppose μ and ν are random vectors with the characteristic functions $\phi_\mu(\cdot)$ and $\phi_\nu(\cdot)$, respectively, and let $\phi_{\mu,\nu}(\cdot)$ be the joint characteristic function of μ and ν. For any complex function $\phi(\cdot)$, let $\|\phi(\cdot)\|^2 = \phi(\cdot)\bar{\phi}(\cdot)$, where $\bar{\phi}(\cdot)$ is the conjugate of $\phi(\cdot)$. The *distance covariance* between μ and ν is defined as

$$\mathrm{dcov}(\mu,\nu) = \int_{\mathbb{R}^{d_\mu+d_\nu}} \|\phi_{\mu,\nu}(r,s) - \phi_\mu(r)\phi_\nu(s)\|^2 \, w(r,s)drds, \tag{4.27}$$

where d_μ and d_ν are dimensions of μ and ν, respectively, and

$$w(r,s) = \left\{ c_{d_\mu} c_{d_\nu} \|r\|_{d_\mu}^{1+d_\mu} \|s\|_{d_\nu}^{1+d_\nu} \right\}^{-1} \tag{4.28}$$

with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$ and $\|a\|_d$ is the Euclidean norm of any vector $a \in \mathbb{R}^d$. Consequently, the DC between μ and ν is defined as

$$\mathrm{dcorr}(\mu,\nu) = \frac{\mathrm{dcov}(\mu,\nu)}{\sqrt{\mathrm{dcov}(\mu,\mu)\mathrm{dcov}(\nu,\nu)}}. \tag{4.29}$$

Székely et al. (2007) showed that random vectors μ and ν are independent if and only if $\mathrm{dcorr}(\mu,\nu) = 0$. This property will be used in an analogeous way to Li et al. (2012) to develop a feature screening procedure and identify covariates associated with the response.

Next, we describe estimation of $\mathrm{dcorr}(\mu,\nu)$ using sample data. Suppose that $\{(\mu_i,\nu_i) : i = 1,\cdots,n\}$ is a random sample and has the same distribution of (μ,ν). Note that by Székely et al. (2007), $\mathrm{dcov}(\mu,\nu)$ can be expressed as

$$\mathrm{dcov}(\mu,\nu) = J_1 + J_2 - 2J_3,$$

where

$$\begin{aligned}
J_1 &= E\left(\|\mu - \widetilde{\mu}\|_{d_\mu} \|\nu - \widetilde{\nu}\|_{d_\nu} \right), \\
J_2 &= E\left(\|\mu - \widetilde{\mu}\|_{d_\mu} \right) E\left(\|\nu - \widetilde{\nu}\|_{d_\nu} \right), \\
J_3 &= E\left\{ E\left(\|\mu - \widetilde{\mu}\|_{d_\mu} \middle| \, \mu \right) E\left(\|\nu - \widetilde{\nu}\|_{d_\nu} \middle| \, \nu \right) \right\},
\end{aligned}$$

and $(\widetilde{\mu}, \widetilde{\nu})$ is an independent copy of (μ, ν). Then J_k with $k = 1, 2, 3$ can be estimated by (Székely et al. 2007)

$$\begin{aligned}
\widehat{J}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mu_i - \mu_j\|_{d_\mu} \|\nu_i - \nu_j\|_{d_\nu}, \\
\widehat{J}_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mu_i - \mu_j\|_{d_\mu} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\nu_i - \nu_j\|_{d_\nu}, \\
\widehat{J}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \|\mu_i - \mu_l\|_{d_\mu} \|\nu_j - \nu_l\|_{d_\nu}.
\end{aligned}$$

As a result, $\mathrm{dcov}(\mu, \nu)$ is estimated by $\widehat{\mathrm{dcov}}(\mu, \nu) = \widehat{J}_1 + \widehat{J}_2 - 2\widehat{J}_3$, and thus, (4.29) can be estimated by

$$\widehat{\mathrm{dcorr}}(\mu, \nu) = \frac{\widehat{\mathrm{dcov}}(\mu, \nu)}{\sqrt{\widehat{\mathrm{dcov}}(\mu, \mu)\widehat{\mathrm{dcov}}(\nu, \nu)}}.$$

4.4.2 Ultrahigh-Dimensional Setting and Feature Selection

To present the idea for the proposed feature screening procedure with measurement error in X, we start with a simpler setting by pretending that X were precisely measured and no censoring exists with $Y = T$. Let

$$\mathcal{I} = \{k : X_k \text{ is dependent on the survival time } T \in [0, \tau]\}$$

denote the *active set* which contains all relevant covariates for the response T with $|\mathcal{I}| = \widetilde{p} < n$, and let $\mathcal{I}^c$ be the complement of $\mathcal{I}$ which contains all irrelevant covariates for the response T.

Let $X_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$ denote the vector containing all the active covariates, and let $X_{\mathcal{I}^c} = \{X_k : k \in \mathcal{I}^c\}$ be the vector containing all the irrelevant covariates. By definition,

$$T \perp\!\!\!\perp X | X_{\mathcal{I}} \text{ or } T \perp\!\!\!\perp X_{\mathcal{I}^c} | X_{\mathcal{I}}. \tag{4.30}$$

For $k = 1, \cdots, p$ and $j = 1, \cdots, d$, let β_{kj} denote the kth component in the vector β_j defined in Section 4.1.1 above (4.1). (4.1) and (4.30) indicate that $\sum_{j=1}^{d} |\beta_{kj}| > 0$ for $k \in \mathcal{I}$

and $\sum_{j=1}^{d} |\beta_{kj}| = 0$ for $k \in \mathcal{I}^c$. Equivalently, if $k \in \mathcal{I}$, then T must depend on X_k through at least one of the d linear combinations of β_j for $j = 1, \cdots, d$; if $k \in \mathcal{I}^c$, then no linear combinations contain k (Yu et al. 2016). As a result, based on those covariates in $\mathcal{I}$, it suffices to consider the SDR problem with

$$T \perp\!\!\!\perp X_{\mathcal{I}} | B_{\mathcal{I}}^{\top} X_{\mathcal{I}},$$

where $B_{\mathcal{I}}$ is the $\widetilde{p} \times d$ matrix, and $\mathcal{S}_{T|X} \subseteq \mathcal{S}(B_{\mathcal{I}})$.

We now need only to estimate the active set $\mathcal{I}$. To address the features of both censored responses and measurement error in covariates, we need to modify the DC method described in Section 4.4.1 to determine the active set $\mathcal{I}$.

Modified Censored Responses:

Since Y and T are not necessarily identical due to censoring, we consider a modified version of Y (Buckley and James 1979):

$$Y^* = \Delta Y + (1 - \Delta) E(T | \Delta = 0), \tag{4.31}$$

which satisfies $E(Y^*) = E(T)$ (Miller 1981, p.151).

To calculate Y^* in (4.31), we need to estimate $E(T|\Delta = 0)$. Now we fix Y at $Y = y$. By Condition (C1) in Appendix C.1 and the spirit of a "Buckley-James-type estimator" (e.g., Buckley and James 1979; Susarla et al. 1984), $E(T|\Delta = 0, Y = y)$ can be re-written as

$$\begin{aligned} E(T|\Delta = 0, Y = y) &= E(T|\tau > T > y, Y = y) \\ &= \int_y^{\tau} t \frac{f_T(t)}{P(\tau > T > y, Y = y)} dt \\ &= \int_y^{\tau} \frac{t f_T(t)}{1 - F_T(y)} dt \\ &= \frac{1}{1 - F_T(y)} \left[\{\tau - y F_T(y)\} - \int_y^{\tau} F_T(t) dt \right], \end{aligned} \tag{4.32}$$

where $f_T(\cdot)$ and $F_T(\cdot)$ are the probability density function and cumulative distribution function of T, respectively, and τ is defined in Section 4.1.3.

Let $G(y) = P(C \geq y)$ for $y > 0$. Note that by derivations similar to (4.15), we can show that

$$F_T(y) = E\{I(T \leq y)\} = E\left\{\frac{\Delta I(Y \leq y)}{G(Y)}\right\},$$

suggesting that $F_T(\cdot)$ can be estimated by

$$\widehat{F}_T(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{G}(Y_i)} I(Y_i \leq y), \tag{4.33}$$

where $\widehat{G}(y)$ is the Kaplan-Meier estimator of $G(y)$. As a result, the estimator of $E(T|\Delta = 0)$, denoted as $\widehat{E}(T|\Delta = 0)$, is determined by (4.32) with $F_T(y)$ replaced by (4.33), and thus, by (4.31), we have an approximate version of Y^*:

$$\widetilde{Y}^* = \Delta Y + (1-\Delta)\widehat{E}(T|\Delta = 0).$$

Feature Screening in the Presence of Measurement Error:

Since our error-prone correction in developing feature screening involves exponential functions, it is generally difficult to recover the function in terms of X if we consider measurement error model (4.4) with a general matrix Γ. However, under the measurement error model (4.4) with $\gamma = 0$ and Γ being the identity matrix, i.e.,

$$X^* = X + \epsilon, \tag{4.34}$$

it is possible to develop a feature screening method.

Following the spirit of the DC method, let

$$\omega_k = \text{dcorr}(Y^*, X_k) \tag{4.35}$$

denote the DC based on Y^* and the k unobserved covariate X_k for $k = 1, \cdots, p$. Now we derive the relationship between Y^* and the surrogate covariate X^*. Let $\phi_{Y^*}(r) = E\{\exp(\mathbf{i}rY^*)\}$ denote the characteristic function of Y^*, where $\mathbf{i}$ is a complex number with $\mathbf{i}^2 = -1$. Define

$$\phi_{X_k^*}(s) = E\{\exp(\mathbf{i}sX_k^*)\}\exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right)$$

and

$$\phi_{Y^*,X_k^*}(r,s) = E\left\{\exp\left(\mathbf{i}rY^* + \mathbf{i}sX_k^*\right)\right\}\exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right)$$

for $k = 1, \cdots, p$, where $\sigma_{\epsilon,kk}$ is the kth diagonal entry of Σ_ϵ. According to (4.27), the *modified* distance covariance between Y^* and X_k^* is defined as

$$\mathrm{dcov}^*(Y^*, X_k^*) = \int_{\mathbb{R}^{1+1}} \left\|\phi_{Y^*,X^*}(r,s) - \phi_{Y^*}(r)\phi_{X_k^*}(s)\right\|^2 w^*(r,s)drds,$$

where $w^*(r,s)$ is determined by (4.28) with $d_\mu = d_\nu = 1$, and thus, the *modified* (or *corrected*) DC between Y^* and X_k^* is given by

$$\mathrm{dcorr}^*(Y^*, X_k^*) = \frac{\mathrm{dcov}^*(Y^*, X_k^*)}{\sqrt{\mathrm{dcov}^*(Y^*, Y^*)\mathrm{dcov}^*(X_k^*, X_k^*)}}. \tag{4.36}$$

As a result, to select the active features for the surrogate covariates, we consider

$$\omega_k^* = \mathrm{dcorr}^*\left(Y^*, X_k^*\right) \tag{4.37}$$

for $k = 1, \cdots, p$, and the corresponding estimator is

$$\begin{aligned}
\widehat{\omega}_k^* &= \widehat{\mathrm{dcorr}^*}\left(\widetilde{Y}^*, X_k^*\right) \\
&= \frac{\widehat{\mathrm{dcov}^*}(\widetilde{Y}^*, X_k^*)}{\sqrt{\widehat{\mathrm{dcov}^*}(\widetilde{Y}^*, \widetilde{Y}^*)\widehat{\mathrm{dcov}^*}(X_k^*, X_k^*)}},
\end{aligned}$$

where $\widehat{\mathrm{dcov}^*}(Y^*, X_k^*) = \widehat{J}_1^* + \widehat{J}_2^* - 2\widehat{J}_3^*$ with

$$\begin{aligned}
\widehat{J}_1^* &= \frac{1}{2n^2\sigma_{\epsilon,kk}}\sum_{i=1}^n\sum_{j=1}^n \left\|\widetilde{Y}_i^* - \widetilde{Y}_j^*\right\|_1 \left\|X_{k,i}^* - X_{k,j}^*\right\|_1, \\
\widehat{J}_2^* &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left\|\widetilde{Y}_i^* - \widetilde{Y}_j^*\right\|_1 \frac{1}{2n^2\sigma_{\epsilon,kk}}\sum_{i=1}^n\sum_{j=1}^n \left\|X_{k,i}^* - X_{k,j}^*\right\|_1, \\
\widehat{J}_3^* &= \frac{1}{2n^3\sigma_{\epsilon,kk}}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n \left\|\widetilde{Y}_i^* - \widetilde{Y}_l^*\right\|_1 \left\|X_{k,j}^* - X_{k,l}^*\right\|_1,
\end{aligned}$$

and $X_{k,i}^*$ is the kth component of X_i^* for ith subject.

As suggested by Li et al. (2012), let the threshold value be $cn^{-\zeta}$ for some constants c and ζ, then the estimated active set is given by

$$\widehat{\mathcal{I}} = \left\{k : \widehat{\omega}_k^* \geq cn^{-\zeta}, k = 1, \cdots, p\right\}. \tag{4.38}$$

In practice, as suggested in Yan et al. (2017), Chen et al. (2019) and among others, we can specify the size of the active set $\mathcal{I}$ to be $\widetilde{p} = \left[\frac{n}{\log(n)}\right]$, where $[\cdot]$ stands for the floor function.

When the active set is determined, then for a subject $i = 1, \cdots, n$, the measurement error model (4.34) based on the active set $\mathcal{I}$ is given by

$$X_{i,\mathcal{I}}^* = X_{i,\mathcal{I}} + \epsilon_{i,\mathcal{I}}, \tag{4.39}$$

where $\epsilon_{i,\mathcal{I}} \sim N\left(0, \Sigma_{\epsilon_\mathcal{I}}\right)$ with $\widetilde{p} \times \widetilde{p}$ covariance matrix $\Sigma_{\epsilon_\mathcal{I}}$, and $X_{i,\mathcal{I}}^*$ and $X_{i,\mathcal{I}}$ are, respectively, $\widetilde{p}$-dimensional vectors of the observed and unobserved covariates based on the active set $\mathcal{I}$ for $i = 1, \cdots, n$.

Since the dimension of $X_{i,\mathcal{I}}^*$ is reduced to be $\widetilde{p} < n$, then $\Sigma_{X_\mathcal{I} X_\mathcal{I}^*} = \text{cov}(X_{i,\mathcal{I}}, X_{i,\mathcal{I}}^*)$ and $\Sigma_{X_\mathcal{I}^*} = \text{var}(X_{i,\mathcal{I}}^*)$ are invertible. Therefore, the "corrected" covariate based on the active set is proposed to be

$$U_{i,\mathcal{I}} = L_\mathcal{I} X_{i,\mathcal{I}}^* \tag{4.40}$$

for $i = 1, \cdots, n$, where $L_\mathcal{I} = \Sigma_{X_\mathcal{I} X_\mathcal{I}^*} \Sigma_{X_\mathcal{I}^*}^{-1}$.

Analogeous to the discussion in Section 4.1.3, we discuss the estimation of $L_\mathcal{I}$ by the three scenarios. When $L_\mathcal{I}$ is known, then we can directly calculate the "corrected" covariates $U_{i,\mathcal{I}}$ by (4.40). When $L_\mathcal{I}$ is unknown and repeated measurements are available, then similar to the discussion of Scenario II in Section 4.1.3, we estimate $L_\mathcal{I}$ by

$$\widehat{L}_\mathcal{I} = \left(\widehat{\Sigma}_{X_\mathcal{I}^*} - \widehat{\Sigma}_{\epsilon_\mathcal{I}}\right) \widehat{\Sigma}_{X_\mathcal{I}^*}^{-1},$$

where $\widehat{\Sigma}_{X_\mathcal{I}^*}$ and $\widehat{\Sigma}_{\epsilon_\mathcal{I}}$ are empirical estimators of $\Sigma_{X_\mathcal{I}^*}$ and $\Sigma_{\epsilon_\mathcal{I}}$, respectively. Finally, when $L_\mathcal{I}$ is unknown and validation data are available, then similar to the discussion of Scenario III in Section 4.1.3, $L_\mathcal{I}$ can be estimated by

$$\widehat{L}_\mathcal{I} = \widehat{\Sigma}_{X_\mathcal{I} X_\mathcal{I}^*} \widehat{\Sigma}_{X_\mathcal{I}^*}^{-1},$$

where $\widehat{\Sigma}_{X_\mathcal{I}^*}^{-1} = \frac{1}{m} \sum_{i=1}^{m} \left(x_{i,\mathcal{I}}^* - \bar{x}_\mathcal{I}^*\right) \left(x_{i,\mathcal{I}}^* - \bar{x}_\mathcal{I}^*\right)^\top$ and $\widehat{\Sigma}_{X_\mathcal{I} X_\mathcal{I}^*} = \frac{1}{m} \sum_{i=1}^{m} \left(x_{i,\mathcal{I}} - \bar{x}_\mathcal{I}\right) \left(x_{i,\mathcal{I}}^* - \bar{x}_\mathcal{I}^*\right)^\top$ with $\bar{x}_\mathcal{I} = \frac{1}{m} \sum_{i=1}^{m} x_{i,\mathcal{I}}$ and $\bar{x}_\mathcal{I}^* = \frac{1}{m} \sum_{i=1}^{m} x_{i,\mathcal{I}}^*$.

4.4.3 Estimation of $(B_{\mathcal{I}}, h_{\mathcal{I}}, d_{\mathcal{I}})$

With the feature screening which reduces X_i to $X_{i,\mathcal{I}}$, let $B_{\mathcal{I}}$, $h_{\mathcal{I}}$, and $d_{\mathcal{I}}$ denote the parameter, bandwidth, and structure dimension based on the active set $\mathcal{I}$, respectively. Since the "corrected" covariates based on the active set $\mathcal{I}$ is obtained in (4.40), we follow the similar procedures in Section 4.2 to determine the estimators of $(B_{\mathcal{I}}, h_{\mathcal{I}}, d_{\mathcal{I}})$. Specifically, replacing U_i in (4.19) by (4.40) gives

$$\widehat{F}_{\mathcal{I}}\left(y, B_{\mathcal{I}}^{\top}u\right) = \frac{\sum_{i=1}^{n} \delta_i I\left(Y_i \leq y\right) \mathcal{K}_h\left(B_{\mathcal{I}}^{\top}U_{i,\mathcal{I}} - B_{\mathcal{I}}^{\top}u\right)}{\sum_{i=1}^{n} \delta_i \mathcal{K}_h\left(B_{\mathcal{I}}^{\top}U_{i,\mathcal{I}} - B_{\mathcal{I}}^{\top}u\right)}. \tag{4.41}$$

Then the CV value based on the active set $\mathcal{I}$ is given by

$$CV_{\mathcal{I}}\left(B_{\mathcal{I}}, d_{\mathcal{I}}, h_{\mathcal{I}}\right) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{I\left(Y_i \leq y\right) - \widehat{F}_{\mathcal{I}}^{(-i)}(y, B_{\mathcal{I}}^{\top}U_{i,\mathcal{I}})\right\}^2 d\widehat{F}_Y(y), \tag{4.42}$$

where $\widehat{F}_Y(\cdot)$ is the empirical distribution function of Y_i and $\widehat{F}_{\mathcal{I}}^{(-i)}(y, B_{\mathcal{I}}^{\top}U_{i,\mathcal{I}})$ is the estimator of (4.41) with the ith subject being deleted.

Therefore, the estimators of $(B_{\mathcal{I}}, d_{\mathcal{I}}, h_{\mathcal{I}})$ can be derived by minimizing (4.42):

$$\left(\widehat{B}_{\mathcal{I}}, \widehat{d}_{\mathcal{I}}, \widehat{h}_{\mathcal{I}}\right) = \underset{B_{\mathcal{I}}, d_{\mathcal{I}}, h_{\mathcal{I}}}{\operatorname{argmin}} CV_{\mathcal{I}}\left(B_{\mathcal{I}}, d_{\mathcal{I}}, h_{\mathcal{I}}\right). \tag{4.43}$$

The computational algorithm in Section 4.2.3 can be applied to the minimization problem (4.43).

4.4.4 Theoretical Results

We first show the validity of feature selection criterion (4.37) in the sense that active features can be selected based on either X_i^* or X_i.

Theorem 4.4.1 *Active features based on X^* and X are the same. That is,*

$$dcorr^*\left(Y^*, X_k^*\right) > 0 \Longleftrightarrow dcorr\left(Y^*, X_k\right) > 0$$

or

$$dcorr^*\left(Y^*, X_k^*\right) = 0 \Longleftrightarrow dcorr\left(Y^*, X_k\right) = 0.$$

Theorem 4.4.1 indicates that the selected variables X_k^* based on (4.37) are equal to the true active variables X_k. For the estimated active set (4.38), we have the following result.

Theorem 4.4.2 *Under regularity conditions (C6) and (C7) in Appendix C.1, as* $n \to \infty$,

$$P\left(\mathcal{I} \subseteq \widehat{\mathcal{I}}\right) \to 1.$$

Theorem 4.4.2 ensures that the important covariates which are associated with the response are not to be screened out with the probability approaching one as the sample size goes to infinity. This property is also called the "sure screening property" by authors such as Fan and Lv (2008) and Li et al. (2012) for the context without measurement error.

Next, we describe the theoretical result of the estimator (4.43), whose proof is similar to the derivation in Section 4.3.

Let

$$U_{\mathcal{I}}(B_0) = \int_0^\tau \left\{ I\left(Y_i \le y\right) - F^{(0)}(y, B_0^\top U_{i,\mathcal{I}}) \right\} F^{(1)}(y, B_0^\top U_{i,\mathcal{I}}) dF_Y(y)$$

and

$$\begin{aligned} \mathcal{A}_{\mathcal{I}} \;=\; & 2E\left(\int_0^\tau \left[\left\{ F^{(1)}\left(y, B_0^\top U_{i,\mathcal{I}}\right) \right\}^{\otimes 2} \right.\right. \\ & \left.\left. -F^{(2)}\left(y, B_0^\top U_{i,\mathcal{I}}\right) \left\{ I\left(Y_i \le y\right) - F^{(0)}\left(y, B_0^\top U_{i,\mathcal{I}}\right) \right\} \right] dF_Y(y) \right). \end{aligned}$$

Define

$$\mathcal{T}_{\mathcal{I}}(B_0) = E\left[\int_0^\tau \left\{ F^{(1)}(y, B_0^\top U_{i,\mathcal{I}}) \right\}^{\otimes 2} dF_Y(y) \right] \Sigma_{X_{\mathcal{I}}^*}.$$

Theorem 4.4.3 *Suppose that regularity conditions in Appendix C.1 hold.*

(a) Let η^* *be any positive number. Then as* $n \to \infty$,

$$\widehat{B}_{\mathcal{I}} \xrightarrow{p} B_0 \text{ and } P\left(\widehat{d}_{\mathcal{I}} = d_0, \left| \frac{\widehat{h}_{\mathcal{I}}}{h_0} - 1 \right| < \eta^* \right) \to 1.$$

(b) Assume that $L_{\mathcal{I}}$ *is known. Then as* $n \to \infty$,

$$\sqrt{n}\left\{ vec(\widehat{B}_{\mathcal{I}}) - vec(B_0) \right\} \xrightarrow{d} N\left(0, \mathcal{A}_{\mathcal{I}}^{-1} \mathcal{B}_{\mathcal{I}} \mathcal{A}_{\mathcal{I}}^{-1}\right),$$

where $\mathcal{B}_{\mathcal{I}} = E\left\{ U_{\mathcal{I}}^{\otimes 2}(B_0) \right\}$.

(c) *Assume that $L_{\mathcal{I}}$ is unknown and estimated based on either repeated measurements or validation data. Let $\Phi_{i,\mathcal{I}}$ be $\left(X^*_{i1,\mathcal{I}} - X^*_{i2,\mathcal{I}}\right)\left(X^*_{i1,\mathcal{I}} - X^*_{i2,\mathcal{I}}\right)^\top - 2\Sigma_{\epsilon_{\mathcal{I}}}$ if $L_{\mathcal{I}}$ is estimated based on repeated measurements, and let $\Phi_{i,\mathcal{I}}$ be $(X_{i,\mathcal{I}} - \mu_{X_{\mathcal{I}}})\left(X^*_{i,\mathcal{I}} - \mu_{X^*_{\mathcal{I}}}\right)^\top - \Sigma_{X_{\mathcal{I}}X^*_{\mathcal{I}}}$ if $L_{\mathcal{I}}$ is estimated from validation data. Then as $n \to \infty$,*

$$\sqrt{n}\left\{vec(\widehat{B}_{\mathcal{I}}) - vec(B_0)\right\} \xrightarrow{d} N\left(0, \mathcal{A}^{-1}_{\mathcal{I};L}\mathcal{B}_{\mathcal{I};L}\mathcal{A}^{-1}_{\mathcal{I};L}\right),$$

where $\mathcal{A}_{\mathcal{I};L} = \mathcal{A}_{\mathcal{I}}$ and $\mathcal{B}_{\mathcal{I};L} = E\left[\{U_{\mathcal{I}}(B_0) + \mathcal{T}_{\mathcal{I}}(B_0)\Phi_{i,\mathcal{I}}\}^{\otimes 2}\right]$.

4.5 Numerical Studies

In this section, we conduct simulation studies to assess the performance of the proposed estimators for a variety of settings. We first design the simulation settings and then present the simulation results. Finally, the methods are implemented to analyze two real datasets.

4.5.1 Simulation Studies

Let $B_0 = (\beta_{10}, \beta_{20})$ be the true value of the $p \times d_0$ matrix with $d_0 = 2$, where $\beta_{10} = (1, 0, 1, 0, 0, \cdots, 0)^\top$ and $\beta_{20} = (0, 1, 0, 1, 0, \cdots, 0)^\top$ are $p \times 1$ vectors with only two elements being 1. We consider cases with $p = 10$ or 1000. The p-dimensional covariates X is generated from the multivariate normal distribution $N(0, \Sigma_X)$, where Σ_X is the covariance matrix with diagonal entries being one and non-diagonal entries being 0.4.

Given the covariates X and B_0, we use three models, the proportional hazards (PH), proportional odds (PO), and additive hazards (AH) models, to generate survival times. Specifically, the corresponding cumulative distribution functions $F(\cdot, \cdot)$ are formulated, respectively, as

$$\begin{aligned} F_{PH}(t, B_0^\top X) &= 1 - \exp\left[-t^2 \exp\left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\}\right], \\ F_{PO}(t, B_0^\top X) &= \frac{\exp\left[t - \left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\}\right]}{1 + \exp\left[t - \left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\}\right]}, \end{aligned}$$

and

$$F_{AH}(t, B_0^\top X) = 1 - \exp\left[-\left\{t^2 + t\left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\}\right\}\right].$$

Let $\mathcal{U}$ be generated from the uniform distribution $U(0,1)$. Then survival times T based on the PH and PO models can be generated from

$$T = \sqrt{\exp\left\{-\left(X^\top\beta_{10}\right)^2 - 2\left(X^\top\beta_{20}\right)\right\}\log\left(1-\mathcal{U}\right)}$$

and

$$T = \log\left\{\left(1+\mathcal{U}\right)^{-1} - 1\right\} + \left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\},$$

respectively, and survival times T based on the AH model can be obtained by solving the following equation

$$T^2 + T\left\{\left(X^\top\beta_{10}\right)^2 + 2\left(X^\top\beta_{20}\right)\right\} + \log\left(1-\mathcal{U}\right) = 0.$$

Let C be the censoring time generated from the uniform distribution $U(0,c)$, where c is a constant that is chosen to yield about 50% censoring rate. Consequently, we calculate $Y = \min\{T, C\}$ and $\Delta = I\left(T \leq C\right)$.

For the measurement error model, we take (4.34) with ϵ_i following the normal distribution $N(0,\Sigma_\epsilon)$, where Σ_ϵ is a diagonal matrix with diagonal entry being $\sigma_\epsilon^2 = 0.15^2, 0.5^2$, or 0.75^2. Furthermore, we consider three scenarios described in Section 4.1.3. If Σ_ϵ is unknown, then the following two scenarios are considered as additional information:

Scenario 1: Validation data

For $i = 1, \cdots, m$ with $m = 100$, X_i and ϵ_i are again be independently generated from $N(0,\Sigma_X)$ and $N(0,\Sigma_\epsilon)$, respectively, and X_i^* is generated from

$$X_i^* = X_i + \epsilon_i$$

for $i = 1, \cdots, m$.

Scenario 2: Repeated measurements

For $i = 1, \cdots, m$ with $m = 100$ and $r = 1, 2$, X_i and ϵ_{ir} are again be independently generated from $N(0,\Sigma_X)$ and $N(0,\Sigma_\epsilon)$, respectively, and X_{ir}^* is generated from

$$X_{ir}^* = X_i + \epsilon_{ir}$$

for $i = 1, \cdots, m$ and $r = 1, 2$.

Let $\{(Y_i, \Delta_i, X_i^*) : i = 1, \cdots, n\}$ denote the sample with size $n = 200, 300$, or 400. We repeat computations 500 times for each setting.

To assess the accuracy of the estimator of B, we consider the Frobenius norm

$$\|\Delta_B\| = \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{d}\left|\widehat{B}_{ij} - B_{0,ij}\right|^2}$$

for $\Delta_B = \widehat{B} - B_0$. To examine the performance of the estimator of $F(\cdot)$, we consider the mean integrand squared error (MISE)

$$MISE(\widehat{F}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\widehat{F}\left(y, \widehat{B}^\top u_i\right) - F(y, B_0^\top u_i)\right\}^2 d\widehat{F}_Y(y).$$

For each setting, we calculate the proportions of $\widehat{d}$ for 500 times simulation, given by

$$\frac{1}{500}\sum_{k=1}^{500} I\left(\widehat{d}_k = d\right) \text{ for } d = 0, 1, 2, 3, \cdots,$$

and determine the estimated dimension as the largest proportion.

For the case with $p = 10$, we implement the proposed method in Section 4.2 directly to the dataset, while for the case with $p = 1000$, we first use the feature screening method to screen the variables, and then estimate B and d using the method in Section 4.4.3. We compare the performance of the proposed methods with the *naive estimators* of $F(\cdot)$ and B, which are derived by directly implementing the observed covariates X_i^* in (4.19) and (4.20). As a reference for comparisons, we also use the true values of X for the estimation, and denote this method as "true".

The results for $p = 10$ with the three scenarios are reported in Tables 4.1-4.3, and the results for $p = 1000$ with the three scenarios are summarized in Tables 4.4-4.6. In terms of estimation of B and $F(\cdot)$, the naive method produces biased results, and the finite sample bias increases as the degree of measurement error increases; the proposed methods greatly outperform the naive approach, yielding results that are fairly close to those produced from the reference method by using the true measurements of the covariates. Agreeing with the phenomenon we observed in the literature of measurement error models, the standard errors associated with the proposed methods are larger than those obtained from the naive method, which is the price paid to correct biases induced from the measurement error in covariates. While the differences for estimation of h and d are not very striking

between the naive method and our proposed methods, our approaches perform better than the naive approach.

In summary, in the presence of measurement error, the naive method yields unsatisfactory results. The proposed methods successfully correct the measurement error effects for various settings.

4.5.2 Analysis of ACTG 175 Dataset

We implement the proposed method to analyze the AIDS Clinical Trials Group (ACTG) 175 data which were discussed by Hammer et al. (1996). The ACTG 175 study was a double-blind randomized clinical trial which evaluated the HIV treatment effects. The dataset is available in R package "`speff2trial`". The dataset contains measurements on 26 variables for 2139 individuals; these variables are `age`, `wtkg`, `hemo`, `homo`, `drugs`, `karnof`, `oprior`, `z30`, `zprior`, `preanti`, `race`, `gender`, `str2`, `strat`, `symptom`, `treat`, `offtrt`, `cd40`, `cd420`, `cd496`, `r`, `cd80`, `cd820`, `cens`, `days` and `arms`. Since the variable `cd496` contains missing values and `r` is its missing indicator, so we remove those two variables. In addition, we remove variables `zprior` and `treat` due to that `zprior` is the constant 1 for all subjects and `treat` indicates whether or not the subject received the zidovudine treatment, overlapping with `arms`. As a result, in addition to the survival time `days` and the censoring indicator `cens`, we have $p = 20$ covariates in the dataset where CD4 is error-prone. The censoring rate of this dataset is approximately 75.6%.

Fourty-four subjects were measured once for the CD4 counts at the baseline, while 2095 subjects had two replicated baseline measurements of CD4 counts. As discussed in Yi (2017, Section 3.6.4), let X denote log(CD4 count + 1). To implement the proposed method, we consider the measurement error model (4.8) due to the availability of repeated measurements. Consequently, based on the discussion of Scenario II in Section 4.1.3, the estimates of Σ_ϵ and Σ_{X^*} are given by $\widehat{\Sigma}_\epsilon = 0.035$ and $\widehat{\Sigma}_{X^*} = 0.114$, respectively, yielding $\widehat{L} = 0.693$ as indicated by (4.12). In this study, we consider the entire data with error correction by $\widehat{L}$. In addition to X, let Z denote the vector of the remaining 19 covariates. Consequently, we let $U = \left(\widehat{L}X, Z^\top\right)^\top$ denote a 20-dimensional vector of covariates to be implemented with the proposed methods.

The naive method and the proposed method give $\widehat{d} = 2$, suggesting that there are two directions in the central space, say β_1 and β_2. We first present the scatter plots of the survival time Y_i and $\widehat{\beta}_k^\top U_i$ with $k = 1, 2$ in Figure 4.1. The naive method and the proposed method show similar patterns but the scatter plot based on naive method seems

more variable than that of the proposed method. We also examine the estimated functions $1 - \widehat{F}(\cdot)$ for subjects $i = 1, 7$ and 23, and the curves are displayed in Figure 4.2. It is seen that the estimated curves based on the naive and the proposed methods have similar patterns.

4.5.3 Analysis of NKI Breast Cancer Data

We now implement our proposed method to analyze the breast cancer data collected by the Netherlands Cancer Institute (NKI) (van de Vijver et al. 2002). Tumor information from 295 women with breast cancer was collected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. Tumors of those patients were primarily invasive breast cancer carcinoma that were about 5 cm in diameter. Patients at diagnosis were 52 years or younger and the diagnosis was done from 1984 to 1995. Of all those patients, 79 patients died before the study ended, yielding approximately the 73.2% censoring rate. For each tumor of a patient, about 25000 gene expressions were collected.

Since measurement error in gene expressions is a typical feature (Rocke and Durbin 2001), it is imperative take into account of the measurement error effects when estimating active set $\mathcal{I}$ and the central space. We treat the log intensities of gene expression values as the covariates and implement the proposed method in Section 4.4 to analyze the data. Because this dataset contains no information to characterize the degree of measurement error that is accompanying with the gene expressions, here we conduct sensitivity analyses to investigate the measurement error effects on analysis results. Let Σ be the covariance matrix of the gene expressions. For sensitivity analyses, we consider $\Sigma + \Sigma_e$ to be the covariance matrix for the measurement error model (4.39), where Σ_e is the diagonal matrix with diagonal elements being a common value σ_e^2, which is specified as $\sigma_e^2 = 0.15^2, 0.55^2$, or 0.75^2 to feature increasing degrees of measurement error in those gene expressions.

We first use (4.37) to determine an estimated active set $\widehat{\mathcal{I}}$ which contains $\widetilde{p} = \left[\frac{79}{\log(79)}\right] =$ 18 response-associated gene expressions, including `NM_016359`, `NM_003748`, `AA555029_RC`, `AL080059`, `AL137718`, `NM_020974`, `NM_002073`, `NM_004994`, `NM_003875`, `NM_015984`, `X05610`, `NM_006931`, `NM_002916`, `NM_001282`, `Contig2399_RC`, `NM_018354`, `NM_003862`, and `NM_000599`. Based on an estimated active set $\widehat{\mathcal{I}}$ and the proposed method in Section 4.4.3, the structural dimension d is suggested to be $\widehat{d}_{\mathcal{I}} = 1$, and the estimate $\widehat{B}_{\mathcal{I}} = \widehat{\beta}_1$ of a basis $B_{\mathcal{I}}$ is obtained to be one direction in the central space. In addition, we apply the naive method to analyze the data. In Figure 4.3, we display four scatter plots of Y and $\beta_1^\top X$ which are obtained from the proposed method with different degrees of measurement error assumed, together with the naive method. The results obtained from the proposed method with dif-

ferent degrees of measurement error show similar patterns of curves, but the curve based on the naive method tends to be linear.

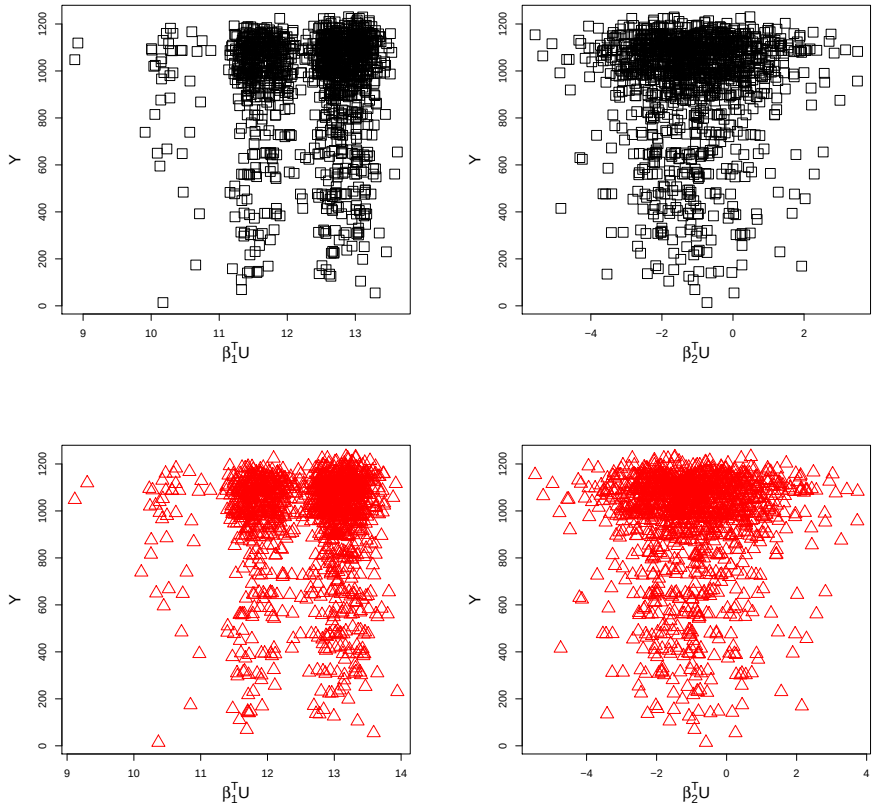
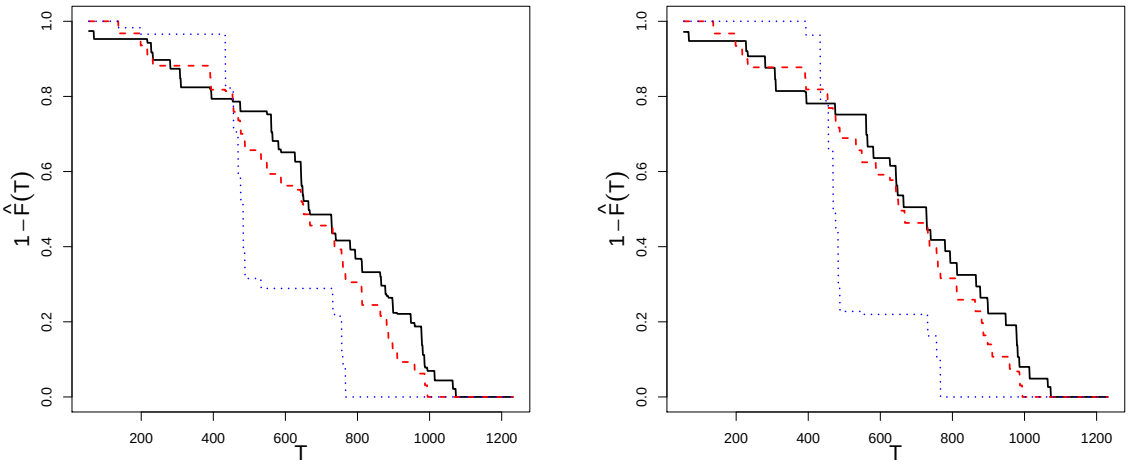
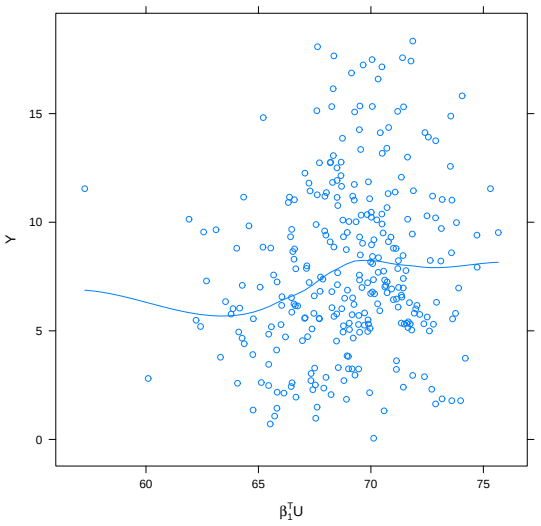


Figure 4.1: Scatter plots of survival time Y and $\beta_j^\top U$ with $j = 1, 2$. The left panel is $\beta_1^\top U$ and the right panel is $\beta_2^\top U$. The first row with black boxes ($\square$) is obtained from the naive approach, and the second row with red triangles ($\triangle$) is obtained from the proposed method.

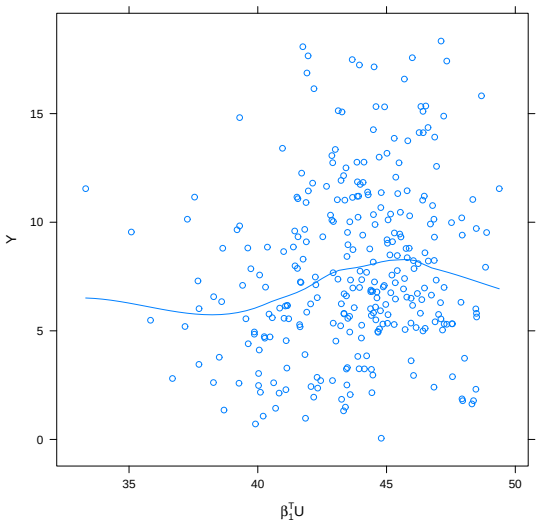


Proposed method *Naive method*

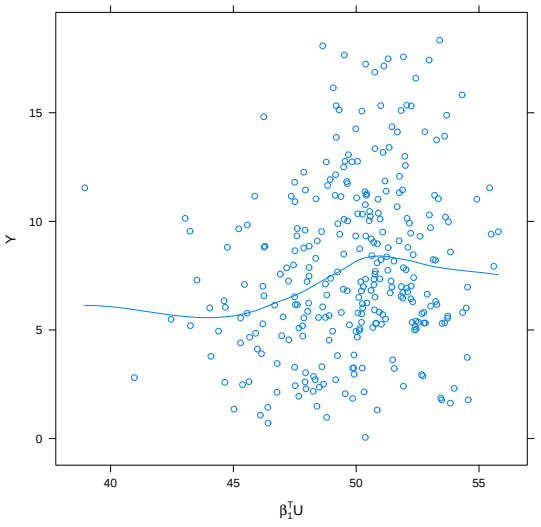
Figure 4.2: Estimated curves of $1 - \widehat{F}(y|U_i)$. The solid curve is for subject $i = 1$, the dash curve is for subject $i = 7$, and the dot curve is for subject $i = 23$.



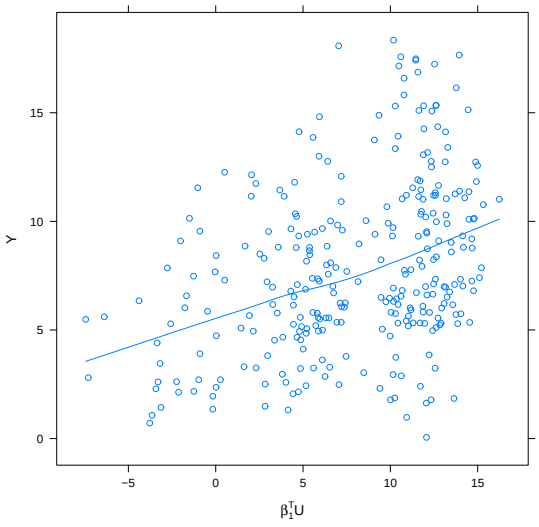
(a) $\sigma_e^2 = 0.15^2$



(b) $\sigma_e^2 = 0.55^2$



(c) $\sigma_e^2 = 0.75^2$



(d) naive

Figure 4.3: Scatter plots of survival time Y and $\beta_1^\top U$. (a)-(c) are based on the proposed method with $\sigma_e^2 = 0.15^2, 0.55^2$, or 0.75^2, (d) is based on the naive estimator.

Table 4.1: Simulation results for the estimators of B wiht known L

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(200, 10)$	0.15	Naive	0.300	0.013	0.188	0.022	0.806	0.015	0.000	0.090	0.910	0.000
			Corrected	0.021	0.020	0.072	0.025	0.802	0.015	0.000	0.000	1.000	0.000
		0.50	Naive	0.315	0.016	0.191	0.018	0.839	0.014	0.000	0.096	0.904	0.000
			Corrected	0.023	0.021	0.073	0.026	0.798	0.015	0.000	0.020	0.980	0.000
		0.75	Naive	0.319	0.016	0.206	0.021	0.824	0.016	0.000	0.103	0.897	0.000
			Corrected	0.033	0.025	0.076	0.024	0.744	0.017	0.000	0.019	0.981	0.000
			True	0.019	0.019	0.062	0.021	0.808	0.016	0.000	0.000	1.000	0.000
PO	$(200, 10)$	0.15	Naive	0.279	0.016	0.197	0.014	0.387	0.017	0.000	0.091	0.909	0.000
			Corrected	0.016	0.022	0.079	0.016	0.544	0.018	0.000	0.000	1.000	0.000
		0.50	Naive	0.297	0.017	0.198	0.017	0.409	0.018	0.000	0.096	0.904	0.000
			Corrected	0.020	0.024	0.086	0.019	0.580	0.017	0.000	0.005	0.995	0.000
		0.75	Naive	0.305	0.018	0.199	0.012	0.377	0.019	0.000	0.096	0.904	0.000
			Corrected	0.028	0.026	0.094	0.022	0.576	0.021	0.000	0.010	0.990	0.000
			True	0.010	0.019	0.068	0.010	0.558	0.018	0.000	0.000	1.000	0.000
AH	$(200, 10)$	0.15	Naive	0.286	0.017	0.184	0.013	0.605	0.020	0.000	0.088	0.912	0.000
			Corrected	0.017	0.028	0.064	0.015	0.743	0.018	0.000	0.000	1.000	0.000
		0.50	Naive	0.295	0.018	0.191	0.014	0.590	0.019	0.000	0.092	0.908	0.000
			Corrected	0.023	0.028	0.061	0.015	0.733	0.017	0.000	0.000	1.000	0.000
		0.75	Naive	0.305	0.018	0.196	0.015	0.568	0.021	0.000	0.092	0.908	0.000
			Corrected	0.028	0.030	0.060	0.018	0.710	0.021	0.000	0.009	0.991	0.000
			True	0.012	0.023	0.037	0.012	0.745	0.017	0.000	0.000	1.000	0.000

Table 4.2: Simulation results for the estimators of B with repeated measurements

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(300, 10)$	0.15	Naive	0.292	0.013	0.193	0.020	0.712	0.015	0.000	0.010	0.990	0.000
			Corrected	0.017	0.021	0.064	0.022	0.759	0.016	0.000	0.000	1.000	0.000
		0.50	Naive	0.295	0.015	0.196	0.017	0.777	0.015	0.000	0.015	0.985	0.000
			Corrected	0.025	0.024	0.070	0.018	0.610	0.018	0.000	0.000	1.000	0.000
		0.75	Naive	0.298	0.016	0.207	0.016	0.765	0.016	0.000	0.016	0.984	0.000
			Corrected	0.029	0.026	0.077	0.018	0.386	0.025	0.000	0.000	1.000	0.000
			True	0.010	0.016	0.058	0.015	0.816	0.015	0.000	0.000	1.000	0.000
PO	$(300, 10)$	0.15	Naive	0.263	0.013	0.189	0.010	0.420	0.018	0.000	0.009	0.991	0.000
			Corrected	0.012	0.026	0.044	0.013	0.514	0.020	0.000	0.000	1.000	0.000
		0.50	Naive	0.305	0.016	0.194	0.011	0.429	0.020	0.000	0.011	0.989	0.000
			Corrected	0.016	0.028	0.053	0.014	0.017	0.021	0.000	0.002	0.997	0.001
		0.75	Naive	0.313	0.016	0.194	0.011	0.492	0.020	0.000	0.013	0.987	0.000
			Corrected	0.024	0.028	0.061	0.019	0.142	0.024	0.000	0.003	0.996	0.001
			True	0.009	0.016	0.041	0.010	0.576	0.020	0.000	0.000	1.000	0.000
AH	$(300, 10)$	0.15	Naive	0.270	0.016	0.176	0.013	0.620	0.020	0.000	0.005	0.994	0.001
			Corrected	0.013	0.023	0.052	0.017	0.652	0.022	0.000	0.000	1.000	0.000
		0.50	Naive	0.275	0.017	0.184	0.015	0.618	0.020	0.000	0.010	0.989	0.001
			Corrected	0.015	0.025	0.055	0.017	0.460	0.022	0.000	0.006	0.994	0.000
		0.75	Naive	0.295	0.018	0.186	0.016	0.638	0.019	0.000	0.011	0.987	0.002
			Corrected	0.016	0.026	0.059	0.018	0.313	0.029	0.000	0.007	0.992	0.001
			True	0.010	0.019	0.036	0.009	0.681	0.019	0.000	0.000	1.000	0.000

Table 4.3: Simulation results for the estimators of B with validation data

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(400, 10)$	0.15	Naive	0.249	0.012	0.156	0.020	0.715	0.012	0.000	0.010	0.985	0.005
			Corrected	0.014	0.021	0.058	0.022	0.820	0.013	0.000	0.000	1.000	0.000
		0.50	Naive	0.251	0.013	0.191	0.016	0.747	0.016	0.000	0.020	0.973	0.007
			Corrected	0.018	0.022	0.063	0.023	0.825	0.014	0.000	0.000	1.000	0.000
		0.75	Naive	0.272	0.016	0.186	0.019	0.778	0.016	0.000	0.017	0.978	0.005
			Corrected	0.017	0.024	0.066	0.025	0.863	0.013	0.000	0.000	1.000	0.000
			True	0.009	0.014	0.040	0.017	0.813	0.016	0.000	0.000	1.000	0.000
PO	$(400, 10)$	0.15	Naive	0.253	0.012	0.164	0.013	0.306	0.019	0.000	0.008	0.992	0.000
			Corrected	0.011	0.023	0.040	0.016	0.506	0.021	0.000	0.002	0.998	0.000
		0.50	Naive	0.259	0.016	0.172	0.014	0.374	0.021	0.000	0.011	0.989	0.000
			Corrected	0.016	0.025	0.042	0.016	0.597	0.021	0.000	0.003	0.997	0.000
		0.75	Naive	0.276	0.016	0.176	0.014	0.415	0.022	0.000	0.013	0.987	0.000
			Corrected	0.021	0.027	0.047	0.017	0.618	0.022	0.000	0.003	0.997	0.000
			True	0.009	0.015	0.030	0.015	0.467	0.019	0.000	0.003	0.997	0.000
AH	$(400, 10)$	0.15	Naive	0.263	0.013	0.159	0.013	0.631	0.018	0.000	0.010	0.990	0.000
			Corrected	0.010	0.022	0.044	0.018	0.742	0.018	0.000	0.003	0.997	0.000
		0.50	Naive	0.261	0.014	0.168	0.015	0.594	0.022	0.000	0.012	0.988	0.000
			Corrected	0.014	0.023	0.046	0.020	0.704	0.019	0.000	0.003	0.997	0.000
		0.75	Naive	0.281	0.016	0.172	0.016	0.652	0.019	0.000	0.011	0.989	0.000
			Corrected	0.015	0.025	0.053	0.023	0.779	0.015	0.000	0.003	0.997	0.000
			True	0.010	0.014	0.034	0.014	0.685	0.018	0.000	0.002	0.998	0.000

Table 4.4: Simulation results for the estimators of B wiht known L and $p \gg n$

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(200, 1000)$	0.15	Naive	0.302	0.013	0.182	0.022	0.949	0.019	0.000	0.012	0.988	0.000
			Corrected	0.026	0.026	0.063	0.029	0.940	0.018	0.000	0.004	0.996	0.000
		0.50	Naive	0.305	0.016	0.191	0.021	0.980	0.016	0.000	0.014	0.986	0.000
			Corrected	0.028	0.029	0.077	0.027	0.975	0.018	0.000	0.005	0.995	0.000
		0.75	Naive	0.308	0.019	0.194	0.024	0.978	0.016	0.000	0.012	0.988	0.000
			Corrected	0.033	0.029	0.079	0.031	0.979	0.016	0.000	0.005	0.994	0.001
			True	0.020	0.017	0.052	0.016	0.942	0.015	0.000	0.003	0.997	0.000
PO	$(200, 1000)$	0.15	Naive	0.297	0.019	0.188	0.012	0.937	0.020	0.000	0.013	0.987	0.000
			Corrected	0.016	0.021	0.059	0.023	0.876	0.019	0.000	0.003	0.997	0.000
		0.50	Naive	0.300	0.019	0.192	0.012	0.993	0.021	0.000	0.014	0.986	0.000
			Corrected	0.024	0.023	0.066	0.029	0.932	0.019	0.000	0.003	0.997	0.000
		0.75	Naive	0.307	0.023	0.199	0.017	1.000	0.023	0.000	0.013	0.987	0.000
			Corrected	0.040	0.028	0.074	0.027	0.996	0.017	0.000	0.004	0.996	0.000
			True	0.025	0.020	0.047	0.019	0.938	0.020	0.000	0.003	0.997	0.000
AH	$(200, 1000)$	0.15	Naive	0.300	0.010	0.185	0.010	1.000	0.016	0.000	0.010	0.990	0.000
			Corrected	0.028	0.016	0.050	0.020	0.997	0.015	0.000	0.003	0.997	0.000
		0.50	Naive	0.313	0.016	0.186	0.010	0.993	0.021	0.000	0.012	0.988	0.000
			Corrected	0.024	0.021	0.059	0.021	0.987	0.017	0.000	0.003	0.997	0.000
		0.75	Naive	0.327	0.020	0.191	0.013	0.937	0.019	0.000	0.014	0.986	0.000
			Corrected	0.032	0.027	0.059	0.028	0.974	0.016	0.000	0.005	0.995	0.000
			True	0.025	0.012	0.044	0.011	0.959	0.016	0.000	0.002	0.998	0.000

Table 4.5: Simulation results for the estimators of B with repeated measurements and $p \gg n$

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(300, 1000)$	0.15	Naive	0.283	0.017	0.173	0.020	0.866	0.019	0.000	0.008	0.990	0.000
			Corrected	0.028	0.018	0.061	0.022	0.999	0.023	0.000	0.003	0.997	0.000
		0.50	Naive	0.297	0.019	0.183	0.023	0.955	0.022	0.000	0.010	0.990	0.000
			Corrected	0.027	0.025	0.067	0.029	1.000	0.021	0.000	0.004	0.996	0.000
		0.75	Naive	0.310	0.018	0.188	0.018	0.990	0.017	0.000	0.013	0.987	0.000
			Corrected	0.028	0.026	0.065	0.027	1.001	0.021	0.000	0.003	0.997	0.000
			True	0.014	0.018	0.049	0.021	0.942	0.018	0.000	0.003	0.997	0.000
PO	$(300, 1000)$	0.15	Naive	0.279	0.014	0.178	0.013	0.956	0.018	0.000	0.013	0.987	0.000
			Corrected	0.032	0.023	0.056	0.022	1.000	0.023	0.000	0.005	0.995	0.000
		0.50	Naive	0.288	0.015	0.185	0.014	0.999	0.016	0.000	0.010	0.990	0.000
			Corrected	0.038	0.024	0.056	0.023	1.000	0.021	0.000	0.004	0.996	0.000
		0.75	Naive	0.288	0.015	0.188	0.016	0.978	0.016	0.000	0.009	0.991	0.000
			Corrected	0.037	0.025	0.062	0.027	1.000	0.021	0.000	0.003	0.997	0.000
			True	0.026	0.014	0.044	0.014	0.973	0.016	0.000	0.002	0.998	0.000
AH	$(300, 1000)$	0.15	Naive	0.288	0.015	0.185	0.013	0.975	0.016	0.000	0.008	0.992	0.000
			Corrected	0.025	0.017	0.049	0.021	0.977	0.016	0.000	0.002	0.998	0.000
		0.50	Naive	0.290	0.016	0.186	0.017	0.991	0.020	0.000	0.012	0.988	0.000
			Corrected	0.027	0.017	0.054	0.023	1.001	0.021	0.000	0.003	0.997	0.000
		0.75	Naive	0.307	0.018	0.184	0.018	0.987	0.021	0.000	0.010	0.990	0.000
			Corrected	0.028	0.020	0.055	0.025	1.000	0.021	0.000	0.002	0.998	0.000
			True	0.022	0.017	0.040	0.015	0.955	0.018	0.000	0.002	0.998	0.000

Table 4.6: Simulation results for the estimators of B with validation data and $p \gg n$

Model	(n, p)	σ_ϵ	Method	Estimator of B		Estimator of $F(\cdot)$		Estimator of h		Estimator of d			
				$\|\Delta_B\|$	S.E.	$MISE(\widehat{F})$	S.E.	$\hat{h}$	S.D.	$\widehat{d}=0$	$\widehat{d}=1$	$\widehat{d}=2$	$\widehat{d}\geq 3$
PH	$(400, 1000)$	0.15	Naive	0.281	0.017	0.167	0.017	0.937	0.018	0.000	0.010	0.990	0.000
			Corrected	0.018	0.018	0.059	0.021	0.800	0.020	0.000	0.003	0.997	0.000
		0.50	Naive	0.295	0.018	0.176	0.017	0.959	0.017	0.000	0.009	0.991	0.000
			Corrected	0.020	0.021	0.060	0.025	0.912	0.021	0.000	0.004	0.996	0.000
		0.75	Naive	0.293	0.018	0.181	0.019	0.990	0.017	0.000	0.012	0.988	0.000
			Corrected	0.023	0.025	0.063	0.028	0.973	0.016	0.000	0.004	0.996	0.000
			True	0.004	0.017	0.044	0.017	0.917	0.020	0.000	0.003	0.997	0.000
PO	$(400, 1000)$	0.15	Naive	0.268	0.019	0.170	0.014	0.954	0.018	0.000	0.013	0.987	0.000
			Corrected	0.028	0.023	0.051	0.023	0.839	0.018	0.000	0.004	0.996	0.000
		0.50	Naive	0.279	0.019	0.178	0.016	0.955	0.018	0.000	0.013	0.987	0.000
			Corrected	0.030	0.023	0.052	0.024	0.916	0.020	0.000	0.005	0.995	0.000
		0.75	Naive	0.281	0.020	0.183	0.018	0.976	0.016	0.000	0.014	0.986	0.000
			Corrected	0.032	0.027	0.060	0.031	0.979	0.016	0.000	0.005	0.995	0.000
			True	0.020	0.018	0.041	0.014	0.991	0.018	0.000	0.004	0.996	0.000
AH	$(400, 1000)$	0.15	Naive	0.279	0.014	0.170	0.016	0.964	0.019	0.000	0.014	0.986	0.000
			Corrected	0.020	0.020	0.042	0.027	0.800	0.023	0.000	0.005	0.995	0.000
		0.50	Naive	0.288	0.015	0.183	0.018	1.010	0.023	0.000	0.014	0.986	0.000
			Corrected	0.023	0.021	0.044	0.026	0.899	0.025	0.000	0.006	0.994	0.000
		0.75	Naive	0.290	0.018	0.142	0.022	0.978	0.026	0.000	0.020	0.980	0.000
			Corrected	0.026	0.023	0.045	0.029	0.978	0.026	0.000	0.006	0.994	0.000
			True	0.019	0.014	0.038	0.018	0.958	0.018	0.000	0.005	0.995	0.000

Chapter 5

Semiparametric Methods for Left-Truncated and Right-Censored Survival Data with Covariate Measurement Error

5.1 Notation and Model

For an individual in the target disease population, let ξ be the calendar time of the recruitment (e.g., the recruitment starts right at the hospital discharge) and let u and r denote the calendar time of the initiating event (e.g., hospital admission) and the failure event (e.g., death), respectively, where $u < r$, and $u < \xi < r$. Let $\widetilde{T} = r - u$ be the failure time (e.g., the time length between the hospital admission and the failure), $\widetilde{A} = \xi - u$ be the truncation time (e.g., the time length between the hospital admission and the hospital discharge). Let $\widetilde{X}$ and $\widetilde{Z}$ be the associated covariates of dimensions $p \times 1$ and $q \times 1$, respectively, and write $\widetilde{V} = (\widetilde{X}^\top, \widetilde{Z}^\top)^\top$. Let $h(a)$ be the probability density function of $\widetilde{A}$ which is unknown, and $H(a) = \int_0^a h(u)du$ be the corresponding distribution function. Let $f(t)$ and $S(t)$ be the density function and the survivor function of the failure time $\widetilde{T}$, respectively.

Consistent with the notation considered by Wu et al. (2018) and Chen (2019a), for an individual with $\widetilde{T} \geq \widetilde{A}$, we let (A, T, V) denote $\left(\widetilde{A}, \widetilde{T}, \widetilde{V}\right)$ to indicate such an individual is eligible for the recruitment so that measuring (A, T, V) is possible. If $\widetilde{T} < \widetilde{A}$, then such an

individual is not included in the study to contribute any information. We define C as the censoring time for a recruited subject. Let $Y = \min\{T, A + C\}$ be the observed time and let $\Delta = I(T \leq A + C)$ be the indicator of a failure event. Figure 1.1 gives an illustration of the relationship among those variables.

5.1.1 Cox Model and Inference

Suppose we have a sample of n subjects where for $i = 1, \cdots, n$, $(Y_i, A_i, \Delta_i, V_i)$ has the same distribution as (Y, A, Δ, V), and $(y_i, a_i, \delta_i, v_i)$ represents realizations of $(Y_i, A_i, \Delta_i, V_i)$. Consider the Cox model for survival times $\widetilde{T}$ with the hazard function

$$\lambda(t|v_i) = \lambda_0(t) \exp(v_i^\top \beta),$$

where $\lambda_0(\cdot)$ is the unknown baseline hazards function, and β is the vector of parameters of primary interest.

Let

$$L_C = \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i} S(y_i|v_i)^{1-\delta_i}}{S(a_i|v_i)} \tag{5.1}$$

be the conditional likelihood of Y_i, given $V_i = v_i$ and $A_i = a_i$, and let

$$L_M = \prod_{i=1}^{n} \frac{S(a_i|v_i) dH(a_i)}{\int_0^\infty S(\alpha|v_i) dH(\alpha)} \tag{5.2}$$

be the marginal likelihood of A_i, given $V_i = v_i$, where $S(t|v_i) = \exp\left\{-\Lambda_0(t) \exp\left(v_i^\top \beta\right)\right\}$, and $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is the cumulative baseline hazards function.

Inference about β is then carried out by maximizing the likelihood function

$$L \propto L_C \times L_M = \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i} S(y_i|v_i)^{1-\delta_i} dH(a_i)}{\int_0^\infty S(\alpha|v_i) dH(\alpha)} \tag{5.3}$$

with respect to the model parameters.

5.1.2 Measurement Error Model

In practice, covariates are often subject to measurement error. For $i = 1, \cdots, n$, suppose that X_i is measured with error with an observed value or surrogate X_i^*, and that Z_i is precisely observed. We first consider the classical additive measurement error model

$$X_i^* = X_i + \epsilon_i, \tag{5.4}$$

where ϵ_i is independent of $\{X_i, Z_i, C_i, A_i, T_i\}$, and $\epsilon_i \sim N(0, \Sigma_\epsilon)$ with covariance matrix Σ_ϵ. Thus, the moment generation function of ϵ_i is given by $m(t) = \exp(\frac{1}{2}t^\top \Sigma_\epsilon t)$, and

$$E\left\{\exp\left(t^\top X_i^*\right)\right\} = m(t)\exp\left(t^\top X_i\right).$$

Model (5.4) has been widely used in the literature (e.g., Carroll et al. 2006; Yi 2017).

In contrast to measurement error model (5.4), we also consider a more flexible model where X_i^* and X_i are characterized by

$$X_i^* = X_i + \Sigma_\epsilon \alpha + \epsilon_i, \tag{5.5}$$

where ϵ_i is characterized in (5.4). Model (5.5) describes a situation where X_i^* and X_i are different not only by a *random* amount ϵ_i but also systematically by a *fixed* amount indicated by $\Sigma_\epsilon \alpha$. Different values of α show various degrees of systematic differences between X_i and X_i^*. When $\alpha = 0$, (5.5) recovers (5.4). Thus, model (5.5) covers a broader class of settings than (5.4) does and also embraces (5.4) as a special case.

In the following two sections, we develop estimation methods with measurement error effects accounted for, where the measurement error model is given by (5.5). To highlight the idea, we assume the parameters in (5.5) are known. Let $W_i^* = X_i^* - \Sigma_\epsilon \alpha$. Then

$$E(W_i^*|X_i) = E(X_i^* - \Sigma_\epsilon \alpha | X_i) = X_i \tag{5.6}$$

and

$$E\left\{\exp\left(t^\top W_i^* - \frac{1}{2}t^\top \Sigma_\epsilon t\right)\middle| X_i\right\} = \exp\left(t^\top X_i\right). \tag{5.7}$$

5.2 Conditional Profile-Likelihood Method

5.2.1 Estimation Method

We begin with a simple perspective by examining the conditional likelihood L_C, determined by (5.1), which allows us to ignore modeling of the truncation times. Let $\ell_C = \log L_C$.

Since ℓ_C contains the X_i whose measurements are unavailable, we want to modify ℓ_C to be a new function, say ℓ_C^*, of the observed measurements and the model parameters so that its conditional expectation equals to ℓ_C:

$$E(\ell_C^*|\mathbb{X},\mathbb{Z},\mathbb{C},\mathbb{A},\mathbb{T}) = \ell_C, \tag{5.8}$$

where the expectation is taken with respective to the conditional distribution of $\mathbb{W}$ given $\{\mathbb{X},\mathbb{Z},\mathbb{C},\mathbb{A},\mathbb{T}\}$, where $\mathbb{X} = \{X_1,\cdots,X_n\}$, $\mathbb{Z} = \{Z_1,\cdots,Z_n\}$, $\mathbb{C} = \{C_1,\cdots,C_n\}$, $\mathbb{A} = \{A_1,\cdots,A_n\}$, $\mathbb{T} = \{T_1,\cdots,T_n\}$, and $\mathbb{W} = \{X_1^*,\cdots,X_n^*\}$. Such a strategy is useful in yielding an unbiased estimating function and is sometimes called the "corrected" likelihood method or the insertion correction approach (e.g., Nakamura 1992; Yi and Lawless 2007; Yi 2017, Chapter 2).

Noticing that the X_i appear in ℓ_C in linear and exponential forms, we define

$$\begin{aligned}\ell_C^* &= \sum_{i=1}^{n}\Big[\delta_i \log\lambda_0(y_i) + \delta_i(w_i^{*\top}\beta_x + z_i^\top\beta_z) \\ &\quad - \{\Lambda_0(y_i) - \Lambda_0(a_i)\}\exp(w_i^{*\top}\beta_x + z_i^\top\beta_z)\{m(\beta_x)\}^{-1}\Big],\end{aligned} \tag{5.9}$$

where w_i^* and z_i represent realizations of W_i^* and Z_i, respectively. It is easily seen that ℓ_C^* satisfies (5.8).

To use (5.9) to derive an estimator of (β_x,β_z), we need to deal with the baseline hazard function $\lambda_0(\cdot)$ and its cumulative function $\Lambda_0(\cdot)$. We discretize $\Lambda_0(\cdot)$ so that $\lambda_0(\cdot)$ has a nonzero value if $t = y_i$ for $i = 1,...,n$; otherwise, $\lambda_0(t) = 0$. Let λ_i denote $\lambda_0(y_i)$ for $i = 1,...,n$. Then $\Lambda_0(t)$ is taken as $\sum_{i=1}^{n} I(y_i \leqslant t)\lambda_i$. Given β_x and β_z, we solve $\frac{\partial\ell_C^*}{\partial\lambda_i} = 0$ for $i = 1,...,n$, which leads to an estimator of λ_i, given by

$$\widehat{\lambda}_i = \frac{\delta_i}{\sum_{k=1}^{n} I(a_k \le y_i \le y_k)\exp\left(w_k^{*\top}\beta_x + z_k^\top\beta_z\right)\{m(\beta_x)\}^{-1}} \quad \text{for } i = 1,...,n; \tag{5.10}$$

and the corresponding estimate of the cumulative baseline hazards function:

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n} I(y_i \le t)\widehat{\lambda}_i\ . \tag{5.11}$$

Plugging (5.10) and (5.11) into (5.9) gives the function

$$\begin{aligned}\widehat{\ell}_C^* &= \sum_{i=1}^{n}\Big[\delta_i \log\widehat{\lambda}_i + \delta_i(w_i^{*\top}\beta_x + z_i^\top\beta_z) \\ &\quad - \Big\{\widehat{\Lambda}_0(y_i) - \widehat{\Lambda}_0(a_i)\Big\}\exp(w_i^{*\top}\beta_x + z_i^\top\beta_z)\{m(\beta_x)\}^{-1}\Big].\end{aligned} \tag{5.12}$$

An estimator of β, called the conditional estimator of β, is then obtained by maximizing $\widehat{\ell}_C^*$:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\, \widehat{\ell}_C^*. \tag{5.13}$$

5.2.2 Asymptotic Results

Let $\beta_0 = (\beta_{x0}^\top, \beta_{z0}^\top)^\top$ denote the true value of β and let Θ denote the parameter space of β. Consistent with others such as Huang et al. (2012), we assume that $\widetilde{T}_i$ has a finite maximal support τ, where $\tau = \sup\left\{t : P(\widetilde{T}_i \le t) < 1\right\} < \infty$, implying that τ is also a maximal support of truncation time. Let $N_i(t) = \Delta_i I(Y_i \le t)$ be the counting process of the observed failure events for subject i. Let $V_i^* = \left(W_i^{*\top}, Z_i^\top\right)^\top$. Define $S^{(k)}(u,\beta) = n^{-1}\sum_{i=1}^n v_i^{*\otimes k}\exp(v_i^{*\top}\beta)I(a_i \le u \le y_i)$ for $k = 0, 1, 2$, where $a^{\otimes 2}$ means $aa^\top$ for the column vector a . Let $\mathcal{S}^{(k)}(u,\beta) = E\left[V_i^{*\otimes k}\exp(V_i^{*\top}\beta)I(A_i \le u \le Y_i)\right]$ be the expectation of $S^{(k)}(u,\beta)$. Using these symbols, we express (5.11) as

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n \exp(v_i^{*\top}\beta)I(a_i \le u \le y_i)\left\{m(\beta_x)\right\}^{-1}}. \tag{5.14}$$

The following theorems, whose proofs are included in Appendix D.3, establish the asymptotic properties of $\widehat{\Lambda}_0(t)$ and $\widehat{\beta}$.

Theorem 5.2.1 *Under regularity conditions in Appendix D.1, we have that as $n \to \infty$,*

$$\sup_{\beta\in\Theta, t\in[0,\tau]} |\widehat{\Lambda}_0(t) - \Lambda_0(t)| \xrightarrow{a.s.} 0,$$

where $\Lambda_0(t) = \int_0^t \left\{\mathcal{S}^{(0)}(u,\beta)\right\}^{-1} m(\beta_x)dP(\Delta_i = 1, Y_i \le u)$.

Theorem 5.2.2 *Under regularity conditions given in Appendix D.1, the estimator $\widehat{\beta}$ obtained from* (5.13) *has the following asymptotic properties:*

(1) $\widehat{\beta} \xrightarrow{p} \beta_0$ as $n \to \infty$;

(2) $\sqrt{n}\left(\widehat{\beta}-\beta_0\right) \xrightarrow{d} N(0, \mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1})$ *as* $n \to \infty$,

where

$$\mathcal{A}_P = \int_0^\tau \left[\left\{\frac{\mathcal{S}^{(2)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} - \left(\frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right)^{\otimes 2}\right\} - \begin{pmatrix} \Sigma_\epsilon & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q} \end{pmatrix}\right] dE\left\{N_i(u)\right\}, \quad (5.15)$$

$$\begin{aligned}\mathcal{B}_P = {} & E\left[\int_0^\tau \left\{\left(V_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right) + \begin{pmatrix} \Sigma_\epsilon\beta_{x0} \\ \mathbf{0}_q \end{pmatrix}\right\} dN_i(u) \right. \\ & \left. - \int_0^\tau \frac{\exp\left(V_i^{*\top}\beta_0\right) I(A_i \le u \le Y_i)}{\mathcal{S}^{(0)}(u,\beta_0)} \left\{V_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right\} dE\left\{N_i(u)\right\}\right]^{\otimes 2},\end{aligned}$$

$\mathbf{0}_{p\times q}$ *represents a* $p \times q$ *matrix with all entries 0, and* $\mathbf{0}_p$ *stands for a* $p \times 1$ *vector with all entries 0.*

5.3 Augmented Pseudo-Likelihood Method

Estimator $\widehat{\beta}$ obtained by (5.13) can be inefficient since it uses only the conditional likelihood L_C with the marginal likelihood L_M ignored, as shown by the likelihood (5.3) formulated in Section 5.1.1. Now we develop an augmented estimator to improve the efficiency of $\widehat{\beta}$ given by (5.13). The basic idea, driven by the form of the likelihood (5.3), is to include the marginal likelihood L_M for the truncation times in the estimation procedure.

5.3.1 Estimation Method

In addition to containing the distribution function $H(\cdot)$ of $\widetilde{A}$, the marginal likelihood L_M in (5.2) involves the unobserved covariate X_i. We first construct a modified version of L_M to address the measurement error effects.

Let μ_X and Σ_X be the mean vector and variance-covariance matrix of X_i, respectively. Let $W_i^* = X_i^* - \Sigma_\epsilon\alpha$ as in (5.6), then model (5.5) gives that $W_i^* = X_i + \epsilon_i$ with $\epsilon_i \sim N\left(0, \Sigma_\epsilon\right)$, yielding that

$$E(X_i|W_i^* = w_i^*) = \mu_X + \left(\Sigma_{W^*} - \Sigma_\epsilon\right)^\top \Sigma_{W^*}^{-1}(w_i^* - \mu_{W^*}), \quad (5.16)$$

where μ_{W^*} and Σ_{W^*} represent the mean and covariance matrix of W_i^*, respectively; we let $\widetilde{x}_{\mathrm{RC},i}$ denote (5.16) for ease of notation. Using the method of moments, (5.16) is estimated by

$$\widehat{x}_i = \widehat{\mu}_{W^*} + \left(\widehat{\Sigma}_{W^*} - \Sigma_\epsilon\right)^\top \widehat{\Sigma}_{W^*}^{-1}(w_i^* - \widehat{\mu}_{W^*}) \tag{5.17}$$

with $\widehat{\mu}_{W^*} = \frac{1}{n}\sum_{i=1}^n w_i^*$ and $\widehat{\Sigma}_{W^*} = \frac{1}{n-1}\sum_{i=1}^n (w_i^* - \widehat{\mu}_{W^*})(w_i^* - \widehat{\mu}_{W^*})^\top$.

As a result, replacing $v_i = \left(x_i^\top, z_i^\top\right)^\top$ with $\left(\widehat{x}_i^\top, z_i^\top\right)^\top$ in likelihood function (5.2) gives

$$L_M^* = \prod_{i=1}^n \frac{S(a_i|\widehat{x}_i, z_i)dH(a_i)}{\int_0^\infty S(\alpha|\widehat{x}_i, z_i)dH(\alpha)}, \tag{5.18}$$

where $S(a_i|\widehat{x}_i, z_i) = \exp\left\{-\Lambda_0(a_i)\exp\left(\widehat{x}_i^\top\beta_x + z_i^\top\beta_z\right)\right\}$.

To use (5.18) for inference about β, we next estimate the distribution function $H(\cdot)$. Directly applying the kernel estimation (Silverman 1978) to the observed truncation times to estimate $dH(\cdot)$ is not suitable way since the observed truncation times form a biased sample. Instead, we use the nonparametric maximum likelihood estimator (NPMLE) (e.g., Wang 1991) to estimate the distribution function of $\widetilde{A}$. For a fixed parameter β, the NPMLE of $H(a)$ in (5.18) is given by

$$\widehat{H}(a) = \left(\sum_{i=1}^n \frac{1}{\widehat{S}(a_i|\widehat{x}_i, z_i)}\right)^{-1} \sum_{i=1}^n \frac{I(a_i \le a)}{\widehat{S}(a_i|\widehat{x}_i, z_i)}, \tag{5.19}$$

where $\widehat{S}(a_i|\widehat{x}_i, z_i) = \exp\left\{-\widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_i^\top\widehat{\beta}_x + z_i^\top\widehat{\beta}_z\right)\right\}$, and $\widehat{\Lambda}_0(\cdot)$ and $\widehat{\beta}$ are consistent estimators of $\Lambda_0(\cdot)$ and β, respectively, proposed in Section 5.2.

Then replacing $H(a)$ by $\widehat{H}(a)$ in (5.18) gives $\widehat{L}_M^*$; let $\widehat{\ell}_M^* = \log \widehat{L}_M^*$, which is given by

$$\begin{aligned}\widehat{\ell}_M^* &= \sum_{i=1}^n \log\left\{d\widehat{H}(a_i)\right\} - \sum_{i=1}^n \widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_i^\top\beta_x + z_i^\top\beta_z\right) \\ &- \sum_{i=1}^n \log\left[\int_0^\infty \exp\left\{-\widehat{\Lambda}_0(\alpha)\exp\left(\widehat{x}_i^\top\beta_x + z_i^\top\beta_z\right)\right\}d\widehat{H}(\alpha)\right].\end{aligned} \tag{5.20}$$

Finally, we consider the pseudo-likelihood function

$$\widehat{\ell}^* = \widehat{\ell}_C^* + \widehat{\ell}_M^*; \tag{5.21}$$

maximizing $\widehat{\ell}^*$ with respect to β gives an estimator of β:

$$\widetilde{\beta} = \underset{\beta}{\text{argmax}}(\widehat{\ell}_C^* + \widehat{\ell}_M^*), \tag{5.22}$$

which is called a pseudo-likelihood estimator of β.

5.3.2 Asymptotic Results

Let

$$\mu\left(\widetilde{x}_{\text{RC},i}, z_i\right) = \int_0^\tau \exp\left\{-\Lambda_0(u)\exp(\widetilde{x}_{\text{RC},i}^\top \beta_{x0} + z_i^\top \beta_{z0})\right\} dH(u). \tag{5.23}$$

Let $\mathcal{N}(t) = P(\Delta_i = 1, Y_i \le t)$, $S(\xi|\widetilde{x}_{\text{RC}}, z) = \exp\left\{-\Lambda_0(\xi)\exp(\widetilde{x}_{\text{RC}}^\top \beta_{x0} + z^\top \beta_{z0})\right\}$, and

$$\begin{aligned}
\psi_i(\beta_0|\widetilde{x}_{\text{RC}}, z) \;=\; & \int_0^\tau \int_0^\tau S(\xi|\widetilde{x}_{\text{RC}}, z) \left\{ \frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} \right. \\
& \left. - \frac{d\mathcal{N}(u)\exp\left(w_i^{*\top}\beta_{x0} + z_i^\top \beta_{z0}\right) I(a_i \le u \le y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2} \right\} m(\beta_{x0}) \\
& \times \exp\left(\widetilde{x}_{\text{RC}}^\top \beta_{x0} + z^\top \beta_{z0}\right) dH(\xi) + o_p(1).
\end{aligned} \tag{5.24}$$

Let $G(a, \widehat{v})$ denote the joint distribution of A_i and $\widehat{V}_i$ where $\widehat{V}_i = \left(\widetilde{X}_{\text{RC},i}^\top, Z_i^\top\right)^\top$. Define

$$\begin{aligned}
& \Psi\left(x_i^*, \widetilde{x}_{\text{RC},i}, z_i, a_i, y_i\right) \\
= \; & \int_0^\tau \left\{ v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} + \begin{pmatrix} \Sigma_\epsilon \beta_{x0} \\ \mathbf{0}_q \end{pmatrix} \right\} dN_i(u) \\
& - \int_0^\tau \frac{\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\mathcal{S}^{(0)}(u,\beta_0)} \left(v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} \right) dE\left\{N_i(u)\right\} \\
& - \left[\int_{-\infty}^{\infty} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u)\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\mathcal{S}^{(0)}(u,\beta_0)^2} \right\} m(\beta_{x0}) \right. \\
& \left. \times \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau) \right] dG(a, \widehat{v}) \\
& + \left[\int_{-\infty}^{\infty} \int_0^\tau \left\{ \frac{1}{\mu\left(\widetilde{x}_{\text{RC}}, z\right)} \frac{\partial}{\partial \beta} \psi_i(\beta_0|\widetilde{x}_{\text{RC}}, z) \right. \right. \\
& \left. \left. - \frac{\partial \mu\left(\widetilde{x}_{\text{RC}}, z\right)}{\partial \beta} \frac{1}{\mu^2\left(\widetilde{x}_{\text{RC}}, z\right)} \psi_i(\beta_0|\widetilde{x}_{\text{RC}}, z) \right\} dG(a, \widehat{v}) \right] \\
& - \frac{\partial}{\partial \beta} \Lambda_0(a_i) \exp\left(\widehat{v}_i^\top \beta_0\right) - \frac{1}{\mu\left(\widetilde{x}_{\text{RC},i}, z_i\right)} \frac{\partial}{\partial \beta} \mu\left(\widetilde{x}_{\text{RC},i}, z_i\right),
\end{aligned} \tag{5.25}$$

and

$$\begin{aligned}\mathcal{A}_M &= E\left[\frac{\partial^2}{\partial\beta\partial\beta^\top}\Lambda_0(A_i)\exp\left(\widehat{V}_i^\top\beta_0\right)\right. \\ &\quad + \left\{\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)\right\}^{-2}\left\{\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)\frac{\partial\left\{\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)\right\}^2}{\partial\beta\partial\beta^\top}\right. \\ &\quad \left.\left.- \left(\frac{\partial\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)}{\partial\beta}\right)^{\otimes 2}\right\}\right]. \end{aligned} \tag{5.26}$$

The following theorem shows the asymptotic results of $\widetilde{\beta}$; the proof is placed in Appendix D.4.1.

Theorem 5.3.1 *Under regularity conditions given in Appendix D.1, estimator $\widetilde{\beta}$ obtained from* (5.22) *has the following properties:*

(1) $\widetilde{\beta} \xrightarrow{p} \beta_0$ *as* $n \to \infty$*;*

(2) $\sqrt{n}\left(\widetilde{\beta} - \beta_0\right) \xrightarrow{d} N(0, \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1})$ *as* $n \to \infty$*;*

where $\mathcal{B} = E(\Psi_i^{\otimes 2})$ *with* $\Psi_i = \Psi\left(X_i^*, \widetilde{X}_{RC,i}, Z_i, A_i, Y_i\right)$*, and* $\mathcal{A} = \mathcal{A}_P + \mathcal{A}_M$ *with* $\mathcal{A}_P$ *and* $\mathcal{A}_M$ *determined by (5.15) and (5.26), respectively.*

The following theorem compares the efficiency between the estimators $\widehat{\beta}$ and $\widetilde{\beta}$ whose proof is given in Appendix D.4.2.

Theorem 5.3.2 *Under regularity conditions given in Appendix D.1, the estimator $\widetilde{\beta}$ obtained from* (5.22) *is more efficient than the estimator $\widehat{\beta}$ determined by* (5.13)*. That is, var$\left(\widehat{\beta}\right)$ − var$\left(\widetilde{\beta}\right)$ is a positive definite matrix.*

5.4 Inference with Main/Validation Data

5.4.1 Estimation of Parameters for Measurement Error Model

In practice, the covariance matrix Σ_ϵ and parameter α for the measurement error model (5.5) are often unknown, and they need to be estimated from additional data sources. First, we comment that model (5.5) and model (5.4) have different requirements of data sources in order to estimate associated model parameters. The availability of additional data for estimation of parameters associated with (5.4) does not necessarily ensure estimability of parameters for model (5.5). To see this, suppose we have replicates of X_i with the W_{ij} being the n_i repeated measurements for $j = 1, \cdots, n_i$ and $i = 1, \cdots, n$. Such data are sufficient for estimation covariance matrix Σ_ϵ if they follow model (5.4):

$$W_{ij} = X_i + \epsilon_{ij},$$

where the ϵ_{ij} are independent of $\{X_i, Z_i, C_i, A_i, T_i\}$ and follow a distribution with mean zero and covariance matrix Σ_ϵ. Using the method of moments, we estimate Σ_ϵ by

$$\widehat{\Sigma}_\epsilon = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(W_{ij} - \bar{W}_{i\cdot}\right) \left(W_{ij} - \bar{W}_{i\cdot}\right)^\top}{\sum_{i=1}^{n} (n_i - 1)}, \tag{5.27}$$

where $\bar{W}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} W_{ij}$. However, if the replicates W_{ij} follow model (5.5) instead, i.e.,

$$W_{ij} = X_i + \Sigma_\epsilon \alpha + \epsilon_{ij},$$

then covariance Σ_ϵ can still be estimated by (5.27), but parameter α is not estimable using the replicate W_{ij}.

To estimate the parameters in model (5.5), we assume the availability of a validation sample. Let $\mathcal{M}$ and $\mathcal{V}$ denote the subject sets for the main study and the external validation study containing n and m subjects, respectively, where $\mathcal{M}$ and $\mathcal{V}$ do not overlap. That is, the available data contain measurements $\{(y_i, a_i, \delta_i, x_i^*, z_i) : i \in \mathcal{M}\}$ from the main study and $\{(x_i^*, z_i, x_i) : i \in \mathcal{V}\}$ from the validation sample. Hence, for the measurement error model, we have

$$X_i^* = X_i + \Sigma_\epsilon \alpha + \epsilon_i$$

for $i \in \mathcal{M} \cup \mathcal{V}$, where the ϵ_i are independent and identically distributed with mean zero and unknown covariance matrix Σ_ϵ, and are independent of $\{X_i, Z_i, C_i, A_i, T_i\}$. We assume that $\lim_{n\to\infty} \frac{m}{n}$ exists and is greater than 0, and let ρ denote this limit.

Estimation of α and Σ_ϵ can be carried out using the least square regression method. Write $\gamma = \Sigma_\epsilon \alpha$ and define

$$Q(\gamma) = \sum_{i\in\mathcal{V}} \|X_i^* - X_i - \gamma\|_2^2 \tag{5.28}$$

where $\|v\|_2^2 = v^\top v$ for a column vector v. Then solving

$$\frac{\partial Q}{\partial \gamma} = 0$$

for γ yields

$$\widehat{\gamma} = \frac{1}{m}\left(\sum_{i\in\mathcal{V}} X_i^* - \sum_{i\in\mathcal{V}} X_i\right). \tag{5.29}$$

For $i \in \mathcal{V}$, let $e_i = X_i^* - X_i - \widehat{\gamma}$ be the residual. Since $E\left(e_i^\top e_i\right) = \frac{m-1}{m}\Sigma_\epsilon$ for $i \in \mathcal{V}$, we obtain that $E\left(\sum_{i\in\mathcal{V}} e_i^\top e_i\right) = (m-1)\,\Sigma_\epsilon$, which yields the unbiased estimator of Σ_ϵ:

$$\widehat{\Sigma}_\epsilon = \frac{1}{m-1}\sum_{i\in\mathcal{V}} e_i^\top e_i. \tag{5.30}$$

Finally, since $\alpha = \Sigma_\epsilon^{-1}\gamma$, we obtain an estimator of α:

$$\widehat{\alpha} = \widehat{\Sigma}_\epsilon^{-1}\widehat{\gamma}.$$

5.4.2 Two-Stage Estimation of Parameter for Survival Model

To estimate the parameter β, we carry out a two-stage estimation procedure. At the first stage, we use (5.29) and (5.30) to, respectively, estimate γ and Σ_ϵ for the measurement error model, as described in Section 5.4.1. At the second stage, we estimate β using a modified version of (5.12) or (5.21), given by

$$\widehat{\ell}_{val}^* = \widehat{\ell}_{val,C}^* + \widehat{\ell}_{val,M}^*, \tag{5.31}$$

where $\widehat{\ell}^*_{val,C}$ and $\widehat{\ell}^*_{val,M}$ are, respectively, $\widehat{\ell}^*_C$ and $\widehat{\ell}^*_M$ with the parameters of the measurement error model (5.5) replaced by their estimates obtained in the first stage. That is,

$$\begin{aligned}\widehat{\ell}^*_{val,C} &= \sum_{i\in\mathcal{M}}\Big[\delta_i \log\widehat{\lambda}_i + \delta_i\left\{(x_i^* - \widehat{\gamma})^\top\beta_x + z_i^\top\beta_z\right\} \\ &\quad - \Big\{\widehat{\Lambda}_0(y_i) - \widehat{\Lambda}_0(a_i)\Big\}\exp\left\{(x_i^* - \widehat{\gamma})^\top\beta_x + z_i^\top\beta_z\right\}\left\{\widehat{m}(\beta_x)\right\}^{-1}\Big] \qquad (5.32)\end{aligned}$$

and

$$\begin{aligned}\widehat{\ell}^*_{val,M} &= \sum_{i\in\mathcal{M}}\log\Big\{d\widehat{H}_{val}(a_i)\Big\} - \sum_{i=1}^n\widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_{val,i}^\top\beta_x + z_i^\top\beta_z\right) \\ &\quad - \sum_{i\in\mathcal{M}}\log\left[\int_0^\infty\exp\Big\{-\widehat{\Lambda}_0(\alpha)\exp\left(\widehat{x}_{val,i}^\top\beta_x + z_i^\top\beta_z\right)\Big\}d\widehat{H}_{val}(\alpha)\right],\end{aligned}$$

where $\widehat{m}(\beta_x) = \exp\left(\frac{1}{2}\beta_x^\top\widehat{\Sigma}_\epsilon\beta_x\right)$, $\widehat{x}_{val,i} = \widehat{\mu}_{W^*} + \left(\widehat{\Sigma}_{W^*} - \widehat{\Sigma}_\epsilon\right)^\top\widehat{\Sigma}_{W^*}^{-1}(w_i^* - \widehat{\mu}_{W^*})$, and

$$\widehat{H}_{val}(a) = \left(\sum_{i=1}^n\frac{1}{\widehat{S}(a_i|\widehat{x}_{val,i},z_i)}\right)^{-1}\sum_{i=1}^n\frac{I(a_i\le a)}{\widehat{S}(a_i|\widehat{x}_{val,i},z_i)}.$$

By analogy to (5.13) and (5.22), two estimators of β can then be obtained by maximizing (5.32) and the pseudo-likelihood (5.31), respectively. That is,

$$\widehat{\beta}_{val} = \underset{\beta}{\text{argmax}}\,\widehat{\ell}^*_{val,C}, \qquad (5.33)$$

and

$$\widetilde{\beta}_{val} = \underset{\beta}{\text{argmax}}(\widehat{\ell}^*_{val,C} + \widehat{\ell}^*_{val,M}) \qquad (5.34)$$

are two estimators of β.

5.4.3 Asymptotic Properties

We now explore the asymptotic results for the two estimators of β described in Section 5.4.2; the proofs are placed in Appendix D.5.

Let ζ_i be the indicator whether or not subject i belongs to the validation sample $\mathcal{V}$, i.e., $\zeta_i = 1$ if $i \in \mathcal{M}$ and $\zeta_i = 0$ if $i \in \mathcal{V}$. Let

$$\begin{aligned}\Phi\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right) &= \int_0^\tau \left\{ v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} + \begin{pmatrix} \Sigma_\epsilon \beta_{x0} \\ \mathbf{0}_q \end{pmatrix} \right\} dN_i(u) \\ &\quad - \int_0^\tau \frac{\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\mathcal{S}^{(0)}(u,\beta_0)} \left(v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} \right) dE\left\{N_i(u)\right\}.\end{aligned}$$

Define

$$\begin{aligned}\mathcal{B}_{val1,i} &= \sqrt{1+\rho}\zeta_i \Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) + \frac{\sqrt{1+\rho}}{\rho} E\left\{N_i(\tau)\right\}(1-\zeta_i) \\ &\quad \times \left[\left\{ \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \frac{1}{\mathcal{S}^{(0)}(u;\beta_0)} \frac{\partial \mathcal{S}^{(1)}(u;\beta_0)}{\partial \gamma} \right\} (X_i^* - X_i) \right. \\ &\quad \left. + \frac{m\beta_x}{m-1} \left\{ \epsilon_i \epsilon_i^\top - (m-1)\Sigma_\epsilon \right\} \right]. \end{aligned} \tag{5.35}$$

Let

$$\begin{aligned}\mathcal{E}_{val,1} &= E\left[\frac{\partial}{\partial \beta} \int_0^\tau d\mathcal{N}(u) \frac{1}{\left\{\mathcal{S}^{(0)}(u;\beta_0)\right\}^2} \frac{\partial \mathcal{S}^{(0)}(u;\beta_0)}{\partial \gamma} \right. \\ &\quad \left. \times m(\beta_{x0}) \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le A_i \le \tau) \right],\end{aligned}$$

$$\begin{aligned}&\Psi_{M1}\left(x_i^*, \widetilde{X}_{\mathrm{RC},i}, z_i, y_i, a_i\right) \\ &= \frac{\partial}{\partial \beta} \left[\int_{-\infty}^{\infty} \int_0^\tau \left\{ \frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u) \exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2} \right\} m(\beta_{x0}) \right. \\ &\quad \left. \times \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau) \right] dG(a, \widehat{v}),\end{aligned}$$

and

$$\begin{aligned}\varphi_{val,i} &= \left[\frac{\sqrt{1+\rho}}{\rho} (X_i^* - X_i) \int_0^\tau \int_0^\tau - \left\{ S(\nu|\widetilde{x}_{\mathrm{RC}}, z) \frac{d\mathcal{N}(t)}{\left\{\mathcal{S}^{(0)}(t;\beta_0)\right\}^2} \frac{\partial \mathcal{S}^{(0)}(t;\beta_0)}{\partial \gamma} m(\beta_{x0}) \right. \right. \\ &\quad \left. \left. \times \exp\left(\widetilde{x}_{\mathrm{RC}}^\top \beta_{x0} + z^\top \beta_{z0}\right) dH(\nu) \right\} + \sqrt{1+\rho} \psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}}, z) \right].\end{aligned}$$

Define

$$\begin{aligned}\mathcal{B}_{val2,i} &= -\sqrt{1+\rho}\zeta_i\Psi_{M1}\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) \\ &\quad +\frac{\sqrt{1+\rho}}{\rho}\mathcal{E}_{val,1}(1-\zeta_i)(X_i^*-X_i), \end{aligned} \tag{5.36}$$

$$\mathcal{B}_{val3,i} = \left[\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu\left(\widetilde{x}_{\mathrm{RC}},z\right)}\frac{\partial}{\partial\beta}\varphi_{val,i} - \frac{\partial\mu\left(\widetilde{x}_{\mathrm{RC}},z\right)}{\partial\beta}\frac{1}{\mu^2\left(\widetilde{x}_{\mathrm{RC}},z\right)}\varphi_{val,i}\right\}\right] dG(a,\widehat{v}),$$

and

$$\widetilde{U}_{M,val,i} = -\frac{\partial}{\partial\beta}\Lambda_0(A_i)\exp\left(\widehat{V}_i^{\top}\beta_{x0}\right) - \frac{1}{\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)}\frac{\partial\mu\left(\widetilde{X}_{\mathrm{RC},i}, Z_i\right)}{\partial\beta}.$$

Theorem 5.4.1 *Under regularity conditions in Appendix D.1, we have that as* $n\to\infty$,

(1) $\widehat{\beta}_{val}\xrightarrow{p}\beta_0$;

(2) $\sqrt{n}\left(\widehat{\beta}_{val}-\beta_0\right)\xrightarrow{d}N\left(0,\mathcal{A}_{P,val}^{-1}\mathcal{B}_{P,val}\mathcal{A}_{P,val}^{-1}\right)$,

where $\mathcal{B}_{P,val}=E\left\{\left(\mathcal{B}_{val1,i}\right)^{\otimes 2}\right\}$, *and*

$$\mathcal{A}_{P,val} = \int_0^{\tau}\left[\left\{\frac{\mathcal{S}^{(2)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} - \left(\frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right)^{\otimes 2}\right\} - \begin{pmatrix}\Sigma_\epsilon & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q}\end{pmatrix}\right] dE\left\{N_i(u)\right\}. \tag{5.37}$$

Theorem 5.4.2 *Under regularity conditions in Appendix D.1, we have that as* $n\to\infty$,

(1) $\widetilde{\beta}_{val}\xrightarrow{p}\beta_0$;

(2) $\sqrt{n}\left(\widetilde{\beta}_{val}-\beta_0\right)\xrightarrow{d}N\left(0,\mathcal{A}_{val}^{-1}\mathcal{B}_{val}\mathcal{A}_{val}^{-1}\right)$,

where

$$\mathcal{B}_{val}=E\left\{\left(\mathcal{B}_{val1,i}+\mathcal{B}_{val2,i}+\mathcal{B}_{val3,i}+\sqrt{1+\rho}\zeta_i\widetilde{U}_{M,val,i}\right)^{\otimes 2}\right\}; \tag{5.38}$$

$$
\begin{aligned}
\mathcal{A}_{val} &= \int_0^\tau \left[\left\{ \frac{\mathcal{S}^{(2)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} - \left(\frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} \right)^{\otimes 2} \right\} - \begin{pmatrix} \Sigma_\epsilon & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q} \end{pmatrix} \right] dE\left\{ N_i(u) \right\} \\
&\quad + E\left[\frac{\partial^2}{\partial\beta\partial\beta^\top} \Lambda_0(A_i) \exp\left(\widehat{V}_i^\top \beta_0 \right) \right. \\
&\quad + \left\{ \mu\left(\widetilde{X}_{RC,i}, Z_i \right) \right\}^{-2} \left\{ \mu\left(\widetilde{X}_{RC,i}, Z_i \right) \frac{\partial \left\{ \mu\left(\widetilde{X}_{RC,i}, Z_i \right) \right\}^2}{\partial\beta\partial\beta^\top} \right. \\
&\quad \left. \left. - \left(\frac{\partial \mu\left(\widetilde{X}_{RC,i}, Z_i \right)}{\partial\beta} \right)^{\otimes 2} \right\} \right]. \qquad (5.39)
\end{aligned}
$$

Theorems 5.4.1 and 5.4.2 establish the asymptotic results for the two estimators $\widehat{\beta}_{val}$ and $\widetilde{\beta}_{val}$. These results offer the basis of conducting inference about β such as calculating confidence intervals or performing hypothesis testing. While both $\widehat{\beta}_{val}$ and $\widetilde{\beta}_{val}$ are consistent estimators of β, their efficiencies are different, as shown in the following theorem.

Theorem 5.4.3 *Under regularity conditions given in Appendix D.1, the estimator* $\widetilde{\beta}_{val}$ *obtained from* (5.34) *is more efficient than the estimator* $\widehat{\beta}_{val}$ *determined by* (5.33)*. That is,* $var\left(\widehat{\beta}_{val}\right) - var\left(\widetilde{\beta}_{val}\right)$ *is a positive definite matrix.*

5.5 Numerical Studies

We conduct simulation studies to assess the finite sample performance of the proposed estimators under a variety of settings.

5.5.1 Design Setup

We consider the setting where the baseline hazards function is set as $\lambda_0(t) = 2t$ and the truncation time $\widetilde{A}$ is generated from the exponential distribution with mean 10. Let $\beta_0 = (\beta_{x0}, \beta_{z0})^\top$ be the true parameters where we set $\beta_0 = (0.3, 1)^\top$. We consider a scenario where $\widetilde{V} = (\widetilde{X}, \widetilde{Z})^\top$ are generated from a bivariate normal distribution with mean zero and

variance-covariance matrix Σ, which is set as $\begin{pmatrix} 4 & 0.5 \\ 0.5 & 36 \end{pmatrix}$. Given $\lambda_0(t)$, $(\widetilde{X}, \widetilde{Z})^\top$ and β_0, the failure time $\widetilde{T}$ is generated from the model:

$$\lambda\left(\widetilde{T}|\widetilde{X}, \widetilde{Z}\right) = 2\widetilde{T}\exp\left(\widetilde{X}\beta_{x0} + \widetilde{Z}\beta_{z0}\right).$$

That is, $\widetilde{T}$ is set as $\sqrt{-\exp\left(\widetilde{X}\beta_{x0} + \widetilde{Z}\beta_{z0}\right)\log(1-U)}$, where U is simulated from the uniform distribution $U(0,1)$. For the measurement error process, we consider model (5.5) with error $\epsilon \sim N(0, \Sigma_\epsilon)$, where variance Σ_ϵ is taken as 0.01, 0.5, and 0.75, respectively, and α is set as 0 or 100. Therefore, the observed data (A, T, V) is collected from $(\widetilde{A}, \widetilde{T}, \widetilde{V})$ by conditioning on that $\widetilde{T} \geq \widetilde{A}$. We repeatedly generate data these steps we obtain a sample of a required size $n = 200$.

We consider three censoring rates, say 0%, 25%, and 50%, and let the censoring time C be generated from the uniform distribution $U(0, c)$, where c is determined by a given censoring rate. Consequently, Y and Δ are determined by $Y = \min\{T, A + C\}$ and $\Delta = I(T \leq A + C)$. 1000 simulations are run for each parameter setting. In Sections 5.5.2 and 5.5.3 we apply the estimation methods in Sections 5.2 and 5.3 with the parameters of the measurement error model (5.5) assumed known. Let $\widehat{\beta} = \left(\widehat{\beta}_x, \widehat{\beta}_z\right)$ denote the estimator derived from (5.13), and let $\widetilde{\beta} = \left(\widetilde{\beta}_x, \widetilde{\beta}_z\right)$ stand for the estimator derived from (5.22).

5.5.2 Performance of Proposed Estimators: α and Σ_ϵ are Known

We report the biases of estimates, the standard error (S.E.), and the mean squared errors (MSE) under the two measurement error models. The results for the classical measurement error model (i.e., model (5.5) with $\alpha = 0$) are reported in Table 5.1, and the results for model (5.5) with $\alpha = 100$ are displayed in Table 5.2.

First, the censoring rate and measurement degree have noticeable impact on each estimation methods. As expected, biases and variance estimates increase as the censoring rate increases. When the measurement degree increases, inference results obtained from the corrected conditional profile likelihood method and the proposed method degrade, and the impact of the measurement error degrees seems more obvious on the conditional profile likelihood approach than our proposed method.

Within a setting with a given censoring rate and a measurement error degree, the three methods perform differently. The naive method performs the worst and the proposed

method performs the best. The naive method produces considerable finite sample biases with coverage rates of 95% confidence intervals significantly departing from the nominal level. Both the conditional profile likelihood approach and the proposed method output satisfactory estimates with small finite sample biases and reasonable coverage rates of 95% confidence intervals. Compared to the variance estimates produced by the naive approach, the two methods which account for measurement error effects yield larger variance estimates, and this is the price paid to remove biases in point estimators. This phenomenon is typical in the literature of measurement error models. However, mean squared errors produced by those two methods tend to be a lot smaller than those obtained from the naive method. Finally, we see that the proposed method is more efficient than the conditional profile likelihood method, which is evident from the comparisons of variance estimates for these two methods. These results confirm theoretical result established by Theorem 5.3.2.

5.5.3 Assessment of Misspecification of Measurement Error Model

We now study the performance of our estimators when the measurement error model is misspecified. Specifically, we consider two scenarios. In Scenario 1, the measurement error model (5.4) is used to generate data, but we use model (5.5) to fit the data; in Scenario 2, we use (5.5) with $\alpha = 100$ as the measurement error model to generate data, but we use model (5.4) to fit the data. We report the average of biases, average of S.E. and mean squared errors (MSE) for estimators $\widehat{\beta}$ and $\widetilde{\beta}$, respectively, obtained from (5.13) and (5.22). The results are displayed in Table 5.3 for Scenario 1 and Table 5.4 for Scenario 2.

Shown in Table 5.3, under Scenario 1 finite sample biases are comparable to those reported in Table 5.1. In addition, S.E.s and MSEs are very close to those obtained from the situation where the fitting model is correctly used. These results are not surprising since the model we used to generate the data is nested in the model we used to fit the data. On the other hand, in Scenario 2 where the model used to fit data differs from the model for generating data, i.e., model misspecification is present, biased results are produced, which is evident from Table 5.4. This simulation study also shows that with model misspecification considered here, the proposed method performs better than the corrected conditional likelihood approach.

5.5.4 Performance with Validation Data

In this subsection, we evaluate the performance of the proposed method in Section 5.4 for situations where the main study and the validation study are available; the data from the

main study are generated as in Section 5.5.1, and the external validation data with size $|\mathcal{V}| = 100$ are also generated independently following the procedure in Section 5.5.1, where the true parameter values of the measurement error model (5.5) are $\alpha = 100$ and $\Sigma_\epsilon = 0.010$, 0.500, or 0.750, respectively, corresponding to increasing degrees of measurement error.

We first apply the estimation procedure described in Section 5.4.1 to estimate α and Σ_ϵ. Corresponding to $\Sigma_\epsilon = 0.010$, 0.500, and 0.750, we obtain estimates of Σ_ϵ: $\widehat{\Sigma}_\epsilon = 0.010$, 0.497, and 0.743, respectively, with the corresponding standard errors 0.001, 0.035 and 0.051; and the corresponding estimates of α are 100.746, 101.154 and 101.492, with the associated standard errors 7.062, 7.029, 7.030, respectively. Then we analyze the data from the main study using the estimators $\widehat{\beta}_{val}$ and $\widetilde{\beta}_{val}$ derived by (5.33) and (5.34), respectively, and present the results in Table 5.5. The results uncover similar findings to those revealed in Section 5.5.2 and demonstrate satisfactory finite sample performance of the proposed estimators $\widehat{\beta}_{val}$ and $\widetilde{\beta}_{val}$. The results also confirm that $\widetilde{\beta}_{val}$ is more efficient than $\widehat{\beta}_{val}$.

5.5.5 Analysis of Worcester Heart Attack Study

In this section, we apply the proposed methods to analyze the data arising from the Worcester Heart Attack Study (WHAS500), which are described in Section 1.6.4. Discussed by Hosmer et al. (2008), a survival time was defined as the time since a subject was admitted to the hospital. We are interested in studying survival times of patients who were discharged alive from the hospital. Hence, a selection criterion was imposed that only those subjects who were discharged alive were eligible to be included in the analysis. That is, individuals were not enrolled in the analysis if they died before discharging from the hospital, hence left truncation occurs. With such a criterion, a sample of size 461 was available. In addition, without imposing such a selection criterion, the sample size in the "original" dataset is 500, yielding a low proportion of truncation $1 - \frac{461}{500} = 7.8\%$. In this data set, the censoring rate is 61.8%. To be more specific, the total length of follow-up (lenfol) is the last event time (i.e., $Y_i = \min(T_i, C_i)$), the length of hospital stay (los) is the truncation time (i.e., A_i), and the vital status at last follow-up (fstat) is δ_i. In our analysis, the covariates include the body mass index (BMI) and the initial heart rate (HR) of a patient. Since BMI is subject to measurement error (e.g., Rothman 2008), we let W denote BMI and consider the measurement error model (5.5). Let Z denote HR.

In this data set, there is no additional data source, such as a validation subsample or replicated measurements which is often required to describe the measurement error process (e.g., Carroll et al. 2006; Yi 2017). To get around this and understand the impact of

measurement error on estimation, we carry out sensitivity analyses. That is, given a range of values for Σ_ϵ and α, we estimate β using $\widehat{\beta}$ and $\widetilde{\beta}$ via (5.13) and (5.22), respectively; and we want to assess how sensitive the results are to different degrees of measurement error. The results for $\alpha = 0$ and 100 are shown in Figures 5.1 and 5.2. Interestingly, under model (5.5) with $\alpha = 100$, the two methods reveals different results. It is seen that while $\widetilde{\beta}$ and $\widehat{\beta}$ are fairly close in values, $\widetilde{\beta}$ is much stabler than $\widehat{\beta}$. For example, $\widetilde{\beta}_x$ and $\widetilde{\beta}_z$ are fairly unchanged as $\Sigma_\epsilon > 0.6$ and $\Sigma_\epsilon < 0.4$; on the contrary, $\widehat{\beta}_x$ has a decreasing trend while $\widehat{\beta}_z$ is fluctuated as Σ_ϵ changes. Relative to the differences between $\widetilde{\beta}_x$ and $\widehat{\beta}_x$, $\widehat{\beta}_z$ and $\widetilde{\beta}_z$ are close to each other and both estimators are more stable than the estimators of β_x.

In Table 5.7, we further report the point estimates (EST), the standard errors and p-values for the estimators $\widehat{\beta}$ and $\widetilde{\beta}$ for the cases with $\Sigma_\epsilon = 0.147$, 0.526 and 0.858, respectively, corresponding to minor, moderate and large measurement error. All the point estimates produced by the two approaches are fairly close as observed from Figures 5.1 and 5.2. For each given method, the results are fairly stable, regardless of the degree of measurement error. The conditional profile likelihood method finds no evidence to support the significance of BMI and HR no matter what value α is specified for model (5.5). The variance estimates of $\widetilde{\beta}_z$ produced by the proposed method in Section 5.3 are noticeably affected by the degree of systematic error, i.e., the value of α in model (5.5), and as a result, the significance of HR is suggested differently by the proposed method in Section 5.3 under different measurement error models.

5.6 Length-Biased Sampling Data with Measurement Error

5.6.1 Length-Biased Sampling

In the foregoing development, we leave the distribution of left-truncation $\widetilde{A}$ discussed in Section 5.1 unspecified. If we impose certain assumptions on $\widetilde{A}$, the preceding development carry through and the new results can then generalize existing work. For instance, considered by Wang (1991) and De Uña-Alvarez (2004), suppose the incidence of disease onset follows a stationary Poisson distribution, then the truncation time follows a uniform distribution. Under this situation, the survival time in the prevalent cohort has a length-biased sampling distribution, because the probability of a survival time is proportional to the length of survival time (e.g., Huang and Qin 2011; Huang et al. 2012).

Consistent with Huang et al. (2012), assume the following conditions for the calendar time of the initial event:

(A1) The variable $(\widetilde{T}, \widetilde{V})$ is independent of u, where u is the time of the occurance of the disease incidence.

(A2) Disease incidence occurs over calendar time at a constant rate.

Then given $V = v$, the conditional density function of (T, A) is

$$\frac{f\left(t|v\right)}{\int_0^\infty \alpha f(\alpha|v)d\alpha}$$

(Lancaster 1990; Huang et al. 2012), and the survival time T has a length-biased conditional density function:

$$\frac{tf\left(t|v\right)}{\int_0^\infty \alpha f(\alpha|v)d\alpha}.$$

Let C_i be the censoring time for subject i. Then we have $Y_i = \min\{T_i, A_i + C_i\}$ and $\Delta_i = \min\{T_i, A_i + C_i\}$. Noting that $\int_0^\infty \alpha f(\alpha|v)d\alpha = \int_0^\infty S\left(\alpha|v\right)d\alpha$, by Assumptions (A1) and (A2) and the independent censoringship, the likelihood function of (Y_i, A_i, Δ_i) given V_i can be constructed as

$$L_{LB} \propto \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i} S(y_i|v_i)^{1-\delta_i}}{\int_0^\infty S(\alpha|v_i)d\alpha}, \tag{5.40}$$

which can be decomposed as the product of

$$L_{C,LB} = \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i} S(y_i|v_i)^{1-\delta_i}}{S(a_i|v_i)}$$

and

$$L_{M,LB} = \prod_{i=1}^{n} \frac{S(a_i|v_i)}{\int_0^\infty S(\alpha|v_i)d\alpha}.$$

Compared with the likelihood function (5.3), the likelihood function (5.40) does not involve the estimation procedure of density function $h(a)$, which can be thought of as a degenerate version of (5.3), agreeing with the standard view that the length-biased sampling

is regarded as a special case of the LTRC data (e.g., Asgharian et al. 2002; Qin and Shen 2010). To develop estimating procedures using (5.40), we need only to deal with $\Lambda_0(\cdot)$ and β but not $h(\cdot)$ as (5.3). In the absence of covariate measurement error, many authors developed methods to handle length-biased data. For example, Qin and Shen (2010) proposed the weighted estimating equation approach, and Huang and Qin (2012) explored a pseudo-profile likelihood method. Here, we further accommodate the feature of covariate measurement error for length-biased data and develop a valid inference method.

5.6.2 Estimation of Parameters for Survival Data

From the decomposition of (5.40), we can see that the conditional likelihood $L_{C,LB}$ is the same as those of (5.3). Hence, the estimator of the conditional likelihood, $\widehat{\beta}_{LB}$, can be derived from (5.13). To emphasize the different setting, let $\widehat{\ell}^*_{C,LB}$ denote the corrected conditional log likelihood under the length-biased sampling, which leads to an estimator of β:

$$\widehat{\beta}_{LB} = \underset{\beta}{\operatorname{argmax}}\, \widehat{\ell}^*_{C,LB}. \tag{5.41}$$

On the other hand, for the marginal likelihood $L_{M,LB}$, there is no density function $h(\cdot)$, so we simply apply the regression calibration (5.17) to replace the error-prone covariate X_i. Hence, the corrected marginal log likelihood, $\widehat{\ell}^*_{M,LB}$, has a similar form to (5.20) except for the estimate of $H(\cdot)$. As a result, an estimator of β is given by

$$\widetilde{\beta}_{LB} = \underset{\beta}{\operatorname{argmax}}(\widehat{\ell}^*_{C,LB} + \widehat{\ell}^*_{M,LB}). \tag{5.42}$$

5.6.3 Asymptotic Results

Let $\mu_{LB}\left(\widetilde{x}_{\mathrm{RC},i}, z_i\right) = \int_0^\tau \exp\left\{-\Lambda_0(u)\exp(\widetilde{x}_{\mathrm{RC},i}^\top \beta_{x0} + z_i^\top \beta_{z0})\right\} du$ be the function of $(\widetilde{x}_{\mathrm{RC},i}, z_i)$. Define

$$
\begin{aligned}
& \Psi_{LB}\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right) \\
= & \int_0^\tau \left\{ v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} + \begin{pmatrix} \Sigma_\epsilon \beta_{x0} \\ \mathbf{0}_q \end{pmatrix} \right\} dN_i(u) \\
& - \int_0^\tau \frac{\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\mathcal{S}^{(0)}(u,\beta_0)} \left(v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} \right) dE\left\{N_i(u)\right\} \\
& - \left[\int_{-\infty}^{\infty} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} + \frac{d\mathcal{N}(u)\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\mathcal{S}^{(0)}(u,\beta_0)^2} \right\} m(\beta_{x0}) \right. \\
& \left. \times \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau) dG(a,\widehat{v}) \right] \\
& + \left[\int_{-\infty}^{\infty} \int_0^\tau \left\{ \frac{1}{\mu_{LB}\left(\widetilde{x}_{\mathrm{RC}}, z\right)} \frac{\partial}{\partial \beta} \psi_{LB,i}(\beta_0 | \widetilde{x}_{\mathrm{RC}}, z) \right. \right. \\
& \left. \left. - \frac{\partial \mu_{LB}\left(\widetilde{x}_{\mathrm{RC}}, z\right)}{\partial \beta} \frac{1}{\mu_{LB}^2\left(\widetilde{x}_{\mathrm{RC}}, z\right)} \psi_{LB,i}(\beta_0 | \widetilde{x}_{\mathrm{RC}}, z) \right\} dG(a,\widehat{v}) \right] \\
& - \frac{\partial}{\partial \beta} \Lambda_0(a_i) \exp\left(\widehat{v}_i^\top \beta_0\right) - \frac{1}{\mu_{LB}\left(\widetilde{x}_{\mathrm{RC},i}, z_i\right)} \frac{\partial \mu_{LB}\left(\widetilde{x}_{\mathrm{RC},i}, z_i\right)}{\partial \beta},
\end{aligned}
$$

where $\psi_{LB,i}(\beta_0|\widetilde{x}_{\mathrm{RC}}, z)$ is similarly defined by (5.24) with the integral relative to $dH(\cdot)$ removed, i.e.,

$$
\begin{aligned}
\psi_{LB,i}(\beta_0|\widetilde{x}_{\mathrm{RC}}, z) = & \int_0^\tau \int_0^\tau S(\xi|\widetilde{x}_{\mathrm{RC}}, z) \left\{ \frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} \right. \\
& \left. - \frac{d\mathcal{N}(u)\exp\left(w_i^{*\top}\beta_{x0} + z_i^\top \beta_{z0}\right) I(a_i \le u \le y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2} \right\} m(\beta_{x0}) \\
& \times \exp\left(\widetilde{x}_{\mathrm{RC}}^\top \beta_{x0} + z^\top \beta_{z0}\right) d\xi + o_p(1).
\end{aligned}
$$

We now establish the following results whose proof is deferred to Appendix D.6.

Theorem 5.6.1 *Under regularity conditions given in Appendix D.1, we have that $n \to \infty$,*

(1) $\widetilde{\beta}_{LB} \xrightarrow{p} \beta_0$;

(2) $\sqrt{n}\left(\widetilde{\beta}_{LB}-\beta_0\right)\xrightarrow{d}N(0,\mathcal{A}_{LB}^{-1}\mathcal{B}_{LB}\mathcal{A}_{LB}^{-1})$,

where $\mathcal{B}_{LB}=E(\Psi_{LB,i}^{\otimes 2})$ *with* $\Psi_{LB,i}=\Psi_{LB}\left(X_i^*,\widetilde{X}_{RC,i},Z_i,Y_i,A_i\right)$, *and*

$$
\begin{aligned}
&\mathcal{A}_{LB}\\
=&\int_0^\tau\left[\left\{\frac{\mathcal{S}^{(2)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}-\left(\frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right)^{\otimes 2}\right\}-\begin{pmatrix}\Sigma_\epsilon & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q}\end{pmatrix}\right]dE\left\{N_i(u)\right\}\\
&+E\left[\frac{\partial^2}{\partial\beta\partial\beta^\top}\Lambda_0(A_i)\exp\left(\widehat{V}_i^\top\beta_0\right)\right.\\
&+\left\{\mu_{LB}\left(\widetilde{X}_{RC,i},Z_i\right)\right\}^{-2}\left\{\mu_{LB}\left(\widetilde{X}_{RC,i},Z_i\right)\frac{\partial\left\{\mu_{LB}\left(\widetilde{X}_{RC,i},Z_i\right)\right\}^2}{\partial\beta\partial\beta^\top}\right.\\
&\left.\left.-\left(\frac{\partial\mu_{LB}\left(\widetilde{X}_{RC,i},Z_i\right)}{\partial\beta}\right)^{\otimes 2}\right\}\right].
\end{aligned}
$$

5.6.4 Simulation Study

To show the numerical performance of estimator $\widetilde{\beta}_{LB}$ in contrast to $\widetilde{\beta}$ which is obtained by (5.22), we conduct a simulation study using the setting in Section 5.5.1 with the distribution of truncation times taken as the uniform distribution UNIF$[0,1]$ and α is set as 0. In addition to $\widetilde{\beta}_{LB}$ from (5.42), we also report the performance of the naive estimator and $\widehat{\beta}_{LB}$ determined by (5.41). The results are reported in Table 5.6.

Simulation results show that for different Σ_ϵ and censoring rates, our proposed methods yield satisfactory results, and $\widetilde{\beta}_{LB}$ is more efficient than $\widehat{\beta}_{LB}$. The naive estimator incurs considerable biases. The results in Table 5.6 are comparable with those reported in Section 5.5.

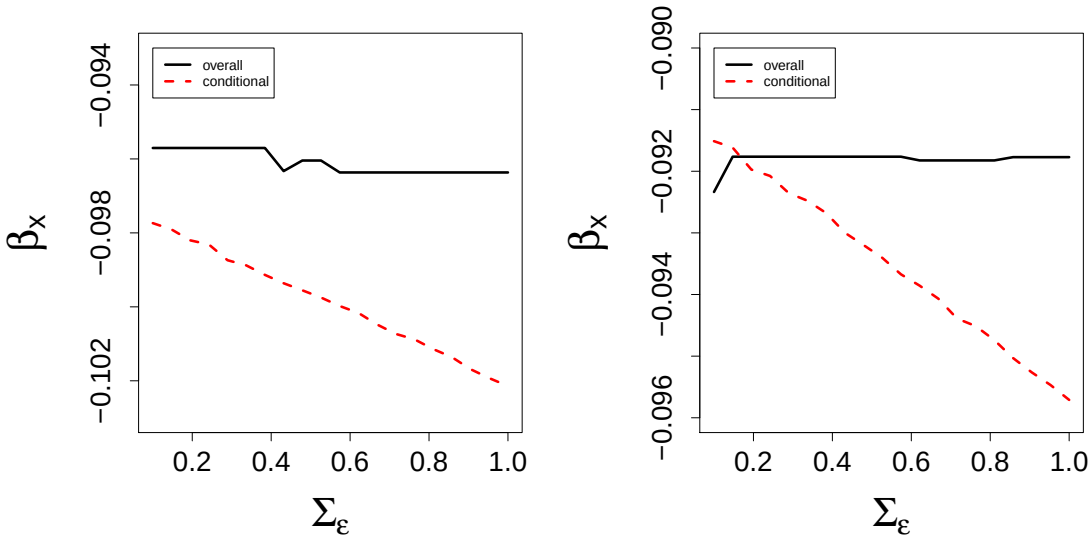


Figure 5.1: The estimator of β_x versus variance Σ_ϵ for sensitivity analysis. Solid line is a curve of $\widetilde{\beta}_x$ from the proposed pseudo-likelihood estimator (5.22); dash line is a curve of $\widehat{\beta}_x$ from the conditional likelihood estimator (5.13). Left panel is $\alpha = 0$ and right panel is $\alpha = 100$.

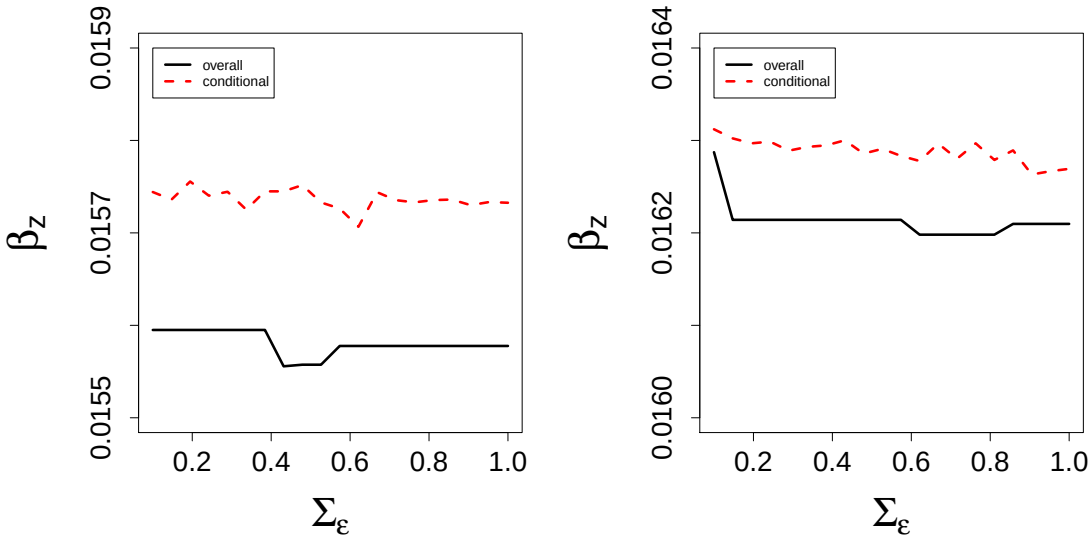


Figure 5.2: The estimator of β_z versus variance Σ_ϵ for sensitivity analysis. Solid line is a curve of $\widetilde{\beta}_z$ from the proposed pseudo-likelihood estimator (5.22); dash line is a curve of $\widehat{\beta}_z$ from the conditional likelihood estimator (5.13). Left panel is $\alpha = 0$ and right panel is $\alpha = 100$.

Table 5.1: Simulation results under measurement error model (5.5) with $\alpha = 0$

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Naive	-0.104	0.041	0.036	0.013	20.2	0.114	0.068	0.049	0.014	60.8
		Conditional $\left(\widehat{\beta}\right)$	0.006	0.068	0.063	0.005	94.0	0.027	0.106	0.105	0.012	92.9
		Full $\left(\widetilde{\beta}\right)$	0.010	0.063	0.062	0.004	94.4	0.029	0.098	0.097	0.010	94.1
	25 %	Naive	0.118	0.035	0.023	0.015	18.0	0.106	0.062	0.032	0.012	52.1
		Conditional $\left(\widehat{\beta}\right)$	0.002	0.047	0.047	0.002	95.3	0.018	0.076	0.076	0.006	94.9
		Full $\left(\widetilde{\beta}\right)$	0.003	0.046	0.045	0.002	94.6	0.018	0.075	0.071	0.006	93.3
	50%	Naive	0.108	0.051	0.033	0.016	8.7	0.150	0.069	0.043	0.025	33.7
		Conditional $\left(\widehat{\beta}\right)$	0.002	0.061	0.058	0.004	94.2	0.017	0.098	0.095	0.010	93.0
		Full $\left(\widetilde{\beta}\right)$	0.007	0.057	0.055	0.003	93.6	0.027	0.097	0.093	0.010	93.3
0.5	0%	Naive	0.124	0.039	0.019	0.018	6.2	0.133	0.061	0.052	0.019	45.1
		Conditional $\left(\widehat{\beta}\right)$	0.024	0.052	0.048	0.003	93.2	0.026	0.076	0.071	0.006	94.6
		Full $\left(\widetilde{\beta}\right)$	0.001	0.045	0.043	0.002	96.9	0.017	0.067	0.057	0.005	95.4
	25 %	Naive	0.124	0.043	0.035	0.018	7.9	0.132	0.066	0.056	0.019	55.2
		Conditional $\left(\widehat{\beta}\right)$	0.026	0.058	0.055	0.004	93.9	0.025	0.083	0.080	0.008	93.7
		Full $\left(\widetilde{\beta}\right)$	0.003	0.054	0.046	0.002	94.6	0.015	0.082	0.079	0.007	94.0
	50%	Naive	0.132	0.053	0.032	0.023	10.1	0.142	0.073	0.043	0.023	25.1
		Conditional $\left(\widehat{\beta}\right)$	0.031	0.072	0.068	0.006	93.4	0.033	0.100	0.099	0.011	93.6
		Full $\left(\widetilde{\beta}\right)$	0.007	0.057	0.054	0.003	94.4	0.026	0.077	0.075	0.009	96.0
0.75	0%	Naive	0.115	0.038	0.037	0.016	10.6	0.112	0.053	0.052	0.014	37.5
		Conditional $\left(\widehat{\beta}\right)$	0.040	0.057	0.056	0.005	93.8	0.036	0.078	0.077	0.007	93.0
		Full $\left(\widetilde{\beta}\right)$	-0.018	0.053	0.041	0.002	95.0	-0.001	0.066	0.064	0.004	94.6
	25 %	Naive	0.126	0.042	0.035	0.020	8.2	0.126	0.061	0.059	0.018	43.0
		Conditional $\left(\widehat{\beta}\right)$	0.042	0.067	0.064	0.006	93.9	0.044	0.090	0.088	0.010	93.7
		Full $\left(\widetilde{\beta}\right)$	-0.015	0.055	0.048	0.003	94.4	0.003	0.077	0.073	0.006	94.2
	50%	Naive	0.135	0.047	0.044	0.020	10.7	0.135	0.056	0.051	0.021	22.4
		Conditional $\left(\widehat{\beta}\right)$	0.040	0.094	0.080	0.008	93.2	0.045	0.106	0.084	0.013	95.0
		Full $\left(\widetilde{\beta}\right)$	-0.013	0.058	0.055	0.003	97.4	0.009	0.088	0.080	0.008	95.8

Table 5.2: Simulation results under measurement error model (5.5) with $\alpha = 100$

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Naive	-0.109	0.040	0.034	0.014	55.4	0.112	0.065	0.052	0.017	57.2
		Conditional $\left(\widehat{\beta}\right)$	0.002	0.043	0.041	0.002	93.8	0.007	0.068	0.066	0.005	93.6
		Full $\left(\widetilde{\beta}\right)$	-0.001	0.042	0.040	0.001	94.5	0.007	0.061	0.053	0.004	94.5
	25%	Naive	0.120	0.038	0.031	0.016	45.0	0.123	0.046	0.044	0.017	55.6
		Conditional $\left(\widehat{\beta}\right)$	-0.003	0.050	0.047	0.002	94.6	0.010	0.075	0.074	0.006	93.2
		Full $\left(\widetilde{\beta}\right)$	0.002	0.047	0.047	0.002	93.0	0.010	0.074	0.055	0.006	94.2
	50%	Naive	0.112	0.056	0.045	0.016	60.4	0.125	0.037	0.035	0.017	47.0
		Conditional $\left(\widehat{\beta}\right)$	0.005	0.064	0.057	0.004	94.6	0.013	0.097	0.084	0.010	93.2
		Full $\left(\widetilde{\beta}\right)$	0.004	0.057	0.050	0.003	94.7	0.021	0.081	0.080	0.007	93.7
0.5	0%	Naive	-0.115	0.041	0.038	0.015	55.8	-0.101	0.064	0.037	0.014	52.0
		Conditional $\left(\widehat{\beta}\right)$	0.028	0.048	0.047	0.003	93.4	0.028	0.074	0.070	0.006	93.0
		Full $\left(\widetilde{\beta}\right)$	0.002	0.046	0.041	0.002	95.0	0.012	0.071	0.053	0.005	96.0
	25%	Naive	-0.115	0.048	0.046	0.019	51.0	-0.117	0.040	0.035	0.015	45.6
		Conditional $\left(\widehat{\beta}\right)$	0.029	0.057	0.054	0.004	93.6	0.027	0.086	0.083	0.008	93.4
		Full $\left(\widetilde{\beta}\right)$	0.005	0.055	0.047	0.002	95.2	0.017	0.080	0.078	0.007	95.4
	50%	Naive	0.129	0.046	0.044	0.019	52.2	0.141	0.094	0.032	0.029	37.4
		Conditional $\left(\widehat{\beta}\right)$	0.034	0.067	0.065	0.006	93.2	0.036	0.111	0.095	0.014	91.0
		Full $\left(\widetilde{\beta}\right)$	0.010	0.057	0.053	0.003	96.0	0.030	0.097	0.094	0.010	93.2
0.75	0%	Naive	0.116	0.037	0.037	0.015	37.3	0.118	0.050	0.032	0.016	29.2
		Conditional $\left(\widehat{\beta}\right)$	0.043	0.061	0.055	0.005	93.2	0.042	0.083	0.076	0.009	92.6
		Full $\left(\widetilde{\beta}\right)$	-0.018	0.041	0.040	0.002	94.0	-0.001	0.067	0.054	0.004	95.8
	25%	Naive	-0.111	0.050	0.049	0.015	56.2	0.151	0.077	0.041	0.029	44.0
		Conditional $\left(\widehat{\beta}\right)$	0.035	0.065	0.062	0.005	93.0	0.034	0.089	0.085	0.009	93.2
		Full $\left(\widetilde{\beta}\right)$	-0.021	0.058	0.056	0.003	94.6	-0.003	0.080	0.071	0.006	93.2
	50%	Naive	0.126	0.055	0.048	0.019	59.0	0.165	0.085	0.079	0.035	25.4
		Conditional $\left(\widehat{\beta}\right)$	0.048	0.080	0.076	0.009	92.6	0.050	0.103	0.103	0.012	94.0
		Full $\left(\widetilde{\beta}\right)$	-0.012	0.060	0.056	0.003	97.0	0.012	0.089	0.082	0.007	94.0

Table 5.3: Simulation results with misspecified measurement error model under Scenario 1

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Conditional $\left(\widehat{\beta}\right)$	0.003	0.043	0.041	0.002	92.2	0.014	0.071	0.066	0.005	93.8
		Full $\left(\widetilde{\beta}\right)$	0.004	0.042	0.040	0.002	93.2	0.014	0.069	0.054	0.005	94.4
	25%	Conditional $\left(\widehat{\beta}\right)$	0.003	0.050	0.047	0.003	92.2	0.023	0.080	0.075	0.007	92.8
		Full $\left(\widetilde{\beta}\right)$	0.006	0.049	0.046	0.002	92.0	0.024	0.076	0.056	0.007	93.4
	50%	Conditional $\left(\widehat{\beta}\right)$	0.008	0.062	0.057	0.004	92.6	0.027	0.095	0.089	0.009	92.8
		Full $\left(\widetilde{\beta}\right)$	0.010	0.057	0.054	0.003	95.4	0.026	0.089	0.083	0.009	93.8
0.5	0%	Conditional $\left(\widehat{\beta}\right)$	0.027	0.050	0.047	0.003	91.6	0.026	0.073	0.070	0.006	94.0
		Full $\left(\widetilde{\beta}\right)$	0.001	0.047	0.043	0.002	93.4	0.012	0.069	0.066	0.005	94.0
	25%	Conditional $\left(\widehat{\beta}\right)$	0.026	0.058	0.054	0.004	91.0	0.028	0.080	0.079	0.007	93.8
		Full $\left(\widetilde{\beta}\right)$	0.002	0.054	0.046	0.002	95.4	0.015	0.073	0.057	0.006	94.4
	50%	Conditional $\left(\widehat{\beta}\right)$	0.024	0.068	0.064	0.005	92.8	0.034	0.104	0.096	0.012	91.8
		Full $\left(\widetilde{\beta}\right)$	0.003	0.057	0.055	0.003	95.8	0.027	0.098	0.094	0.010	93.6
0.75	0%	Conditional $\left(\widehat{\beta}\right)$	0.044	0.055	0.055	0.005	90.4	0.045	0.078	0.076	0.008	92.0
		Full $\left(\widetilde{\beta}\right)$	-0.017	0.052	0.049	0.002	94.0	0.001	0.068	0.064	0.005	93.4
	25%	Conditional $\left(\widehat{\beta}\right)$	0.040	0.066	0.062	0.006	91.8	0.044	0.090	0.085	0.010	92.6
		Full $\left(\widetilde{\beta}\right)$	-0.019	0.052	0.049	0.003	93.6	0.004	0.079	0.076	0.006	93.8
	50%	Conditional $\left(\widehat{\beta}\right)$	0.044	0.083	0.075	0.009	92.0	0.039	0.111	0.102	0.014	93.4
		Full $\left(\widetilde{\beta}\right)$	-0.012	0.058	0.055	0.003	96.0	0.005	0.090	0.088	0.008	95.0

Table 5.4: Simulation results with misspecified measurement error model under Scenario 2

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Conditional $\left(\widehat{\beta}\right)$	0.041	0.056	0.044	0.005	78.4	0.057	0.076	0.064	0.009	53.2
		Full $\left(\widetilde{\beta}\right)$	0.038	0.044	0.043	0.003	86.6	0.060	0.071	0.060	0.009	85.0
	25%	Conditional $\left(\widehat{\beta}\right)$	0.050	0.049	0.047	0.005	80.0	0.069	0.076	0.075	0.010	87.8
		Full $\left(\widetilde{\beta}\right)$	0.048	0.049	0.046	0.005	82.6	0.069	0.075	0.074	0.010	87.6
	50%	Conditional $\left(\widehat{\beta}\right)$	0.060	0.062	0.058	0.007	83.0	0.103	0.096	0.091	0.019	82.0
		Full $\left(\widetilde{\beta}\right)$	0.061	0.061	0.056	0.007	72.0	0.101	0.092	0.090	0.019	80.6
0.5	0%	Conditional $\left(\widehat{\beta}\right)$	0.075	0.052	0.047	0.008	63.8	0.077	0.072	0.070	0.011	83.6
		Full $\left(\widetilde{\beta}\right)$	0.033	0.048	0.043	0.003	90.4	0.053	0.070	0.065	0.008	87.8
	25%	Conditional $\left(\widehat{\beta}\right)$	0.079	0.054	0.053	0.009	70.4	0.082	0.082	0.079	0.013	84.8
		Full $\left(\widetilde{\beta}\right)$	0.040	0.048	0.045	0.004	91.4	0.070	0.080	0.075	0.012	85.6
	50%	Conditional $\left(\widehat{\beta}\right)$	0.084	0.067	0.065	0.011	78.8	0.166	0.111	0.096	0.040	61.4
		Full $\left(\widetilde{\beta}\right)$	0.049	0.056	0.056	0.006	77.8	0.100	0.098	0.095	0.020	64.8
0.75	0%	Conditional $\left(\widehat{\beta}\right)$	0.089	0.061	0.055	0.012	63.2	0.089	0.076	0.076	0.014	82.0
		Full $\left(\widetilde{\beta}\right)$	-0.049	0.050	0.042	0.004	80.4	0.059	0.067	0.061	0.008	88.0
	25%	Conditional $\left(\widehat{\beta}\right)$	0.093	0.066	0.062	0.013	70.0	0.093	0.088	0.085	0.016	83.8
		Full $\left(\widetilde{\beta}\right)$	-0.050	0.050	0.047	0.005	82.4	0.067	0.074	0.072	0.011	87.8
	50%	Conditional $\left(\widehat{\beta}\right)$	0.098	0.080	0.076	0.016	77.2	0.101	0.103	0.103	0.021	89.0
		Full $\left(\widetilde{\beta}\right)$	0.053	0.055	0.045	0.006	78.2	0.090	0.089	0.087	0.016	87.8

Table 5.5: Simulation results under measurement error model (5.5) with $\alpha = 100$ in the presence of validation data

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Naive	0.069	0.033	0.030	0.006	12.5	0.107	0.045	0.045	0.013	15.4
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.003	0.043	0.041	0.002	93.8	0.017	0.070	0.066	0.005	93.6
		Full $\left(\widetilde{\beta}_{val}\right)$	0.004	0.042	0.040	0.002	93.4	0.016	0.068	0.065	0.005	94.2
	25%	Naive	0.061	0.039	0.037	0.005	16.4	0.094	0.048	0.045	0.011	17.9
		Conditional $\left(\widehat{\beta}_{val}\right)$	-0.003	0.050	0.047	0.003	92.6	0.008	0.075	0.074	0.006	94.2
		Full $\left(\widetilde{\beta}_{val}\right)$	0.002	0.045	0.043	0.002	95.6	0.015	0.074	0.055	0.006	94.2
	50%	Naive	0.117	0.040	0.040	0.015	15.0	0.106	0.050	0.048	0.014	11.2
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.009	0.064	0.057	0.004	91.2	0.013	0.097	0.089	0.010	93.2
		Full $\left(\widetilde{\beta}_{val}\right)$	0.005	0.057	0.057	0.003	97.0	0.019	0.090	0.085	0.008	92.4
0.5	0%	Naive	0.145	0.020	0.021	0.022	3.4	-0.067	0.058	0.008	0.056	43.8
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.028	0.048	0.047	0.003	93.4	0.027	0.073	0.070	0.006	93.0
		Full $\left(\widetilde{\beta}_{val}\right)$	0.002	0.041	0.037	0.002	95.0	0.012	0.071	0.069	0.005	95.4
	25%	Naive	0.147	0.025	0.022	0.022	8.8	-0.103	0.060	0.059	0.014	58.0
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.029	0.057	0.054	0.004	93.6	0.027	0.086	0.080	0.008	93.6
		Full $\left(\widetilde{\beta}_{val}\right)$	0.005	0.047	0.045	0.002	95.0	0.017	0.080	0.078	0.007	95.4
	50%	Naive	0.142	0.046	0.040	0.022	24.4	-0.190	0.059	0.057	0.040	36.4
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.033	0.066	0.065	0.006	93.4	0.035	0.111	0.095	0.014	91.0
		Full $\left(\widetilde{\beta}_{val}\right)$	0.010	0.057	0.053	0.003	95.8	0.029	0.097	0.094	0.010	93.2
0.75	0%	Naive	-0.133	0.026	0.023	0.018	8.2	-0.113	0.053	0.053	0.016	46.0
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.042	0.061	0.055	0.006	93.2	0.041	0.083	0.076	0.009	92.8
		Full $\left(\widetilde{\beta}_{val}\right)$	-0.017	0.051	0.041	0.002	97.6	-0.001	0.067	0.054	0.004	94.0
	25%	Naive	-0.137	0.025	0.024	0.019	10.6	-0.147	0.058	0.054	0.025	33.2
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.035	0.065	0.062	0.006	92.0	0.033	0.089	0.085	0.009	92.8
		Full $\left(\widetilde{\beta}_{val}\right)$	-0.021	0.047	0.057	0.002	93.8	-0.003	0.078	0.078	0.006	95.6
	50%	Naive	-0.144	0.032	0.034	0.022	18.6	-0.239	0.055	0.051	0.060	18.4
		Conditional $\left(\widehat{\beta}_{val}\right)$	0.048	0.080	0.076	0.009	92.6	0.050	0.103	0.103	0.013	94.2
		Full $\left(\widetilde{\beta}_{val}\right)$	-0.012	0.053	0.064	0.003	95.0	0.012	0.085	0.082	0.007	94.8

Table 5.6: Simulation results with length-biased sampling

Σ_ϵ	cr	Method	Estimator of β_x					Estimator of β_z				
			Bias	S.E.	MVE	MSE	CP (%)	Bias	S.E.	MVE	MSE	CP (%)
0.01	0%	Naive	0.085	0.037	0.018	0.010	19.4	0.084	0.055	0.053	0.008	66.1
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.000	0.044	0.041	0.002	93.4	0.007	0.071	0.071	0.005	93.6
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.000	0.041	0.040	0.002	94.8	0.007	0.071	0.069	0.005	93.2
	25%	Naive	0.094	0.042	0.018	0.012	14.1	0.094	0.056	0.052	0.011	57.2
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.003	0.057	0.054	0.003	93.0	0.009	0.093	0.090	0.009	94.4
		Full $\left(\widetilde{\beta}_{LB}\right)$	0.006	0.051	0.050	0.003	94.0	0.010	0.092	0.088	0.009	94.4
	50%	Naive	0.112	0.051	0.023	0.018	12.7	0.116	0.072	0.046	0.016	38.7
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.002	0.081	0.077	0.007	93.0	0.025	0.136	0.126	0.019	93.6
		Full $\left(\widetilde{\beta}_{LB}\right)$	0.003	0.067	0.060	0.005	94.0	0.020	0.116	0.114	0.014	94.4
0.5	0%	Naive	0.090	0.027	0.017	0.011	13.6	0.098	0.054	0.051	0.011	66.1
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.029	0.048	0.047	0.003	93.6	0.030	0.075	0.073	0.007	93.0
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.000	0.038	0.036	0.001	97.0	0.004	0.071	0.069	0.005	96.0
	25%	Naive	0.100	0.041	0.029	0.013	11.8	0.108	0.056	0.053	0.013	54.1
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.024	0.065	0.062	0.005	93.2	0.036	0.098	0.094	0.011	94.2
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.002	0.050	0.049	0.002	95.4	0.008	0.089	0.084	0.008	94.6
	50%	Naive	0.108	0.051	0.024	0.017	11.7	0.116	0.069	0.047	0.016	40.8
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.036	0.096	0.090	0.010	93.6	0.051	0.140	0.136	0.022	94.0
		Full $\left(\widetilde{\beta}_{LB}\right)$	0.006	0.068	0.061	0.005	94.2	0.015	0.118	0.113	0.014	95.0
0.75	0%	Naive	0.108	0.026	0.018	0.014	9.1	0.113	0.054	0.047	0.014	53.5
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.044	0.056	0.056	0.005	92.4	0.044	0.087	0.079	0.010	92.0
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.019	0.037	0.037	0.002	95.4	-0.011	0.072	0.070	0.005	93.0
	25%	Naive	0.116	0.041	0.021	0.016	6.6	0.126	0.054	0.054	0.018	43.4
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.042	0.073	0.073	0.007	93.4	0.045	0.105	0.102	0.013	94.6
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.017	0.046	0.038	0.002	95.0	-0.011	0.085	0.084	0.007	94.2
	50%	Naive	0.115	0.049	0.023	0.017	8.4	0.123	0.063	0.047	0.017	34.5
		Conditional $\left(\widehat{\beta}_{LB}\right)$	0.065	0.128	0.121	0.021	93.8	0.081	0.176	0.168	0.038	94.2
		Full $\left(\widetilde{\beta}_{LB}\right)$	-0.008	0.067	0.066	0.005	94.0	0.004	0.119	0.113	0.014	96.4

Table 5.7: Sensitivity analyses result of Worcester Heart Attack Study Data

			Estimator $\widehat{\beta}$			Estimator $\widetilde{\beta}$		
α	Σ_ϵ		EST	S.E	p-value	EST	S.E	p-value
0	0.147	β_x	-0.098	0.081	0.229	-0.096	0.041	0.024
		β_z	0.016	0.018	0.201	0.016	0.017	0.398
	0.526	β_x	-0.099	0.084	0.232	-0.096	0.042	0.025
		β_z	0.016	0.018	0.202	0.016	0.017	0.401
	0.858	β_x	-0.101	0.085	0.235	-0.096	0.044	0.029
		β_z	0.016	0.018	0.203	0.016	0.017	0.401
100	0.147	β_x	-0.092	0.063	0.149	-0.092	0.019	2e-06
		β_z	0.016	0.011	0.123	0.016	0.003	4e-07
	0.526	β_x	-0.093	0.065	0.150	-0.092	0.020	3e-06
		β_z	0.016	0.011	0.123	0.016	0.003	4e-07
	0.858	β_x	-0.095	0.066	0.152	-0.092	0.020	4e-06
		β_z	0.016	0.011	0.124	0.016	0.003	4e-07

Chapter 6

Model Selection and Model Averaging for Analysis of Truncated and Censored Data with Measurement Error

6.1 Notation and Model

In this chapter, we adopt similar notation which has been defined in Chapter 5. Specifically, as defined in Section 5.1, let $\widetilde{T}$, $\widetilde{A}$, and $\widetilde{V} = \left(\widetilde{X}^\top, \widetilde{Z}^\top\right)^\top$ denote the failure time, the truncation time, and the $(p+q)$-dimensional vector of covariates, respectively. Based on the discussion in Section 5.1, for an individual with $\widetilde{T} \geq \widetilde{A}$, we let (A, T, V) with $V = \left(X^\top, Z^\top\right)^\top$ denote $\left(\widetilde{A}, \widetilde{T}, \widetilde{V}\right)$ to indicate such an individual is eligible for the recruitment so that measuring (A, T, V) is possible. If $\widetilde{T} < \widetilde{A}$, then such an individual is not included in the study to contribute any information. We define C as the censoring time for a recruited subject. Let $Y = \min\{T, A + C\}$ be the observed time and let $\Delta = I(T \leq A + C)$ be the indicator of a failure event. Figure 1.1 gives an illustration of the relationship among those variables.

6.1.1 Cox Model and Inference

Suppose that we have a sample of n subjects and that for $i = 1, \cdots, n$, $(Y_i, A_i, \Delta_i, V_i)$ has the same distribution as (Y, A, Δ, V) and $(y_i, a_i, \delta_i, v_i)$ represents realizations of $(Y_i, A_i, \Delta_i, V_i)$. Consider the Cox model for survival times $\widetilde{T}$ whose hazard function is modeled as

$$\lambda(t|v_i) = \lambda_0(t) \exp(v_i^\top \beta), \tag{6.1}$$

where $\lambda_0(\cdot)$ is an unknown baseline hazards function, and β is the vector of the parameters that are of interest.

Let $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ be the cumulative baseline hazards function. Let $\mathcal{F}(t|v_i) = \exp\left\{-\Lambda_0(t)\exp\left(v_i^\top \beta\right)\right\}$ denote the survivor function of $\widetilde{T}$ given the covariates and let $f(t|v_i) = -\frac{d}{dt}\mathcal{F}(t|v_i)$.

By Assumptions (C5) and (C6) in Appendix E.1, the likelihood function is given by

$$L \propto \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i}\mathcal{F}(y_i|v_i)^{1-\delta_i}dH(a_i)}{\int_0^\infty \mathcal{F}(u|v_i)dH(u)},$$

which can be equivalently re-written as the product of the conditional likelihood

$$L_C = \prod_{i=1}^{n} \frac{f(y_i|v_i)^{\delta_i}\mathcal{F}(y_i|v_i)^{1-\delta_i}}{\mathcal{F}(a_i|v_i)} \tag{6.2}$$

and the marginal likelihood

$$L_M = \prod_{i=1}^{n} \frac{\mathcal{F}(a_i|v_i)dH(a_i)}{\int_0^\infty \mathcal{F}(u|v_i)dH(u)}. \tag{6.3}$$

Discussion on this can be found in Wang et al. (1993), Huang et al. (2012) and Chen (2019). In principle, estimation of the model parameters may proceed with maximizing $L \propto L_C \times L_M$ with respect to the model parameters.

6.1.2 Framework of Submodels

In specifying the model (6.1) we include all the covariates in the model without discretion; irrelevant or unimportant covariates may be included in the model. To feature this, we

consider a framework initiated by Hjort and Claeskens (2006), the so-called *local model misspecification* framework.

Let Z_i represent the vector of important covariates that are always being included in the model, and let X_i represent the vector of covariates which may be subject to exclusion when building a model. Write $X_i = (X_{i1}, \ldots, X_{ip})^\top$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})^\top$. Let $\beta = (\beta_x^\top, \beta_z^\top)^\top$ be a vector with dimension $d = p + q$, where β_x is the parameter vector for which we are unsure whether or not all of its components should be included in the model and β_z is the parameter vector which should be used in the model. Let the true value of β be represented by $\beta_0 = \left(\frac{\eta^\top}{\sqrt{n}}, \beta_{z0}^\top\right)^\top$, where η is a parameter, and $\frac{\eta}{\sqrt{n}}$ represents the degree of the departure of the corresponding model from the null model $\beta_0 = (0, \beta_{z0})$ (Wang et al., 2012; Wang et al. 2015).

Let $\mathcal{S}$ be the class of all subsets of $\{1, 2, \cdots, p\}$ which are increasingly ordered. For any $S \in \mathcal{S}$, let $|S|$ denote the number of the elements in S. If $|S| = 0$, then we say that the set S is *null*; if $|S| = p$, then such S is called the *full* set. Let $\beta_S = \left(\beta_{x,S}^\top, \beta_{z,S}^\top\right)^\top$ denote the parameter vector for the candidate model which corresponds to the covariates indexed by S, with $\beta_{x,S}$ being an $|S|$-subvector of β_x. Although covariate Z_i is always included in the model, the subscript S in $\beta_{z,S}$ is used to emphasize that this is the parameter under the candidate model associated with S.

We now define a projection operator. For any S, let π_S be an $|S| \times p$ matrix with element 0 or 1; in each row there is one and only one element which takes value 1 and in each column there is at most one element taking value 1. More specifically, if $S = \left(j_1, j_2, \cdots, j_{|S|}\right)$ with $1 \leqslant j_1 < j_2 < \cdots < j_{|S|} \leqslant p$, then the (k, j_k) element of π_S takes value 1 for $k = 1, \cdots, |S|$; other elements of π_S take value 0. Let $\Pi_S = \begin{pmatrix} \pi_S & \mathbf{0}_{|S|\times q} \\ \mathbf{0}_{q\times p} & I_{q\times q} \end{pmatrix}$, where $\mathbf{0}_{p\times q}$ is the $p \times q$ matrix with entries zero, and $I_{q\times q}$ is the $q \times q$ identity matrix. Then applying Π_S to $\left(X_i^\top, Z_i^\top\right)^\top$ gives us the $(|S| + q) \times 1$ vector, $\Pi_S \left(X_i^\top, Z_i^\top\right)^\top$, which includes the covariates in the candidate model S.

6.1.3 Measurement Error Model

In practice, covariates are often subject to measurement error. Suppose that covariate X_i is measured with error, and X_i^* is an observed value, or surrogate, of X_i. Suppose that covariate Z_i is precisely observed. We consider the widely considered measurement error

model

$$X_i^* = X_i + \epsilon_i, \tag{6.4}$$

where ϵ_i is independent of all other variables (e.g., Carroll et al. 2006; Yi 2017), and $\epsilon_i \sim N(0, \Sigma_\epsilon)$ with covariance matrix Σ_ϵ. To highlight the idea, we assume that Σ_ϵ is known for now. Thus, the moment generation function of ϵ_i is $m(t) = \exp(\frac{1}{2}t^\top \Sigma_\epsilon t)$, and

$$E\left\{\exp\left(t^\top X_i^*\right)\right\} = m(t)\exp\left(t^\top X_i\right).$$

6.2 Methodology for the Correction of Measurement Error Effects

6.2.1 Correction for Conditional Log-Likelihood Function

For any candidate model S, let $L_{C,S}$ denote the derived conditional likelihood function, which, similar to the expression of L_C in (6.2), leads to the conditional log-likelihood function

$$\begin{aligned}\ell_{C,S} \;=\; & \sum_{i=1}^{n}\Big[\delta_i \log\lambda_0(y_i) + \delta_i\left\{(\pi_S x_i)^\top\beta_x + z_i^\top\beta_z\right\} \\ & - \left\{\Lambda_0(y_i) - \Lambda_0(a_i)\right\}\exp\left\{(\pi_S x_i)^\top\beta_x + z_i^\top\beta_z\right\}\Big],\end{aligned}$$

showing that $(\pi_S X_i)^\top\beta_x$ and $\exp\left\{(\pi_S X_i)^\top\beta_x\right\}$ are the only terms involving error-prone covariates.

To correct for the measurement error effects, we first manipulate the measurement error model (6.4) as

$$\pi_S X_i^* = \pi_S X_i + \pi_S\epsilon_i, \tag{6.5}$$

where $\pi_S\epsilon_i \sim N\left(0, \pi_S\Sigma_\epsilon\pi_S^\top\right)$, yielding the moment generating function

$$m_S(t) = E\left\{\exp\left(t^\top\pi_S\epsilon_i\right)\right\} = \exp\left(\frac{1}{2}t^\top\pi_S\Sigma_\epsilon\pi_S^\top t\right).$$

Consequently,

$$E\left(\pi_S X_i^* | X_i\right) = \pi_S X_i \tag{6.6}$$

and

$$E\left\{\exp\left(\beta_{x,S}^{\top}\pi_S X_i^* - \frac{\beta_{x,S}^{\top}\pi_S\Sigma_\epsilon\pi_S^{\top}\beta_{x,S}}{2}\right)\middle| X_i\right\} = \exp\left(\beta_{x,S}^{\top}\pi_S X_i\right). \tag{6.7}$$

Define

$$\begin{aligned}\ell_{C,S}^* &= \sum_{i=1}^{n}\Big[\delta_i\log\lambda_0(y_i) + \delta_i\left(\left(\pi_S x_i^*\right)^{\top}\beta_{x,S} + z_i^{\top}\beta_{z,S}\right)\\ &- \left\{\Lambda_0(y_i) - \Lambda_0(a_i)\right\}\exp\left\{\left(\pi_S x_i^*\right)^{\top}\beta_{x,S} + z_i^{\top}\beta_{z,S} - \frac{\beta_{x,S}^{\top}\pi_S\Sigma_\epsilon\pi_S^{\top}\beta_{x,S}}{2}\right\}\Big]. \end{aligned}\tag{6.8}$$

Then combining (6.6) and (6.7) yields that

$$E\left(\ell_{C,S}^*|X_i, Z_i\right) = \ell_{C,S};$$

this property ensures that working with the function $\ell_{C,S}^*$ allows us to recover the information carried by $\ell_{C,S}$; $\ell_{C,S}^*$ is computable since all the relevant variables have available measurements but $\ell_{C,S}$ is not due to its dependence on X.

Furthermore, with $(\beta_{x,S}^{\top}, \beta_{z,S}^{\top})^{\top}$ fixed, maximizing (6.8) with respect to $\lambda_0(y_i)$, we derive the estimated cumulative baseline function as

$$\widehat{\Lambda}_{0,S}(t) = \int_0^t \frac{\frac{1}{n}\sum_{i=1}^{n} dN_i(u)}{m_S^{-1}(\beta_{x,S})G_S^{(0)}\left(u, \beta_{x,S}, \beta_{z,S}\right)}, \tag{6.9}$$

where $N_i(u) = I\left(Y_i \leqslant u\right)$, $Y_i(u) = I\left(A_i \leqslant u \leqslant Y_i\right)$, and

$$G_S^{(0)}\left(u, \beta_{x,S}, \beta_{z,S}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i(u)\exp\left\{\left(\pi_S x_i^*\right)^{\top}\beta_{x,S} + z_i^{\top}\beta_{z,S}\right\}. \tag{6.10}$$

Finally, combining (6.9) and (6.8), we define

$$\begin{aligned}\widehat{\ell}_{C,S}^* &= \sum_{i=1}^{n}\Big[\delta_i\log\widehat{\lambda}_{0,S}(y_i) + \delta_i\left(\left(\pi_S x_i^*\right)^{\top}\beta_{x,S} + z_i^{\top}\beta_{z,S}\right)\\ &- \left(\widehat{\Lambda}_{0,S}(y_i) - \widehat{\Lambda}_{0,S}(a_i)\right)\exp\left\{\left(\pi_S x_i^*\right)^{\top}\beta_{x,S} + z_i^{\top}\beta_{z,S} - \frac{\beta_{x,S}^{\top}\pi_S\Sigma_\epsilon\pi_S^{\top}\beta_{x,S}}{2}\right\}\Big], \end{aligned}\tag{6.11}$$

on which the inference is based, as discussed in the following subsection.

6.2.2 Augmented Pseudo-Likelihood Estimation

Similarly, the formulation of the marginal likelihood (6.3) for the candidate model S is

$$L_{M,S} = \prod_{i=1}^{n} \frac{\mathcal{F}(a_i|\pi_S x_i, z_i) dH(a_i)}{\int_0^\infty \mathcal{F}(\alpha|\pi_S x_i, z_i) dH(\alpha)}, \tag{6.12}$$

where $\mathcal{F}(a_i|\pi_S x_i, z_i) = \exp\left[-\Lambda_0(a_i) \exp\left\{(\pi_S x_i)^\top \beta_{x,S} + z_i^\top \beta_{z,S}\right\}\right]$. Noting that the marginal likelihood (6.12) involves the unobserved covariate X_i, we now construct a modified version of (6.12) to address the measurement error effects.

Let μ_X and Σ_X be the mean vector and variance-covariance matrix of X_i, respectively. Let $X_{iS}^* = \pi_S X_i^*$ as in (6.6), then model (6.5) gives that $X_{iS}^* = \pi_S X_i + \pi_S \epsilon_i$ with $\pi_S \epsilon_i \sim N\left(0, \pi_S \Sigma_\epsilon \pi_S^\top\right)$, yielding that

$$E(\pi_S X_i | X_{i,S}^* = x_{i,S}^*) = \pi_S \mu_X + \left(\Sigma_{X_S^*} - \Sigma_{\epsilon,S}\right)^\top \Sigma_{X_S^*}^{-1} (x_{i,S}^* - \mu_{X_S^*}), \tag{6.13}$$

where $\Sigma_{\epsilon,S} = \pi_S \Sigma_\epsilon \pi_S^\top$, and $\mu_{X_S^*}$ and $\Sigma_{X_S^*}$ represent the mean and covariance matrix of $X_{i,S}^*$, respectively.

We let $\widetilde{x}_{i,S}$ denote (6.13) for ease of notation. Using the method of moments, (6.13) is estimated by

$$\widehat{x}_{i,S} = \widehat{\mu}_{X_S^*} + \left(\widehat{\Sigma}_{X_S^*} - \Sigma_{\epsilon,S}\right)^\top \Sigma_{X_S^*}^{-1} \left(x_{i,S}^* - \widehat{\mu}_{X_S^*}\right) \tag{6.14}$$

with $\widehat{\mu}_{X_S^*} = \frac{1}{n} \sum_{i=1}^{n} x_{i,S}^*$ and $\widehat{\Sigma}_{X_S^*} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,S}^* - \widehat{\mu}_{X_S^*})(x_{i,S}^* - \widehat{\mu}_{X_S^*})^\top$.

As a result, replacing $\pi_S x_i$ with $\widehat{x}_{i,S}$ in likelihood function (6.12) gives

$$L_{M,S}^* = \prod_{i=1}^{n} \frac{\mathcal{F}(a_i|\widehat{x}_{i,S}, z_i) dH(a_i)}{\int_0^\infty \mathcal{F}(\alpha|\widehat{x}_{i,S}, z_i) dH(\alpha)}, \tag{6.15}$$

where $\mathcal{F}(a_i|\widehat{x}_{i,S}, z_i) = \exp\left\{-\Lambda_0(a_i) \exp\left(\widehat{x}_{i,S}^\top \beta_{x,S} + z_i^\top \beta_{z,S}\right)\right\}$.

To use (6.15) for inference about β, we need to estimate the distribution function $H(\cdot)$. Directly applying the kernel estimation (Silverman 1978) to the observed truncation times to estimate $dH(\cdot)$ is not suitable since the observed truncation times form a biased sample. Instead, we use the nonparametric maximum likelihood estimator (NPMLE) (e.g., Wang 1991) to estimate the distribution function $H(\cdot)$ of $\widetilde{A}$.

For a fixed parameter β, the NPMLE of $H(a)$ based on a candidate model S in (6.15) is given by

$$\widehat{H}_S(a) = \left(\sum_{i=1}^{n} \frac{1}{\widehat{\mathcal{F}}(a_i|\widehat{x}_{i,S}, z_i)} \right)^{-1} \sum_{i=1}^{n} \frac{I(a_i \le a)}{\widehat{\mathcal{F}}(a_i|\widehat{x}_{i,S}, z_i)},$$

where $\widehat{\mathcal{F}}(a_i|\widehat{x}_{i,S}, z_i) = \exp\left\{-\widehat{\Lambda}_{0,S}(a_i) \exp\left(\widehat{x}_{i,S}^\top \widehat{\beta}_{x,CS} + z_i^\top \widehat{\beta}_{z,CS}\right)\right\}$, $\widehat{\Lambda}_{0,S}(\cdot)$ is given by (6.9), and $\widehat{\beta}_{CS} = \left(\widehat{\beta}_{x,CS}^\top, \widehat{\beta}_{z,CS}^\top\right)^\top$ is determined by $\widehat{\beta}_{CS} = \underset{\beta}{\operatorname{argmax}}\, \widehat{\ell}^*_{C,S}$.

Then replacing $H(a)$ in (6.15) with $\widehat{H}_S(a)$ gives the $\widehat{L}^*_{M,S}$, and letting $\widehat{\ell}^*_{M,S} = \log\left(\widehat{L}^*_{M,S}\right)$ gives

$$\begin{aligned} \widehat{\ell}^*_{M,S} &= \sum_{i=1}^{n} \log\left\{d\widehat{H}_S(a_i)\right\} - \sum_{i=1}^{n} \widehat{\Lambda}_{0,S}(a_i) \exp\left(\widehat{x}_{i,S}^\top \beta_x + z_i^\top \beta_z\right) \\ &- \sum_{i=1}^{n} \log\left[\int_0^\infty \exp\left\{-\widehat{\Lambda}_{0,S}(\alpha) \exp\left(\widehat{x}_{i,S}^\top \beta_x + z_i^\top \beta_z\right)\right\} d\widehat{H}_S(\alpha)\right]. \end{aligned} \tag{6.16}$$

Finally, the model parameter β_S for the candidate model S can be estimated by

$$\widehat{\beta}_S = \left(\widehat{\beta}_{x,S}^\top, \widehat{\beta}_{z,S}^\top\right)^\top = \underset{\beta_S}{\operatorname{argmax}} (\widehat{\ell}^*_{C,S} + \widehat{\ell}^*_{M,S}).$$

Immediately, when $|S| = p$, i.e., all the variables $\{X_i, Z_i\}$ are included in the model, we have that $\pi_S = I_{p\times p}$ and hence $\pi_S X_i^* = X_i^*$, and therefore, (6.11) and (6.16), respectively, become

$$\begin{aligned} \widehat{\ell}^*_C &= \sum_{i=1}^{n} \Big[\delta_i \log \widehat{\lambda}_0(y_i) + \delta_i \left(x_i^{*\top} \beta_x + z_i^\top \beta_z\right) \\ &- \left\{\widehat{\Lambda}_0(y_i) - \widehat{\Lambda}_0(a_i)\right\} \exp\left\{x_i^{*\top} \beta_x + z_i^\top \beta_z - \frac{\beta_x^\top \Sigma_\epsilon \beta_x}{2}\right\}\Big], \end{aligned} \tag{6.17}$$

and

$$\begin{aligned} \widehat{\ell}^*_M &= \sum_{i=1}^{n} \log\left\{d\widehat{H}(a_i)\right\} - \sum_{i=1}^{n} \widehat{\Lambda}_0(a_i) \exp\left(\widehat{x}_i^\top \beta_x + z_i^\top \beta_z\right) \\ &- \sum_{i=1}^{n} \log\left[\int_0^\infty \exp\left\{-\widehat{\Lambda}_0(\alpha) \exp\left(\widehat{x}_i^\top \beta_x + z_i^\top \beta_z\right)\right\} d\widehat{H}(\alpha)\right], \end{aligned} \tag{6.18}$$

where

$$\widehat{H}(a) = \left(\sum_{i=1}^{n} \frac{1}{\widehat{\mathcal{F}}(a_i|\widehat{x}_i, z_i)} \right)^{-1} \sum_{i=1}^{n} \frac{I(a_i \leq a)}{\widehat{\mathcal{F}}(a_i|\widehat{x}_i, z_i)},$$

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\frac{1}{n}\sum_{i=1}^{n} dN_i(u)}{m^{-1}(\beta_x) G^{(0)}(u, \beta_x, \beta_z)} \tag{6.19}$$

with

$$G^{(0)}(u, \beta_x, \beta_z) = \frac{1}{n} \sum_{i=1}^{n} Y_i(u) \exp\left\{ \left(x_i^{*\top}, z_i^{\top} \right) \begin{pmatrix} \beta_x \\ \beta_z \end{pmatrix} \right\}, \tag{6.20}$$

and

$$\widehat{x}_i = \widehat{\mu}_{X^*} + \left(\widehat{\Sigma}_{X^*} - \Sigma_\epsilon \right)^{\top} \Sigma_{X^*}^{-1} (x_i^* - \widehat{\mu}_{X^*}) \tag{6.21}$$

with $\widehat{\mu}_{X^*} = \frac{1}{n}\sum_{i=1}^{n} x_i^*$ and $\widehat{\Sigma}_{X^*} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i^* - \widehat{\mu}_{X^*})(x_i^* - \widehat{\mu}_{X^*})^{\top}$. Consequently, the estimator of β based on the full dataset is

$$\widehat{\beta}_{full} = \left(\widehat{\beta}_{x,full}^{\top}, \widehat{\beta}_{z,full}^{\top} \right)^{\top} = \underset{\beta}{\operatorname{argmax}} \left(\widehat{\ell}_C^* + \widehat{\ell}_M^* \right).$$

6.3 Focused Information Criterion and Model Averaging

In this section, we first examine the asymptotic properties for the estimators derived from different candidate models. We then define the *focus parameter* and base on it to introduce the selection criterion for a suitable model. Finally, we establish large sample properties of model averaging estimators.

6.3.1 Asymptotic Results for A Candidate Model

Given a candidate model S, we define

$$\begin{aligned}\ell^*_{P,S} &= \sum_{i=1}^{n}\Bigg[\delta_i\left\{(\pi_S x_i^*)^\top\beta_{x,S}+z_i^\top\beta_{z,S}\right\}+\frac{1}{2}\delta_i\log\left\{m_S(\beta_{x,S})\right\}\\ &\quad -\delta_i\log\left\{\sum_{j=1}^{n}\exp\left((\pi_S x_i^*)^\top\beta_{x,S}+z_i^\top\beta_{z,S}\right)I(a_j\le y_i\le y_j)\right\}\Bigg],\end{aligned}\tag{6.22}$$

which is related to $\ell^*_{C,S}$ in (6.8) so that $\ell^*_{C,S}-\ell^*_{P,S}$ is free of β. We let $\ell^*_{R,S}$ denote this difference, i.e., $\ell^*_{C,S}=\ell^*_{P,S}+\ell^*_{R,S}$; the relevant detail is available in Wang et al. (1993). Similarly, the partial log-likelihood function under the full model can be derived as

$$\begin{aligned}\ell^*_{P} &= \sum_{i=1}^{n}\Bigg[\delta_i\left\{x_i^{*\top}\beta_x+z_i^\top\beta_z\right\}+\frac{1}{2}\delta_i\log\left\{m(\beta_x)\right\}\\ &\quad -\delta_i\log\left\{\sum_{j=1}^{n}\exp\left(x_i^{*\top}\beta_x+z_i^\top\beta_z\right)I(a_j\le y_i\le y_j)\right\}\Bigg],\end{aligned}\tag{6.23}$$

which is related to (6.17) in that their difference is free of β.

Let $U_P(\beta_x,\beta_z)=\frac{\partial\ell^*_P}{\partial\beta}$, $U_M(\beta_x,\beta_z)=\frac{\partial\widehat{\ell}^*_M}{\partial\beta}$, $U_{P,S}(\beta_x,\beta_z)=\frac{\partial\ell^*_{P,S}}{\partial\beta}$ and $U_{M,S}(\beta_x,\beta_z)=\frac{\partial\widehat{\ell}^*_{M,S}}{\partial\beta}$ be pseudo-score functions, where ℓ^*_P, $\widehat{\ell}^*_M$, $\ell^*_{P,S}$ and $\widehat{\ell}^*_{M,S}$ are determined by (6.23), (6.18), (6.22) and (6.16), respectively. The following two lemmas present the relationship between the candidate model S and the full model.

Lemma 6.3.1 *For any candidate model S, let $\Sigma_{X^*_S}$ be the covariance matrix of X^*_S and let Σ_{X^*} be the covariance matrix of X^*. Then*

$$\pi_S^\top\Sigma^{-1}_{X^*_S}\pi_S=\Sigma^{-1}_{X^*}.$$

Lemma 6.3.2 *Under regularity conditions in Appendix E.1, the following results hold for any candidate model S,*

(a) $U_{P,S}(0,\beta_{z0})=\Pi_S U_P(0,\beta_{z0})$;

(b) $U_{M,S}(0,\beta_{z0})=\Pi_S U_M(0,\beta_{z0})$.

These two lemmas are useful for providing the following asymptotic results. In particular, Lemma 6.3.2 describes the connection of the pseudo-score functions under candidate model S and the full model; the pseudo-score function under the candidate model S can be expressed as the product of the projection matrix Π_S and the pseudo-score functions under the full model. Thus, Lemma 6.3.2 allows us to focus on deriving the asymptotic results for the full model (i.e., Theorem 6.3.1 (a)); the asymptotic results for the candidate model S (i.e., Theorem 6.3.1 (b)) can then be immediately derived from Lemma 6.3.2.

Theorem 6.3.1 *Under regularity conditions in Appendix E.1, we have that as $n \to \infty$,*

(a) under the full model,

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,full} \\ \widehat{\beta}_{z,full} - \beta_{z0} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \eta \\ 0 \end{pmatrix},\ \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right);$$

(b) under the candidate model S,

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} \xrightarrow{d} N\left(\mathcal{A}_S^{-1}\Pi_S\mathcal{A}\begin{pmatrix} \eta \\ 0 \end{pmatrix},\ \mathcal{A}_S^{-1}\mathcal{B}_S\mathcal{A}_S^{-1}\right),$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{A}_S$ and $\mathcal{B}_S$ are defined in Appendix E.2.3.

Theorem 6.3.2 *Under regularity conditions in Appendix E.1, we have that under the candidate model S, as $n \to \infty$,*

$$\sqrt{n}\left\{\widehat{\Lambda}_{0,S}(t) - \Lambda_0(t)\right\} \xrightarrow{d} \mathcal{V}(t) - \begin{pmatrix} F_{x,S}(t) \\ F_z(t) \end{pmatrix}^\top W_S + F_x(t)^\top \eta,$$

where $\mathcal{V}(t)$, $F_{x,S}(t)$,$F_x(t)$ and $F_z(t)$ are given in Appendix E.2.4, η is the parameter defined in Section 6.1.2, and W_S represents a random variable whose distribution is identical to the limiting distribution described in Theorem 6.3.1 (b).

6.3.2 Focused Parameter and Asymptotic Results

Rather than examining the model parameters individually, in applications we are often interested in their combined forms or functions of those parameters. To facilitate such

settings, we let $\mu = \mu(\beta_x, \beta_z, \Lambda_0(\cdot))$ be a scalar function of parameter $\beta = \left(\beta_x^\top, \beta_z^\top\right)^\top$ and function $\Lambda_0(t)$. The new parameter μ plays the role of using a simple *scalar* measure to express certain combined information of the original multi-dimensional parameters; it is called the *focus* parameter (Claeskens and Hjort 2003, 2008; Hjort and Claeskens 2006). The choice of the function $\mu(\cdot)$ is often driven by the nature of individual problems (to be discussed in Section 6.3.3). In contrast to the notation $\beta_0 = (\frac{\eta^\top}{\sqrt{n}}, \beta_{z0}^\top)$ defined in Section 6.1.2, we let $\mu_{true} = \mu\left(\frac{\eta}{\sqrt{n}}, \beta_{z0}, \Lambda_0(\cdot)\right)$ denote the true value of the focus parameter μ. By the invariance property of the maximum likelihood estimator, $\widehat{\mu}_S = \mu\left(\widehat{\beta}_{x,S}, \widehat{\beta}_{z,S}, \widehat{\Lambda}_0(\cdot)\right)$ can be taken as the estimated focus parameter corresponding to the candidate model S. We comment that although the density function $h(\cdot)$ of left truncation time is unknown, we do not include it when defining the focus parameter.

For $\mathcal{A}$ and $\mathcal{B}$ in Theorem 6.3.1, we express them as block matrices according to the dimension of the covariates X_i and Z_i: $\mathcal{A} = \begin{pmatrix} A_{xx} & A_{xz} \\ A_{zx} & A_{zz} \end{pmatrix}$ and $\mathcal{B} = \begin{pmatrix} B_{xx} & B_{xz} \\ B_{zx} & B_{zz} \end{pmatrix}$. Let $\mathcal{A}^{-1} = \begin{pmatrix} A^{xx} & A^{xz} \\ A^{zx} & A^{zz} \end{pmatrix}$ denote the inverse matrix of $\mathcal{A}$. We now present the asymptotic properties of the focus parameters whose proof is placed in Appendix E.2.5.

Theorem 6.3.3 *Assume that the conditions in Theorem 6.3.1 hold and consider the candidate model S.*

(a) if the focus parameter $\mu = \mu(\beta_x, \beta_z)$ is the function of parameter β alone, then as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right) \xrightarrow{d} \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M + \omega^\top \left\{\eta - (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\};$$

(b) if the focus parameter $\mu = \mu(\beta_x, \beta_z, \Lambda_0(t))$ is the function of parameter β and cumulative baseline hazard function, then as $n \to \infty$,

$$\begin{aligned} \sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right) \xrightarrow{d} & \frac{\partial \mu}{\partial \Lambda_0} \mathcal{V}(t) + \left(\frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t)\right)^\top A_{zz}^{-1} M \\ & + (\omega + \kappa)^\top \left\{\eta - (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\}, \end{aligned}$$

where $\mathbb{H}_S = (A^{xx})^{-1/2} \pi_S^\top \left\{\pi_S (A^{xx})^{-1} \pi_S^\top\right\}^{-1} \pi_S (A^{xx})^{-1/2}$, $\omega = \frac{\partial \mu}{\partial \beta_x} - A_{zx}^\top A_{zz}^{-1} \frac{\partial \mu}{\partial \beta_z}$, $\kappa = \frac{\partial \mu}{\partial \Lambda_0} F_x(t) - A_{zx}^\top A_{zz}^{-1} \frac{\partial \mu}{\partial \Lambda_0} F_z(t)$, $\mathcal{U} = \eta + \mathcal{W}$, $\mathcal{W} = A^{xx} J - A^{xx} A_{xz} A_{zz}^{-1} M$, and M and J are random variables having the distributions $N(0, B_{zz})$ and $N(0, B_{xx})$, respectively.

6.3.3 Practical Settings and Focus Information Criterion

In this subsection, we illustrate the choice of the focus parameters using examples which are pertinent to the hazard ratio and the survivor function, the quantities commonly used in survival analysis. We further use these examples to present the focus information criterion (FIC) for model selection.

Setting 1: The hazard ratio.
Under the Cox model (6.1), the hazard ratio for $V = v_0$ to $V = v_0 + \mathbf{1}_d$ is

$$\frac{\lambda\left(t|v_0 + \mathbf{1}_d\right)}{\lambda\left(t|v_0\right)} = \exp\left(\mathbf{1}_d^\top \beta\right),$$

where $\mathbf{1}_d$ is the d-dimensional unit vector with $d = p + q$, and v_0 is a value of V. In this case, the focus parameter can be taken as

$$\mu = \mu\left(\beta\right) = \exp\left(\mathbf{1}_d^\top \beta\right). \tag{6.24}$$

The focus parameter μ gives us a single-valued measure which describes the change of the hazard function if every covariate is changed by 1 unit. We now discuss the FIC based on the focus parameter (6.24). The main idea of the FIC is to first work out the mean squared error (MSE) for the estimator of the focus parameter derived from each candidate model, and then determine the final model by the smallest MSE. To use Theorem 6.3.3 (a) for this purpose, we present the following lemma.

Lemma 6.3.3 *Under the conditions in Appendix E.1, we have*

$$\mathcal{W} \sim N\left(0, \sigma_{xx}\right),$$

where $\mathcal{W}$ is given in Theorem 6.3.3 (b), and σ_{xx} is the asymptotic covariance matrix of $\widehat{\beta}_{x,full}$ in Theorem 6.3.1 (a) whose expression is left in Appendix E.2.7.

Combining Theorem 6.3.3 (a) and Lemma 6.3.3 gives the bias and the variance of $\widehat{\mu}_S$:

$$\widehat{\mu}_S - \mu_{true} = \omega^\top \left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\eta$$

and

$$\text{var}\left(\widehat{\mu}_S\right) = \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} B_{zz} A_{zz}^{-1}\left(\frac{\partial \mu}{\partial \beta_z}\right) + \omega^\top (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\sigma_{xx}(A^{xx})^{-1/2}\mathbb{H}_S(A^{xx})^{1/2}\omega.$$

Let $\Phi_S = (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}$. Then the MSE of $\widehat{\mu}_S$ is derived as

$$\begin{aligned} E\left[(\widehat{\mu}_S - \mu_{true})^2\right] &= \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} B_{zz} A_{zz}^{-1} \left(\frac{\partial \mu}{\partial \beta_z}\right) \\ &\quad + \omega^\top \left\{ (I_{p\times p} - \Phi_S)\, \eta\eta^\top \, (I_{p\times p} - \Phi_S)^\top + \Phi_S \sigma_{xx} \Phi_S^\top \right\} \omega. \end{aligned} \tag{6.25}$$

The first term of (6.25) does not depend on the candidate model S, so to make the comparison among different candidate models be focused, we drop this term and simply let the second term of (6.25) reflect the MSE of $\widehat{\mu}_S$; we then define

$$FIC_S = \omega^\top \left\{ (I_{p\times p} - \Phi_S)\, \eta\eta^\top \, (I_{p\times p} - \Phi_S)^\top + \Phi_S \sigma_{xx} \Phi_S^\top \right\} \omega \tag{6.26}$$

to be the focus information criterion (FIC) for model S and the focus parameter (6.24).

To use (6.26), $\eta\eta^\top$ needs to be estimated. By Theorem 6.3.1 and Lemma 6.3.3, $\sqrt{n}\widehat{\beta}_{x,full} = \widehat{\eta} \sim N(\eta, \sigma_{xx})$, thus $E\left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top\right) = \sigma_{xx} + \eta\eta^\top$, suggesting that $n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx}$ is an asymptotically unbiased estimator of $\eta\eta^\top$. Consequently, the FICs in (6.26) is estimated by

$$\widehat{FIC}_S = \widehat{\omega}_1^\top \left\{ \left(I_{p\times p} - \widehat{\Phi}_S\right) \left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx}\right) \left(I_{p\times p} - \widehat{\Phi}_S\right)^\top + \widehat{\Phi}_S \widehat{\sigma}_{xx} \widehat{\Phi}_S^\top \right\} \widehat{\omega}_1, \tag{6.27}$$

where $\widehat{\Phi}_S = \widehat{A}_{xx}^{-1/2} \widehat{H}_S \widehat{A}_{xx}^{1/2}$, $\widehat{\omega}_1 = \frac{\partial \mu(\widehat{\beta}_S)}{\partial \beta_x} - \widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \frac{\partial \mu(\widehat{\beta}_S)}{\partial \beta_z}$ and $\widehat{\sigma}_{xx}$ is the estimated asymptotic covariance matrix of $\widehat{\beta}_{x,full}$.

Setting 2: Covariate effects either β_x or β_z but not both.
In contrast to the hazard ratio, sometimes our interest focuses on either β_x or β_z but not both. When we are interested in β_x alone, the focus parameter is set as $\mu = \beta_x$ and when β_z is of prime interest, we take μ as β_z.

Similar to the derivations for Setting 1, we derive that the estimated FIC for $\mu = \beta_x$ under the candidate model S is given by

$$\widehat{FIC}_S = \mathbf{1}_p^\top \left\{ \left(I_{p\times p} - \widehat{\Phi}_S\right) \left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx}\right) \left(I_{p\times p} - \widehat{\Phi}_S\right)^\top + \widehat{\Phi}_S \widehat{\sigma}_{xx} \widehat{\Phi}_S^\top \right\} \mathbf{1}_p,$$

and

$$\begin{aligned} \widehat{FIC}_S &= \left(\widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \mathbf{1}_q\right)^\top \left\{ \left(I_{p\times p} - \widehat{\Phi}_S\right) \left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx}\right) \left(I_{p\times p} - \widehat{\Phi}_S\right)^\top \right. \\ &\quad \left. + \widehat{\Phi}_S \widehat{\sigma}_{xx} \widehat{\Phi}_S^\top \right\} \left(\widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \mathbf{1}_q\right), \end{aligned}$$

for $\mu = \beta_z$ under the candidate model S.

Setting 3: The cumulative baseline hazard function.
In some applications, as discussed by Hjort and Claeskens (2006), the cumulative baseline hazard function $\Lambda_0(\cdot)$ is of prime interest, and in this case the focus parameter μ is set as $\Lambda_0(t_0)$ for some time point, say t_0, of interest.

Applying Theorem 6.3.3 (b) with $\omega = 0$ and $\kappa = F_x(t_0) - A_{zx}^\top A_{zz}^{-1} F_z(t_0)$, we can work out the MSE of $\widehat{\mu}_S$ for the candidate model S. Similar to (6.27), the FIC for μ under the candidate model S is estimated by

$$\widehat{FIC}_S = \widehat{\kappa}_2^\top \left\{ \left(I_{p\times p} - \widehat{\Phi}_S \right) \left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx} \right) \left(I_{p\times p} - \widehat{\Phi}_S \right)^\top + \widehat{\Phi}_S \widehat{\sigma}_{xx} \widehat{\Phi}_S^\top \right\} \widehat{\kappa}_2, \quad (6.28)$$

where $\widehat{\kappa}_2 = \frac{\partial \mu(\widehat{\Lambda}_{0,S}(t_0))}{\partial \Lambda_0} \widehat{F}_x(t_0) - \widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \frac{\partial \mu(\widehat{\Lambda}_{0,S}(t_0))}{\partial \Lambda_0} \widehat{F}_z(t_0)$, and $\widehat{\Phi}_S$ and $\widehat{\sigma}_{xx}$ are the same as described in Setting 1.

Setting 4: The survivor function.
Under the Cox model (6.1), the survivor function

$$\mathcal{F}(t|v) = \exp\left\{ -\Lambda_0(t) \exp\left(v^\top \beta \right) \right\}$$

is a semi-parametric function since it involves both parameter β and the unspecified function $\Lambda_0(\cdot)$. In applications, we are often interested in the survival information at certain time point, say t_0. In this situation, we take the focus parameter to be

$$\mu = \mu(\beta, \Lambda_0(t_0)) = \exp\left\{ -\Lambda_0(t_0) \exp\left(v_0^\top \beta \right) \right\}$$

for some given covariate value v_0.

Again, Theorem 6.3.3 (b) can be used to derive the MSE of $\widehat{\mu}_S$. Specifically, the bias and the MSE of $\widehat{\mu}_S$ are

$$\widehat{\mu}_S - \mu_{true} = (\omega + \kappa)^\top \left\{ I_{p\times p} - (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \right\} \eta$$

and

$$\begin{aligned} E\left[(\widehat{\mu}_S - \mu_{true})^2 \right] &= \left(\frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t_0) \right)^\top A_{zz}^{-1} B_{zz} A_{zz}^{-1} \left(\frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t_0) \right) \\ &\quad + (\omega + \kappa)^\top \left\{ (I_{p\times p} - \Phi_S) \eta\eta^\top (I_{p\times p} - \Phi_S)^\top + \Phi_S \sigma_{xx} \Phi_S^\top \right\} (\omega + \kappa), \end{aligned}$$

respectively. Similar to the discussion for (6.27), we drop those quantities which are unrelated to S and replace $\eta\eta^\top$ by its asymptotically unbiased estimator, and then we obtain an estimate of the FIC for the candidate model S:

$$\begin{aligned}\widehat{FIC}_S &= (\widehat{\omega}_3 + \widehat{\kappa}_3)^\top \left\{ \left(I_{p\times p} - \widehat{\Phi}_S \right) \left(n\widehat{\beta}_{x,full}\widehat{\beta}_{x,full}^\top - \widehat{\sigma}_{xx} \right) \left(I_{p\times p} - \widehat{\Phi}_S \right)^\top \right. \\ &\quad \left. + \widehat{\Phi}_S \widehat{\sigma}_{xx} \widehat{\Phi}_S^\top \right\} (\widehat{\omega}_3 + \widehat{\kappa}_3), \end{aligned} \tag{6.29}$$

where $\widehat{\Phi}_S$ and $\widehat{\sigma}_{xx}$ are the same as described in Setting 1, $\widehat{\kappa}_3 = \frac{\partial \mu(\widehat{\beta}_S, \widehat{\Lambda}_{0,S}(t_0))}{\partial \Lambda_0} \widehat{F}_x(t_0) - \widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \frac{\partial \mu(\widehat{\beta}_S, \widehat{\Lambda}_{0,S}(t_0))}{\partial \Lambda_0} \widehat{F}_z(t_0)$, and $\widehat{\omega}_3 = \frac{\partial \mu(\widehat{\beta}_S), \widehat{\Lambda}_{0,S}(t_0))}{\partial \beta_x} - \widehat{A}_{zx}^\top \widehat{A}_{zz}^{-1} \frac{\partial \mu(\widehat{\beta}_S), \widehat{\Lambda}_{0,S}(t_0))}{\partial \beta_z}$.

These settings cover the scenarios we usually encounter in survival analysis. Different FIC measures are used to reflect different focuses on the performance of various candidate models. The basic principle is to determine the final model based on the smallest MSE, or equivalently, the smallest $\widehat{FIC}_S$ among all the candidate models S. It is expected that with different focus parameters, the resultant final models are usually different from each other.

Alternatively, using AIC or BIC to select the variables for the different focus parameters is also discussed by some authors, e.g., Claeskens and Hjort (2003), Hjort and Claeskens (2006), and Wang et al. (2012), among others. For the candidate model S, the AIC and BIC are defined as

$$AIC_S = 2\left\{ \widehat{\ell}^*_{C,S}(\widehat{\beta}_S) + \widehat{\ell}^*_{M,S}(\widehat{\beta}_S) \right\} - 2|S|$$

and

$$BIC_S = 2\left\{ \widehat{\ell}^*_{C,S}(\widehat{\beta}_S) + \widehat{\ell}^*_{M,S}(\widehat{\beta}_S) \right\} - \log(n)|S|,$$

respectively. The final models are selected by choosing the maximizer of AIC_S and BIC_S, respectively. Apparently, the best candidate model resulted from using AIC_S or BIC_S is the same regardless of different forms of the focus parameters, since both criteria are based on the likelihood function instead of the focus parameters. On the contrary, the FIC method allows us to select the most suitable model with our focus of interest capitalized on. We will numerically compare the performance among AIC, BIC and FIC by simulation studies in Section 6.4.

6.3.4 Frequentist Model Averaging

The preceding development suggests that differently best candidate models can be yielded from different selection criteria with different focus parameters. As discussed by Clyde and George (2004) and Wang et al. (2009), conducting parameter estimation using a specifically selected model is not ideal since the associated uncertainty is ignored. To circumvent this issue, we employ the frequentist model averaging (FMA) method to construct an estimator of μ. The idea is to use the estimators derived from different candidate models to work out a suitable linear combination of them, given by

$$\widehat{\mu}_{ave} = \sum_{S \in \mathcal{S}} w(S|\widehat{\eta})\widehat{\mu}_S,$$

where $\widehat{\mu}_S$ is the estimator of μ derived from the candidate model S, $\widehat{\eta} = \sqrt{n}\widehat{\beta}_{x,full}$, $w(S|\widehat{\eta})$, to be discussed at the end of this section, is a positive random weight which corresponds to the candidate model S and is data-driven (Claeskens and Hjort 2008, p.195), and all the weights are constrained by $\sum_{S \in \mathcal{S}} w(S|\widehat{\eta}) = 1$. Based on Theorem 6.3.3, we derive the asymptotic distributions of $\widehat{\mu}_{ave}$.

Theorem 6.3.4 *Under regularity conditions in Appendix E.1,*

(a) if the focus parameter $\mu = \mu(\beta_x, \beta_z)$ is the function of parameter β_x and β_z alone, then as $n \to \infty$,

$$\sqrt{n}\,(\widehat{\mu}_{ave} - \mu_{true}) \xrightarrow{d} \left(\frac{\partial \mu}{\partial \beta_z}\right)^{\top} A_{zz}^{-1} M + \omega^{\top} \left\{ \mathcal{U} - \sum_{S \in \mathcal{S}} w(S|\mathcal{U})(A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U} \right\};$$

(b) if the focus parameter $\mu = \mu(\beta_x, \beta_z, \Lambda_0(t))$ is the function of parameter β and the cumulative baseline hazard function, then as $n \to \infty$,

$$\begin{aligned} \sqrt{n}\,(\widehat{\mu}_{ave} - \mu_{true}) \quad \xrightarrow{d} \quad & \frac{\partial \mu}{\partial \Lambda_0}\mathcal{V}(t) + \left\{ \frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t) \right\}^{\top} A_{zz}^{-1} M \\ & + (\omega + \kappa)^{\top} \left\{ \mathcal{U} - \sum_{S \in \mathcal{S}} w(S|\mathcal{U})(A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U} \right\}, \end{aligned}$$

where $w(S|\mathcal{U})$ represents the weight to which $w(S|\widehat{\eta})$ converges in distribution.

Several conventional choices of the weights $w(S|\widehat{\eta})$ are available. For instance, using the AIC, Buckland et al. (1997) suggested the so-called *smooth AIC* weight which is proportional to $\exp\left(\frac{1}{2}\text{AIC}_S\right)$, given by

$$w_{aic,S} = \frac{\exp\left(\frac{1}{2}\text{AIC}_S\right)}{\sum_{S'\in\mathcal{S}} \exp\left(\frac{1}{2}\text{AIC}_{S'}\right)}, \tag{6.30}$$

where AIC_S is the AIC score for the candidate model S. Claeskens and Hjort (2008) suggested to replace AIC_S in (6.30) by $\Delta_{\text{AIC},S} = \text{AIC}_S - \max_{S'\in\mathcal{S}} \text{AIC}_{S'}$ to avoid the numeric problem that the denominator of (6.30) can be quite close to zero, a phenomenon that we also observed in our numerical studies, i.e., the weight is

$$w_{aic,S} = \frac{\exp\left(\frac{1}{2}\Delta_{\text{AIC},S}\right)}{\sum_{S'\in\mathcal{S}} \exp\left(\frac{1}{2}\Delta_{\text{AIC},S'}\right)}. \tag{6.31}$$

In contrast, we can use the following weight

$$w_{bic,S} = \frac{\exp\left(\frac{1}{2}\Delta_{\text{BIC},S}\right)}{\sum_{S'\in\mathcal{S}} \exp\left(\frac{1}{2}\Delta_{\text{BIC},S'}\right)}, \tag{6.32}$$

and call it the *smooth BIC* weight. Similarly, with the FIC, the weight is defined as

$$w_{fic,S} = \frac{\exp\left(\frac{1}{2}\frac{\text{FIC}_S}{(\widehat{\omega}-\widehat{\kappa})^\top \widehat{\sigma}_{xx}(\widehat{\omega}-\widehat{\kappa})}\right)}{\sum_{S'\in\mathcal{S}} \exp\left(\frac{1}{2}\frac{\text{FIC}_{S'}}{(\widehat{\omega}-\widehat{\kappa})^\top \widehat{\sigma}_{xx}(\widehat{\omega}-\widehat{\kappa})}\right)}, \tag{6.33}$$

as suggested by Hjort and Claeskens (2006) and Claeskens and Hjort (2008).

6.4 Numerical Studies

In this section, we first conduct simulation studies to assess the performance of the proposed estimators, and then implement the methods to analyze a real dataset.

6.4.1 Simulation Studies

For each setting, we run 500 simulations. We examine the cases with the sample size $n = 100$ and $n = 200$, respectively. For the covariates, we generate $\widetilde{X}$ from $N(\mathbf{0}_6, \Sigma_X)$ and $\widetilde{Z}$ from $N(\mathbf{0}_2, \Sigma_Z)$ independently, where

$$\Sigma_X = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}_{6\times 6} \quad \text{and} \quad \Sigma_Z = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}_{2\times 2}.$$

The goal here is to select the important variables in $\widetilde{X}$ and always retain the covariate $\widetilde{Z}$ in the models. Consequently, for the variable selection, there are $2^6 = 64$ candidate models.

The survival time is generated using model (6.1) where the baseline hazard function is set as $\lambda_0(t) = 2t$. More specifically, the failure time is generated by

$$\widetilde{T} = \sqrt{-\exp\left(\widetilde{X}^\top \beta_{x0} + \widetilde{Z}^\top \beta_{z0}\right) \log(1-U)},$$

where U is simulated from the uniform distribution $U(0,1)$, and the true parameter $\beta_0 = \left(\beta_{x0}^\top, \beta_{z0}^\top\right)^\top$ is set as $\beta_{x0} = \frac{\eta}{\sqrt{n}}$ and $\beta_{z0} = (0.6, 0.6)^\top$. We consider three cases with

$$(1)\ \eta = (0,0,0,0,0,0)^\top, \quad (2)\ \eta = (1,1,1,0,0,0)^\top, \quad \text{and } (3)\ \eta = (1,1,1,1,1,1)^\top.$$

Case (1) gives a *null* model, Case (2) indicates that some covariates are not included in the true model, and Case (3) says that the *full* model contains all the covariates.

Let the truncation time $\widetilde{A}$ be generated from the exponential distribution with mean 10. The observed data (A, T, V) are then obtained from $(\widetilde{A}, \widetilde{T}, \widetilde{V})$ using the condition $\widetilde{T} \geq \widetilde{A}$. Independently repeat this data simulation step n times to generate a sample of size n. The censoring variable C is generated from the uniform distribution $U(0, c)$ where c is a constant that is chosen to yield about 50% censoring rate. Consequently, Y and Δ are determined by $Y = \min\{T, A + C\}$ and $\Delta = I(T \leq A + C)$. Therefore, $(Y_i, A_i, \Delta_i, V_i)$ with $i = 1, \cdots, n$ is the sample with size n in the dataset.

Consistent with Section 6.1.2, X_i is the error-prone covariates and X_i^* is the observed variable which is generated from

$$X_i^* = X_i + \epsilon_i,$$

where $\epsilon_i \sim N(\mathbf{0}_6, \Sigma_\epsilon)$, and Σ_ϵ is a diagonal matrix whose diagonal elements are all specified as 0.1 or 0.5.

For the focus parameter, we consider three forms, given by

$$\text{(a) } \mu_{10}(\beta) = \exp\left(\mathbf{1}_6^\top \beta\right), \text{ (b) } \mu_{20}(\Lambda_0) = \Lambda_0(1), \text{ and}$$
$$\text{(c) } \mu_{30}(\beta, \Lambda_0) = \exp\left\{-\Lambda_0(1)\exp\left(\mathbf{1}_6^\top \beta\right)\right\},$$

respectively.

Our interest is to select variables based on different forms of focus parameters using the proposed FIC method. As comparisons, we also apply AIC or BIC to select variables, then use the plug-in method to obtain estimates of the focus parameters. For each simulation setting and different focus parameters, we first estimate the parameters and determine the best model by different selection criteria, AIC, BIC and FIC. Then we compute the estimated focus parameter using the best selected model. Let $\widehat{\mu}_j$ denote the resulting estimated focus parameter for simulation j, where $j = 1, \cdots, 500$; and we compute the square root mean squared error (RMSE) as $\sqrt{500^{-1}\sum_{j=1}^{500}(\widehat{\mu}_j - \mu_0)^2}$, which is used to report the accuracy of the estimated focus parameters. In contrast, we also report the results obtained from the naive method which ignores the measurement error in covariates. The results for the sample size $n = 100$ are reported in Table 6.1 and the results with $n = 200$ are summarized in Table 6.2. Furthermore, model averaging estimators with weights (6.31), (6.32) and (6.33) are also investigated, and the results are displayed under the headings sAIC, sBIC, and sFIC in Tables 6.1 and 6.2.

As expected, the RMSEs for the proposed estimators are smaller than those for the naive estimators regardless of the selection criteria; and the differences become more noticeable as measurement error is more substantial. No matter what estimation method is, either the naive approach or the proposed approach, RMSEs using FIC tends to result in smaller RMSEs than using AIC or BIC under our simulation settings. Furthermore, the model averaging estimators, sAIC, sBIC and sFIC, are comparable to their counterparts, AIC, BIC and FIC, respectively, and sFIC outperforms both sAIC and sBIC under the settings we consider.

6.4.2 Analysis of Worcester Heart Attack Study Data

In this section, we use our methods to analyze the data arising from the Worcester Heart Attack Study (WHAS500). Data were collected over thirteen 1-year periods beginning in

1975 and extending in 2001 on all patients with acute myocardial infarction (MI) admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. Discussed by Hosmer et al. (2008), a survival time was defined as the time since a subject was admitted to the hospital. We are interested in studying survival times of patients who were discharged alive from the hospital. Hence, a selection criterion was imposed that only those subjects who were discharged alive were eligible to be included in the analysis. That is, individuals were not enrolled in the analysis if they died before discharging from the hospital, hence left truncation occurs. With such a criterion, a sample of size 461 was available. In this data set, the censoring rate is 61.8%.

The following covariates are included in our analysis: initial heart rate (X_1), initial systolic blood pressure (X_2), initial diastolic blood pressure (X_3), body mass index (X_4), age (Z_1) and gender (Z_2), and we let $\beta = (\beta_{x_1}, \beta_{x_2}, \beta_{x_3}, \beta_{x_4}, \beta_{z_1}, \beta_{z_2})^\top$ denote the vector of the corresponding parameters formed by model (6.1). Covariates X_1, X_2, X_3 and X_4 are error-prone due to the reasons including inaccurate measurement devices and/or procedures, the biological variability, and temporal variations. Using the notation in Section 6.1.2, we have $p = 4$ and $q = 2$, thus, the number of all possible candidate models is $2^p = 16$. Similar to the settings in Section 6.4.1, we discuss three focus parameters: the hazard ratio (μ_1) with coefficient $\mathbf{1}_6$, the cumulative baseline hazard function (μ_2) at time t_0, and the survivor function (μ_3) at time t_0 with coefficients being empirical means of variables, where t_0 is the median of all the observed values Y_i. Our goal is to select important variables from X_1 to X_4 for different focus parameters, with Z_1 and Z_2 always retained.

We first present the estimators of β under the full model using both the proposed approach discussed in Sections 6.2.1 and 6.2.2, and the naive approach which ignores measurement error. Since this dataset contains no additional information, such as repeated measurement or validation data, for the characterization of the measurement error process, we conduct sensitivity analyses to investigate the measurement error effects. Specifically, let Σ be the sample covariance matrix, and for sensitivity analyses we consider $\Sigma + \Sigma_e$ to be the covariance matrix for the measurement error model (6.4), where Σ_e is the diagonal matrix with diagonal elements being a common value $\sigma_e^2 \in [0, 1]$. The estimation results are shown in Figure 6.1. We can see that as the degree of measurement error changes, the patterns of $\widehat{\beta}_{x_j} (j = 1, 2, 3, 4)$ are fluctuated while $\widehat{\beta}_{z_1}$ and $\widehat{\beta}_{z_2}$ are fairly stable .

To examine the proposed estimators more closely, we focus on $\sigma_e^2 = 0.1^2, 0.5^2$ and 1^2 which represent a minor, moderate and substantial measurement error effect, respectively. Table 6.3 summarizes the estimates, the standard errors (SE) and the p-values of both the proposed and the naive methods. As expected, SEs of the naive estimator are generally smaller than those of the proposed estimator. Both the naive and the proposed methods suggest all the covariates are significant, regardless of measurement error degrees.

Next, we report the variable selection results based on AIC, BIC, and FIC for the three focus parameters discussed in Section 6.4.1 and present the best five candidate models in Table 6.4. Here we use a label, such as "134" to represent that the variables X_1, X_3, and X_4 selected for the model. First, we observe that the best model selected by AIC contains more variables than those by BIC. It is interesting to see that selection results by the naive approach and the proposed approach are similar. Secondly, we report the FIC results for different focus parameters. The FIC approach results in the relatively more parsimonious models, regardless of using both the proposed method or the naive method. In addition, we can see that with a given focus parameter, the variable selection results change as σ_e^2 changes. Variables are differently selected if using different criteria. For example, initial heart rate (X_1) and initial diastolic blood pressure (X_3) are frequently selected by both the AIC and the BIC; while initial systolic blood pressure (X_2), initial diastolic blood pressure (X_3) and body mass index (X_4) are frequently selected by the FIC approach with different focus parameters.

Finally, in Table 6.5 we summarize the results for the estimates of the focus parameters based on the best models and for the model averaging estimators. We can see that the estimated focus parameters by the proposed approach and the naive approach are different even the best candidate models are the same under some specific criteria. The proposed approach yields different estimates of the focus parameters when the measurement error degree σ_e^2 varies.

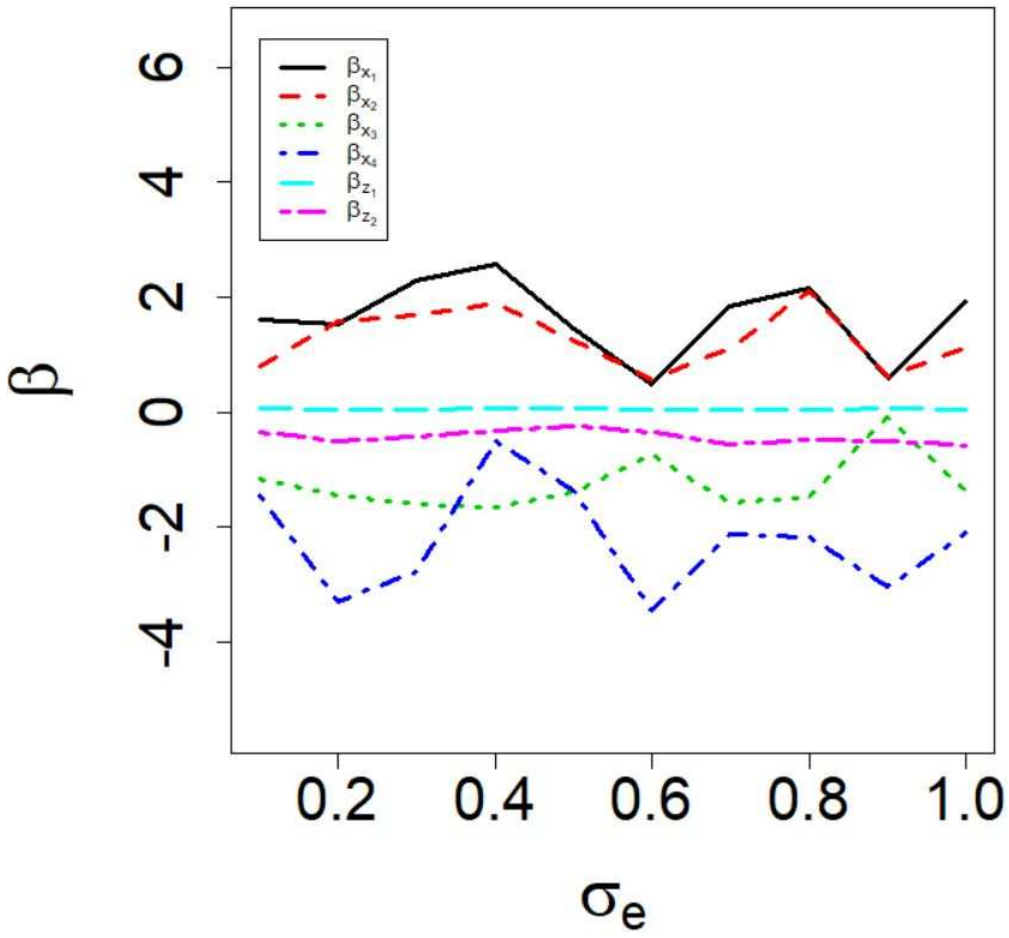


Figure 6.1: Sensitivity of the estimates obtained for Worcester Heart Attack Study Data.

Table 6.1: Simulation results: RMSE of the estimators for focus parameters with $n = 100$

Method	σ_ϵ		β_0	AIC	BIC	FIC	sAIC	sBIC	sFIC
Proposed	0.1	μ_1	(1)	1.576	1.657	1.118	1.287	1.422	1.156
			(2)	1.794	1.527	1.374	1.747	1.261	1.102
			(3)	2.158	1.789	1.406	1.906	1.393	1.392
		μ_2	(1)	0.287	0.267	0.268	0.280	0.236	0.229
			(2)	0.286	0.285	0.272	0.256	0.235	0.231
			(3)	0.262	0.262	0.254	0.261	0.242	0.226
		μ_3	(1)	0.056	0.056	0.048	0.052	0.054	0.046
			(2)	0.036	0.036	0.032	0.033	0.035	0.032
			(3)	0.022	0.022	0.020	0.020	0.022	0.014
	0.5	μ_1	(1)	1.330	1.343	0.889	1.147	1.126	0.929
			(2)	1.882	1.696	1.326	1.276	1.562	1.062
			(3)	1.910	1.774	1.471	1.407	1.665	1.373
		μ_2	(1)	0.304	0.297	0.290	0.297	0.255	0.224
			(2)	0.365	0.365	0.290	0.320	0.280	0.259
			(3)	0.305	0.306	0.302	0296	0.254	0.247
		μ_3	(1)	0.052	0.052	0.046	0.047	0.050	0.045
			(2)	0.033	0.033	0.032	0.033	0.033	0.032
			(3)	0.017	0.017	0.010	0.020	0.020	0.010
Naive	0.1	μ_1	(1)	2.130	1.892	1.744	2.101	1.755	1.697
			(2)	1.920	2.426	1.629	1.840	2.004	1.587
			(3)	2.894	2.256	1.843	2.779	1.805	1.623
		μ_2	(1)	0.306	0.306	0.283	0.333	0.260	0.244
			(2)	0.330	0.329	0.309	0.303	0.267	0.258
			(3)	0.317	0.296	0.305	0.296	0.267	0.259
		μ_3	(1)	0.073	0.073	0.066	0.071	0.072	0.066
			(2)	0.062	0.062	0.055	0.058	0.056	0.053
			(3)	0.063	0.063	0.047	0.056	0.060	0.046
	0.5	μ_1	(1)	1.952	1.768	1.567	1.764	1.519	1.408
			(2)	2.357	1.857	1.478	2.162	1.621	1.335
			(3)	2.329	2.014	1.605	2.196	1.817	1.420
		μ_2	(1)	0.370	0.340	0.311	0.341	0.289	0.277
			(2)	0.417	0.417	0.328	0.381	0.322	0.302
			(3)	0.381	0.340	0.349	0.340	0.283	0.281
		μ_3	(1)	0.067	0.067	0.057	0.064	0.065	0.055
			(2)	0.081	0.081	0.048	0.076	0.081	0.047
			(3)	0.046	0.046	0.041	0.045	0.046	0.041

Table 6.2: Simulation results: RMSE of the estimators for focus parameters with $n = 200$

Method	σ_ϵ		β_0	AIC	BIC	FIC	sAIC	sBIC	sFIC
Proposed	0.1	μ_1	(1)	0.825	0.747	0.727	0.757	0.718	0.686
			(2)	1.156	0.942	0.724	0.973	0.862	0.723
			(3)	1.708	1.477	1.183	1.494	1.356	1.178
		μ_2	(1)	0.210	0.210	0.169	0.208	0.192	0.153
			(2)	0.188	0.180	0.160	0.180	0.167	0.147
			(3)	0.191	0.190	0.171	0.185	0.173	0.152
		μ_3	(1)	0.032	0.032	0.026	0.030	0.030	0.025
			(2)	0.017	0.017	0.014	0.014	0.017	0.014
			(3)	0.025	0.025	0.014	0.022	0.025	0.017
	0.5	μ_1	(1)	0.798	0.768	0.637	0.736	0.713	0.667
			(2)	0.952	0.917	0.817	0.744	0.837	0.733
			(3)	1.317	1.303	1.288	1.250	1.282	1.228
		μ_2	(1)	0.234	0.234	0.179	0.234	0.203	0.156
			(2)	0.203	0.203	0.161	0.203	0.176	0.177
			(3)	0.201	0.201	0.168	0.205	0.184	0.157
		μ_3	(1)	0.028	0.028	0.022	0.026	0.028	0.020
			(2)	0.017	0.017	0.026	0.017	0.017	0.014
			(3)	0.010	0.010	0.010	0.010	0.010	0.010
Naive	0.1	μ_1	(1)	1.276	1.175	1.138	1.180	1.128	1.082
			(2)	1.829	1.187	1.122	1.398	1.494	1.037
			(3)	2.605	1.943	1.249	2.173	1.831	1.249
		μ_2	(1)	0.293	0.276	0.193	0.276	0.241	0.175
			(2)	0.224	0.205	0.185	0.205	0.189	0.184
			(3)	0.216	0.214	0.197	0.214	0.192	0.184
		μ_3	(1)	0.036	0.036	0.032	0.033	0.033	0.030
			(2)	0.035	0.035	0.030	0.032	0.032	0.030
			(3)	0.037	0.036	0.033	0.035	0.035	0.028
	0.5	μ_1	(1)	1.301	1.205	1.197	1.247	1.133	1.107
			(2)	1.214	1.238	0.985	0.932	0.936	0.822
			(3)	1.902	1.431	1.354	1.551	1.338	1.321
		μ_2	(1)	0.293	0.276	0.193	0.276	0.242	0.175
			(2)	0.298	0.268	0.196	0.268	0.235	0.196
			(3)	0.251	0.235	0.187	0.235	0.214	0.179
		μ_3	(1)	0.044	0.044	0.040	0.042	0.044	0.039
			(2)	0.044	0.044	0.039	0.042	0.041	0.039
			(3)	0.041	0.041	0.032	0.040	0.041	0.030

Table 6.3: Sensitivity analyses for Worcester Heart Attack Study Data: estimation results

Method	Variable	Estimate	SE	p-value
Proposed ($\sigma_e^2 = 0.1^2$)	Initial Heart Rate (X_1)	1.538	0.157	1.169e-22
	Initial Systolic Blood Pressure (X_2)	1.570	0.074	6.765e-100
	Initial Diastolic Blood Pressure (X_3)	-1.448	0.174	8.661e-17
	Body Mass Index (X_4)	-3.297	0.011	0.000
	Age (Z_1)	0.039	0.013	0.003
	Gender (Z_2)	-0.521	0.242	0.031
Proposed ($\sigma_e^2 = 0.5^2$)	Initial Heart Rate (X_1)	1.444	0.120	2.374e-33
	Initial Systolic Blood Pressure (X_2)	1.250	0.104	2.816e-33
	Initial Diastolic Blood Pressure (X_3)	-1.405	0.179	4.188e-15
	Body Mass Index (X_4)	-1.370	0.011	0.000
	Age (Z_1)	0.070	0.017	3.828e-05
	Gender (Z_2)	-0.247	0.124	0.046
Proposed ($\sigma_e^2 = 1^2$)	Initial Heart Rate (X_1)	1.933	0.122	1.540e-56
	Initial Systolic Blood Pressure (X_2)	1.102	0.062	1.120e-70
	Initial Diastolic Blood Pressure (X_3)	-1.382	0.152	9.714e-20
	Body Mass Index (X_4)	-2.091	0.016	0.000
	Age (Z_1)	0.049	0.018	0.006
	Gender (Z_2)	-0.577	0.149	0.000
Naive	Initial Heart Rate (X_1)	0.894	0.110	4.391e-16
	Initial Systolic Blood Pressure (X_2)	0.326	0.047	4.029e-12
	Initial Diastolic Blood Pressure (X_3)	-0.834	0.109	1.988e-14
	Body Mass Index (X_4)	-1.694	0.010	0.000
	Age (Z_1)	0.054	0.007	1.217e-14
	Gender (Z_2)	-0.379	0.116	0.001

Table 6.4: Sensitivity analyses for Worcester Heart Attack Study Data: variable selection results

Method	AIC		BIC		FIC - μ_1		FIC - μ_2		FIC - μ_3	
	Variables	Values	Variables	Values	Variables	Values	Variables	Values	Variables	Values
Proposed	1234	-7729.171	13	-7745.705	24	0.054	14	1.244	2	1.091
($\sigma_e^2 = 0.1^2$)	134	-7742.008	134	-7754.408	1234	0.112	134	2.135	23	1.127
	123	-7743.861	123	-7756.261	123	0.465	13	2.414	3	1.371
	13	-7749.721	14	-7757.979	34	0.499	34	2.798	34	1.804
	14	-7751.449	1234	-7761.236	124	0.794	24	3.598	1	2.744
Proposed	1234	-7696.241	13	-7704.508	3	0.335	24	1.389	4	2.330
($\sigma_e^2 = 0.5^2$)	134	-7701.565	134	-7713.966	2	0.574	2	1.487	14	2.765
	123	-7709.508	123	-7717.775	23	1.091	124	2.028	1234	3.095
	13	-7710.878	14	-7723.278	12	3.882	13	2.612	34	3.453
	14	-7716.625	1234	-7729.025	123	3.989	3	2.872	3	3.861
Proposed	1234	-7653.490	13	-7670.024	24	2.310	12	1.041	24	1.057
($\sigma_e^2 = 1^2$)	134	-7659.390	134	-7671.791	34	3.414	34	1.564	3	1.251
	123	-7669.290	123	-7681.066	1234	7.621	124	2.139	34	1.518
	13	-7672.799	14	-7681.691	234	8.271	234	2.289	1	1.791
	14	-7674.820	1234	-7687.220	123	8.544	134	2.804	1234	2.502
Naive	134	-7755.097	134	-7762.497	24	0.709	234	2.601	2	1.228
	123	-7755.013	13	-7764.221	1234	1.304	23	2.632	24	1.840
	13	-7755.954	14	-7766.855	123	1.330	24	2.650	1	2.004
	1234	-7756.657	123	-7767.413	234	1.332	123	2.885	1234	2.175
	14	-7758.589	1	-7768.988	23	1.342	124	3.067	4	3.042

Table 6.5: Sensitivity analyses for Worcester Heart Attack Study Data: estimates of the focus parameters

μ	Method	AIC	sAIC	BIC	sBIC	FIC	sFIC
μ_1	Proposed ($\sigma_e^2 = 0.1^2$)	0.720	0.626	0.257	0.166	0.444	0.347
	Proposed ($\sigma_e^2 = 0.5^2$)	0.772	0.626	0.327	0.351	0.459	0.352
	Proposed ($\sigma_e^2 = 1^2$)	0.781	0.671	0.377	0.327	0.463	0.357
	Naive	0.496	0.692	0.496	0.958	0.245	0.660
μ_2	Proposed ($\sigma_e^2 = 0.1^2$)	0.031	0.051	0.025	0.084	0.037	0.102
	Proposed ($\sigma_e^2 = 0.5^2$)	0.044	0.052	0.015	0.053	0.039	0.104
	Proposed ($\sigma_e^2 = 1^2$)	0.037	0.040	0.036	0.079	0.045	0.109
	Naive	0.018	0.011	0.018	0.008	0.007	0.007
μ_3	Proposed ($\sigma_e^2 = 0.1^2$)	0.887	0.886	0.656	0.883	0.469	0.681
	Proposed ($\sigma_e^2 = 0.5^2$)	0.664	0.721	0.659	0.765	0.496	0.698
	Proposed ($\sigma_e^2 = 1^2$)	0.683	0.692	0.684	0.732	0.539	0.815
	Naive	0.603	0.552	0.603	0.452	0.461	0.407

Chapter 7

Summary and Discussion

In this section, we present the summaries for the previous chapters.

Chapter 2 :

In Chapter 2, we discuss the analysis of graphical models with mismeasurement in variables. We consider three scenarios where error-contaminated variables are only discrete, or continuous, or mixed with both types. We employ the exponential family distribution to facilitate the joint distribution of the variables, which gives a broad class of models useful for many applications. To understand the mismeasurement effects, we derive a lower bound of the asymptotic bias. To correct for the mismeasurement effects, we propose a simulation-based method to derive valid estimation of graphs. The theoretical and the numerical results demonstrate that the proposed methods perform satisfactorily and that the naive analysis commonly yields misleading results.

To highlight the key ideas, we focus our attention on estimation of the network structure and assume the parameters for the mismeasurement models (2.6) and (2.10) to be known. Such an assumption is typically feasible in two circumstances: (i) prior studies provide the information on the degree of mismeasurement, and (ii) we are interested in conducting sensitivity analyses to understand how mismeasurement effects may affect inference results.

In situations where the parameters for the mismeasurement models (2.6) and (2.10) must be estimated, we may utilize the information carried with additional data sources such as repeated measurements or validation subsamples. For instance, with the availability of repeated measurements, estimation of misclassification probabilities can proceed in the same manner that discussed by Yi and He (2017, Section 4),

and Σ_ϵ for the measurement error model (2.6) can be estimated by

$$\widehat{\Sigma}_\epsilon = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(X^{*\mathrm{C}(ij)} - \bar{X}^{*\mathrm{C}(i)}\right)\left(X^{*\mathrm{C}(ij)} - \bar{X}^{*\mathrm{C}(i)}\right)^\top}{\sum_{i=1}^{n}(n_i - 1)},$$

where $X^{*\mathrm{C}(ij)}$ denotes the jth replicate of $X^{\mathrm{C}(i)}$ with $j = 1, \cdots, n_i$, n_i is the number of the replicates for subject i, and $\bar{X}^{*\mathrm{C}(i)} = n_i^{-1}\sum_{j=1}^{n_i} X^{*\mathrm{C}(ij)}$ for $i = 1, \cdots, n$. When validation data are available, one may adapt the discussion of Yi et al. (2015, 2019) to incorporate estimation of the parameters for the mismeasurement models (2.6) and (2.10) into inferential procedures.

Chapter 3 :

While high-dimensional survival data become more accessible and methods of variable selection for survival data have been developed, we still face the challenges induced from survival data with network structured covariates subject to measurement error. Handling such data does not only require more complex modeling but also involve more complicated technical derivations of theoretical results. In Chapter 3, we explore this important problem and propose graphical proportional hazards measurement error models to accommodate high-dimensional survival data with both the network structure and measurement error. We utilize exponential family graphical models to characterize covariate network structures and examine mismeasurement in both continuous and discrete covariates. Our developed inferential methods are justified theoretically, and their finite sample performance is demonstrated to be satisfactory through simulation studies.

Although the development is based on the Cox regression model, extensions to other survival models, such as the accelerated failure time model, the additive hazards model, and transformation models, are possible. The development can be carried out in the same manner with suitable modifications with the partial likelihood score functions replaced by unbiased estimating functions.

Chapter 4 :

Sufficient dimension reduction (SDR) is a useful tool in regression models, which mainly reduces the dimension of variables without losing information of variables. Even though many inference methods have been developed, research gaps still exist. Censored responses and mismeasurement in covariates are ubiquitous and need to be

properly addressed in survival analysis. In Chapter 4, we deal with high-dimensional censored data with measurement error using cumulative distribution models which cover frequently used models in survival analysis. We propose valid inferential procedures to correct the measurement error and estimate the central space. Our developed inferential methods are justified theoretically, and their finite sample performance is demonstrated to be satisfactory through simulation studies.

Several possible extensions are worth exploring. For example, as presented by Li and Yin (2007), the condition of the normal distribution for the covariates X is required when replacing X^* by U, and only such a condition makes the invariance law hold. However, if X does not follow the normal distribution, it is unclear whether or not the invariance law still holds. In Chapter 4, we mainly focus on the case where the covariate X is continuous. If the covariate is discrete, then applying the proposed feature screening method can obtain the active set, but how to correct the measurement error for the discrete variables in the SDR method is still unknown. It is interesting to explore these important research topics.

Chapter 5 :

Although survival analysis has proven useful and many methods have been developed for analyzing survival data with individual features, there has been little work of addressing these features simultaneously in inferential procedures, as noted by Yi and Lawless (2007). In Chapter 5, we develop two estimation methods to handle left-truncated right-censored survival data with measurement error in covariates. We establish asymptotic results for the proposed methods rigorously and explore the issues of robustness and efficiency of the proposed methods. We further demonstrate satisfactory finite sample performance of our methods using simulation studies.

The proposed methods can also accommodate length-biased survival data with covariate measurement error. Length-biased data arise commonly from many fields including epidemiological studies, cancer research, and etiology studies, and many methods have been developed for analysis of such data. However, the validity of these methods is limited due to the key assumption that data must be accurately collected. In application, measurements of the variables usually are error-contaminated. Accommodating the feature of measurement error, our proposed methods generalize the scope of usual methods of handling length-biased survival data.

Chapter 6 :

Left-truncated and right-censored data arise commonly from studies of survival information. Analysis of such survival data is further complicated by other common

features. Typically, error-prone covariates and the presence of unimportant covariates make the analysis more difficult than without these features. In Chapter 6, we develop estimation methods using the FIC criterion to handle left-truncated and right-censored survival data with measurement error in covariates. We implement the model averaging technique to derive more efficient estimators of the focus parameters and establish asymptotic results of the proposed estimators. Numerical studies confirm the satisfactory performance of our proposed methods.

Although our development is carried out for the useful focus parameters described in Section 6.3.3, the scope of our methods is broad. For instance, if the focus parameters are percentiles (e.g., median), one can adapt the development here to accommodate such settings. In our development here, we focus on the case where continuous covariates are subject to measurement error. In some applications, discrete covariates may be subject to measurement error, or there is a mix of error-prone discrete and continuous covariates, it is interesting to further extend our methods to address such problems, and this is our future work.

Two general comments raised by a committee member are (a) if the dimension of parameter is allowed to diverge with the sample size, and (b) if there is the rate of convergence in the theoretical results. The development in this thesis does not consider this two issues. Instead, a fixed dimension p is considered in this thesis. Undoubtedly, it will also be interesting to consider such a problem by allowing p approaches ∞ as $n \to \infty$.

This thesis focuses on developing valid estimation procedures with the emphasis of establishing asymptotic results of the developed estimators, including the consistency and asymptotic distributions. Establishing the convergence rate is worth being explored in the future along the same lines of the current development.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd *International Symposium on Information Theory*, eds by Petrov, N. and Czaki, F., 267-281. Akademiai Kaido, Bydapest.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10, 1100-1120.

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, 21, 1494-1542.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*, 97, 201-209.

Augustin, T. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, 31, 43-50.

Bandara, S., Schlöder, J. P., Eils, R., Bock, H. G., and Meyer, T. (2009). Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Computational Biology*, 5, e1000558. doi:10.1371/journal.pcbi.1000558

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2, 273-277.

Bertrand, A., Legrand, C., Carroll, R. J., De Meester, C., and Van Keilegom, I. (2017). Inference in a survival cure model with mismeasured covariates using a simulation-extrapolation approach. *Biometrika*, 104, 31-50.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (1991). *Measurement Error in Surveys*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, 53, 603-618.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66, 429-436.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC.

Buzas, J. F. (1998). Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference*, 67, 247-257.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, 92, 303-316.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075-1093.

Carroll, R. J., Delaigle, A., and Hall, P. (2007). Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *Journal of the Royal Statistical Society, Series B*, 69, 859-878.

Carroll, R. J., Delaigle, A., and Hall, P. (2009). Nonparametric prediction in measurement error models (with discussion). *Journal of the American Statistical Association*, 104, 993-1014.

Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *The Annals of Statistics*, 23, 161-181.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error model. *Journal of the American Statistical Association*, 91, 242-250.

Carroll, R. J. and Li, K.-C. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*, 87, 1040-1050.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, 99, 736-750.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Model*. CRC Press, New York.

Chen, L.-P. (2018). Semiparametric estimation for the accelerated failure time model with length-biased sampling and covariate measurement error. *Stat*, 7: e209. DOI: 10.1002/sta4.209.

Chen, L.-P. (2019a). Pseudo likelihood estimation for the additive hazards model with data subject to left-truncation and right-censoring. *Statistics and Its Interface*, 12, 135-148.

Chen, L.-P. (2019b). Semiparametric estimation for cure survival model with left-truncated and right-censored data and covariate measurement error. *Statistics and Probability Letters*, 154. DOI: 10.1016/j.spl.2019.06.023

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102, 47-64.

Chen, X., Sheng, W., and Yin, X. (2018). Efficient sparse estimate of sufficient dimension reduction in high dimension. *Technometrics*, 60, 161-168.

Chen, X., Zhang, Y., Chen, X., and Liu, Y. (2019). A simple model-free survival conditional feature screening. *Statistics and Probability Letters*, 146, 156-160.

Cheng, J., Li, T., Levina, E., and Zhu, J. (2017). High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26, 367-378.

Claeskens, G. and Carroll, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 94, 249-265.

Claeskens, G. and Hjort, N. L.(2003). The focused information criterion. *Journal of the American Statistical Association*, 98, 900-945.

Claeskens, G. and Hjort, N. L.(2008). *Model Selection and Model Averaging*. Cambridge University Press, New York.

Clyde, M and George, E. I. (2004). Model uncertainty. *Statistical Science*, 19, 81-94.

Cook, J. R. and Stefaski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314-1328.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89, 177-189.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cook, R. D., Forzani, L., and Rothman, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40, 353-384.

Cook, R. J. and Lawless, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. CRC Press, New York.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. CRC Press.

Dalal, O. and Rajaratnam, B. (2017). Sparse Gaussian graphical model estimation via alternating minimization. *Biometrika*, 104, 379-395.

De Uña-Alvarez, J. (2004). Nonparametric estimation under length-biased sampling and Type I censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, 56, 667-681.

Du, J., Zhang, Z., and Xie, T. (2017). Focus information criterion and model averaging in censored quantile regression. *Metrika*, 80, 547-570.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer, New York.

Fan, J., Feng, Y., and Wu, Y. (2010). Ultrahigh dimensional variable selection for Cox's proportional hazards model. *IMS Collect*, 6, 70-86.

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

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

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society. Series B*, 79, 405-421.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of Royal Statistical Society, Series B*, 70, 849-911.

Fan, J. and Song, R. (2010). Sure independent screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38, 3567-3604.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432-441.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Greene, W. F. and Cai, J. (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 60, 987-996.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré*, 38, 907-921.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C., for the Aids Clinical Trials Group Study 175 Study Team (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335, 1081-1090.

Hansen, B. E. (2008). Least square forecast averaging. *Journal of Econometrics*, 146, 342-350.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167, 38-46.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

He, W. and Yi, G. Y. (2009). Survival prediction with gene expression profiles. *JP Journal of Biostatistics*, 3, 17-39.

Hilafu, H. and Yin, X. (2017). Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics*, 26, 26-34.

Hjort, N. L. and Claeskens, G. (2003). Frequestist model average estimators. *Journal of the American Statistical Association*, 98, 879-899.

Hjort, N. L. and Claeskens, G.(2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101, 1449-1464.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19, 293-325.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14, 382-417.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis : Regression Modeling of Time to Event Data*. Wiley, New York.

Hu, C. and Lin, D. Y. (2002). Cox regression with covariate measurement error. *Scandnavian Journal of Statistics*, 29, 637-655.

Huang, C. Y., Qin J., and Follmann, D. A. (2012). A maximum pseudo-profile likelihood estimator for the Cox model under length-biased sampling. *Biometrika*, 99, 199-210.

Huang, C. Y. and Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100, 877-888.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *The Annals of Statistics*, 41, 1142-1165.

Huang, J., Liu, L., Liu, Y., and Zhao, X. (2014). Group selection in the Cox model with a diverging number of covariates. *Statistica Sinica*, 24, 1787-1810.

Huang, M.-Y. and Chiang, C.-T. (2017). An effective semiparametric estimation approach for the sufficient dimension reduction model. *Journal of the American Statistical Association*, 112, 1296-1310.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric correction approach. *Journal of the American Statistical Association*, 95, 1209-1219.

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer, New York.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics*, 18, 191-219.

Küchenoff, H., Bender, R., and Langner, I. (2007). Effect of Berkson measurement error on parameter estimates in Cox regression models. *Lifetime Data Analysis*, 13, 261-272.

Küchenhoff, H., Mwalili, S. M., and Leasaffre, E. (2006). A general method for dealing with misclassification regression: the misclassification SIMEX. *Biometrics*, 62, 85-96.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge University Press.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lee, J. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24, 230-253.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111, 241-255.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press.

Li, B., Chun, H., and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107, 152-167.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102, 997-1008.

Li, B. and Yin, X. (2007). On surrogate dimension reduction for measurement error regression: an invariance law. *The Annals of Statistics*, 35, 2143-2172.

Liang, H. and Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Computational Statistics & Data Analysis*, 52, 2538-2548.

Li, J. and Ma, S. (2013). *Survival Analysis in Medicine and Genetics*, CRC Press.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86, 316-327.

Li, K.-C., Wang, J.-L., and Chen, C.-H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics*, 27, 1-23.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107, 1129-1139.

Li, Y. and Lin, X. (2003). Functional inference in frailty measurement error models for clustered survival data using the simex approach. *Journal of the American Statistical Association*, 98, 191-203.

Li, Y. and Ryan, L. (2006). Inference on survival data with covariate measurement error - An imputation-based approach. *Scandinavian Journal of Statistics*, 33, 169-190.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of American Statistical Association*, 84, 1074-1078.

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

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

Lin, W. and Lv, J. (2013). High-dimensional additive hazards regression. *Journal of American Statistical Association*, 108, 247-264.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regression. *Biometrics*, 67, 513-523.

Lue, H.-H., Chen, C.-H., and Chang, W.-H. (2011). Dimension reduction in survival regressions with censored data via an imputed spline approach. *Biometrical Journal*, 53, 426-443.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least square. *The Annals of Statistics*, 37, 3498-3528.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli*, 16, 274-300.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107, 168-179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*, 41, 250-268.

Mallow, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28, 863-884.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34, 1436-1462.

Miller, R. G. (1981). *Survival Analysis*. Wiley, New York.

Nadkarni, N. V., Zhao, Y., and Kosorok, M. R. (2011). Inverse regression estimation for censored data. *Journal of the American Statistical Association*, 106, 178-190.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, 48, 829-838.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331-342.

Qin, J. and Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under Cox model. *Biometrics*, 66, 382-392.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*, 106, 1434-1449.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of American Statistical Association*, 92, 179-191.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38, 1287-1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935-980.

Resnick, S. I. (2013). *A Probability Path*. Birkhäuser.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8, 557-569.

Rothman K. J. (2008). BMI-related errors in the measurement of obesity. *International Journal of Obesity*, 32, 56-59.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 523-529.

Schwarz, G. (1978). Estimating the dimension of model. *The Annals of Statistics*, 6, 461-464.

Shaw, P. A. and Prentice, R. L. (2012). Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics*, 68, 397-407.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivative, *The Annals of Statistics*, 6, 177-184.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Song, R., Lu, W., Ma, S., and Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101, 799-814.

Song, X. and Huang, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, 61, 702-714.

Su, Y. and Wang, J. (2012). Modeling left-truncated and right-censored survival data with longitudinal covariates. *The Annals of Statistics*, 40, 1465-1488.

Susarla, V., Tsai, W. Y., and Ryzin, J. V. (1984). A Buckley-James-type estimator for the mean with censored data. *Biometrika*, 71, 624-625.

Sun, H. and Li, H. (2012). Robust Gaussian graphical modeling via ℓ_1 penalization. *Biometrics*, 68, 1197-1206.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35, 2769-2794.

Tan, K. M., Ning, Y., Witten, D. M., and Liu, H. (2016). Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103, 761-777.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Tsai, W. Y., Jewell, N. P., and Wang, M. C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74, 883-886.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A.M., Voskuil, D. W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347, 1999-2009.

Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, 53, 131-145.

Wang, C. Y. (1999). Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error. *Statistics and Probability Letters*, 45, 371-378.

Wang, H., Chen, X., and Flournoy, N. (2016). The focused information criterion for varying-coefficient partially linear measurement error models. *Statistical Papers*, 57, 99-113.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wang, H., Li, Y., and Sun, J. (2015). Focused and model average estimation for regression analysis of panel count data. *Scandinavian Journal of Statistics*, 42, 732-745.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103, 811-821.

Wang, H., Zhang, X., and Zou, G. (2009). Frequentist model averaging estimation: A review. *Journal of Systems Science and Complexity*, 22, 732-748.

Wang, H., Zou, G., and Wan, A. T. K. (2012). Model averaging for varying-coefficient partially linear measurement error models. *Electronic Journal of Statistics*, 6, 1017-1039.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, 25, 831-851.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86, 130-143.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* , 83, 343-354.

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, 49, 1-11.

Wu, F., Kim, S., Qin, J., Saran, R., and Li, Y. (2018). A pairwise likelihood augmented Cox estimator for left-truncated data. *Biometrics*, 74, 100-108.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35, 2654-2690.

Xia, Y., Zhang, D., and Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*, 105, 278-290.

Xie, S. H., Wang, C. Y., and Prentice, R. L. (2001). A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society, Series B*, 63, 855-870.

Xu, G., Wang, S., and Huang, J. Z. (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, 41, 365-381.

Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying coefficients, *Biometrics*, 68, 419-428.

Yan, X., Tang, N., and Zhao, X. (2017). The Spearman rank correlation screening for ultrahigh dimensional censored data. arXiv:1702.02708v1

Yang, E., Baker, Y., Ravikumar, P., Allen, G. I., and Liu, Z. (2014). Mixed graphical models via exponential families. *Journal of Machine Learning Research*, 33, 1042-1050.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distribution. *Journal of Machine Learning Research*, 16, 3813-3847.

Yang, L., Fang, Y., Wang, J., and Shao, Y. (2017). Variable selection for partially linear models via learning gradients. *Electronic Journal of Statistics*, 11, 2907-2930.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error and Misclassification: Strategy, Method and Application*. Springer, New York.

Yi, G. Y. and He, W. (2017). Analysis of case-control data with interacting misclassified covariates. *Journal of Statistical Distributions and Application*, 4:16. DOI: 10.1186/s40488-017-0069-0

Yi, G. Y. and Lawless, J. F. (2007). A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*, 137, 1816-1828.

Yi, G. Y. and Lawless, J. F. (2012). Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error. *The Canadian Journal of Statistics*, 40, 530-549.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99, 151-165.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015). Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association*, 110, 681-696.

Yi, G. Y., Yan, Y., Liao, X., and Spiegelman, D. (2019). Parametric regression analysis with covariate misclassification in main study/validation study designs. *The International Journal of Biostatistics*, 15. DOI: 10.1515/ijb-2017-0002

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society, Series B*, 77, 879-892.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, 39, 3392-3416.

Yörük, E., Ochs, M. F., Geman, D., and Younes, L. (2011). A comprehensive statistical model for cell signaling and protein activity inference. *IEEE/ACM Trans Comput Biol Bioinform*, 8, 592-606.

Yu, Z., Zhu, L., Peng, H., and Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika*, 100, 641-654.

Yu, Z., Dong, Y., and Shao, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *The Annals of Statistics*, 44, 2594-2623.

Yuan, M. and Lin, Yi. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94, 19-35

Zeng, P., and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101, 271-290.

Zhang, C. - H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 28, 894-942.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94, 691-703.

Zhang, J., Zhu, L., and Zhu, L. (2014). Surrogate dimension reduction in measurement error regressions. *Statistica Sinica*, 24, 1341-1363.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39, 174-200.

Zhang, Y., Ouyang, Z., and Zhao, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *Annals of Applied Statistics*, 11, 161-184.

Zhao, G., Ma, Y., and Lu, W. (2017). Efficient estimation for dimension reduction with censored data. arXiv:1710.05377.

Zhao, S. and Prentice, R. L. (2014). Covariate measurement error correction methods in mediation analysis with failure time data. *Biometrics*, 70, 835-844.

Zhao, X. and Zhou, X. (2014). Sufficient dimension reduction on marginal regression for gaps of recurrent events. *Journal of Multivariate Analysis*, *127*, 56-71.

Zhou, S., van de Geer, S., and Bühlmann, P. (2009). Adaptive lasso for high-dimensional regression and Gaussian graphical modeling. arXiv:0903.2515

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

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101, 630-643.

Zhu, L., Zhu, L., and Feng, Z. (2010). Dimension reduction in regression through cumulative slicing estimation. *Journal of the American Statistical Association*, 105, 1455-1466.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106, 1464-1475.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101, 1638-1651.

APPENDICES

Appendix A

Proofs for the Results in Chapter 2

A.1 Regularity Conditions

(A1) There exists a positive number $\alpha \in (0, 1)$ such that

$$\left\| Q_{\mathcal{S}_r^c \mathcal{S}_r} \left(Q_{\mathcal{S}_r \mathcal{S}_r}^{-1} \right) \right\|_\infty \leq 1 - \alpha.$$

(A2) There exists $\rho_1 > 0$ such that the smallest eigenvalue $\Lambda_{\min}(Q_{\mathcal{S}_r \mathcal{S}_r}) > \rho_1$. Besides, there also exists $\rho_2 < \infty$ such that $\Lambda_{\max} \left(\sum_{i=1}^{n} W_b^{(i)}(\zeta) W_b^{(i)^\top}(\zeta) \right) < \rho_2$ for all $b = 1, \cdots, B$ and $\zeta \in \mathcal{Z}$, where $W_b^{(i)}(\zeta) = \left(W_{b,r}^{(i)}(\zeta), W_{b,V \setminus \{r\}}^{(i)}(\zeta) \right)$.

(A3) The function $D(\cdot)$ is third-order differentiable, and there exist η_1 and η_2 such that $|D''(y)| < \eta_1$ and $|D'''(y)| < \eta_2$ for every y.

(A4) The extrapolation function is theoretically exact.

(A5) For every $b = 1, \cdots, B$, $\zeta \in \mathcal{Z}$ and node $r = 1, \cdots, p$, there exist common κ_1, κ_2 such that

(i) $E\{W_{b,r}(\zeta)\} < \kappa_1$,

(ii) $E\{W_{b,r}(\zeta)^2 - (1 + \zeta)\Sigma_{\epsilon;r,r}\} < \kappa_2$, where $\Sigma_{\epsilon;r,r}$ is entry (r, r) in Σ_ϵ.

(A6) Denote $\lambda_n = \lambda_{n1} = \lambda_{n2}$ and $\lambda_n = \lambda_{n3} = \lambda_{n2}$.

Here we briefly comment assumptions described above. Assumptions (A1) and (A2) are also called *mutual incoherence* and *dependency condition*, respectively. Those two conditions are frequently assumed in the neighbourhood approach (e.g., Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Yang et al., 2015; Chen et al. 2015). (A4) is a regular condition in SIMEX method. We make bound condition on the expectation of the working data in (A5). Specifically, (i) implies $E\{W_{b,r}(\zeta)\} = E\{E(W_{b,r}(\zeta)|X_r)\} = E(X_r) < \kappa_1$; (ii) yields $E\{E(W_{b,r}(\zeta)^2|X_r)\} = E\{X_r^2 + (1+\zeta)\Sigma_{\epsilon;r,r} - (1+\zeta)\Sigma_{\epsilon;r,r}\} = E(X_r^2) < \kappa_2$. The implication of those two conditions matches conditions in Yang et al. (2015) and Chen et al. (2015). Assumption (A6) is frequently used in mixed graphical model (e.g., Lee and Hastie 2015; Wang et al. 2015).

A.2 Technical Lemmas

In this section, we mainly present some lemmas which will be used in the proof of the main theorems.

Lemma A.2.1 *Let X_i, $i = 1, \cdots, n$, be the i.i.d. random variables. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Suppose that $E\{\exp(aX_i)\}$ exists and is free of the index i for $a > 0$, then for any $\delta > 0$,*

$$P\left(\bar{X} > \delta\right) \leq \frac{\exp\left(n \log\left[E\left\{\exp(aX_i)\right\}\right]\right)}{\exp\left(na\delta\right)}. \tag{A.1}$$

Proof:

Let $X = \sum_{i=1}^{n} X_i$, then by the Markov's inequality, for any $a > 0$, we have

$$\begin{aligned} P\left(\bar{X} > \delta\right) &= P\left(X > n\delta\right) \\ &\leq \frac{E\left\{\exp\left(aX\right)\right\}}{\exp\left(na\delta\right)} \\ &= \frac{E\left\{\prod_{i=1}^{n} \exp\left(aX_i\right)\right\}}{\exp\left(na\delta\right)}. \end{aligned} \tag{A.2}$$

Noting that $E\left\{\prod_{i=1}^{n}\exp\left(aX_i\right)\right\}$ in (A.2) can be written as

$$\begin{aligned} E\left\{\prod_{i=1}^{n}\exp\left(aX_i\right)\right\} &= \prod_{i=1}^{n}E\left\{\exp\left(aX_i\right)\right\} \\ &= \exp\left(\log\left[\prod_{i=1}^{n}E\left\{\exp\left(aX_i\right)\right\}\right]\right) \\ &= \exp\left(\sum_{i=1}^{n}\log\left[E\left\{\exp\left(aX_i\right)\right\}\right]\right) \\ &= \exp\left(n\log\left[E\left\{\exp\left(aX_i\right)\right\}\right]\right). \end{aligned} \tag{A.3}$$

As a result, combining (A.2) and (A.3) gives (A.1). □

Lemma A.2.2 *Under regularity conditions (A1) - (A4), we have*

$$\begin{aligned} P\left(\left\|\nabla_{\theta(r)}\ell_{b,\zeta}(\theta(r))\right\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right) &< 2\exp\left\{\exp(c_1p') - c_2n + (1+\zeta)\Sigma_{\epsilon;r,t}\right\} \\ &\quad +2\exp\left(-\frac{3}{2}\kappa_2\log(p)\right). \end{aligned}$$

Proof:

Noting that $\nabla_{\theta(r)}\ell_{b,\zeta}(\theta(r)) = \left(\nabla_{\theta_r}\ell_{b,\zeta}(\theta(r)), \nabla_{\theta_{\setminus r}}\ell_{b,\zeta}^{\top}(\theta(r))\right)^{\top}$. Then according to the definition (2.17), we have

$$\nabla_{\theta_r}\ell_{b,\zeta}(\theta(r)) = -\frac{1}{n}\sum_{i=1}^{n}U_r^{(i)}$$

and

$$\nabla_{\theta_{\setminus r}}\ell_{b,\zeta}(\theta(r)) = \left(-\frac{1}{n}\sum_{i=1}^{n}U_t^{(i)} : t \in V\setminus\{r\}\right),$$

where

$$U_r^{(i)} = \left\{W_{b,r}^{(i)}(\zeta) - D'\left(\theta_r + W_{b,V\setminus\{r\}}^{(i)\top}(\zeta)\theta_{\setminus r}\right)\right\}$$

and

$$U_t^{(i)} = W_{b,r}^{(i)}(\zeta)W_{b,t}^{(i)}(\zeta) - W_{b,t}^{(i)}(\zeta)D'\left(\theta_r + W_{b,V\setminus\{r\}}^{(i)\top}(\zeta)\theta_{\setminus r}\right).$$

for any node $t \in V \setminus \{r\}$. To show the desired result, we focus X on the continuous random vector and divide the remaining proof into three steps.

Step 1: *Examine $U_t^{(i)}$ and show that*

$$\begin{aligned} & E\left\{\exp\left(aU_t^{(i)}\right)\middle| X_{V\setminus\{r\}}^{(i)}\right\} \\ \approx\ & \exp\left\{\frac{a^2}{2}\left(X_r^{(i)}\right)^2 D''\left(\theta_r + vaX_t^{(i)} + X_{V\setminus\{r\}}^{(i)}\theta_{\setminus r}\right)\right\} \times \exp\left\{(1+\zeta)\Sigma_{\epsilon;r,t}\right\} \end{aligned}$$

for some constant a, where $\Sigma_{\epsilon;r,t}$ is the (r,t) entry of Σ_ϵ.

For any constant a, we have

$$\begin{aligned} & E\left\{\exp\left(aU_t^{(i)}\right)\middle| X_{V\setminus\{r\}}^{(i)}\right\} \\ =\ & E\left(\exp\left[a\left\{W_{b,r}^{(i)}(\zeta)W_{b,t}^{(i)}(\zeta) - W_{b,t}^{(i)}(\zeta)D'\left(\theta_r + W_{b,V\setminus\{r\}}^{(i)\top}(\zeta)\theta_{\setminus r}\right)\right\}\right]\middle| X_{V\setminus\{r\}}^{(i)}\right) \\ =\ & E\left\{E\left(\exp\left[a\left\{W_{b,r}^{(i)}(\zeta)W_{b,t}^{(i)}(\zeta)\right.\right.\right. \qquad \text{(A.4)} \\ & \left.\left.\left. - W_{b,t}^{(i)}(\zeta)D'\left(\theta_r + W_{b,V\setminus\{r\}}^{(i)\top}(\zeta)\theta_{\setminus r}\right)\right\}\right]\middle| X_r^{(i)}, X_{V\setminus\{r\}}^{(i)}\right)\middle| X_{V\setminus\{r\}}^{(i)}\right\}. \qquad \text{(A.5)} \end{aligned}$$

Since $W_b^{(i)}(\zeta)|X^{(i)} \sim N\left(X^{(i)}, (1+\zeta)\Sigma_\epsilon\right)$, then for any $r,t \in V$ and $r \neq t$, we have $E\left\{W_{b,r}^{(i)}(\zeta)\middle| X_r^{(i)}\right\} = X_r^{(i)}$ and $\operatorname{cov}\left\{W_{b,r}^{(i)}(\zeta), W_{b,t}^{(i)}(\zeta)\middle| X^{(i)}\right\} = (1+\zeta)\Sigma_{\epsilon;r,t}$, where $\Sigma_{\epsilon;r,t}$ is the (r,t) entry of Σ_ϵ. Then $E\left\{W_{b,r}^{(i)}(\zeta)W_{b,t}^{(i)}(\zeta)\middle| X^{(i)}\right\} = X_r^{(i)}X_t^{(i)} + (1+\zeta)\Sigma_{\epsilon;r,t}$. Therefore, we have an approximation

$$\begin{aligned} & E\left(\exp\left[a\left\{W_{b,r}^{(i)}(\zeta)W_{b,t}^{(i)}(\zeta) - W_{b,t}^{(i)}(\zeta)D'\left(\theta_r + W_{b,V\setminus\{r\}}^{(i)\top}(\zeta)\theta_{\setminus r}\right)\right\}\right]\middle| X_r^{(i)}, X_{V\setminus\{r\}}^{(i)}\right) \\ \approx\ & \exp\left[a\left\{X_r^{(i)}X_t^{(i)} + (1+\zeta)\Sigma_{\epsilon;r,t} - X_t^{(i)}D'\left(\theta_r + X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\right], \end{aligned}$$

and (A.4) becomes

$$
\begin{aligned}
& E\left\{\exp\left(aU_t^{(i)}\right)\middle| X_{V\setminus\{r\}}^{(i)}\right\} \\
\approx\ & E\left(\exp\left[a\left\{X_r^{(i)}X_t^{(i)}+(1+\zeta)\Sigma_{\epsilon;r,t}-X_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\right]\middle| X_{V\setminus\{r\}}^{(i)}\right) \\
=\ & \int \exp\left\{aX_r^{(i)}X_t^{(i)}+(1+\zeta)\Sigma_{\epsilon;r,t}\right\}P\left(X_r^{(i)}|X_{V\setminus\{r\}}^{(i)}\right)dX_r^{(i)} \\
& \times\exp\left\{-aX_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\} \\
=\ & \int\left[\exp\left\{aX_r^{(i)}X_t^{(i)}+(1+\zeta)\Sigma_{\epsilon;r,t}\right\}\right. \\
& \times\exp\left\{\theta_rX_r^{(i)}+X_r^{(i)}X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}+\mathbb{C}(X_r^{(i)})-D\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\} \\
& \left.\times\exp\left\{-aX_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\right]dX_r^{(i)} \\
=\ & \int\left[\exp\left\{\theta_rX_r^{(i)}+(1+\zeta)\Sigma_{\epsilon;r,t}+X_r^{(i)}\left(X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}+aX_t^{(i)}\right)\right.\right. \\
& \left.+\mathbb{C}(X_r^{(i)})-D\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\} \\
& \left.\times\exp\left\{-aX_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\right]dX_r^{(i)}, \qquad \text{(A.6)}
\end{aligned}
$$

where the second step is due to the implementation of (2.5). Furthermore, by adding and subtracting an additional term $D\left(\theta_r+aX_t^{(i)}+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)$, (A.6) can be written as

$$
\begin{aligned}
& E\left\{\exp\left(aU_t^{(i)}\right)\middle| X_{V\setminus\{r\}}^{(i)}\right\} \\
\approx\ & \int\exp\left\{\theta_rX_r^{(i)}+X_r^{(i)}\left(X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}+aX_t^{(i)}\right)\right. \\
& \left.+\mathbb{C}(X_r^{(i)})-D\left(\theta_r+aX_t^{(i)}+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}dX_r^{(i)} \\
& \times\exp\left\{D\left(\theta_r+aX_t^{(i)}+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)-D\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\} \\
& \times\exp\left\{-aX_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\times\exp\left\{(1+\zeta)\Sigma_{\epsilon;r,t}\right\} \\
=\ & \exp\left\{D\left(\theta_r+aX_t^{(i)}+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)-D\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\} \\
& \times\exp\left\{-aX_t^{(i)}D'\left(\theta_r+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\times\exp\left\{(1+\zeta)\Sigma_{\epsilon;r,t}\right\} \\
=\ & \exp\left\{\frac{a^2}{2}\left(X_r^{(i)}\right)^2D''\left(\theta_r+vaX_t^{(i)}+X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)\right\}\times\exp\left\{(1+\zeta)\Sigma_{\epsilon;r,t}\right\} \qquad \text{(A.7)}
\end{aligned}
$$

where the second step holds since the integration is one, and the third step is due to the second order Taylor series expansion on $D\left(\theta_r + aX_t^{(i)} + X_{V\setminus\{r\}}^{(i)\top}\theta_{\setminus r}\right)$ around $a = 0$ and $v \in (0,1)$.

Step 2: *Examine* $\frac{1}{n}\sum_{i=1}^{n} U_t^{(i)}$ *and show that*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_t^{(i)}\right| > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \;\middle|\; \mathcal{E}_1, \mathcal{E}_2\right) < 2\exp\left\{-\frac{n}{2\eta_1\kappa_2}\left(\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right)^2 + (1+\zeta)\Sigma_{\epsilon;r,t}\right\}, \quad \text{(A.8)}$$

where $\mathcal{E}_1 = \left\{\max_{i,r} X_r^{(i)} \le 4\log p'\right\}$ *and* $\mathcal{E}_2 = \left\{\max_{t\in V}\frac{1}{n}\sum_{i=1}^{n}\left(X_t^{(i)}\right)^2 \le \kappa_2\right\}$.

By the derivations of Proposition 3 and Lemma 9 in Yang et al. (2015), we have $P(\mathcal{E}_1^c) \le c_1 p'$ and $P(\mathcal{E}_2^c) \le \exp(-c_2 n)$ for some constants c_1 and c_2, where $\mathcal{E}_1^c$ and $\mathcal{E}_2^c$ are complement sets of $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively. Therefore, by condition (A3) and any $\delta > 0$, applying Lemma A.2.1 gives

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_t^{(i)}\right| > \delta \;\middle|\; \mathcal{E}_1, \mathcal{E}_2\right) < 2\exp\left\{n\left(\frac{\eta_1\kappa_2 a^2}{2} - \delta a\right) + (1+\zeta)\Sigma_{\epsilon;r,t}\right\},$$

and specifying $a = \frac{\delta}{\eta_1\kappa_2}$ yields

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_t^{(i)}\right| > \delta \;\middle|\; \mathcal{E}_1, \mathcal{E}_2\right) < 2\exp\left\{-\frac{n\delta^2}{2\eta_1\kappa_2} + (1+\zeta)\Sigma_{\epsilon;r,t}\right\}.$$

Finally, specifying $\delta = \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}$ gives (A.8).

Step 3: *Examine* $U_r^{(i)}$ *and show*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_r^{(i)}\right| > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \;\middle|\; \mathcal{E}_1, \mathcal{E}_2\right) < 2\exp\left\{-\frac{3n}{4\eta_1}\left(\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right)^2\right\}. \quad \text{(A.9)}$$

Indeed, by the derivations similar to Step 1, we can show that for any constant $\widetilde{a}$,

$$E\left\{\exp\left(\widetilde{a}U_r^{(i)}\right)\middle|\, X_{V\setminus\{r\}}^{(i)}\right\} \approx \exp\left\{-\frac{\widetilde{a}^2}{2}D''\left((\widetilde{v}\widetilde{a} + \theta_r) + X_{V\setminus r}^{(i)\top}\theta_{\setminus r}\right)\right\}$$

for some constant $\widetilde{v} \in (0,1)$. Then for some constant $\widetilde{\delta}$, by the derivations similar to Step 2 with Condition (A3) and $\widetilde{a}$ replaced by $\frac{\widetilde{\delta}}{\eta_1}$, we can show that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_r^{(i)}\right| > \widetilde{\delta} \,\middle|\, \mathcal{E}_1, \mathcal{E}_2\right) < 2\exp\left\{-\frac{3n\widetilde{\delta}^2}{4\eta_1}\right\}. \tag{A.10}$$

Finally, replacing $\widetilde{\delta}$ by $\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}$ gives (A.9).

Step 4: *Examine $\nabla_{\theta(r)}\ell_{b,\zeta}(\theta(r))$ and show the final result.*

Recall that $\nabla_{\theta(r)}\ell_{b,\zeta}(\theta(r)) = \left(\nabla_{\theta_r}\ell_{b,\zeta}(\theta(r)), \nabla_{\theta_{\setminus r}}\ell_{b,\zeta}^{\top}(\theta(r))\right)^{\top}$. Then by (A.8) and (A.9), we have

$$\begin{aligned} & P\left(\left\|\nabla_{\theta(r)}\ell_{b,\zeta}(\theta(r))\right\|_{\infty} > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \,\middle|\, \mathcal{E}_1, \mathcal{E}_2\right) \\ < \;& 2\exp\left\{-\frac{n}{2\eta_1\kappa_2}\left(\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right)^2 + \log p + (1+\zeta)\Sigma_{\epsilon;r,t}\right\} \\ & +2\exp\left\{-\frac{3n}{4\eta_1}\left(\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right)^2\right\}. \end{aligned}$$

As a result, provided that $\lambda_n > \sqrt{\frac{32\eta_1\kappa_2\log(p)}{n}}\left(\frac{2-\alpha}{\alpha}\right)$, we have

$$\begin{aligned} P\left(\left\|\nabla\ell_{b,\zeta}(\theta(r))\right\|_{\infty} > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\right) \;<\; & 2\exp\left\{\exp(c_1 p') - c_2 n + (1+\zeta)\Sigma_{\epsilon;r,t}\right\} \\ & +2\exp\left(-\frac{3}{2}\kappa_2\log(p)\right), \end{aligned}$$

which holds due to the inequality $P(A) \le P(\mathcal{E}_1^c) + P(\mathcal{E}_2^c) + P(A|\,\mathcal{E}_1, \mathcal{E}_2)$ for an event A (Yang et al. 2015, p.29). □

Lemma A.2.3 *Let $\widehat{\theta}(r;\zeta,b) = \left(\widehat{\theta}_{\mathcal{S}_r}^{\top}(r;\zeta,b), \widehat{\theta}_{\mathcal{S}_r^c}^{\top}(\zeta,b)\right)^{\top}$ with*

$$\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) = \left(\widehat{\theta}_r(\zeta,b), \widehat{\theta}_{\mathcal{S}_r}^{\top}(\zeta,b)\right)^{\top}.$$

Under regularity conditions (A1) - (A4), we have

$$\left\|\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) - \theta_{0;\mathcal{S}_r}(r)\right\|_2 \le \frac{6\sqrt{d_r}\lambda_n}{\rho_1}. \tag{A.11}$$

Proof:

By the definition of $\mathcal{S}_r^c$, we have

$$\widehat{\theta}\left(r;\zeta,b\right)=\left(\widehat{\theta}_{\mathcal{S}_r}^{\top}\left(r;\zeta,b\right),\widehat{\theta}_{\mathcal{S}_r^c}^{\top}\left(\zeta,b\right)\right)^{\top}=\left(\widehat{\theta}_{\mathcal{S}_r}^{\top}\left(r;\zeta,b\right),0_{(p-d_r-1)}^{\top}\right)^{\top},$$

where 0_d stands for the d-dimensional zero vector. According, we write the true value of $\theta(r)$ as $\theta_0\left(r\right)=\left(\theta_{0;\mathcal{S}_r}^{\top}\left(r\right),\theta_{0;\mathcal{S}_r^c}^{\top}\right)^{\top}=\left(\theta_{0;\mathcal{S}_r}^{\top}\left(r\right),0_{(p-d_r-1)}^{\top}\right)^{\top}$ with $\theta_{0;\mathcal{S}_r}(r)=\left(\theta_{0;r},\theta_{0;\mathcal{S}_r}^{\top}\right)^{\top}$.

Claim: *For $\zeta\in\mathcal{Z}$ and $b=1,\cdots,B$, let $\widehat{u}_{\mathcal{S}_r}=\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b)-\theta_{0;\mathcal{S}_r}(r)$. Show that*

$$\left\|\widehat{u}_{\mathcal{S}_r}\right\|_2\leq\frac{6\sqrt{d_r}\lambda_n}{\rho_1}. \tag{A.12}$$

We define the function $\Phi:\mathbb{R}^{d_r+1}\to\mathbb{R}$ by

$$\Phi(u)=\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+u\right)-\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)+\lambda_n\left(\left\|\theta_{0;\mathcal{S}_r}+u\right\|_1-\left\|\theta_{0;\mathcal{S}_r}\right\|_1\right), \tag{A.13}$$

where we express any parameter value $\theta_{\mathcal{S}_r}(r)$ by $u+\theta_{0;\mathcal{S}_r}(r)$.

Note that $\Phi(u)$ is a convex function since $\ell_{b,\zeta}(\cdot)$ defined in (2.17) and the lasso function $\|\cdot\|_1$ defined in (2.13) are both convex functions. Similar to the derivations for Lemma of Ravikumar et al. (2010), to show (A.12), it suffices to show that

$$\Phi(u)>0\ \text{ for any }\ u\ \text{ with }\ \left\|u\right\|_2=\mathcal{B}, \tag{A.14}$$

where $\mathcal{B}=\frac{6\sqrt{d_r}\lambda_n}{\rho_1}$.

By the second order Taylor series expansion on $\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+u\right)-\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)$ around $u=0$, (A.13) becomes

$$\Phi(u)\ =\ T_1+T_2+T_3, \tag{A.15}$$

where

$$T_1\ =\ \nabla_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)u; \tag{A.16a}$$

$$T_2\ =\ \frac{1}{2}u^{\top}\nabla^2_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+vu\right)u; \tag{A.16b}$$

$$T_3\ =\ \lambda_n\left(\left\|\theta_{0;\mathcal{S}_r}+u\right\|_1-\left\|\theta_{0;\mathcal{S}_r}\right\|_1\right), \tag{A.16c}$$

and v is some constant in $(0,1)$.

We first specify $\mathcal{B}$ in (A.14) by $\mathcal{M}\lambda_n\sqrt{d_r}$ for some $\mathcal{M}>0$. The remaining task is to individually examine T_1, T_2 and T_3 for their bound when $\|u\|_2 = \mathcal{M}\lambda_n\sqrt{d_r}$. We proceed with the following four steps.

Step 1: *Show that*

$$\|T_1\|_1 \;<\; \frac{(\lambda_n\sqrt{d_r})^2}{4}\mathcal{M} \quad \textit{for} \quad \|u\|_2 = \mathcal{M}\lambda_n\sqrt{d_r}. \tag{A.17}$$

For the first term T_1 in (A.15), by the result in Lemma A.2.2, we have

$$\begin{aligned}
\|T_1\|_1 &= \left\|\nabla_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)u\right\|_1 \\
&\leq \left\|\nabla_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)\right\|_\infty \|u\|_1 \\
&\leq \left\|\nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)\right)\right\|_\infty \sqrt{d_r}\,\|u\|_2 \\
&< \frac{(\lambda_n\sqrt{d_r})^2}{4}\mathcal{M}.
\end{aligned}$$

Step 2: *Show that*

$$T_2 \;\geq\; \frac{\rho_1(\lambda_n\sqrt{d_r})^2\mathcal{M}^2}{2} \quad \textit{for} \quad \|u\|_2 = \mathcal{M}\lambda_n\sqrt{d_r}. \tag{A.18}$$

Note that $\nabla^2_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+vu\right)$ can be expressed as

$$\begin{aligned}
&\nabla^2_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+vu\right) \\
=\;& \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D''\left(\theta_{0r} + W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\left(\theta_{0;\mathcal{S}_r}+vu\right)\right).
\end{aligned} \tag{A.19}$$

Then applying the Taylor series expansion on $D''(\cdot)$ around $\theta_{0;\mathcal{S}_r}=0$, then (A.19) can be re-written as

$$\begin{aligned}
\nabla^2_{\theta_{\mathcal{S}_r}(r)}\ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r)+vu\right) &= \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D''\left(\theta_{0r} + W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\,\theta_{0;\mathcal{S}_r}\right) \\
&\quad + \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D'''(\bar{\eta})\left(vuW^{(i)}_{b,V\setminus\{r\}}(\zeta)\right),
\end{aligned}$$

where $\bar{\eta}$ lies on the "line segment" between $\theta_{0;\mathcal{S}_r}$ and $\theta_{0;\mathcal{S}_r} + vu$. Then by conditions (A2), (A3), and (A5), we have

$$\begin{aligned}
T_2 &= u^\top \nabla^2_{\theta_{\mathcal{S}_r}(r)} \ell_{b,\zeta}\left(\theta_{0;\mathcal{S}_r}(r) + vu\right) u \\
&\geq \min_{u:\|u\|_2=\mathcal{B}} \left[u^\top \left\{ \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D''\left(\theta_{0r} + W^{(i)}_{b,V\setminus\{r\}}(\zeta)\,\theta_{0;\mathcal{S}_r}\right) \right\} u \right] \\
&\quad + \min_{u:\|u\|_2=\mathcal{B}} \left[u^\top \left\{ \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D'''(\bar{\eta}) \left(vuW^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\right) \right\} u \right] \\
&\geq \mathcal{B}^2 \Lambda_{\min}(Q_{\mathcal{S}_r\mathcal{S}_r}) \\
&\quad - \max_{u:\|u\|_2=\mathcal{B}} \left[u^\top \left\{ \sum_{i=1}^{n} W^{(i)}_{b,V\setminus\{r\}}(\zeta)\, W^{(i)\top}_{b,V\setminus\{r\}}(\zeta)\, D'''(\bar{\eta}) \left(vuW^{(i)}_{b,V\setminus\{r\}}(\zeta)\right) \right\} u \right] \\
&\geq \mathcal{B}^2\rho_1 - \mathcal{B}^3\rho_2\eta_2\kappa_1 \\
&\geq \frac{\rho_1(\lambda_n\sqrt{d_r})^2\mathcal{M}^2}{2}.
\end{aligned}$$

Step 3: *Show that*

$$T_3 \geq -(\lambda_n\sqrt{d_r})^2\mathcal{M}^2 \quad \textit{for} \quad \|u\|_2 = \mathcal{M}\lambda_n\sqrt{d_r}. \tag{A.20}$$

Finally, for the last term T_3 in (A.15), applying the triangle inequality gives

$$\|\theta_{0;\mathcal{S}_r}\|_1 = \|\theta_{0;\mathcal{S}_r} + u - u\|_1 \leq \|\theta_{0;\mathcal{S}_r} + u\|_1 + \|u\|_1,$$

which implies

$$\|\theta_{0;\mathcal{S}_r} + u\|_1 - \|\theta_{0;\mathcal{S}_r}\|_1 \geq -\|u\|_1.$$

Therefore, we have

$$T_3 = \lambda_n\left(\|\theta_{0;\mathcal{S}_r} + u\|_1 - \|\theta_{0;\mathcal{S}_r}\|_1\right) \geq -(\lambda_n\sqrt{d_r})^2\mathcal{M}^2.$$

Step 4: *Establish (A.12).*

Therefore, combining (A.17), (A.18), and (A.20) with (A.15) gives

$$\Phi(u) \geq (\lambda_n\sqrt{d_r})^2\mathcal{M}\left(\frac{-1}{4} + \frac{\rho_1}{4}\mathcal{M} - 1\right). \tag{A.21}$$

To ensure the right-hand-side of (A.21) be bounded below by zero, we must have

$$\frac{-1}{4}+\frac{\rho_1}{4}\mathcal{M}-1>0,$$

which is equivalent to $\mathcal{M}>\frac{5}{\rho_1}$. We take $\mathcal{M}^*=\frac{6}{\rho_1}$, and thus, $\mathcal{B}^*=\mathcal{M}^*\lambda_n\sqrt{d_r}=\frac{6\sqrt{d_r}\lambda_n}{\rho_1}$ and (A.14) holds. As a result, (A.12) is shown. □

Lemma A.2.4 *Let*

$$R_n=\left\{\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)-\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\right\}\left\{\widehat{\theta}(r;\zeta,b)-\theta_0(r)\right\}, \tag{A.22}$$

where $\bar{\theta}$ lies on the "line segment" between $\widehat{\theta}(r;\zeta,b)$ and $\theta_0(r)$. Then under regularity conditions (A1) - (A4), we have

$$\left\|R_n\right\|_\infty\leq\frac{72\eta_1\rho_2 d_r\lambda_n^2}{\rho_1^2}.$$

Proof:

Since $\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta(r)\right)=\sum_{i=1}^{n}W^{(i)}_{b,V\setminus\{r\}}\left(\zeta\right)W^{(i)^\top}_{b,V\setminus\{r\}}\left(\zeta\right)D''\left(\theta_r+W^{(i)\top}_{b,V\setminus\{r\}}\left(\zeta\right)\theta_{\setminus r}\right)$, then

$$\begin{aligned}
&\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)-\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\\
=\;&\sum_{i=1}^{n}W^{(i)}_{b,V\setminus\{r\}}\left(\zeta\right)W^{(i)^\top}_{b,V\setminus\{r\}}\left(\zeta\right)\left\{D''\left(\theta_r+W^{(i)\top}_{b,V\setminus\{r\}}\left(\zeta\right)\bar{\theta}\right)-D''\left(\theta_r+W^{(i)\top}_{b,V\setminus\{r\}}\left(\zeta\right)\theta_{\setminus r}\right)\right\}.
\end{aligned}$$

By conditions (A2) and (A3), the maximum eigenvalue of $\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)-\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)$ is

$$\begin{aligned}
&\Lambda_{\max}\left\{\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)-\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\right\}\\
=\;&\max_{\xi:\|\xi\|_2=1}\xi^\top\left\{\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)-\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\right\}\xi\\
\leq\;&\max_{\xi:\|\xi\|_2=1}\xi^\top\left(\sum_{i=1}^{n}W^{(i)}_{b,V\setminus\{r\}}\left(\zeta\right)W^{(i)^\top}_{b,V\setminus\{r\}}\left(\zeta\right)\right)\xi\\
&\times\xi^\top\left|D''\left(\theta_r+W^{(i)\top}_{b,V\setminus\{r\}}\left(\zeta\right)\bar{\theta}\right)-D''\left(\theta_r+W^{(i)\top}_{b,V\setminus\{r\}}\left(\zeta\right)\theta_{\setminus r}\right)\right|\xi\\
\leq\;&2\eta_1\rho_2.
\end{aligned}\tag{A.23}$$

As a result, by Lemma A.2.3 and (A.23), we have

$$\begin{aligned}
\|R_n\|_1 &\leq \|R_n\|_2^2 \\
&\leq \Lambda_{\max}\left\{\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right) - \nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\right\} \times \left\|\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\|_2^2 \\
&\leq 2\eta_1\rho_2\left(\frac{6\sqrt{d_r}\lambda_n}{\rho_1}\right)^2 \\
&= \frac{72\eta_1\rho_2 d_r\lambda_n^2}{\rho_1^2},
\end{aligned}$$

and thus the proof is completed. □

A.3 Proof of Theorem 2.2.1

In contrast to the naive log likelihood function, we first consider the log likelihood function based on true random variables:

$$\ell(\theta(r)) = -\frac{1}{n}\sum_{i=1}^{n}\log\left\{P\left(X_r|X_{V\setminus\{r\}}\right)\right\},$$

where $P\left(X_r|X_{V\setminus\{r\}}\right)$ is defined in (2.5). Similar to (2.12), the estimator based on true random variables is given by

$$\widetilde{\theta}(r) = \underset{\theta(r)}{\operatorname{argmin}}\left\{\ell(\theta(r)) + \lambda_n\left\|\theta_{\setminus r}\right\|_1\right\} \tag{A.24}$$

with $\widetilde{\theta}(r) = \left(\widetilde{\theta}_r, \widetilde{\theta}_{\setminus r}^\top\right)^\top$. To ease the notation, let $\widetilde{\theta}$, $\widehat{\theta}_{nv}$, θ and θ_0 denote $\widetilde{\theta}(r)$, $\widehat{\theta}_{nv}(r)$, $\theta(r)$ and $\theta_0(r)$, respectively.

Let $\widetilde{\theta}_{rt}$ denote the tth component in $\widetilde{\theta}_{\setminus r}$. Let $\widetilde{z} = \left(\widetilde{z}_r, \widetilde{z}_{\setminus r}^\top\right)^\top$ be a p-dimensional vector with the tth component in $\widetilde{z}_{\setminus r}$ being $\widetilde{z}_t = \operatorname{sign}\left(\widetilde{\theta}_{rt}\right)$ if $\widetilde{\theta}_{rt} \neq 0$ and $|\widetilde{z}_t| \leq 1$ otherwise, while $\widetilde{z}_r$, corresponding to θ_r, is set to zero since the nodewise term θ_r is not penalized in (A.24). In addition, let $\widehat{z}_{nv}$ denote a p-dimensional vector which is defined similar to $\widetilde{z}$ but corresponds to $\widehat{\theta}_{nv}$. Then by the KKT conditions, we have

$$\frac{\partial\ell_{nv}\left(\widehat{\theta}_{nv}\right)}{\partial\theta} + \lambda_n\widehat{z}_{nv} = 0 \tag{A.25}$$

and

$$\frac{\partial \ell\left(\widetilde{\theta}\right)}{\partial \theta} + \lambda_n \widetilde{z} = 0. \tag{A.26}$$

By the first order Taylor series expansion on $\frac{\partial \ell_{nv}\left(\widehat{\theta}_{nv}\right)}{\partial \theta}$ and $\frac{\partial \ell\left(\widetilde{\theta}\right)}{\partial \theta}$ around θ_0, we have

$$\frac{\partial \ell_{nv}\left(\widehat{\theta}_{nv}\right)}{\partial \theta} \approx \frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta} + \frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\left(\widehat{\theta}_{nv} - \theta_0\right) \tag{A.27}$$

and

$$\frac{\partial \ell\left(\widetilde{\theta}\right)}{\partial \theta} \approx \frac{\partial \ell\left(\theta_0\right)}{\partial \theta} + \frac{\partial^2 \ell\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\left(\widetilde{\theta} - \theta_0\right) \tag{A.28}$$

Combining (A.27) and (A.28) yields

$$\begin{aligned} \frac{\partial \ell_{nv}\left(\widehat{\theta}_{nv}\right)}{\partial \theta} - \frac{\partial \ell\left(\widetilde{\theta}\right)}{\partial \theta} \approx & \left(\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta} - \frac{\partial \ell\left(\theta_0\right)}{\partial \theta}\right) + \frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\widehat{\theta}_{nv} - \frac{\partial^2 \ell\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\widetilde{\theta} \\ & - \left(\frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial \theta \partial \theta^\top} - \frac{\partial^2 \ell\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\right)\theta_0 \end{aligned} \tag{A.29}$$

The second order derivative of $\ell_{nv}(\theta_0)$ and $\ell(\theta_0)$ can be, respectively, expressed as

$$\frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial \theta \partial \theta^\top} = \frac{1}{n}\sum_{i=1}^{n} X^{*(i)}_{V\setminus\{r\}} X^{*(i)\top}_{V\setminus\{r\}} D''\left(\theta_{0;r} + X^{*(i)\top}_{V\setminus\{r\}}\theta_{0;\setminus r}\right)$$

and

$$\frac{\partial^2 \ell\left(\theta_0\right)}{\partial \theta \partial \theta^\top} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}_{V\setminus\{r\}} X^{(i)\top}_{V\setminus\{r\}} D''\left(\theta_{0;r} + X^{(i)\top}_{V\setminus\{r\}}\theta_{0;\setminus r}\right).$$

Since $X^{*(i)}_{V\setminus\{r\}}\left|X^{(i)}\right. \sim N\left(X^{(i)}_{V\setminus\{r\}}, \Sigma_{\epsilon;\setminus r}\right)$, where $\Sigma_{\epsilon;\setminus r}$ is the covariance matrix Σ_ϵ with deleted rth row and rth column. As a result, we have $E\left(X^{*(i)}_{V\setminus\{r\}} X^{*(i)\top}_{V\setminus\{r\}}\right) = X^{(i)}_{V\setminus\{r\}} X^{(i)^\top}_{V\setminus\{r\}} + \Sigma_{\epsilon;\setminus r}$. Hence, we have approximations

$$\begin{aligned} & E\left(\frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial \theta \partial \theta^\top}\middle| X_{V\setminus\{r\}}\right) \\ \approx & \frac{1}{n}\sum_{i=1}^{n}\left\{\left(X^{(i)}_{V\setminus\{r\}} X^{(i)^\top}_{V\setminus\{r\}} + \Sigma_{\epsilon;\setminus r}\right) D''\left(\theta_{0;r} + X^{(i)\top}_{V\setminus\{r\}}\theta_{0;\setminus r}\right)\right\} \end{aligned} \tag{A.30}$$

and

$$E\left(\frac{\partial^2 \ell\left(\theta_0\right)}{\partial\theta\partial\theta^\top}\middle| X_{V\setminus\{r\}}\right) \approx \frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_{V\setminus\{r\}}^{(i)}X_{V\setminus\{r\}}^{(i)\top}\right)D''\left(\theta_{0;r}+X_{V\setminus\{r\}}^{(i)\top}\theta_{0;\setminus r}\right)\right\}. \tag{A.31}$$

Therefore, by the Law of Large Numbers with (A.30) and (A.31), we have that as $n\to\infty$,

$$\frac{\partial^2 \ell_{nv}\left(\theta_0\right)}{\partial\theta\partial\theta^\top} \xrightarrow{p} \mathcal{Q}_{nv} \quad \text{and} \quad \frac{\partial^2 \ell\left(\theta_0\right)}{\partial\theta\partial\theta^\top} \xrightarrow{p} \mathcal{Q}_r, \tag{A.32}$$

where

$$\mathcal{Q}_{nv} = E\left\{\left(X_{V\setminus\{r\}}^{(i)}X_{V\setminus\{r\}}^{(i)\top}+\Sigma_{\epsilon;\setminus r}\right)D''\left(\theta_{0;r}+X_{V\setminus\{r\}}^{(i)\top}\theta_{0;\setminus r}\right)\right\} \tag{A.33}$$

and

$$\mathcal{Q}_r = E\left\{\left(X_{V\setminus\{r\}}^{(i)}X_{V\setminus\{r\}}^{(i)\top}\right)D''\left(\theta_{0;r}+X_{V\setminus\{r\}}^{(i)\top}\theta_{0;\setminus r}\right)\right\}. \tag{A.34}$$

Then the relationship between (A.33) and (A.34) is determined by

$$\mathcal{Q}_{nv} = \mathcal{Q}_r + \Sigma_{\epsilon;\setminus r}\mathcal{D}_r, \tag{A.35}$$

where $\mathcal{D}_r = E\left\{D''\left(\theta_{0;r}+X_{V\setminus\{r\}}^{(i)\top}\theta_{0;\setminus r}\right)\right\}$. On the other hand, by (A.25) and (A.26), we have

$$\frac{\partial \ell_{nv}\left(\widehat{\theta}_{nv}\right)}{\partial\theta} - \frac{\partial \ell\left(\widetilde{\theta}\right)}{\partial\theta} = -\lambda_n\left(\widehat{z}_{nv}-\widetilde{z}\right). \tag{A.36}$$

Thus, combining (A.32), (A.35), and (A.36) with (A.29) gives

$$\begin{aligned} -\lambda_n\left(\widehat{z}_{nv}-\widetilde{z}\right) &\approx \left(\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial\theta}-\frac{\partial \ell\left(\theta_0\right)}{\partial\theta}\right)+\left(\mathcal{Q}_r+\Sigma_{\epsilon;\setminus r}\mathcal{D}_r\right)\widehat{\theta}_{nv}-\mathcal{Q}_r\widetilde{\theta}-\Sigma_{\epsilon;\setminus r}\mathcal{D}_r\theta_0 \\ &= \left(\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial\theta}-\frac{\partial \ell\left(\theta_0\right)}{\partial\theta}\right)+\mathcal{Q}_r\left(\widehat{\theta}_{nv}-\widetilde{\theta}\right)+\Sigma_{\epsilon;\setminus r}\mathcal{D}_r\left(\widehat{\theta}_{nv}-\theta_0\right). \end{aligned} \tag{A.37}$$

By the triangle inequality, $\|\widehat{z}_{nv}-\widetilde{z}\|_\infty \le \|\widehat{z}_{nv}\|_\infty + \|\widetilde{z}\|_\infty < 2$. Besides, by (A.37), we have

$$\left\|\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial\theta}-\frac{\partial \ell\left(\theta_0\right)}{\partial\theta}\right\|_\infty \le 2\lambda_n + \|\mathcal{Q}_r\|_\infty\left\|\widehat{\theta}_{nv}-\widetilde{\theta}\right\|_\infty + \left\|\Sigma_{\epsilon;\setminus r}\mathcal{D}_r\right\|_\infty\left\|\widehat{\theta}_{nv}-\theta_0\right\|_\infty, \tag{A.38}$$

and thus rearranging (A.38) gives

$$\begin{aligned}\left\|\widehat{\theta}_{nv}-\widetilde{\theta}\right\|_{\infty} &\geq \left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left(\left\|\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta}-\frac{\partial \ell\left(\theta_0\right)}{\partial \theta}\right\|_{\infty}-2\lambda_n\right) \\ &\quad -\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty}. \end{aligned} \tag{A.39}$$

Noting that based on true random variables, Lemmas 9 and 10 in Yang et al. (2015) show that there exist some constants $\widetilde{\alpha}\in(0,1)$ and $\widetilde{\rho}>0$, such that

$$\left\|\frac{\partial \ell\left(\theta_0\right)}{\partial \theta}\right\|_{\infty}\leq \frac{\lambda_n\widetilde{\alpha}}{4\left(2-\widetilde{\alpha}\right)} \tag{A.40}$$

and

$$\left\|\widetilde{\theta}-\theta_0\right\|_{\infty}\leq 5\widetilde{\rho}\lambda_n \tag{A.41}$$

with large probabilities.

Finally, applying the triangle inequality on $\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty}$, we have

$$\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty} \geq \left\|\widehat{\theta}_{nv}-\widetilde{\theta}\right\|_{\infty}-\left\|\widetilde{\theta}-\theta_0\right\|_{\infty},$$

and thus, implementing (A.39), (A.40) and (A.41) gives

$$\begin{aligned}\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty} &\geq \left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left(\left\|\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta}\right\|_{\infty}-\frac{\lambda_n\widetilde{\alpha}}{4\left(2-\widetilde{\alpha}\right)}-2\lambda_n\right) \\ &\quad -\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty}-5\widetilde{\rho}\lambda_n.\end{aligned}$$

Consequently, we have

$$\begin{aligned}\left\|\widehat{\theta}_{nv}-\theta_0\right\|_{\infty} &\geq \left\{1+\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\right\}^{-1}\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left(\left\|\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta}\right\|_{\infty}-\frac{\lambda_n\widetilde{\alpha}}{4\left(2-\widetilde{\alpha}\right)}-2\lambda_n\right) \\ &\quad -\left\{1+\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\right\}^{-1}5\widetilde{\rho}\lambda_n \\ &= \left\{\left\|\mathcal{Q}_r\right\|_{\infty}+\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\right\}^{-1}\left(\left\|\frac{\partial \ell_{nv}\left(\theta_0\right)}{\partial \theta}\right\|_{\infty}-\frac{\lambda_n\widetilde{\alpha}}{4\left(2-\widetilde{\alpha}\right)}-2\lambda_n\right) \\ &\quad -\left\{1+\left\|\mathcal{Q}_r\right\|_{\infty}^{-1}\left\|\Sigma_{\epsilon;\backslash r}\mathcal{D}_r\right\|_{\infty}\right\}^{-1}5\widetilde{\rho}\lambda_n. \end{aligned} \tag{A.42}$$

To ensure that $\left\|\widehat{\theta}_{nv} - \theta_0\right\|_\infty$ is bounded below by a positive constant, we must have that the right-hand side in (A.42) is larger than zero, yielding that

$$\lambda_n < \left\{ \frac{16 - 7\widetilde{\alpha}}{4(2 - \widetilde{\alpha})} + 5\widetilde{\rho} \left\| \mathcal{Q}_r \right\|_\infty \right\}^{-1} \left\| \frac{\partial \ell_{nv}(\theta_0)}{\partial \theta} \right\|_\infty . \tag{A.43}$$

As a result, provided (A.43), the right-hand side of (A.42) is larger than zero. Thus, desired inequality is obtained, and the proof is completed. □

A.4 Proof of Theorem 2.3.1

A.4.1 Proof of Part (a)

The following derivations consist of three parts.

Step 1: *Let $\widehat{\theta}_{rt}(\zeta, b)$ denote the tth component of $\widehat{\theta}_{\setminus r}(\zeta, b)$. Examine*

$$\widehat{\mathcal{N}}_b(r; \zeta) = \left\{ t \in V \setminus \{r\} : \widehat{\theta}_{rt}(\zeta, b) \neq 0 \right\}$$

and show that

$$\widehat{\mathcal{N}}_b(r; \zeta) = \mathcal{N}(r) \tag{A.44}$$

with a large probability, where $\mathcal{N}(r)$ is the neighbourhood defined in (2.4).

In the proof of Lemma A.2.3, we write $\widehat{\theta}(r; \zeta, b) = \left(\widehat{\theta}_{\mathcal{S}_r}^\top (r; \zeta, b), \widehat{\theta}_{\mathcal{S}_r^c}^\top (\zeta, b) \right)^\top$ with $\widehat{\theta}_{\mathcal{S}_r}(r; \zeta, b) = \left(\widehat{\theta}_r(\zeta, b), \widehat{\theta}_{\mathcal{S}_r}^\top (\zeta, b) \right)^\top$. Let $\widehat{z} = \left(\widehat{z}_r, \widehat{z}_{\setminus r}^\top \right)^\top$ be a p-dimensional vector with the tth component in $\widehat{z}_{\setminus r}$ being $\widehat{z}_t = \text{sign}\left(\widehat{\theta}_{rt}(\zeta, b) \right)$ if $\widehat{\theta}_{rt}(\zeta, b) \neq 0$ and $|\widehat{z}_t| \leq 1$ otherwise, while $\widehat{z}_r$, corresponding to θ_r, is set to zero since the nodewise term θ_r is not penalized in (2.18). To show the sparsity recovery, we consider the primal dual witness (PDW) method (e.g., Hastie et al. 2015, p.307). The strategy of the PDW method is to

(i) $\widehat{\theta}_{\mathcal{S}_r^c}(\zeta, b) = 0_{p - d_r - 1}$ and $\widehat{\theta}_{\mathcal{S}_r}(r; \zeta, b) = \underset{\theta_{\mathcal{S}_r}(r)}{\text{argmin}} \left\{ \ell_{b,\zeta}\left(\theta(r)\right) + \lambda_n \left\| \theta_{\mathcal{S}_r} \right\|_1 \right\}$;

(ii) write $\widehat{z} = \left(\widehat{z}_{\mathcal{S}_r}^\top, \widehat{z}_{\mathcal{S}_r^c}^\top \right)^\top$ corresponding to the components of $\widehat{\theta}_{\mathcal{S}_r}(r; \zeta, b)$ and $\widehat{\theta}_{\mathcal{S}_r^c}(\zeta, b)$;

(iii) show that

$$\left\|\widehat{z}_{\mathcal{S}_r^c}\right\|_\infty < 1. \tag{A.45}$$

Indeed, as discussed in Lemma 11.2 of Hastie et al. (2015, p.307), if (A.45) is true, then $\widehat{\theta}(r;\zeta,b) = \left(\widehat{\theta}_{\mathcal{S}_r}^\top(r;\zeta,b), 0_{p-d_r-1}^\top\right)^\top$ is an optimal solution of (2.18), and thus, (A.44) holds with a large probability (e.g., Hastie et al. 2015, Theorem 11.3). So, the remaining task is to show (A.45).

By the KKT conditions, we have

$$\nabla_{\theta(r)}\ell_{b,\zeta}\left(\widehat{\theta}(r;\zeta,b)\right) + \lambda_n\widehat{z} = 0. \tag{A.46}$$

Adding $-\nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)$ to the both sides of (A.46) gives

$$\nabla_{\theta(r)}\ell_{b,\zeta}\left(\widehat{\theta}(r;\zeta,b)\right) - \nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right) = -\lambda_n\widehat{z} - \nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right). \tag{A.47}$$

By the Mean Value Theorem (MVT), there exists $\bar{\theta}$ which lies on the "line segment" between $\widehat{\theta}(r;\zeta,b)$ and $\theta_0(r)$, such that

$$\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)\left\{\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\} = -\lambda_n\widehat{z} - \nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right).$$

Adding $\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\left\{\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\}$ to both sides of (A.47) yields

$$\begin{aligned}
& \nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\left\{\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\} \\
= \; & -\lambda_n\widehat{z} - \nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right) - \left[\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\bar{\theta}\right)\left\{\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\}\right. \\
& \left. -\nabla^2_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)\left\{\widehat{\theta}(r;\zeta,b) - \theta_0(r)\right\}\right] \\
\triangleq \; & -\lambda_n\widehat{z} - V - R_n,
\end{aligned} \tag{A.48}$$

where R_n is defined (A.22), and $V = \nabla_{\theta(r)}\ell_{b,\zeta}\left(\theta_0(r)\right)$.

Let $V = \left(V_{\mathcal{S}_r}^\top, V_{\mathcal{S}_r^c}^\top\right)^\top$ and $R_n = \left(R_{n\mathcal{S}_r}^\top, V_{n\mathcal{S}_r^c}^\top\right)^\top$ Now, by (A.48) and (i), we have

$$\begin{pmatrix} Q_{\mathcal{S}_r\mathcal{S}_r} & Q_{\mathcal{S}_r\mathcal{S}_r^c} \\ Q_{\mathcal{S}_r^c\mathcal{S}_r} & Q_{\mathcal{S}_r^c\mathcal{S}_r^c} \end{pmatrix}\begin{pmatrix} \widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) - \theta_{0;\mathcal{S}_r}(r) \\ 0 \end{pmatrix} = -\lambda_n\begin{pmatrix} \widehat{z}_{\mathcal{S}_r} \\ \widehat{z}_{\mathcal{S}_r^c} \end{pmatrix} - \begin{pmatrix} V_{\mathcal{S}_r} \\ V_{\mathcal{S}_r^c} \end{pmatrix} - \begin{pmatrix} R_{n\mathcal{S}_r} \\ R_{n\mathcal{S}_r^c} \end{pmatrix},$$

and it implies that

$$Q_{\mathcal{S}_r\mathcal{S}_r}\left\{\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b)-\theta_{0;\mathcal{S}_r}(r)\right\} = \lambda_n\widehat{z}_{\mathcal{S}_r}-V_{\mathcal{S}_r}-R_{n\mathcal{S}_r}, \tag{A.49a}$$

$$Q_{\mathcal{S}_r^c\mathcal{S}_r}\left\{\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b)-\theta_{0;\mathcal{S}_r}(r)\right\} = \lambda_n\widehat{z}_{\mathcal{S}_r^c}-V_{\mathcal{S}_r^c}-R_{n\mathcal{S}_r^c}. \tag{A.49b}$$

Combining (A.49a) and (A.49b) yields

$$Q_{\mathcal{S}_r^c\mathcal{S}_r}Q_{\mathcal{S}_r\mathcal{S}_r}^{-1}\left(-\lambda_n\widehat{z}_{\mathcal{S}_r}-V_{\mathcal{S}_r}-R_{n\mathcal{S}_r}\right)=-\lambda_n\widehat{z}_{\mathcal{S}_r^c}-V_{\mathcal{S}_r^c}-R_{n\mathcal{S}_r^c} \tag{A.50}$$

and thus our target $\widehat{z}_{\mathcal{S}_r^c}$ can be expressed as

$$\widehat{z}_{\mathcal{S}_r^c}=\frac{1}{\lambda_n}\left\{Q_{\mathcal{S}_r^c\mathcal{S}_r}Q_{\mathcal{S}_r\mathcal{S}_r}^{-1}\left(\lambda_n\widehat{z}_{\mathcal{S}_r}+V_{\mathcal{S}_r}+R_{n\mathcal{S}_r}\right)-V_{\mathcal{S}_r^c}-R_{n\mathcal{S}_r^c}\right\}. \tag{A.51}$$

We now show (A.45). Given

$$\lambda_n<\frac{\rho_1^2}{288\eta_1\rho_2 d_r}, \tag{A.52}$$

then (A.51) gives that

$$\begin{aligned}
\|\widehat{z}_{\mathcal{S}_r^c}\|_\infty &\le \frac{1}{\lambda_n}\left(\left\|Q_{\mathcal{S}_r^c\mathcal{S}_r}Q_{\mathcal{S}_r\mathcal{S}_r}^{-1}\right\|_\infty\|\lambda_n\widehat{z}_{\mathcal{S}_r}+V_{\mathcal{S}_r}+R_{n\mathcal{S}_r}\|_\infty+\left\|V_{\mathcal{S}_r^c}\right\|_\infty+\left\|V_{n\mathcal{S}_r^c}\right\|_\infty\right)\\
&\le \frac{1}{\lambda_n}\left\{\left\|Q_{\mathcal{S}_r^c\mathcal{S}_r}Q_{\mathcal{S}_r\mathcal{S}_r}^{-1}\right\|_\infty\left(\lambda_n\|\widehat{z}_{\mathcal{S}_r}\|_\infty+\|V_S\|_\infty+\|R_{n\mathcal{S}_r}\|_\infty\right)+\left\|V_{\mathcal{S}_r^c}\right\|_\infty+\left\|R_{n\mathcal{S}_r^c}\right\|_\infty\right\}\\
&\le \frac{1}{\lambda_n}\left\{\left\|Q_{\mathcal{S}_r^c\mathcal{S}_r}Q_{\mathcal{S}_r\mathcal{S}_r}^{-1}\right\|_\infty\left(\lambda_n\|\widehat{z}_{\mathcal{S}_r}\|_\infty+\|V\|_\infty+\|R_n\|_\infty\right)+\|V\|_\infty+\|R_n\|_\infty\right\}\\
&\le \frac{1}{\lambda_n}\left\{(1-\alpha)\left(\lambda_n+\frac{\alpha\lambda_n}{8-4\alpha}+\frac{72\eta_1\rho_2 d_r\lambda_n^2}{\rho_1^2}\right)+\frac{\alpha\lambda_n}{8-4\alpha}+\frac{72\eta_1\rho_2 d_r\lambda_n^2}{\rho_1^2}\right\}\\
&= (1-\alpha)+(2-\alpha)\left(\frac{\alpha}{8-4\alpha}+\frac{\alpha}{8-4\alpha}\right)\\
&= (1-\alpha)+\frac{\alpha}{2}\\
&= 1-\frac{\alpha}{2}\\
&\le 1,
\end{aligned}$$

where the third step is due to that $\|V_{\mathcal{S}_r}\|_\infty \le \|V\|_\infty$, $\left\|V_{\mathcal{S}_r^c}\right\|_\infty \le \|V\|_\infty$, $\|R_{n\mathcal{S}_r}\|_\infty \le \|R_n\|_\infty$ and $\|R_{n\mathcal{S}_r{}^c}\|_\infty \le \|R_n\|_\infty$, the fourth step comes from Condition (A1), Lemmas A.2.2 and A.2.4, and $\|\widehat{z}_{\mathcal{S}_r}\|_\infty \le 1$ by the construction of $\widehat{z}_{\mathcal{S}_r}$, and the fifth step is due to (A.52).

Hence, by PDW approach, we have (A.44) for every $b = 1, \cdots, B$ and $\zeta \in \mathcal{Z}$ with a large probability.

Step 2: *Let* $\widehat{\theta}_{rt}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{rt}(\zeta, b)$ *and* $\widehat{\mathcal{N}}(r;\zeta) = \left\{t \in V \setminus \{r\} : \widehat{\theta}_{rt}(\zeta) \neq 0\right\}$. *Show that*

$$\widehat{\mathcal{N}}(r;\zeta) = \mathcal{N}(r) \tag{A.53}$$

with a large probability.

Since $\widehat{\theta}(r;\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}(r;\zeta,b)$ and B is a fixed finite number, it implies that $\widehat{\theta}_{rt}(\zeta) \neq 0$ as $\widehat{\theta}_{rt}(\zeta, b) \neq 0$. Then by (A.44), we have (A.53) with a large probability.

Step 3: *Establish the desired result.*

Finally, since $\widehat{\theta}(r;\zeta) \xrightarrow{p} \widehat{\theta}(r)$ as $\zeta \to -1$, then $\widehat{\mathcal{N}}(r;\zeta) \xrightarrow{p} \widehat{\mathcal{N}}(r)$. As a result, by (A.53), we conclude that $\widehat{\mathcal{N}}(r) = \mathcal{N}(r)$ with a large probability and $\zeta \to -1$. □

A.4.2 Proof of Part (b)

By Lemma A.2.3 and the fact that $\|\cdot\|_\infty \leq \|\cdot\|_2$, we have

$$\left\|\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) - \theta_{0;\mathcal{S}_r}(r)\right\|_\infty \leq \frac{6\sqrt{d_r}\lambda_n}{\rho_1}.$$

By the definition similar to (2.19), we have $\widehat{\theta}_{\mathcal{S}_r}(r;\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b)$ for a fixed number B. Then for any $\zeta \in \mathcal{Z}$, we have

$$\begin{aligned}
\left\|\widehat{\theta}_{\mathcal{S}_r}(r;\zeta) - \theta_{0;\mathcal{S}_r}(r)\right\|_\infty &= \left\|\frac{1}{B}\sum_{b=1}^{B}\left\{\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) - \theta_{0;\mathcal{S}_r}(r)\right\}\right\|_\infty \\
&\leq \frac{1}{B}\sum_{b=1}^{B}\left\|\widehat{\theta}_{\mathcal{S}_r}(r;\zeta,b) - \theta_{0;\mathcal{S}_r}(r)\right\|_\infty \\
&< \frac{1}{B}\sum_{b=1}^{B}\left(\frac{6\sqrt{d_r}\lambda_n}{\rho_1}\right) \\
&= \frac{6\sqrt{d_r}\lambda_n}{\rho_1}.
\end{aligned} \tag{A.54}$$

Let $\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) = \left(\widehat{\theta}_{\mathcal{S}_r}(r;\zeta) : \zeta \in \mathcal{Z}\right)$ and $\theta_{0;\mathcal{S}_r}(r;\mathcal{Z}) = (\theta_{0;\mathcal{S}_r}(r) : \zeta \in \mathcal{Z})$. Since (A.54) holds for all $\zeta \in \mathcal{Z}$, then we have

$$\left\|\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \theta_{0;\mathcal{S}_r}(r;\mathcal{Z})\right\|_\infty < \frac{6d_r^{\frac{3}{2}}\lambda_n}{\rho_1}. \tag{A.55}$$

Let $\mathcal{R}(\Gamma) = \widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \mathcal{G}(\mathcal{Z},\Gamma)$. By the least squares method, the estimator $\widehat{\Gamma}$ is obtained by solving

$$\mathcal{G}_\Gamma^\top \mathcal{R}(\Gamma) = 0.$$

Then we have

$$\begin{aligned}
\mathcal{G}_\Gamma^\top \left\{\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \theta_{0;\mathcal{S}_r}(r;\mathcal{Z})\right\} &= \mathcal{G}_\Gamma^\top \left\{\mathcal{G}(\mathcal{Z},\widehat{\Gamma}) - \mathcal{G}(\mathcal{Z},\Gamma)\right\} \\
&= \mathcal{G}_\Gamma^\top \mathcal{G}_\Gamma \left(\widehat{\Gamma} - \Gamma\right) + o_p(1),
\end{aligned} \tag{A.56}$$

where the second equality is due to the Mean Value Theorem with respect to Γ and consistency of the least squares estimator. (A.56) further gives

$$\left(\widehat{\Gamma} - \Gamma\right) = \left(\mathcal{G}_\Gamma^\top \mathcal{G}_\Gamma\right)^{-1} \mathcal{G}_\Gamma^\top \left\{\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \theta_{0;\mathcal{S}_r}(r;\mathcal{Z})\right\} + o_p(1). \tag{A.57}$$

Furthermore, since $\widehat{\theta}_{\mathcal{S}_r}(r) = \mathcal{G}(-1,\widehat{\Gamma})$, then

$$\begin{aligned}
\widehat{\theta}_{\mathcal{S}_r}(r) - \theta_{0;\mathcal{S}_r}(r) &= \mathcal{G}(-1,\widehat{\Gamma}) - \mathcal{G}(-1,\Gamma) \\
&= \mathcal{G}'(-1,\Gamma)\left(\widehat{\Gamma} - \Gamma\right).
\end{aligned} \tag{A.58}$$

As a result, combining (A.55) and (A.57) with (A.58) yields

$$\begin{aligned}
\left\|\widehat{\theta}_{\mathcal{S}_r}(r) - \theta_{0;\mathcal{S}_r}(r)\right\|_\infty &= \left\|\mathcal{G}'(-1,\Gamma)\left(\widehat{\Gamma} - \Gamma\right)\right\|_\infty \\
&= \left\|\mathcal{G}'(-1,\Gamma)\left(\mathcal{G}_\Gamma^\top \mathcal{G}_\Gamma\right)^{-1} \mathcal{G}_\Gamma^\top \left\{\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \theta_{0;\mathcal{S}_r}(r;\mathcal{Z})\right\}\right\|_\infty \\
&\leq \left\|\mathcal{G}'(-1,\Gamma)\left(\mathcal{G}_\Gamma^\top \mathcal{G}_\Gamma\right)^{-1} \mathcal{G}_\Gamma^\top\right\|_1 \left\|\left\{\widehat{\theta}_{\mathcal{S}_r}(r;\mathcal{Z}) - \theta_{0;\mathcal{S}_r}(r;\mathcal{Z})\right\}\right\|_\infty \\
&\leq \left\|\mathcal{G}'(-1,\Gamma)\left(\mathcal{G}_\Gamma^\top \mathcal{G}_\Gamma\right)^{-1} \mathcal{G}_\Gamma^\top\right\|_1 \frac{6d_r^{\frac{3}{2}}\lambda_n}{\rho_1} \\
&\triangleq \mathcal{A}\frac{6d_r^{\frac{3}{2}}\lambda_n}{\rho_1}.
\end{aligned}$$

Hence, we complete the proof. □

A.4.3 Proof of Part (c)

As discussed in Ravikumar et al. (2010, p. 1301), to show the correctness of sign recovery, i.e., $\operatorname{sign}\left(\widehat{\theta}_{\mathcal{S}_r}(r)\right) = \operatorname{sign}\left(\theta_{\mathcal{S}_r}(r)\right)$, it suffices to check the boundness of $\left\|\widehat{\theta}_{\mathcal{S}_r}(r) - \theta_{\mathcal{S}_r}(r)\right\|_\infty$. Since Theorem 2.3.1 (b) holds, then by the fact that $\|a\|_\infty < \|a\|_2$ for every nonzero vector a, we directly obtain the desired result. □

A.5 Proof of Theorem 2.4.1

Let

$$\ell_{\mathrm{C};b,\zeta}\left(\theta_{\mathrm{C}}(r)\right) = -\frac{1}{n}\sum_{i=1}^{n}\log\left\{P\left(W_{b,r}^{\mathrm{C(i)}}\left(\zeta\right)\middle| W_{b,V_{\mathrm{C}}\setminus\{r\}}^{\mathrm{C}(i)}\left(\zeta\right), W_{b}^{\mathrm{D}(i)}\left(\zeta\right)\right)\right\}$$

and

$$\ell_{\mathrm{D};b,\zeta}\left(\theta_{\mathrm{D}}(r')\right) = -\frac{1}{n}\sum_{i=1}^{n}\log\left\{P\left(W_{b,r'}^{\mathrm{D}(i)}\left(\zeta\right)\middle| W_{b,V_{\mathrm{D}}\setminus\{r'\}}^{\mathrm{D}(i)}\left(\zeta\right), W_{b}^{\mathrm{C}(i)}\left(\zeta\right)\right)\right\}.$$

Similar to the derivations in proof of Lemma A.2.2, there exists a constant $\alpha \in (0,1)$, such that

$$\left\|\nabla_{\theta_{\mathrm{C}}(r)}\ell_{\mathrm{C};b,\zeta}\left(\theta_{\mathrm{C}}(r)\right)\right\|_\infty \leq \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \tag{A.59}$$

and

$$\left\|\nabla_{\theta_{\mathrm{D}}(r')}\ell_{\mathrm{D};b,\zeta}\left(\theta_{\mathrm{D}}(r')\right)\right\|_\infty \leq \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \tag{A.60}$$

with large probabilities.

In addition, let $\widehat{\theta}_{\mathrm{C}}(r;\zeta,b)$ and $\widehat{\theta}_{\mathrm{D}}(r';\zeta,b)$ denote the estimators determined by (2.26) and (2.28), respectively. According to two sets $\mathcal{S}_{\mathrm{C},r}$ and $\mathcal{S}_{\mathrm{D},r'}$, we write $\widehat{\theta}_{\mathrm{C}}(r;\zeta,b) = \left(\widehat{\theta}_{r}^{\mathrm{C}}(\zeta,b), \widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}^{\top}(\zeta,b), \widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}^{c}}^{\top}(\zeta,b)\right)^{\top}$ and $\widehat{\theta}_{\mathrm{D}}(r';\zeta,b) = \left(\widehat{\theta}_{r'}^{\mathrm{D}}(\zeta,b), \widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}^{\top}(\zeta,b), \widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}^{c}}^{\top}(\zeta,b)\right)^{\top}$, where $\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(\zeta,b)$ and $\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(\zeta,b)$ are subvectors of

$$\left(\widehat{\theta}_{r}^{\mathrm{C}\top}(\zeta,b), \widehat{\theta}_{\setminus r}^{\mathrm{C}\top}(\zeta,b)\right)^{\top} \quad \text{and} \quad \left(\widehat{\theta}_{r'}^{\mathrm{D}\top}(\zeta,b), \widehat{\theta}_{\setminus r'}^{\mathrm{D}\top}(\zeta,b)\right)^{\top}$$

containing nonzero elements, respectively.

Let

$$\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta,b)=\left(\widehat{\theta}^{\mathrm{C}}_{r}(\zeta,b),\widehat{\theta}^{\top}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(\zeta,b)\right)^{\top}\quad\text{and}\quad\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta,b)=\left(\widehat{\theta}^{\mathrm{D}}_{r'}(\zeta,b),\widehat{\theta}^{\top}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(\zeta,b)\right)^{\top}.$$

Then similar to the derivations in proof of Lemma A.2.3, we can show that

$$\left\|\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta,b)-\theta_{0;\mathrm{C},\mathcal{S}_{\mathrm{C},r}}(r)\right\|_2\leq\frac{6\sqrt{d_{\mathrm{C},r}}\lambda_n}{\rho_1}\tag{A.61}$$

and

$$\left\|\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta,b)-\theta_{0;\mathrm{D},\mathcal{S}_{\mathrm{D},r'}}(r')\right\|_2\leq\frac{6\sqrt{d_{\mathrm{D},r'}}\lambda_n}{\rho_1}.\tag{A.62}$$

Finally, let

$$R_{n;\mathrm{C}}=\left\{\nabla^2_{\theta_{\mathrm{C}}(r)}\ell_{\mathrm{C};b,\zeta}\left(\bar{\theta}_{\mathrm{C}}\right)-\nabla^2_{\theta_{\mathrm{C}}(r)}\ell_{\mathrm{C};b,\zeta}\left(\theta_{0;\mathrm{C}}(r)\right)\right\}\left\{\widehat{\theta}_{\mathrm{C}}(r;\zeta,b)-\theta_{0;\mathrm{C}}(r)\right\}$$

and

$$R_{n;\mathrm{D}}=\left\{\nabla^2_{\theta_{\mathrm{D}}(r')}\ell_{\mathrm{D};b,\zeta}\left(\bar{\theta}_{\mathrm{D}}\right)-\nabla^2_{\theta_{\mathrm{D}}(r')}\ell_{\mathrm{D};b,\zeta}\left(\theta_{0;\mathrm{D}}(r')\right)\right\}\left\{\widehat{\theta}_{\mathrm{D}}(r';\zeta,b)-\theta_{0;\mathrm{D}}(r')\right\},$$

where $\bar{\theta}_{\mathrm{C}}$ lies on the "line segment" between $\widehat{\theta}_{\mathrm{C}}(r;\zeta,b)$ and $\theta_{0;\mathrm{C}}(r)$ and $\bar{\theta}_{\mathrm{D}}$ lies on the "line segment" between $\widehat{\theta}_{\mathrm{D}}(r';\zeta,b)$ and $\theta_{0;\mathrm{D}}(r')$. Then by the derivations similar to proof of Lemma A.2.4, we have

$$\left\|R_{n;\mathrm{C}}\right\|_\infty\leq\frac{72\eta_1\rho_2 d_{\mathrm{C},r}\lambda_n^2}{\rho_1^2}\quad\text{and}\quad\left\|R_{n;\mathrm{D}}\right\|_\infty\leq\frac{72\eta_1\rho_2 d_{\mathrm{D},r'}\lambda_n^2}{\rho_1^2}.\tag{A.63}$$

A.5.1 Proof of Part (a)

We first examine the neibourhood sets associated with $r\in V_{\mathrm{C}}$. Let $\widehat{\theta}_{\mathrm{C},k}$ denote the kth component in $\widehat{\theta}_{\mathrm{C}}(r;\zeta,b)$ and let $\widehat{z}_{\mathrm{C}}$ be a p-dimensional vector with kth component being $\widehat{z}_{\mathrm{C},k}=\mathrm{sign}(\widehat{\theta}_{\mathrm{C},k})$ if $\widehat{\theta}_{\mathrm{C},k}\neq 0$ and $|\widehat{z}_{\mathrm{C},k}|\leq 1$ otherwise. According to the PDW strategy in Appendix A.4.1,

(i) we define $\widehat{\theta}_{\mathrm{C};\mathcal{S}^c_{\mathrm{C},r}}(\zeta, b) = 0_{p-d_{\mathrm{C},r}-1}$ and

$$\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta,b) = \underset{\theta_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r)}{\operatorname{argmin}} \left\{ \ell_{\mathrm{C};b,\zeta}\left(\theta_{\mathrm{C}}(r)\right) + \lambda_n \left(\left\|\theta^{\mathrm{C}}_{\setminus r}\right\|_1 + \left\|\theta^{\mathrm{CD}}_{r}\right\|_1 \right) \right\};$$

(ii) write $\widehat{z}_{\mathrm{C}} = \left(\widehat{z}^{\top}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}, \widehat{z}^{\top}_{\mathrm{C};\mathcal{S}^c_{\mathrm{C},r}} \right)^{\top}$ corresponding to the components of $\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta,b)$ and $\widehat{\theta}_{\mathrm{C};\mathcal{S}^c_{\mathrm{C},r}}(\zeta,b) = 0_{p-d_{\mathrm{C},r}-1}$;

(iii) by the similar derivations in Appendix A.4.1 and results (A.59), (A.61), and (A.63), we can show that

$$\left\| \widehat{z}_{\mathrm{C};\mathcal{S}^c_{\mathrm{C},r}} \right\|_{\infty} \leq 1.$$

Therefore, by the derivations similar to Steps 2 and 3 in Appendix A.4.1, we can show that

$$\widehat{\mathcal{N}}_{\mathrm{C}}(r) = \mathcal{N}_{\mathrm{C}}(r) \;\text{ and }\; \widehat{\mathcal{N}}_{\mathrm{CD}}(r) = \mathcal{N}_{\mathrm{CD}}(r)$$

with a large probability.

We next examine the neibourhood sets associated with $r' \in V_{\mathrm{D}}$. Let $\widehat{\theta}_{\mathrm{C},k'}$ denote the k'th component in $\widehat{\theta}_{\mathrm{D}}(r';\zeta,b)$ and let $\widehat{z}_{\mathrm{D}}$ be a p-dimensional vector with k'th component being $\widehat{z}_{\mathrm{D},k'} = \operatorname{sign}(\widehat{\theta}_{\mathrm{D},k'})$ if $\widehat{\theta}_{\mathrm{D},k'} \neq 0$ and $|\widehat{z}_{\mathrm{D},k'}| \leq 1$ otherwise. According to the PDW strategy in Appendix A.4.1,

(i) we define $\widehat{\theta}_{\mathrm{D};\mathcal{S}^c_{\mathrm{D},r'}}(\zeta, b) = 0_{p-d_{\mathrm{D},r'}-1}$ and

$$\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r;\zeta,b) = \underset{\theta_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r')}{\operatorname{argmin}} \left\{ \ell_{\mathrm{D};b,\zeta}\left(\theta_{\mathrm{D}}(r')\right) + \lambda_n \left(\left\|\theta^{\mathrm{D}}_{\setminus r'}\right\|_1 + \left\|\theta^{\mathrm{DC}}_{r'}\right\|_1 \right) \right\};$$

(ii) write $\widehat{z}_{\mathrm{D}} = \left(\widehat{z}^{\top}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}, \widehat{z}^{\top}_{\mathrm{D};\mathcal{S}^c_{\mathrm{D},r'}} \right)^{\top}$ corresponding to the components of $\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta,b)$ and $\widehat{\theta}_{\mathrm{D};\mathcal{S}^c_{\mathrm{D},r'}}(\zeta,b) = 0_{p-d_{\mathrm{D},r'}-1}$;

(iii) by the similar derivations in Appendix A.4.1 and results (A.60), (A.62), and (A.63), we can show that

$$\left\| \widehat{z}_{\mathrm{D};\mathcal{S}^c_{\mathrm{D},r'}} \right\|_{\infty} \leq 1.$$

Therefore, by the derivations similar to Steps 2 and 3 in Appendix A.4.1, we can show that

$$\widehat{\mathcal{N}}_{\mathrm{D}}(r') = \mathcal{N}_{\mathrm{D}}(r') \;\text{ and }\; \widehat{\mathcal{N}}_{\mathrm{DC}}(r') = \mathcal{N}_{\mathrm{DC}}(r')$$

with a large probability.

A.5.2 Proof of Part (b)

Let $\theta_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta) = \frac{1}{B}\sum_{b=1}^{B}\theta_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta,b)$ and $\theta_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta) = \frac{1}{B}\sum_{b=1}^{B}\theta_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta,b)$. Then by (A.61) and (A.62), we have that for every $\zeta \in \mathcal{Z}$,

$$\left\|\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r;\zeta) - \theta_{0;\mathrm{C},\mathcal{S}_{\mathrm{C},r}}(r)\right\|_2 \leq \frac{6\sqrt{d_{\mathrm{C},r}}\lambda_n}{\rho_1}$$

and

$$\left\|\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r';\zeta) - \theta_{0;\mathrm{D},\mathcal{S}_{\mathrm{D},r'}}(r')\right\|_2 \leq \frac{6\sqrt{d_{\mathrm{D},r'}}\lambda_n}{\rho_1},$$

respectively. Finally, let $\zeta \to -1$, we obtain

$$\left\|\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r) - \theta_{0;\mathrm{C},\mathcal{S}_{\mathrm{C},r}}(r)\right\|_2 \leq \frac{6\sqrt{d_{\mathrm{C},r}}\lambda_n}{\rho_1} \tag{A.64}$$

and

$$\left\|\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r') - \theta_{0;\mathrm{D},\mathcal{S}_{\mathrm{D},r'}}(r')\right\|_2 \leq \frac{6\sqrt{d_{\mathrm{D},r'}}\lambda_n}{\rho_1}. \tag{A.65}$$

A.5.3 Proof of Part (c)

Similar to the derivations in Appendix A.4.3, (A.64) and (A.65) indicate that both

$$\left\|\widehat{\theta}_{\mathrm{C};\mathcal{S}_{\mathrm{C},r}}(r) - \theta_{0;\mathrm{C},\mathcal{S}_{\mathrm{C},r}}(r)\right\|_\infty \quad \text{and} \quad \left\|\widehat{\theta}_{\mathrm{D};\mathcal{S}_{\mathrm{D},r'}}(r') - \theta_{0;\mathrm{D},\mathcal{S}_{\mathrm{D},r'}}(r')\right\|_\infty$$

are bounded. As a result, the sign recovery is shown. □

Appendix B

Proofs for the Results in Chapter 3

B.1 Regularity Conditions

(C1) $P\left(Y_i(\tau)=1\right)>0$, where τ is an upper bound of failure times which is assumed to be finite.

(C2) $\int_0^\tau \lambda_0(t)dt<\infty$.

(C3) The $\{N_i(t), Y_i(t), X_i^*\}$ are independent and identically distributed for $i=1,\cdots,n$.

(C4) Censoring time is non-informative. That is, the failure time and the censoring time are independent, given the covariate.

(C5) There exists γ_1, γ_2, such that $E(X_r)<\gamma_1$ and $E(X_r^2)<\gamma_2$, where X_r is the rth element of X. Furthermore, $E(X_sX_\nu)$ is also bounded for any $s\neq\nu$.

(C6) There exist η_1 and η_2, such that $\Lambda_{\min}\left(\sum_{j=1}^n X_jX_j^\top\right)>\eta_1$ and $\Lambda_{\max}\left(\sum_{j=1}^n X_jX_j^\top\right)<\eta_2$, where $\Lambda_{\max}(A)$ is the maximum eigenvalue of the matrix A.

(C7) $E\left\{I_{\beta;b,\zeta}(\beta,\Theta)\right\}$ and $E\left\{I_{\Theta;b,\zeta}(\beta,\Theta)\right\}$ are positive definite matrices for every β and Θ. Moreover, there exist κ_1 and κ_2 such that

$$\Lambda_{min}\left\{I_{\beta;b,\zeta}(\beta,\Theta)\right\}>\kappa_1 \quad \text{and} \quad \Lambda_{min}\left\{I_{\Theta;b,\zeta}(\beta,\Theta)\right\}>\kappa_2,$$

where $\Lambda_{min}(A)$ represents the minimum eigenvalue of the matrix A, and $I_{\beta;b,\zeta}(\cdot)$ and $I_{\Theta;b,\zeta}(\cdot)$ are given in (3.24).

(C8) There exists a positive number $\alpha \in (0,1)$ such that

$$\left\| I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}\left(\beta_0,\Theta_0\right) I^{-1}_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left(\beta_0,\Theta_0\right)\right\|_\infty \leq 1-\alpha$$

and

$$\left\| I_{\Theta,\mathcal{S}_2^c\mathcal{S}_2;b,\zeta}\left(\beta_0,\Theta_0\right) I^{-1}_{\Theta,\mathcal{S}_2\mathcal{S}_2;b,\zeta}\left(\beta_0,\Theta_0\right)\right\|_\infty \leq 1-\alpha,$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are defined in Section 3.3, $\mathcal{S}_k^c$ is the complement of $\mathcal{S}_k$ with $k=1,2$, $I_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left(\beta_0,\Theta_0\right)$ and $I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}\left(\beta_0,\Theta_0\right)$ are $d_\beta \times d_\beta$ and $(p-d_\beta)\times d_\beta$ sub-matrices of $I_{\beta;b,\zeta}\left(\beta_0,\Theta_0\right)$, respectively, and $I_{\Theta,\mathcal{S}_2\mathcal{S}_2;b,\zeta}\left(\beta_0,\Theta_0\right)$ and $I_{\Theta,\mathcal{S}_2^c\mathcal{S}_2;b,\zeta}\left(\beta_0,\Theta_0\right)$ are $d_\Theta \times d_\Theta$ and $(p^2-p-d_\Theta)\times d_\Theta$ sub-matrices of $I_{\Theta;b,\zeta}\left(\beta_0,\Theta_0\right)$, respectively.

(C9) The extrapolant function is assumed known and is differentially continuous.

Conditions (C1) to (C4) are regular assumptions in survival analysis for the establishment of the asymptotic properties (e.g., Andersen and Gill 1982). Conditions (C5) and (C6) come from the requirements for covariates in the graphical model theory (e.g., Chen et al. 2015; Yang et al. 2015). Condition (C7) ensures the information matrices to be positive definite. Condition (C8), also called *mutual incoherence*, is often assumed for variable selection (Yang et al. 2015). Condition (C9) is the a requirement for the SIMEX method (Carroll et al. 2006, Chapter 5).

B.2 Some Lemmas

In this section, we present key lemmas to be used in the proofs of the theorems. First, we let $U_{\beta,\mathcal{S}_1;b,\zeta}(\beta,\Theta)$ and $U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta,\Theta)$ be the sub-vectors of $U_{\beta;b,\zeta}(\beta,\Theta)$, defined by (3.22), which are indexed by $\mathcal{S}_1$ and $\mathcal{S}_1^c$, respectively. That is, $U_{\beta,\mathcal{S}_1;b,\zeta}(\beta,\Theta)=\left(U^{(j)}_{\beta;b,\zeta}(\beta,\Theta): j\in\mathcal{S}_1\right)$ and $U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta,\Theta)=\left(U^{(j)}_{\beta;b,\zeta}(\beta,\Theta): j\in\mathcal{S}_1^c\right)$, where $U^{(j)}_{\beta;b,\zeta}(\beta,\Theta)$ is the jth element of $U_{\beta;b,\zeta}(\beta,\Theta)$.

Next, we let $U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta,\Theta)$ and $U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta,\Theta)$ be the sub-vectors of $U_{\Theta;b,\zeta}(\beta,\Theta)$, defined by (3.22), which are indexed by $\mathcal{S}_2$ and $\mathcal{S}_2^c$, respectively. That is, $U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta,\Theta)=\left(U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta): j\in\mathcal{S}_2\right)$ and $U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta,\Theta)=\left(U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta): j\in\mathcal{S}_2^c\right)$, where $U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)$ is the jth element of $U_{\Theta;b,\zeta}(\beta,\Theta)$.

In the first two lemmas, we establish bounds for the score functions, respectively, correspond to β and Θ.

Lemma B.2.1 *Under Conditions (C3) and (C5) in Appendix B.1, we have*

$$\begin{aligned} P\left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_{n1}}{4}\right) &\leq d_\beta\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right)+c_1p^{-2}\right\}; \\ P\left(\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha\lambda_{n1}}{4}\right) &\leq (p-d_\beta)\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right)+c_1p^{-2}\right\}, \end{aligned}$$

where c_1 is a positive constant, $\|a\|_\infty$ is the infinity norm defined as $\max_i |a_i|$ for a vector a with elements a_i's, λ_{n1} is the tuning parameter in (3.14), and d_β is defined after Theorem 3.3.1 in Section 3.3.

Proof:

The proof consists of the following four steps.

Step 1: *Show that* $\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} = \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} + o_p(1)$.

Let

$$\mathcal{G}_{b,\zeta}(u;\beta,\Theta) = E\left\{Y_i(u)\exp\left(\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,t}(\zeta)\theta_{s\nu}\right)\right\} \tag{B.1}$$

denote the expectation of each term in $G_{b,\zeta}(u;\beta,\Theta)$. Differentiating (3.25) with respect to β yields

$$\begin{aligned} G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta) &= \sum_{i=1}^n\left[Y_i(u)\left\{W^{(i)}_b(\zeta)\right\}\exp\left(\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r\right.\right. \\ &\quad \left.\left.+\sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu}\right)\right]. \end{aligned} \tag{B.2}$$

Let

$$\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta) = E\left[Y_i(u)\left\{W^{(i)}_b(\zeta)\right\}\exp\left(\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r+\sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,t}(\zeta)\theta_{s\nu}\right)\right]. \tag{B.3}$$

Now we examine the asymptotic behavior of $G_{b,\zeta}(u;\beta,\Theta)$ and $G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)$.

Since $\{N_i(t) : t \in [0,\tau]\}$ and $\{Y_i(t) : t \in [0,\tau]\}$ are Glivenko-Cantelli classes (van der Vaart and Wellner 1996, Example 2.4.2), then by the Glivenko-Cantelli Theorem (Resnick 2013, Theorem 7.5.2), we have that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} dN_i(t) \xrightarrow{a.s.} d\mathbf{N}(u), \tag{B.4}$$

$$\frac{1}{n}G_{b,\zeta}(t;\beta,\Theta) \xrightarrow{a.s.} \mathcal{G}_{b,\zeta}(t;\beta,\Theta), \tag{B.5}$$

and

$$\frac{1}{n}G^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta) \xrightarrow{a.s.} \mathcal{G}^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta) \tag{B.6}$$

uniformly for $t \in [0,\tau]$, thus,

$$\frac{G^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta)}{G_{b,\zeta}(t;\beta,\Theta)} \xrightarrow{a.s.} \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(t;\beta,\Theta)} \tag{B.7}$$

uniformly for $t \in [0,\tau]$, where $\mathbf{N}(t) = E\{N_i(t)\}$.

For any $t > 0$, (B.4) and (B.7) are written as

$$d\bar{N}(t) = d\mathbf{N}(t) + o_p(1) \tag{B.8}$$

and

$$\frac{G^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta)}{G_{b,\zeta}(t;\beta,\Theta)} = \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(t;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(t;\beta,\Theta)} + o_p(1), \tag{B.9}$$

where $d\bar{N}(t) = \frac{1}{n}\sum_{i=1}^{n} dN_i(t)$.

Step 2: *Examine $U_{\beta;b,\zeta}(\beta,\Theta)$ given in (3.22) to show that*

$$U_{\beta;b,\zeta}(\beta,\Theta) = \frac{1}{n}\sum_{i=1}^{n} U_{\beta,i} + o_p(1).$$

Differentiating $\ell_{b,\zeta}(\beta,\Theta)$ in (3.11) with respect to β gives

$$U_{\beta;b,\zeta}(\beta,\Theta)=\frac{-1}{n}\sum_{i=1}^{n}\int\left\{W_b^{(i)}(\zeta)-\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right\}dN_i(u). \tag{B.10}$$

Adding and subtracting common terms each related to (B.8) or (B.9), we write (B.10) as

$$\begin{aligned}
U_{\beta;b,\zeta}(\beta,\Theta) &= \frac{-1}{n}\sum_{i=1}^{n}\int W_b^{(i)}(\zeta)dN_i(u)+\int\left\{\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right\}d\bar{N}(u)\\
&\quad-\int\left\{\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right\}d\mathbf{N}(u)+\int\left\{\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right\}d\mathbf{N}(u)\\
&\quad-\int\left\{\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}d\mathbf{N}(u)+\int\left\{\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}d\mathbf{N}(u)\\
&\quad-\int\left\{\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}d\bar{N}(u)+\int\left\{\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}d\bar{N}(u)\\
&= \frac{-1}{n}\sum_{i=1}^{n}\int W_b^{(i)}(\zeta)dN_i(u)\\
&\quad+\int\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\left\{d\bar{N}(u)-d\mathbf{N}(u)\right\}+\int\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}d\mathbf{N}(u)\\
&\quad+\int\left\{\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}-\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}\left\{d\bar{N}(u)-d\mathbf{N}(u)\right\}\\
&= \frac{-1}{n}\sum_{i=1}^{n}\int W_b^{(i)}(\zeta)dN_i(u)\\
&\quad+\int\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\left\{d\bar{N}(u)-d\mathbf{N}(u)\right\}+\int\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}d\mathbf{N}(u)+o_p(1)\\
&= \frac{-1}{n}\sum_{i=1}^{n}\int\left\{W_b^{(i)}(\zeta)-\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}dN_i(u)\\
&\quad+\int\left\{\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}-\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}d\mathbf{N}(u)+o_p(1),
\end{aligned} \tag{B.11}$$

where the second step is due to the combinations of terms, and the third step comes from

that $\int \left\{ \frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} - \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} \left\{ d\bar{N}(u) - d\mathbf{N}(u) \right\} = o_p(1)$ which is due to (B.8) and (B.9).

Examining the last integral in (B.11) gives that

$$
\begin{aligned}
& \int \left\{ \frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} - \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} d\mathbf{N}(u) \\
= & \int \left\{ \frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)\,\mathcal{G}_{b,\zeta}(u;\beta,\Theta) - G_{b,\zeta}(u;\beta,\Theta)\,\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)\,\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} d\mathbf{N}(u) \\
= & \frac{1}{n} \int \left[\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)\,\mathcal{G}_{b,\zeta}(u;\beta,\Theta) - G_{b,\zeta}(u;\beta,\Theta)\,\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\left\{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)\right\}^2} \right] d\mathbf{N}(u) + o_p(1) \\
= & \frac{1}{n} \sum_{i=1}^{n} \int \left[\frac{Y_i(u) \exp\left(\sum_{r\in V} W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E} W^{(i)}_{b,s}(\zeta) W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu} \right)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right. \\
& \left. \times \left\{ W^{(i)}_{b}(\zeta) - \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} \right] d\mathbf{N}(u) + o_p(1), \qquad \text{(B.12)}
\end{aligned}
$$

where the second equality is due to (B.5) and the last step is by (3.25) and (B.2).

Combining (B.11) and (B.12), we obtain that

$$
U_{\beta;b,\zeta}(\beta,\Theta) = \frac{1}{n} \sum_{i=1}^{n} U_{\beta,i} + o_p(1), \qquad \text{(B.13)}
$$

which, by Condition (C3), (3.23) and (B.9), is an i.i.d. sum of random variables with mean

zero , where

$$
\begin{aligned}
U_{\beta,i} &= \int - \left\{ W_b^{(i)}(\zeta) - \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} dN_i(u) \\
&\quad - \int \left[\frac{Y_i(u) \exp\left(\sum_{r\in V} W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E} W^{(i)}_{b,s}(\zeta) W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu} \right)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right. \\
&\quad \left. \times \left\{ W_b^{(i)}(\zeta) - \frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right\} \right] d\mathbf{N}(u). \qquad \text{(B.14)}
\end{aligned}
$$

Step 3: *We examine the bound of $U^{(j)}_{\beta;b,\zeta}(\beta,\Theta)$, the jth entry of $U_{\beta;b,\zeta}(\beta,\Theta)$ in (B.13), and show that there exist positive constants L and c_1 such that*

$$
P\left(\left| U^{(j)}_{\beta;b,\zeta}(\beta,\Theta) \right| \geq D \right) \;<\; 2\exp\left(\frac{-n^2D^2}{L^2} \right) + c_1 p^{-2} \quad \textit{for any} \quad D > 0. \qquad \text{(B.15)}
$$

Let $\mathcal{E} = \left\{ \max_{i,j} \left| W^{(i)}_{b,j}(\zeta) \right| \leq 4\ \log(p) \right\}$. Following the similar derivations for Proposition 4 of Yang et al. (2015), we can show that there exists a positive constant c_1, such that

$$
P(\mathcal{E}) \geq 1 - c_1 p^{-2}. \qquad \text{(B.16)}
$$

Finally, to show (B.15), we need Hoeffding's inequality which is included here for completeness.

Proposition B.2.1 *(Hoeffding's inequality)*
Let Z_i, $i = 1, \cdots, n$, be the i.i.d. random variables with the support $[a_i, b_i]$ where a_i and b_i are finite numbers for $i = 1, \cdots, n$. Let $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then for any $D > 0$,

$$
P\left(\left| \bar{Z} - E(\bar{Z}) \right| \geq D \right) \leq 2\exp\left(-\frac{n^2D^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right).
$$

For $j = 1, \cdots, p$, we write $U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)$ as $\frac{1}{n}\sum_{i=1}^{n} U^{(j)}_{\beta,i}$, with $U^{(j)}_{\beta,i}$ representing the jth entry of $U_{\beta,i}$ in (B.14). (B.16) shows that $W^{(i)}_{b,j}(\zeta)$ is bounded for all i, j, b, and ζ with a high probability, which suggests that by (B.14) conditional on the event $\mathcal{E}$, $\left|U^{(j)}_{\beta,i}\right| < L$ for some constant $L > 0$.

By (3.23), $E\left\{U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right\} = 0$. Then by Hoeffding's inequality with replacing $\bar{Z}$ and $E\left(\bar{Z}\right)$ by $U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)$ and $E\left\{U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right\}$, respectively, we have that for any $D > 0$,

$$P\left(\left|U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right| \geq D \;\middle|\; \mathcal{E}\right) < 2\exp\left(\frac{-n^2D^2}{L^2}\right). \tag{B.17}$$

Combining (B.16) and (B.17), we obtain that for any $D > 0$,

$$\begin{aligned} P\left(\left|U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right| \geq D\right) &\leq P\left(\left|U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right| \geq D \;\middle|\; \mathcal{E}\right) + P\left(\mathcal{E}^c\right) \\ &< 2\exp\left(\frac{-n^2D^2}{L^2}\right) + c_1p^{-2}. \end{aligned}$$

Step 4: *We examine $U_{\beta,\mathcal{S}_1;b,\zeta}(\beta, \Theta)$ and $U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta, \Theta)$, defined in Lemma B.2.1, and show that*

$$\begin{aligned} P\left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta, \Theta)\right\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_{n1}}{4}\right) &\leq d_\beta\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1p^{-2}\right\}; \\ P\left(\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta, \Theta)\right\|_\infty > \frac{\alpha\lambda_{n1}}{4}\right) &\leq (p - d_\beta)\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1p^{-2}\right\}. \end{aligned}$$

First, for any constant α in $(0, 1)$, we have that

$$\begin{aligned} P\left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta, \Theta)\right\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_{n1}}{4}\right) &\leq \sum_{j\in\mathcal{S}_1} P\left(\left|U^{(j)}_{\beta;b,\zeta}(\beta, \Theta)\right| > \frac{\alpha}{2-\alpha}\frac{\lambda_{n1}}{4}\right) \\ &\leq \sum_{j\in\mathcal{S}_1}\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1p^{-2}\right\} \\ &= d_\beta\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1p^{-2}\right\}, \end{aligned}$$

where the first step is due to the fact that $\|a\|_\infty \leq \|a\|_1$ for a vector a, the second step applies (B.15) and that $0 < \frac{\alpha}{2-\alpha} < 1$, and the last step is due to the definition $d_\beta = |\mathcal{S}_1|$,

Similarly,

$$\begin{aligned} P\left(\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha\lambda_{n1}}{4}\right) &\leq \sum_{j\in\mathcal{S}_1^c} P\left(\left|U^{(j)}_{\beta;b,\zeta}(\beta,\Theta)\right| > \frac{\alpha\lambda_{n1}}{4}\right) \\ &\leq \sum_{j\in\mathcal{S}_1^c}\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1 p^{-2}\right\} \\ &= (p-d_\beta)\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n1}^2\right) + c_1 p^{-2}\right\}, \end{aligned}$$

where the second step is due to (B.15) and $0 < \alpha < 1$. Thus, the proof of Lemma B.2.1 completes. □

Lemma B.2.2 *Under Conditions (C3) and (C5) in Appendix B.1, we have*

$$\begin{aligned} P\left(\|U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta,\Theta)\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_{n2}}{4}\right) &\leq d_\Theta\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n2}^2\right) + c_1 p^{-2}\right\}; \\ P\left(\left\|U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha\lambda_{n2}}{4}\right) &\leq \left(p^2-p-d_\Theta\right)\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda_{n2}^2\right) + c_1 p^{-2}\right\}, \end{aligned}$$

where c_1 is a positive constant, λ_{n2} is the tuning parameter in (3.14), and d_Θ is defined after Theorem 3.3.1 in Section 3.3.

Proof:

The proof of Lemma B.2.2 follows the steps similar to those for Lemma B.2.1; the only differences are caused from differentiating with respect to different parameters β and Θ. For completeness, here we outline key steps.

Let $\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu} = \mathrm{vec}\left(W^{(i)}_b(\zeta)W^{(i)\top}_b(\zeta) - \mathrm{diag}\left\{W^{(i)}_b(\zeta)W^{(i)\top}_b(\zeta)\right\}\right)$. Differentiating (3.25) with respect to Θ yields

$$\begin{aligned} G^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta) &= \sum_{i=1}^n\left[Y_i(u)\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}\right\}\right. \\ &\quad \left.\times\exp\left\{\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu}\right\}\right]. \end{aligned} \tag{B.18}$$

Let

$$\begin{aligned}\mathcal{G}^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta) &= E\left[Y_i(u)\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}\right\}\right.\\ &\quad\left.\times\exp\left\{\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r+\sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu}\right\}\right].\end{aligned}$$

. Then similar to (B.9), we can show that

$$\frac{G^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}=\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}+o_p(1). \tag{B.19}$$

On the other hand, differentiating $\ell_{b,\zeta}(\beta,\Theta)$ in (3.11) with respect to Θ gives

$$U_{\Theta;b,\zeta}(\beta,\Theta) = -\sum_{i=1}^{n}\int\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}-\frac{G^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right\}dN_i(u). \tag{B.20}$$

By the similar derivations for (B.11) and (B.12), we can show that

$$U_{\Theta;b,\zeta}(\beta,\Theta)=\frac{1}{n}\sum_{i=1}^{n}U_{\Theta,i}+o_p(1), \tag{B.21}$$

which, by Condition (C3), (3.23) and (B.19), is an i.i.d. sum of random variables with mean zero, where

$$\begin{aligned}U_{\Theta,i} &= \int\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}-\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}dN_i(u)\\ &\quad-\int\left[\frac{Y_i(u)\exp\left(\sum_{r\in V}W^{(i)}_{b,r}(\zeta)\beta_r+\sum_{(s,\nu)\in E}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu}\right)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right.\\ &\quad\left.\times\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}-\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right\}\right]d\mathbf{N}(u).\end{aligned} \tag{B.22}$$

For $j=1,\cdots,p^2-p$, let $U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)$ denote the jth entry of the (p^2-p)-dimensional vector $U_{\Theta;b,\zeta}(\beta,\Theta)$ in (B.21). We write $U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)$ as $\frac{1}{n}\sum_{i=1}^{n}U^{(j)}_{\Theta,i}$, with $U^{(j)}_{\Theta,i}$ representing the

jth entry of $U_{\Theta,i}$ in (B.22). Similar to the argument in Step 3 of the proof in Lemma B.2.1, we have that in the presence of $\mathcal{E}$, $\left|U^{(j)}_{\Theta,i}\right| < L$ for some constant $L > 0$.

Since by (3.23), $E\left\{U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right\} = 0$, then by Hoeffding's inequality in Proposition B.1 with replacing $\bar{Z}$ and $E\left(\bar{Z}\right)$ by $U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)$ and $E\left\{U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right\}$, respectively, we have that for any $D > 0$,

$$\begin{aligned} P\left(\left|U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right| \geq D\right) &\leq P\left(\left\{\left|U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right| \geq D\right\} \cap \mathcal{E}\right) + P\left(\mathcal{E}^c\right) \\ &< 2\exp\left(\frac{-n^2D^2}{L^2}\right) + c_1p^{-2}. \end{aligned} \tag{B.23}$$

As a result, for any constant α in $(0,1)$, we have that

$$\begin{aligned} P\left(\left\|U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_{n2}}{4}\right) &\leq \sum_{j=1}^{|\mathcal{S}_2|} P\left(\left|U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right| > \frac{\alpha}{2-\alpha}\frac{\lambda_{n2}}{4}\right) \\ &\leq \sum_{j=1}^{|\mathcal{S}_2|}\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda^2_{n2}\right) + c_1p^{-2}\right\} \\ &= d_\Theta\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda^2_{n2}\right) + c_1p^{-2}\right\}, \end{aligned}$$

where the second step applies (B.23) and that $0 < \frac{\alpha}{2-\alpha} < 1$, and the last step is due to the definition $d_\Theta = |\mathcal{S}_2|$.

Similarly,

$$\begin{aligned} P\left(\left\|U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta,\Theta)\right\|_\infty > \frac{\alpha\lambda_{n2}}{4}\right) &\leq \sum_{j=1}^{|\mathcal{S}_2^c|} P\left(\left|U^{(j)}_{\Theta;b,\zeta}(\beta,\Theta)\right| > \frac{\alpha\lambda_{n2}}{4}\right) \\ &\leq \sum_{j=1}^{|\mathcal{S}_2^c|}\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda^2_{n2}\right) + c_1p^{-2}\right\} \\ &= \left(p^2 - p - d_\Theta\right)\left\{2\exp\left(\frac{-1}{L^2}n^2\lambda^2_{n2}\right) + c_1p^{-2}\right\}. \end{aligned}$$

Therefore, the proof of Lemma B.2.2 completes. □

Next, we consider the differences between the parameters and their estimators and establish their upper bounds.

Lemma B.2.3 *Under regularity conditions in Appendix B.1, we have that*

$$\left\|\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}\right\|_2 \leq \lambda_{n1}\sqrt{d_\beta}\frac{1}{\kappa_1}\left\{1 + 2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\} \tag{B.24}$$

and

$$\left\|vec\left(\widehat{\Theta}_{\mathcal{S}_2}\right) - vec\left(\Theta_{0;\mathcal{S}_2}\right)\right\|_2 \leq \lambda_{n2}\sqrt{d_\Theta}\frac{1}{\kappa_2}\left\{1 + 2\left(\sum_{s\neq\nu} v_{s\nu}^2\right)^{2}\right\}, \tag{B.25}$$

where $\|\cdot\|_2$ represents the L_2-norm, κ_1 and κ_2 are defined in Condition (C7), λ_{n1} and λ_{n2} are defined in (3.14), and w_r and $v_{s\nu}$ are defined in (3.12) and (3.13), respectively.

Proof of (B.24):

For $\widehat{\beta}_b(\zeta)$ defined by (3.14), we write $\widehat{\beta}_b(\zeta) = \left(\widehat{\beta}_{b;\mathcal{S}_1}^\top(\zeta), \widehat{\beta}_{b;\mathcal{S}_1^c}^\top(\zeta)\right)^\top = \left(\widehat{\beta}_{b;\mathcal{S}_1}^\top(\zeta), 0_{(p-d_\beta)}^\top\right)^\top$, and accordingly, we write the true value of β as $\beta_0 = \left(\beta_{0;\mathcal{S}_1}^\top, \beta_{0;\mathcal{S}_1^c}^\top\right)^\top = \left(\beta_{0;\mathcal{S}_1}^\top, 0_{(p-d_\beta)}^\top\right)^\top$, where 0_b stands for the b-dimensional zero vector.
Part 1: *For $\zeta \in \mathcal{Z}$ and $b = 1, \cdots, B$, let $\widehat{u}_\beta(b,\zeta) = \widehat{\beta}_{b;\mathcal{S}_1}(\zeta) - \beta_{0;\mathcal{S}_1}$. Show that*

$$\|\widehat{u}_\beta(b,\zeta)\|_2 < \lambda_{n1}\sqrt{d_\beta}\frac{1}{\kappa_1}\left\{1 + 2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\}. \tag{B.26}$$

We note that by the coordinate-descent method, once Θ is given, then $\widehat{\beta}_b(\zeta)$ can be obtained by (3.18). In this sense, when implementing (3.14) for obtaining $\widehat{\beta}_b(\zeta)$, we fix $\Theta = \Theta_0$ and consider the difference related to (3.14)

$$\Psi_\beta(u) = \left\{\ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1} + u, \Theta_0\right) - \ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\right\} + \lambda_{n1}\left\{\rho_1\left(\beta_{0;\mathcal{S}_1} + u\right) - \rho_1\left(\beta_{0;\mathcal{S}_1}\right)\right\}, \tag{B.27}$$

where we express any parameter value $\beta_{\mathcal{S}_1}$ as $\beta_{0;\mathcal{S}_1} + u$.

Note that $\Psi_\beta(u)$ is a convex function since $\ell_{b,\zeta}(\beta, \Theta)$ defined in (3.11) and $\rho_1(\cdot)$ defined in (3.12) are both convex functions. Similar to the derivations for Lemma 3 of Ravikumar et al. (2010), to show (B.26), it suffices to show that

$$\Psi_\beta(u) > 0 \text{ for any } u \text{ with } \|u\|_2 = \mathcal{B}, \tag{B.28}$$

where $\mathcal{B} = \lambda_{n1}\sqrt{d_\beta}\frac{1}{\kappa_1}\left\{1 + 2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\}$.

To see why this is true, we apply the argument of contradiction. Suppose (B.28) is true but $\|\widehat{u}_\beta(b,\zeta)\|_2 > \mathcal{B}$. Then there exists a constant $\xi \in (0,1)$ such that $\|\xi\widehat{u}_\beta(b,\zeta)\|_2 = \mathcal{B}$. By (B.28), $\Psi_\beta\left(\xi\widehat{u}_\beta(b,\zeta)\right) > 0$, i.e.,

$$\Psi_\beta\left(\xi\widehat{u}_\beta(b,\zeta) + (1-\xi)0_{d_\beta}\right) > 0. \tag{B.29}$$

By (3.14) and (3.18), $\widehat{u}_\beta(b,\zeta)$ minimizes (B.27). Because $\Psi_\beta(0_{d_\beta}) = 0$, we obtain that $\Psi_\beta\left(\widehat{u}_\beta(b,\zeta)\right) < 0$. Since $\Psi_\beta\left(u\right)$ is a convex function, we have that

$$\begin{aligned}\Psi_\beta\left(\xi\widehat{u}_\beta(b,\zeta) + (1-\xi)0_{d_\beta}\right) &\leq \xi\Psi_\beta\left(\widehat{u}_\beta(b,\zeta)\right) + (1-\xi)\Psi_\beta\left(0_{d_\beta}\right)\\ &= \xi\Psi_\beta\left(\widehat{u}_\beta(b,\zeta)\right)\\ &\leq 0,\end{aligned}$$

thus, contradicting (B.29).

As a result, in the following development, we need only to show (B.28) by the following two steps.

Step 1: *Show that for any $\mathcal{M} > 0$, if $\|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}$, then*

$$\Psi_\beta(u) > \left(\lambda_{n1}\sqrt{d_\beta}\right)^2 \mathcal{M}\left\{-\frac{1}{4} + \frac{\kappa_1}{2}\mathcal{M} - \left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\}. \tag{B.30}$$

Examining the first difference of (B.27) using the second order Taylor series expansion of $\ell_{b,\zeta}(\beta_{0;\mathcal{S}_1} + u, \Theta_0)$ around $u = 0_{d_\beta}$, we have

$$\begin{aligned}&\ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1} + u, \Theta_0\right) - \ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\\ =\ & U_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)u + \frac{1}{2}u^\top I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1} + ku, \Theta_0\right)u,\end{aligned}$$

where k is some constant in $(0,1)$, and $U_{\beta;b,\zeta}(\cdot)$ and $I_{\beta;b,\zeta}(\cdot)$ are defined in (3.22) and (3.24), respectively. Then (B.27) is written as

$$\Psi_\beta(u) \;=\; T_1 + T_2 + T_3, \tag{B.31}$$

where

$$\begin{aligned} T_1 &= U_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right) u; && \text{(B.32a)} \\ T_2 &= \frac{1}{2} u^\top I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1} + ku, \Theta_0\right) u; && \text{(B.32b)} \\ T_3 &= \lambda_{n1}\left\{\rho_1\left(\beta_{0;\mathcal{S}_1} + u\right) - \rho_1\left(\beta_{0;\mathcal{S}_1}\right)\right\}. && \text{(B.32c)} \end{aligned}$$

The remaining task is to individually examine T_1, T_2 and T_3 for their bounds when $\|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}$. We proceed with the following three steps.

Step 1.1: *Show that*

$$\|T_1\|_1 < \frac{1}{4}\left(\sqrt{d_\beta}\lambda_{n1}\right)^2 \mathcal{M} \quad \textit{for} \quad \|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}. \tag{B.33}$$

By taking the L_1-norm, (B.32a) becomes

$$\begin{aligned} \|T_1\|_1 &= \left\|U_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right) u\right\|_1 \\ &\leq \left\|U_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\right\|_\infty \|u\|_1 \\ &\leq \left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\right\|_\infty + \left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\right\|_\infty\right)\sqrt{d_\beta}\,\|u\|_2 \\ &< \frac{3\alpha - \alpha^2}{1-\alpha}\frac{\lambda_{n1}}{4}\sqrt{d_\beta}\mathcal{B} \\ &< \frac{1}{4}\left(\sqrt{d_\beta}\lambda_{n1}\right)^2 \mathcal{M}, \end{aligned}$$

where the second inequality is due to that $U_{\beta,\mathcal{S}_1;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)$ and $U_{\beta,\mathcal{S}_1;b,\zeta}\left(\beta_{0;\mathcal{S}_1^c}, \Theta_0\right)$ are sub-vectors of $U_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)$, the third inequality is due to Lemma B.2.1 and the Cauchy-Schwarz inequality on $\|u\|_1$, and the fourth inequality is due to $\mathcal{B} = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}$ and $0 < \alpha < 1$.

Step 1.2: *Show that*

$$T_2 > \frac{1}{2}\kappa_1\left(\sqrt{d_\beta}\lambda_{n1}\right)^2 \mathcal{M}^2 \quad \textit{for} \quad \|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}. \tag{B.34}$$

Since $I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)$ is a positive definite matrix and $\Lambda_{min}\left(I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_0\right)\right) > \kappa_1$ by Condition (C7), then we have that for any u with $\|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}$,

$$\begin{aligned} T_2 &= \frac{1}{2}u^\top I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1} + ku, \Theta_0\right) u \\ &\geq \min_{u:\|u\|_2=\sqrt{d_\beta}\lambda_{n1}\mathcal{M}} \left\{\frac{1}{2}u^\top I_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1} + ku, \Theta_0\right) u\right\} \\ &> \frac{1}{2}\kappa_1\left(\sqrt{d_\beta}\lambda_{n1}\right)^2\mathcal{M}^2. \end{aligned}$$

Step 1.3: *Show that*

$$T_3 \geq -\left(\sqrt{d_\beta}\lambda_{n1}\right)^2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\mathcal{M} \quad \textit{for} \quad \|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}. \tag{B.35}$$

For $r \in \mathcal{S}_1$, let $z_r = w_r u_r$ and $\widetilde{\beta}_{0r} = w_r\beta_{0r}$, where u_r and β_{0r} represent the rth component of u and β_0, respectively. Then by (3.12), we have

$$\begin{aligned} \rho_1\left(\beta_{0;\mathcal{S}_1} + u\right) - \rho_1\left(\beta_{0;\mathcal{S}_1}\right) &= \sum_{r\in\mathcal{S}_1} w_r\left|\beta_{0r} + u_r\right| - \sum_{r\in V\mathcal{S}_1} w_r\left|\beta_{0r}\right| \\ &= \sum_{r\in\mathcal{S}_1}\left|\widetilde{\beta}_{0r} + z_r\right| - \sum_{r\in\mathcal{S}_1}\left|\widetilde{\beta}_{0r}\right| \\ &\geq -\sum_{r\in\mathcal{S}_1}|z_r| \\ &= -\|z\|_1 \\ &\geq -\sqrt{d_\beta}\,\|z\|_2^2 \\ &= -\sqrt{d_\beta}\sum_{r\in\mathcal{S}_1} z_r^2, \end{aligned} \tag{B.36}$$

where the third step is due to the triangle inequality, and the second last step is due to the Cauchy-Schwarz inequality.

Furthermore, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sum_{r\in\mathcal{S}_1} z_r^2 &= \sum_{r\in\mathcal{S}_1} w_r^2 u_r^2 \\ &\leq \left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2} \left(\sum_{r\in\mathcal{S}_1} u_r^2\right)^{1/2} \\ &= \left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2} \|u\|_2 . \end{aligned} \tag{B.37}$$

Therefore, combining (B.32c), (B.36) and (B.37), we obtain that for any u with $\|u\|_2 = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}$,

$$\begin{aligned} T_3 &= \lambda_{n1}\left\{\rho_1\left(\beta_{0;\mathcal{S}_1}+u\right)-\rho_1\left(\beta_{0;\mathcal{S}_1}\right)\right\} \\ &\geq -\lambda_{n1}\sqrt{d_\beta}\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\|u\|_2 \\ &= -\left(\sqrt{d_\beta}\lambda_{n1}\right)^2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\mathcal{M}. \end{aligned}$$

Combining (B.31), (B.33), (B.34) and (B.35) gives (B.30).

Step 2: *Show (B.28).*

To ensure the right-hand-side of (B.30) be bounded below by zero, we must have

$$-\frac{1}{4}+\frac{\kappa_1}{2}\mathcal{M}-\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2} > 0,$$

which is equivalent to requiring

$$\mathcal{M} > \frac{1}{\kappa_1}\left\{\frac{1}{2}+2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\}. \tag{B.38}$$

Hence, setting $\mathcal{M}^* = \frac{1}{\kappa_1}\left\{1+2\left(\sum_{r\in\mathcal{S}_1} w_r^2\right)^{1/2}\right\}$, the right-hand-side of (B.38), we see that $\mathcal{B}^* = \sqrt{d_\beta}\lambda_{n1}\mathcal{M}^*$ and that (B.28) holds. Therefore, (B.26) is shown.

Part 2: *Show (B.24).*

By (3.15) and (B.26), we obtain that

$$\begin{aligned}\left\|\widehat{\beta}_{\mathcal{S}_1}(\zeta)-\beta_{0;\mathcal{S}_1}\right\|_2 &\leq \frac{1}{B}\sum_{b=1}^{B}\left\|\widehat{\beta}_{b;\mathcal{S}_1}(\zeta)-\beta_{0;\mathcal{S}_1}\right\|_2 \\ &< \lambda_{n1}\sqrt{d_\beta}\frac{1}{\kappa_1}\left\{1+2\left(\sum_{r\in\mathcal{S}_1}w_r^2\right)^{1/2}\right\}, \qquad \text{(B.39)}\end{aligned}$$

and let $\zeta \to -1$, (B.39) becomes

$$\left\|\widehat{\beta}_{\mathcal{S}_1}-\beta_{0;\mathcal{S}_1}\right\|_2 \leq \lambda_{n1}\sqrt{d_\beta}\frac{1}{\kappa_1}\left\{1+2\left(\sum_{r\in\mathcal{S}_1}w_r^2\right)^{1/2}\right\}.$$

Proof of (B.25):

Let $\operatorname{vec}\left(\widehat{\Theta}_b(\zeta)\right)=\left(\operatorname{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)^\top,\operatorname{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2^c}(\zeta)\right)^\top\right)^\top=\left(\operatorname{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)^\top,0^\top_{(p^2-p-d_\Theta)}\right)^\top$ and $\operatorname{vec}(\Theta_0)=\left(\operatorname{vec}(\Theta_{0;\mathcal{S}_2})^\top,\operatorname{vec}\left(\Theta_{0;\mathcal{S}_2^c}\right)^\top\right)^\top=\left(\operatorname{vec}(\Theta_{0;\mathcal{S}_2})^\top,0^\top_{(p^2-p-d_\Theta)}\right)^\top$.

Similar to the proof of (B.24), we now fix $\beta=\beta_0$, and consider the difference related to (3.14)

$$\begin{aligned}\Psi_\Theta(v) &= \ell_{b,\zeta}\left(\beta_0,\operatorname{vec}(\Theta_{0;\mathcal{S}_2})+v\right)-\ell_{b,\zeta}\left(\beta_0,\operatorname{vec}(\Theta_{0;\mathcal{S}_2})\right) \\ &\quad +\lambda_{n2}\left\{\rho_2\left(\operatorname{vec}(\Theta_{0;\mathcal{S}_2})+v\right)-\rho_2\left(\operatorname{vec}(\Theta_{0;\mathcal{S}_2})\right)\right\},\end{aligned}$$

for where we write $\operatorname{vec}(\Theta_{\mathcal{S}_2})$ as $\operatorname{vec}(\Theta_{0;\mathcal{S}_2})+v$. The main purpose is to find the bound for $\widehat{v}_\Theta(b,\zeta)=\operatorname{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)-\operatorname{vec}(\Theta_{0;\mathcal{S}_2})$.

By the similar derivations to the proof of (B.24), we obtain

$$\left\|\operatorname{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)-\operatorname{vec}(\Theta_{0;\mathcal{S}_2})\right\|_2<\lambda_{n2}\sqrt{d_\Theta}\frac{1}{\kappa_2}\left\{1+2\left(\sum_{s\neq\nu}v_{s\nu}^2\right)^2\right\}.$$

Therefore, taking average with respect to b and taking $\zeta=-1$ gives the desired result, (B.25). □

Lemma B.2.4 *Let*

$$R_{n;\beta} = \left\{ I_{\beta;b,\zeta}\left(\bar{\beta}, \Theta_0\right) - I_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right) \right\} \left(\widehat{\beta}_b(\zeta) - \beta_0 \right)$$

with $\bar{\beta}$ *being a vector which lies on the "line segment" between* $\widehat{\beta}_b(\zeta)$ *and* β_0*, and let*

$$R_{n;\Theta} = \left\{ I_{\Theta;b,\zeta}\left(\beta_0, \bar{\Theta}\right) - I_{\Theta;b,\zeta}\left(\beta_0, \Theta_0\right) \right\} \left\{ vec(\widehat{\Theta}_b(\zeta)) - vec(\Theta_0) \right\}$$

with $\bar{\Theta}$ *being a matrix whose vectorization* $vec(\bar{\Theta})$ *lies on the "line segment" between between* $vec(\widehat{\Theta}_b(\zeta))$ *and* $vec(\Theta_0)$*. Then under regularity conditions in Appendix B.1, we have*

$$\left\| R_{n;\beta} \right\|_\infty \ < \ \frac{\alpha}{2-\alpha} \frac{\lambda_{n1}}{4} \tag{B.40}$$

and

$$\left\| R_{n;\Theta} \right\|_\infty \ < \ \frac{\alpha}{2-\alpha} \frac{\lambda_{n2}}{4}. \tag{B.41}$$

Proof of (B.40):

The proof consists of the following two steps.
Step 1: *Show that*

$$P\left(\left\| I_{\beta;b,\zeta}\left(\bar{\beta}, \Theta_0\right) - I_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right) \right\|_\infty > \frac{\alpha}{2-\alpha} \frac{\left(\mathcal{M}_\beta^*\right)^{-1}}{4} \frac{1}{d_\beta} \ \middle| \ \mathcal{E} \right) < 2n \exp\left(\frac{-n}{4L^2} \right).$$

First, we write

$$\begin{aligned}
& I_{\beta;b,\zeta}\left(\bar{\beta}, \Theta_0\right) - I_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right) \\
= \ & \sum_{i=1}^n \int \left\{ \frac{G^{(2)}_{\beta;b,\zeta}\left(u; \bar{\beta}, \Theta_0\right)}{G_{b,\zeta}\left(u; \bar{\beta}, \Theta_0\right)} - \left(\frac{G^{(1)}_{\beta;b,\zeta}\left(u; \bar{\beta}, \Theta_0\right)}{G_{b,\zeta}\left(u; \bar{\beta}, \Theta_0\right)} \right)^{\otimes 2} \right\} dN_i(u) \\
& - \sum_{i=1}^n \int \left\{ \frac{G^{(2)}_{\beta;b,\zeta}\left(u; \beta_0, \Theta_0\right)}{G_{b,\zeta}\left(u; \beta_0, \Theta_0\right)} - \left(\frac{G^{(1)}_{\beta;b,\zeta}\left(u; \beta_0, \Theta_0\right)}{G_{b,\zeta}\left(u; \beta_0, \Theta_0\right)} \right)^{\otimes 2} \right\} dN_i(u) \\
\triangleq \ & \sum_{i=1}^n T_{3i} - \sum_{i=1}^n T_{4i},
\end{aligned} \tag{B.42}$$

where $G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta) = \nabla^2_\beta G_{b,\zeta}(u;\beta,\Theta)$ and $a^{\otimes 2} = aa^\top$ for any nonzero vector a. Let

$$\begin{aligned}
&\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta) \\
&= E\left[Y_i(u)\left\{W^{(i)}_b(\zeta)\right\}^{\otimes 2}\exp\left(\sum_{r\in V} W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E} W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,t}(\zeta)\theta_{s\nu}\right)\right]. \qquad \text{(B.43)}
\end{aligned}$$

By the derivation similar to (B.6), we have that as $n \to \infty$,

$$\frac{1}{n}G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta) \xrightarrow{a.s.} \mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta) \qquad \text{(B.44)}$$

uniformly in $u \in [0,\tau]$. By (B.5), (B.6), and (B.44), we have that as $n \to \infty$,

$$\sup_{u\in[0,\tau]}\left|\frac{G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} - \frac{\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right| \to 0$$

and

$$\sup_{u\in[0,\tau]}\left|\left(\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2} - \left(\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}\right| \to 0.$$

Then there are some constants K_2, $K_3 > 0$, such that

$$\sup_{u\in[0,\tau]}\left|\frac{G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} - \frac{\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right| < K_2 \qquad \text{(B.45a)}$$

$$\sup_{u\in[0,\tau]}\left|\left(\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2} - \left(\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}\right| < K_3. \qquad \text{(B.45b)}$$

Combining (B.45a) and (B.45b) gives

$$\begin{aligned}
& \sup_{u\in[0,\tau]}\left|\left\{\frac{G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}-\left(\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}\right\}\right. \\
& \left.-\left\{\frac{\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}-\left(\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}\right\}\right| \\
\leq\ & \sup_{u\in[0,\tau]}\left|\frac{G^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}-\frac{\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right| \\
& +\sup_{u\in[0,\tau]}\left|\left(\frac{G^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}-\left(\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)}\right)^{\otimes 2}\right| \\
<\ & K_2+K_3\triangleq K_1. \qquad \text{(B.46)}
\end{aligned}$$

Since the result (B.46) satisfies the requirement of the proof in Lemma A.3 of Lin and Lv (2013), then by the derivations similar to Lemma A.3 of Lin and Lv (2013), we can show that $E(T_{ki})<C_1n^{-1/2}$ for a positive constant C_1 and $k=3,4$.

Hence, by Theorem 9 in Massart (2000), we have

$$P\left\{|T_{ki}|>C_1n^{-1/2}\left(\frac{1}{2}+\frac{D}{2}\right)\ \middle|\ \mathcal{E}\right\}<\exp\left(\frac{-D^2}{4L^2}\right) \qquad \text{(B.47)}$$

for $k=3,4$. Hence, by (B.47), we have

$$\begin{aligned}
P\left\{|T_{3i}-T_{4i}|>C'n^{-1/2}(1+D)\ \middle|\ \mathcal{E}\right\} &\leq 2P\left\{|T_{3i}|>C'n^{-1/2}\left(\frac{1}{2}+\frac{D}{2}\right)\ \middle|\ \mathcal{E}\right\} \\
&< 2\exp\left(\frac{-D^2}{4L^2}\right). \qquad \text{(B.48)}
\end{aligned}$$

Finally, by (B.48), we can show that

$$\begin{aligned}
& P\left(\left\|I_{\beta;b,\zeta}\left(\bar{\beta},\Theta_0\right)-I_{\beta;b,\zeta}\left(\beta_0,\Theta_0\right)\right\|_\infty>\frac{\alpha}{2-\alpha}\frac{\left(\mathcal{M}^*_\beta\right)^{-1}}{4}\frac{1}{d_\beta}\ \middle|\ \mathcal{E}\right) \\
\leq\ & \sum_{i=1}^{n}P\left(|T_{3i}-T_{4i}|>\frac{\alpha}{2-\alpha}\frac{\left(\mathcal{M}^*_\beta\right)^{-1}}{4}\frac{1}{d_\beta}\ \middle|\ \mathcal{E}\right) \\
<\ & 2n\exp\left(\frac{-n}{4L^2}\right). \qquad \text{(B.49)}
\end{aligned}$$

Step 2: *Examine* $\|R_{n;\beta}\|_\infty$*, the infinity norm of* $R_{n;\beta}$*, and show (B.40).*

By the definition of $R_{n;\beta}$ in Lemma B.2.4, the bound of $\|R_{n;\beta}\|_\infty$ can be determined by

$$\begin{aligned}
\|R_{n;\beta}\|_\infty &\leq \left\|I_{\beta;b,\zeta}\left(\bar{\beta},\Theta_0\right) - I_{\beta;b,\zeta}\left(\beta_0,\Theta_0\right)\right\|_\infty \left\|\widehat{\beta}_b(\zeta) - \beta_0\right\|_1 \\
&\leq \left\|I_{\beta;b,\zeta}\left(\bar{\beta},\Theta_0\right) - I_{\beta;b,\zeta}\left(\beta_0,\Theta_0\right)\right\|_\infty \sqrt{d_\beta}\left\|\widehat{\beta}_b(\zeta) - \beta_0\right\|_2 \\
&< \frac{\alpha}{2-\alpha}\frac{\lambda_{n1}}{4},
\end{aligned}$$

where the second step is due to the Cauchy-Schwarz inequality, and the third step is due to (B.26) and (B.49).

Proof of (B.41):

This proof is similar to the proof of (B.40) except for the consideration of $I_{\Theta;b,\zeta}(\cdot,\cdot)$. Specifically, write $I_{\Theta;b,\zeta}\left(\beta_0,\bar{\Theta}\right) - I_{\Theta;b,\zeta}\left(\beta_0,\Theta_0\right)$ as

$$\begin{aligned}
& I_{\Theta;b,\zeta}\left(\beta_0,\bar{\Theta}\right) - I_{\Theta;b,\zeta}\left(\beta_0,\Theta_0\right) \\
= & \sum_{i=1}^n \int \left\{ \frac{G^{(2)}_{\Theta;b,\zeta}\left(u;\beta_0,\bar{\Theta}\right)}{G_{b,\zeta}\left(u;\beta_0,\bar{\Theta}\right)} - \left(\frac{G^{(1)}_{\Theta;b,\zeta}\left(u;\beta_0,\bar{\Theta}\right)}{G_{b,\zeta}\left(u;\beta_0,\bar{\Theta}\right)} \right)^{\otimes 2} \right\} dN_i(u) \\
& - \sum_{i=1}^n \int \left\{ \frac{G^{(2)}_{\Theta;b,\zeta}\left(u;\beta_0,\Theta_0\right)}{G_{b,\zeta}\left(u;\beta_0,\Theta_0\right)} - \left(\frac{G^{(1)}_{\Theta;b,\zeta}\left(u;\beta_0,\Theta_0\right)}{G_{b,\zeta}\left(u;\beta_0,\Theta_0\right)} \right)^{\otimes 2} \right\} dN_i(u),
\end{aligned}$$

where $G^{(2)}_{\Theta;b,\zeta}\left(u;\beta,\Theta\right) = \nabla^2_\Theta G_{b,\zeta}\left(u;\beta,\Theta\right)$. Let

$$\begin{aligned}
\mathcal{G}^{(2)}_{\Theta;b,\zeta}\left(u;\beta,\Theta\right) &= E\left[Y_i(u) \left\{ \left(W^{(i)}_{b,s}(\zeta) W^{(i)}_{b,\nu}(\zeta) \right)^{\otimes 2}_{s\neq\nu} \right\} \right. \\
&\quad \left. \times \exp\left\{ \sum_{r\in V} W^{(i)}_{b,r}(\zeta)\beta_r + \sum_{(s,\nu)\in E} W^{(i)}_{b,s}(\zeta) W^{(i)}_{b,\nu}(\zeta)\theta_{s\nu} \right\} \right].
\end{aligned}$$

.

By the Glivenko-Cantelli Theorem (e.g., Resnick 2013, Theorem 7.5.2), we have that as $n \to \infty$,

$$\frac{1}{n} G^{(2)}_{\Theta;b,\zeta}\left(u;\beta,\Theta\right) \xrightarrow{a.s.} \mathcal{G}^{(2)}_{\Theta;b,\zeta}\left(u;\beta,\Theta\right)$$

uniformly in $u \in [0, \tau]$. By the derivation similar to (B.46), we can show that

$$\begin{aligned} & \sup_{u\in[0,\tau]} \left| \left\{ \frac{G^{(2)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} - \left(\frac{G^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{G_{b,\zeta}(u;\beta,\Theta)} \right)^{\otimes 2} \right\} \right. \\ & \left. - \left\{ \frac{\mathcal{G}^{(2)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} - \left(\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(u;\beta,\Theta)}{\mathcal{G}_{b,\zeta}(u;\beta,\Theta)} \right)^{\otimes 2} \right\} \right| \\ < \; & K_2^* \end{aligned}$$

for some constant $K_2^* > 0$. As a result, by the derivations similar to (B.49), we have

$$\begin{aligned} & P\left(\left\| I_{\Theta;b,\zeta}\left(\beta_0, \bar{\Theta}\right) - I_{\Theta;b,\zeta}\left(\beta_0, \Theta_0\right) \right\|_\infty > \frac{\alpha}{2-\alpha} \frac{(\mathcal{M}^*_\Theta)^{-1}}{4} \frac{1}{d_\Theta} \;\middle|\; \mathcal{E} \right) \\ < \; & 2n \exp\left(\frac{-n}{4L^2} \right), \end{aligned}$$

where $\mathcal{M}^*_\Theta = \frac{1}{\kappa_2} \left\{ 1 + 2 \left(\sum_{s\neq\nu} v_{s\nu}^2 \right)^{1/2} \right\}$. Therefore, we conclude that

$$\begin{aligned} \left\| R_{n;\Theta} \right\|_\infty &\leq \left\| I_{\Theta;b,\zeta}\left(\beta_0, \bar{\Theta}\right) - I_{\Theta;b,\zeta}\left(\beta_0, \Theta_0\right) \right\|_\infty \left\| \mathrm{vec}(\widehat{\Theta}_b(\zeta)) - \mathrm{vec}(\Theta_0) \right\|_1 \\ &< \frac{\alpha}{2-\alpha} \frac{\lambda_{n2}}{4}, \end{aligned}$$

and thus (B.41) follows. □

B.3 Proof of Theorem 3.3.1

To show Theorem 3.3.1, our strategy is first to derive the result for given b and ζ. We then use the equations (3.15) to establish the results for $\widehat{\beta}(\zeta)$ and $\widehat{\Theta}(\zeta)$. Finally, under Condition (C9), we extrapolate the estimators and obtain the desired result. The proof consists of the following three steps.

Step 1: *We claim that*

$$\left\| \left(\widehat{\beta}_b(\zeta)^\top, vec\left(\widehat{\Theta}_b(\zeta) \right)^\top \right)^\top - \left(\beta_0^\top, vec\left(\Theta_0 \right)^\top \right)^\top \right\|_2 = O_p\left(\frac{1}{\sqrt{n}} \right). \tag{B.50}$$

Let

$$\psi_{b,\zeta}\left(\beta,\Theta\right)=\ell_{b,\zeta}\left(\beta,\Theta\right)+\lambda_{n1}\rho_1(\beta)+\lambda_{n2}\rho_2(\Theta). \tag{B.51}$$

To show (B.50), as described in Fan and Li (2001, 2002), it suffices to prove that for any $1>\epsilon>0$ and constant $\widetilde{\mathcal{B}}>0$,

$$P\left\{\inf_{\|\mathbf{U}\|=\widetilde{\mathcal{B}}}\psi_{b,\zeta}\left(\beta_0+\frac{u_n}{\sqrt{n}},\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)>\psi_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)\right)\right\}>1-\epsilon, \tag{B.52}$$

where $\mathbf{U}=\left(u_n^\top,v_n^\top\right)^\top$, $u_n=\sqrt{n}\left(\widehat{\beta}_b(\zeta)-\beta_0\right)$, and $v_n=\sqrt{n}\left(\mathrm{vec}\left(\widehat{\Theta}_b(\zeta)\right)-\mathrm{vec}\left(\Theta_0\right)\right)$.

We now write

$$\psi_{b,\zeta}\left(\beta_0+\frac{u_n}{\sqrt{n}},\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\psi_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)\right)=J_1+J_2+J_3, \tag{B.53}$$

where

$$\begin{aligned}
J_1 &= \ell_{b,\zeta}\left(\beta_0+\frac{u_n}{\sqrt{n}},\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\ell_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)\right); && \text{(B.54a)}\\
J_2 &= \lambda_{n1}\left\{\rho_1\left(\beta_0+\frac{u_n}{\sqrt{n}}\right)-\rho_1\left(\beta_0\right)\right\}; && \text{(B.54b)}\\
J_3 &= \lambda_{n2}\left\{\rho_2\left(\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\rho_2\left(\mathrm{vec}\left(\Theta_0\right)\right)\right\}. && \text{(B.54c)}
\end{aligned}$$

Step 1.1: *Show that*

$$J_1>0. \tag{B.55}$$

In (B.54a), adding and subtracting an additional term gives

$$\begin{aligned}
J_1 &= \ell_{b,\zeta}\left(\beta_0+\frac{u_n}{\sqrt{n}},\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\ell_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)\\
&\quad+\ell_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\ell_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)\right),
\end{aligned}$$

and using the second order Taylor series expansion yields

$$\begin{aligned} J_1 &= \frac{u_n^\top}{\sqrt{n}} U_{\beta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right) + \frac{1}{2!}\frac{u_n^\top}{\sqrt{n}} I_{\beta;b,\zeta}\left(\beta_0 + \frac{h_1 u_n}{\sqrt{n}}, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right)\frac{u_n}{\sqrt{n}} \\ &\quad + \frac{v_n^\top}{\sqrt{n}} U_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right) + \frac{1}{2!}\frac{v_n^\top}{\sqrt{n}} I_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{h_2 v_n}{\sqrt{n}}\right)\frac{v_n}{\sqrt{n}} \\ &= \left\{\frac{u_n^\top}{\sqrt{n}} U_{\beta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right) + \frac{v_n^\top}{\sqrt{n}} U_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right)\right\} \\ &\quad + \left\{\frac{1}{2!}\frac{u_n^\top}{\sqrt{n}} I_{\beta;b,\zeta}\left(\beta_0 + \frac{h_1 u_n}{\sqrt{n}}, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right)\frac{u_n}{\sqrt{n}}\right. \\ &\quad \left. + \frac{1}{2!}\frac{v_n^\top}{\sqrt{n}} I_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{h_2 v_n}{\sqrt{n}}\right)\frac{v_n}{\sqrt{n}}\right\} \\ &\triangleq J_{1,1} + J_{1,2}, \end{aligned} \tag{B.56}$$

where $h_1, h_2 \in (0,1)$, and the second step is obtained by combining the first order terms and the second order terms, respectively.

By (3.23), we have $E\left\{U_{\beta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right)\right\} = 0$ and $E\left\{U_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right)\right\} = 0$, which implies that

$$\frac{1}{n} U_{\beta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right) = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{B.57}$$

and

$$\frac{1}{n} U_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right) = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{B.58}$$

Multiplying $\sqrt{n}u_n^\top$ and $\sqrt{n}v_n^\top$ to (B.57) and (B.58), respectively, gives

$$\frac{u_n^\top}{\sqrt{n}} U_{\beta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right) = u_n^\top\, O_p(1); \tag{B.59a}$$

$$\frac{v_n^\top}{\sqrt{n}} U_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0)\right) = v_n^\top\, O_p(1). \tag{B.59b}$$

Note that $\operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}$ approaches $\operatorname{vec}(\Theta_0)$ in probability as $n \to \infty$. Therefore, the order of $J_{1,1}$ is $\widetilde{\mathcal{B}}$, if $\|\mathbf{U}\| = \widetilde{\mathcal{B}}$.

On the other hand, by the Law of Large Numbers, we have that as $n \to \infty$,

$$\begin{aligned}\frac{1}{n} I_{\beta;b,\zeta}\left(\beta_0 + \frac{h_1 u_n}{\sqrt{n}}, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right) &\xrightarrow{p} \mathcal{I}_{\beta;b,\zeta}(\beta_0, \Theta_0);\\ \frac{1}{n} I_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{h_2 v_n}{\sqrt{n}}\right) &\xrightarrow{p} \mathcal{I}_{\Theta;b,\zeta}(\beta_0, \Theta_0),\end{aligned}$$

where $I_{\Theta;b,\zeta}(\cdot)$ and $I_{\Theta;b,\zeta}(\cdot)$ are defined in (3.24),

$$\mathcal{I}_{\beta;b,\zeta}(\beta_0, \Theta_0) = E\left\{\frac{1}{n} I_{\beta;b,\zeta}(\beta_0, \operatorname{vec}(\Theta_0))\right\}$$

and

$$\mathcal{I}_{\Theta;b,\zeta}(\beta_0, \Theta_0) = E\left\{\frac{1}{n} I_{\Theta;b,\zeta}(\beta_0, \operatorname{vec}(\Theta_0))\right\}.$$

Therefore, we can equivalently write

$$\frac{1}{n} I_{\beta;b,\zeta}\left(\beta_0 + \frac{h_1 u_n}{\sqrt{n}}, \operatorname{vec}(\Theta_0) + \frac{v_n}{\sqrt{n}}\right) = \mathcal{I}_{\beta;b,\zeta}(\beta_0, \Theta_0) + o_p(1); \tag{B.60a}$$

$$\frac{1}{n} I_{\Theta;b,\zeta}\left(\beta_0, \operatorname{vec}(\Theta_0) + \frac{h_2 v_n}{\sqrt{n}}\right) = \mathcal{I}_{\Theta;b,\zeta}(\beta_0, \Theta_0) + o_p(1). \tag{B.60b}$$

As a result, we have

$$J_{1,2} = \frac{1}{2} u_n^\top \left\{\mathcal{I}_{\beta;b,\zeta}(\beta_0, \Theta_0) + o_p(1)\right\} u_n + \frac{1}{2} v_n^\top \left\{\mathcal{I}_{\Theta;b,\zeta}(\beta_0, \Theta_0) + o_p(1)\right\} v_n. \tag{B.61}$$

Note that the event $\mathcal{E}$ defined in Step 3 of the proof of Lemma B.2.1 restricts us to consider bounded $W_b^{(i)}(\zeta)$, thus, $\mathcal{G}_{b,\zeta}(u;\beta,\Theta)$, $\mathcal{G}^{(1)}_{\beta;b,\zeta}(u;\beta,\Theta)$ and $\mathcal{G}^{(2)}_{\beta;b,\zeta}(u;\beta,\Theta)$, defined in (B.1), (B.3) and (B.43), respectively, are bounded, showing that by (B.60a) and (B.42), $\mathcal{I}_{\beta;b,\zeta}(\beta_0, \Theta_0)$ is bounded elementwisely. Similarly, $\mathcal{I}_{\Theta;b,\zeta}(\beta_0, \Theta_0)$ is bounded elementwisely. Therefore, (B.61) shows that the order of $J_{1,2}$ is $\widetilde{\mathcal{B}}^2$ if $\|\mathbf{U}\| = \widetilde{\mathcal{B}}$.

In addition, by Condition (C7), both $\mathcal{I}_{\beta;b,\zeta}(\cdot,\cdot)$ and $\mathcal{I}_{\Theta;b,\zeta}(\cdot,\cdot)$ are positive definite matrices, so for any non-zero vectors u and v and by (B.60a) and (B.60b), we obtain that

$$J_{1,2} > 0. \tag{B.62}$$

As a result, by (B.59a), (B.59b) and (B.61) together with (B.56), we conclude that when $\widetilde{\mathcal{B}}$ is sufficiently large, $J_{1,2}$ dominates $J_{1,1}$, and (B.62) ensures that J_1 is bounded below by

a positive constant. Thus, $J_1 > 0$ when if $\|\mathbf{U}\| = \widetilde{\mathcal{B}}$ for a sufficiently large $\widetilde{\mathcal{B}}$.
Step 1.2: *Show that*

$$J_2 = o_p(1) \quad as \quad \frac{\lambda_{n1}}{\sqrt{n}} \to 0. \tag{B.63}$$

.

For $\rho_1(\cdot)$ in (B.54b), we have

$$\begin{aligned}
\rho_1\left(\beta_0 + \frac{u_n}{\sqrt{n}}\right) - \rho_1\left(\beta_0\right) &= \sum_{r\in\mathcal{S}_1} w_r \left|\beta_{0r} + \frac{u_{nr}}{\sqrt{n}}\right| - \sum_{r\in\mathcal{S}_1} w_r \left|\beta_{0r}\right| \\
&= \sum_{r\in\mathcal{S}_1} w_r \left(\left|\beta_{0r} + \frac{u_{nr}}{\sqrt{n}}\right| - \left|\beta_{0r}\right|\right) \\
&= \sum_{r\in\mathcal{S}_1} w_r \left\{\frac{\beta_{0r}}{|\beta_{0r}|}\frac{u_{nr}}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right)\right\} \\
&= \frac{1}{\sqrt{n}} \sum_{r\in\mathcal{S}_1} \left\{\text{sign}\left(\beta_{0r}\right) u_{nr} w_r + o_p(1)\right\},
\end{aligned}$$

where the third step is because of the Taylor series expansion of $\left|\beta_{0r} + \frac{u_{nr}}{\sqrt{n}}\right|$ around $u_{nr} = 0$. Then by the similar derivations of Lemma 3 in Lee and Liu (2012), as $\frac{\lambda_{n1}}{\sqrt{n}} \to 0$, we have

$$\lambda_{n1}\left\{\rho_1\left(\beta_0 + \frac{u_n}{\sqrt{n}}\right) - \rho_1\left(\beta_0\right)\right\} = o_p(1),$$

i.e., (B.63) follows.
Step 1.3: *Show that*

$$J_3 = o_p(1) \quad as \quad \frac{\lambda_{n2}}{\sqrt{n}} \to 0. \tag{B.64}$$

For $\rho_2(\cdot)$ in (B.54b), we have

$$\begin{aligned}
\rho_2\left(\text{vec}\left(\Theta_0\right) + \frac{v_n}{\sqrt{n}}\right) - \rho_2\left(\text{vec}\left(\Theta_0\right)\right) &= \sum_{(s,\nu)\in\mathcal{S}_2} \left|\theta_{0s\nu} + \frac{v_{n,s\nu}}{\sqrt{n}}\right| - \sum_{(s,\nu)\in\mathcal{S}_2} \left|\theta_{0s\nu}\right| \\
&= \frac{1}{\sqrt{n}} \sum_{(s,\nu)\in\mathcal{S}_2} \left\{\text{sign}\left(\theta_{0s\nu}\right) v_{n,s\nu} + o_p(1)\right\}.
\end{aligned}$$

Hence, as $\frac{\lambda_{n2}}{\sqrt{n}} \to 0$, we have

$$\lambda_{n2}\left\{\rho_2\left(\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\rho_2\left(\mathrm{vec}\left(\Theta_0\right)\right)\right\}=o_p(1),$$

i.e., (B.64) follows.

Therefore, together with (B.53), (B.55), (B.63) and (B.64), for sufficiently large $\widetilde{\mathcal{B}}$ and $\|\mathbf{U}\|_2=\widetilde{\mathcal{B}}$, we have

$$\psi_{b,\zeta}\left(\beta_0+\frac{u_n}{\sqrt{n}},\mathrm{vec}\left(\Theta_0\right)+\frac{v_n}{\sqrt{n}}\right)-\psi_{b,\zeta}\left(\beta_0,\mathrm{vec}\left(\Theta_0\right)\right)>0. \tag{B.65}$$

Therefore, as described by Fan and Li (2001, 2002), under $\|\mathbf{U}\|_2=\widetilde{\mathcal{B}}$, (B.65) implies that (B.52) holds. Consequently, the result (B.50) is shown.

Step 2: *Taking average on (B.50) with respect to b.*

By the formulation of the SIMEX algorithm, for any given $\zeta\in\mathcal{Z}$, we have

$$\widehat{\beta}(\zeta)=\frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_b(\zeta)\ \text{ and }\ \mathrm{vec}\left(\widehat{\Theta}(\zeta)\right)=\frac{1}{B}\sum_{b=1}^{B}\mathrm{vec}\left(\widehat{\Theta}_b(\zeta)\right).$$

Hence, (B.50) gives that

$$\begin{aligned}
&\left\|\left(\widehat{\beta}(\zeta)^\top,\mathrm{vec}\left(\widehat{\Theta}(\zeta)\right)^\top\right)^\top-\left(\beta_0^\top,\mathrm{vec}\left(\Theta_0\right)^\top\right)^\top\right\|_2^2\\
=&\left\|\frac{1}{B}\sum_{b=1}^{B}\left\{\left(\widehat{\beta}_b(\zeta)^\top,\mathrm{vec}\left(\widehat{\Theta}_b(\zeta)\right)^\top\right)^\top-\left(\beta_0^\top,\mathrm{vec}\left(\Theta_0\right)^\top\right)^\top\right\}\right\|_2^2\\
\le&\frac{1}{B}\sum_{b=1}^{B}\left\|\left(\widehat{\beta}_b(\zeta)^\top,\mathrm{vec}\left(\widehat{\Theta}_b(\zeta)\right)^\top\right)^\top-\left(\beta_0^\top,\mathrm{vec}\left(\Theta_0\right)^\top\right)^\top\right\|_2^2\\
=&O_p\left(\frac{1}{n}\right),
\end{aligned} \tag{B.66}$$

where the last step is due to (B.50).

Step 3: *Establish the final result.*

Let $\zeta \to -1$ on (B.66) gives

$$\left\| \left(\widehat{\beta}^\top, \mathrm{vec}(\widehat{\Theta})^\top \right) - \left(\beta_0^\top, \mathrm{vec}\left(\Theta_0\right)^\top \right) \right\|_2 = O_p\left(\frac{1}{\sqrt{n}} \right),$$

and thus the proof completes. □

B.4 Proof of Theorem 3.3.2

B.4.1 Proof of Part (a)

In this part, we first study the property of the estimator of β, and then examine the estimator of Θ. Finally, combining those two results yields the desired theorem.
Part 1: *Inference for* $\widehat{\mathcal{S}}_1$

The following derivations consist of three steps.
Step 1: *Examine* $\widehat{\mathcal{S}}_1(b,\zeta) = \left\{ j : \textit{the } j\textit{th entry of } \widehat{\beta}_b(\zeta) \textit{ is non-zero} \right\}$ *and show that*

$$\widehat{\mathcal{S}}_1(b,\zeta) = \mathcal{S}_1 \tag{B.67}$$

with a large probability.

In the proof of Lemma B.2.3, we define $\widehat{\beta}_b(\zeta) = \left(\widehat{\beta}^\top_{b;\mathcal{S}_1}(\zeta), \widehat{\beta}^\top_{b;\mathcal{S}_1^c}(\zeta) \right)^\top$ and $\beta_0 = \left(\beta^\top_{0;\mathcal{S}_1}, \beta^\top_{0;\mathcal{S}_1^c} \right)^\top$, and let $\widehat{\beta}_{b,r}(\zeta)$ denote the rth component of $\widehat{\beta}_b(\zeta)$. Let $\widehat{z}$ be the p-dimensional vector with the rth component being $\widehat{z}_r = \mathrm{sign}\left(\widehat{\beta}_{b,r}(\zeta) \right)$ if $\widehat{\beta}_{b,r}(\zeta) \neq 0$ and $|\widehat{z}_r| \leq 1$ otherwise. To show the sparsity recovery, we consider the primal dual witness (PDW) method (e.g., Hastie et al. 2015, p.307). The strategy of the PDW method is to

(i) set $\widehat{\beta}_{b;\mathcal{S}_1^c}(\zeta) = 0_{(p-d_\beta)}$ and $\widehat{\beta}_{b;\mathcal{S}_1}(\zeta) = \underset{\beta_{\mathcal{S}_1}}{\mathrm{argmin}} \left\{ \ell_{b,\zeta}\left(\beta_{\mathcal{S}_1}, \Theta_0\right) + \lambda_{n1}\rho_1\left(\beta_{\mathcal{S}_1}\right) \right\}$;

(ii) write $\widehat{z} = \left(\widehat{z}^\top_{\mathcal{S}_1}, \widehat{z}^\top_{\mathcal{S}_1^c} \right)^\top$ corresponding to the components of $\widehat{\beta}_{b;\mathcal{S}_1}(\zeta)$ and $\widehat{\beta}_{b;\mathcal{S}_1^c}(\zeta)$;

(iii) then show that

$$\left\| \widehat{z}_{\mathcal{S}_1^c} \right\|_\infty < 1. \tag{B.68}$$

Indeed, as discussed in Lemma 11.2 of Hastie et al. (2015, p.307), if (B.68) is true, then $\widehat{\beta}_b(\zeta) = \left(\widehat{\beta}_{b;\mathcal{S}_1}^\top(\zeta), 0_{(p-d_\beta)}^\top\right)^\top$ is an optimal solution of (3.14), and thus, (B.67) holds with a large probability (e.g., Hastie et al. 2015, Theorem 11.3). So the remaining task is to show (B.68).

By the KKT conditions and Theorem 3.3.1, we have

$$U_{\beta;b,\zeta}\left(\widehat{\beta}_b(\zeta), \Theta_0\right) + \lambda_{n1}\widehat{z} = 0. \tag{B.69}$$

Adding $-U_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right)$ to the both sides of (B.69) gives

$$U_{\beta;b,\zeta}\left(\widehat{\beta}_b(\zeta), \Theta_0\right) - U_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right) = -\lambda_{n1}\widehat{z} - U_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right). \tag{B.70}$$

Applying the Mean Valued Theorem to the left-hand side of (B.70), we obtain that

$$I_{\beta;b,\zeta}\left(\overline{\beta}, \Theta_0\right)\left(\widehat{\beta}_b(\zeta) - \beta_0\right) = -\lambda_{n1}\widehat{z} - U_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right), \tag{B.71}$$

where $\overline{\beta}$ is a vector which lies on the "line segment" between $\widehat{\beta}_b(\zeta)$ and β_0.

Adding $I_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right)\left(\widehat{\beta}_b(\zeta) - \beta_0\right)$ to the both sides of (B.71) yields

$$I_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right)\left(\widehat{\beta}_b(\zeta) - \beta_0\right) = -\lambda_{n1}\widehat{z} - U_{\beta;b,\zeta}\left(\beta_0, \Theta_0\right) - R_{n;\beta}, \tag{B.72}$$

where $R_{n;\beta}$ is defined in Lemma B.2.4.

We write $R_{n;\beta} = \left(R_{n,\mathcal{S}_1;\beta}^\top, R_{n,\mathcal{S}_1^c;\beta}^\top\right)^\top$ according to the components of $\mathcal{S}_1$ and $\mathcal{S}_1^c$. (i) of the PDW method indicates that $\widehat{\beta}_b(\zeta) = \left(\widehat{\beta}_{b;\mathcal{S}_1}^\top(\zeta), 0_{(p-d_\beta)}^\top\right)^\top$ and $\beta_0 = \left(\beta_{0;\mathcal{S}_1}^\top, 0_{(p-d_\beta)}^\top\right)^\top$, and thus, by the matrix algebra, (B.72) can be written as

$$I_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left(\beta_0, \Theta_0\right)\left(\widehat{\beta}_{b;\mathcal{S}_1}(\zeta) - \beta_{0;\mathcal{S}_1}\right) = -U_{\beta,\mathcal{S}_1;b,\zeta}\left(\beta_0, \Theta_0\right) - \lambda_{n1}\widehat{z}_{\mathcal{S}_1} - R_{n,\mathcal{S}_1;\beta} \tag{B.73a}$$

$$I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}\left(\beta_0, \Theta_0\right)\left(\widehat{\beta}_{b;\mathcal{S}_1}(\zeta) - \beta_{0;\mathcal{S}_1}\right) = -U_{\beta,\mathcal{S}_1^c;b,\zeta}\left(\beta_0, \Theta_0\right) - \lambda_{n1}\widehat{z}_{\mathcal{S}_1^c} - R_{n,\mathcal{S}_1^c;\beta} \tag{B.73b}$$

For ease of notation, let $I_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}$ and $I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}$ denote $I_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left(\beta_0, \Theta_0\right)$ and $I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}\left(\beta_0, \Theta_0\right)$, respectively. Combining (B.73a) and (B.73b) gives

$$\begin{aligned} & I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta} I_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}^{-1}\left\{U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0, \Theta_0) + \lambda_{n1}\widehat{z}_{\mathcal{S}_1} + R_{n,\mathcal{S}_1;\beta}\right\} \\ = \; & U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0, \Theta_0) + \lambda_{n1}\widehat{z}_{\mathcal{S}_1^c} + R_{n,\mathcal{S}_1^c;\beta}, \end{aligned} \tag{B.74}$$

yielding that

$$\begin{aligned}\widehat{z}_{\mathcal{S}_1^c} &= \frac{1}{\lambda_{n1}}\left[I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}I^{-1}_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left\{U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0,\Theta_0)+\lambda_{n1}\widehat{z}_{\mathcal{S}_1}+R_{n,\mathcal{S}_1;\beta}\right\}\right] \\ &\quad -\frac{1}{\lambda_{n1}}\left\{U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0,\Theta_0)+R_{n,\mathcal{S}_1^c;\beta}\right\}. \end{aligned} \tag{B.75}$$

Now we are ready to show (B.68). (B.75) gives that

$$\begin{aligned}\left\|\widehat{z}_{\mathcal{S}_1^c}\right\|_\infty &\leq \frac{1}{\lambda_{n1}}\left\|I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}I^{-1}_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\left(U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0,\Theta_0)+\lambda_{n1}\widehat{z}_{\mathcal{S}_1}+R_{n,\mathcal{S}_1;\beta}\right)\right\|_\infty \\ &\quad +\frac{1}{\lambda_{n1}}\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0,\Theta_0)+R_{n,\mathcal{S}_1^c;\beta}\right\|_\infty \\ &\leq \frac{1}{\lambda_{n1}}\left\|I_{\beta,\mathcal{S}_1^c\mathcal{S}_1;b,\zeta}I^{-1}_{\beta,\mathcal{S}_1\mathcal{S}_1;b,\zeta}\right\|_\infty\left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty+\lambda_{n1}\left\|\widehat{z}_{\mathcal{S}_1}\right\|_\infty+\left\|R_{n,\mathcal{S}_1;\beta}\right\|_\infty\right) \\ &\quad +\frac{1}{\lambda_{n1}}\left(\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty+\left\|R_{n,\mathcal{S}_1^c;\beta}\right\|_\infty\right) \\ &\leq \frac{1}{\lambda_{n1}}(1-\alpha)\left(\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty+\lambda_{n1}+\left\|R_{n,\mathcal{S}_1;\beta}\right\|_\infty\right) \\ &\quad +\frac{1}{\lambda_{n1}}\left(\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty+\left\|R_{n,\mathcal{S}_1^c;\beta}\right\|_\infty\right) \\ &\leq (1-\alpha)+\frac{1}{\lambda_{n1}}(1-\alpha)\left\|U_{\beta,\mathcal{S}_1;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty \\ &\quad +\frac{1}{\lambda_{n1}}\left\|U_{\beta,\mathcal{S}_1^c;b,\zeta}(\beta_0,\Theta_0)\right\|_\infty+\frac{1}{\lambda_{n1}}(2-\alpha)\left\|R_{n;\beta}\right\|_\infty \\ &< 1-\alpha+\frac{\alpha}{4}+\frac{\alpha}{4}+\frac{\alpha}{4} \\ &= 1-\frac{\alpha}{4}<1, \end{aligned}$$

where the third step is because Condition (C8) and $\left\|\widehat{z}_{\mathcal{S}_1}\right\|_\infty \leq 1$ by the construction of $\widehat{z}_{\mathcal{S}_1}$, the fourth step is due to that $\left\|R_{n,\mathcal{S}_1;\beta}\right\|_\infty \leq \left\|R_{n;\beta}\right\|_\infty$ and $\left\|R_{n,\mathcal{S}_1^c;\beta}\right\|_\infty \leq \left\|R_{n;\beta}\right\|_\infty$, the second last step comes from Lemmas B.2.1 and B.2.4. Hence, we have $\widehat{\mathcal{S}}_1(b,\zeta)=\mathcal{S}_1$ for every $b=1,\cdots,B$ and $\zeta\in\mathcal{Z}$ with a large probability.

Step 2:

Let $\widehat{\mathcal{S}}_1(\zeta)=\left\{j\in V:\widehat{\beta}_j(\zeta)\neq 0\right\}$. By (3.15) and (B.67), we have that $\widehat{\mathcal{S}}_1(\zeta)=\mathcal{S}_1(\zeta)$ with a large probability.

Step 3:

Finally, since $\widehat{\beta}(\zeta) \xrightarrow{p} \widehat{\beta}$ as $\zeta \to -1$, then $\widehat{\mathcal{S}}_1(\zeta) \xrightarrow{p} \widehat{\mathcal{S}}_1$. As a result, $\widehat{\mathcal{S}}_1 = \mathcal{S}_1$ with a large probability and $\zeta \to -1$.

Part 2: *Inference for* $\widehat{\mathcal{S}}_2$

The proof of this part follows similar to that for Part 1; the only differences here are to replace the quantities for β in Part 1 with the corresponding versions for Θ, as outlined in the following three steps.

Step 1: *Examine* $\widehat{\mathcal{S}}_2(b,\zeta) = \left\{(s,\nu) : \textit{entry } (s,\nu) \textit{ of } \widehat{\Theta}_b(\zeta) \textit{ is non-zero}\right\}$ *and show that*

$$\widehat{\mathcal{S}}_2(b,\zeta) = \mathcal{S}_2$$

with a large probability.

In the proof of Lemma B.2.3, we define

$$\text{vec}\left(\widehat{\Theta}_b(\zeta)\right) = \left(\text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)^\top, \text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2^c}(\zeta)\right)^\top\right)^\top,$$

and let $\widehat{\Theta}_{b,s\nu}(\zeta)$ denote the component (s,ν) of $\widehat{\Theta}_b(\zeta)$. Let $\widehat{\mu}$ be the (p^2-p)-dimensional vector with the $s\nu$th component being $\widehat{\mu}_{s\nu} = \text{sign}\left(\widehat{\Theta}_{b,s\nu}(\zeta)\right)$ if $\widehat{\Theta}_{b,s\nu}(\zeta) \neq 0$ and $|\widehat{\mu}_{s\nu}| \leq 1$ otherwise. To show the sparsity recovery, we consider the primal dual witness (PDW) method (e.g., Hastie et al. 2015, p.307). The strategy of the PDW method is to

(i) set $\text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2^c}(\zeta)\right) = 0_{(p^2-p-d_\Theta)}$ and

$$\text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right) = \underset{\Theta_{\mathcal{S}_2}}{\text{argmin}}\left\{\ell_{b,\zeta}\left(\beta_0, \Theta_{\mathcal{S}_2}\right) + \lambda_{n2}\rho_2\left(\Theta_{\mathcal{S}_2}\right)\right\};$$

(ii) write $\widehat{\mu} = \left(\widehat{\mu}_{\mathcal{S}_2}^\top, \widehat{\mu}_{\mathcal{S}_2^c}^\top\right)^\top$ corresponding to the components of $\text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2}(\zeta)\right)$ and $\text{vec}\left(\widehat{\Theta}_{b;\mathcal{S}_2^c}(\zeta)\right)$;

(iii) then show that

$$\left\|\widehat{\mu}_{\mathcal{S}_2^c}\right\|_\infty < 1.$$

By the derivation similar to (B.74), we have

$$\begin{aligned} & I_{\Theta,\mathcal{S}_2^c\mathcal{S}_2;b,\zeta} I^{-1}_{\Theta,\mathcal{S}_2\mathcal{S}_2;b,\zeta} \left\{U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta_0,\Theta_0) + \lambda_{n2}\widehat{\mu}_{\mathcal{S}_2} + R_{n,\mathcal{S}_2;\Theta}\right\} \\ = \; & U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta_0,\Theta_0) + \lambda_{n2}\widehat{\mu}_{\mathcal{S}_2^c} + R_{n,\mathcal{S}_2^c;\Theta}, \end{aligned}$$

yielding that

$$\begin{aligned} \widehat{\mu}_{\mathcal{S}_2^c} \; = \; & \frac{1}{\lambda_{n2}}\left[I_{\Theta,\mathcal{S}_2^c\mathcal{S}_2;b,\zeta} I^{-1}_{\Theta,\mathcal{S}_2\mathcal{S}_2;b,\zeta}\left\{U_{\Theta,\mathcal{S}_2;b,\zeta}(\beta_0,\Theta_0) + \lambda_{n2}\widehat{\mu}_{\mathcal{S}_2} + R_{n,\mathcal{S}_2;\Theta}\right\}\right] \\ & -\frac{1}{\lambda_{n2}}\left\{U_{\Theta,\mathcal{S}_2^c;b,\zeta}(\beta_0,\Theta_0) + R_{n,\mathcal{S}_2^c;\Theta}\right\} \end{aligned}$$

and $\left\|\widehat{\mu}_{\mathcal{S}_2^c}\right\|_\infty < 1$. As a result, we have $\widehat{\mathcal{S}}_2(b,\zeta) = \mathcal{S}_2$ for every $b = 1,\cdots,B$ and $\zeta \in \mathcal{Z}$ with a large probability.

Step 2:

Let $\widehat{\mathcal{S}}_2(\zeta) = \left\{(s,\nu) \in E : \widehat{\Theta}_{s\nu}(\zeta) \neq 0\right\}$, then we have $\widehat{\mathcal{S}}_2(\zeta) = \mathcal{S}_2(\zeta)$ with a large probability by the relationship (3.15) and the result in Step 2.

Step 3:

Finally, since $\widehat{\Theta}(\zeta) \xrightarrow{p} \widehat{\Theta}$ as $\zeta \to -1$, then $\widehat{\mathcal{S}}_2(\zeta) \xrightarrow{p} \widehat{\mathcal{S}}_2$. As a result, $\widehat{\mathcal{S}}_2 = \mathcal{S}_2$ with a large probability and $\zeta \to -1$.

Part 3: *Inference for* $\widehat{\mathcal{N}}$

Since $\widehat{\mathcal{S}}_1 = \mathcal{S}_1$ and $\widehat{\mathcal{S}}_2 = \mathcal{S}_2$ with a large probability. Then by the definition of $\widehat{\mathcal{N}}$ and $\mathcal{N}$, we conclude that $\widehat{\mathcal{N}} = \mathcal{N}$ with a large probability.

B.4.2 Proof of Part (b)

To show the sign recovery, as described in Ravikumar et al. (2010, p.1301), it suffices to prove that $\left\|\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}\right\|_\infty$ and $\left\|\mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0,\mathcal{S}_2})\right\|_\infty$ are bounded. Noting that

$$\left\|\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}\right\|_\infty \; \leq \; \left\|\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}\right\|_2$$

and

$$\left\|\mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0,\mathcal{S}_2})\right\|_\infty \; \leq \; \left\|\mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0,\mathcal{S}_2})\right\|_2,$$

then Lemma B.2.3 shows that $\left\|\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}\right\|_\infty$ and $\left\|\mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0,\mathcal{S}_2})\right\|_\infty$ are bounded.□

B.5 Proof of Theorem 3.3.3

Define

$$\Phi_{b,\zeta}\left(\widetilde{u},\widetilde{v}\right)=\psi_{b,\zeta}\left(\beta_0+\frac{\widetilde{u}}{\sqrt{n}},\mathrm{vec}(\Theta_0)+\frac{\widetilde{v}}{\sqrt{n}}\right)-\psi_{b,\zeta}\left(\beta_0,\mathrm{vec}(\Theta_0)\right), \tag{B.76}$$

where $\psi_{b,\zeta}\left(\cdot,\cdot\right)$ is defined in (B.51), and $\beta_0+\frac{\widetilde{u}}{\sqrt{n}}$ and $\mathrm{vec}(\Theta_0)+\frac{\widetilde{v}}{\sqrt{n}}$, respectively, express parameter values of β and $\mathrm{vec}(\Theta)$ that are of interest.

As derived in Appendix B.3, $u_n=\sqrt{n}\left(\widehat{\beta}_b(\zeta)-\beta_0\right)$ and $v_n=\sqrt{n}\left(\mathrm{vec}\left(\widehat{\Theta}_b(\zeta)\right)-\mathrm{vec}(\Theta_0)\right)$ satisfy

$$(u_n,v_n)^\top=\underset{\widetilde{u},\widetilde{v}}{\mathrm{argmin}}\,\Phi_{b,\zeta}(\widetilde{u},\widetilde{v}). \tag{B.77}$$

Recall that $\mathcal{S}_1=\{j\in V:\beta_j\neq 0\}$, $\mathcal{S}_2=\{(s,\nu)\in E:\theta_{s\nu}\neq 0\}$. By Theorem 3.3.2 (a), we have $\widehat{\beta}_{\mathcal{S}_1^c}=0$ and $\widehat{\Theta}_{\mathcal{S}_2^c}=0$. Therefore, we can express $u_n=\left(\widehat{c}_b^\top(\zeta),0_{(p-d_\beta)}^\top\right)^\top$ and $v_n=\left(\widehat{d}_b^\top(\zeta),0_{(p^2-p-d_\Theta)}^\top\right)^\top$, where

$$\widehat{c}_b(\zeta)=\sqrt{n}\left(\widehat{\beta}_{b,\mathcal{S}_1}(\zeta)-\beta_{0;\mathcal{S}_1}\right)\quad\text{and}\quad\widehat{d}_b(\zeta)=\sqrt{n}\left\{\mathrm{vec}(\widehat{\Theta}_{b,\mathcal{S}_2}(\zeta))-\mathrm{vec}(\Theta_{0;\mathcal{S}_2})\right\}. \tag{B.78}$$

Furthermore, let parameter values $\beta_{\mathcal{S}_1}$ and $\mathrm{vec}(\Theta_{\mathcal{S}_2})$ be expressed as $\beta_{0;\mathcal{S}_1}+\frac{c}{\sqrt{n}}$ and $\mathrm{vec}(\Theta_{0;\mathcal{S}_2})+\frac{d}{\sqrt{n}}$, respectively. Then (B.77) is re-written as

$$\left(\widehat{c}_b(\zeta),\widehat{d}_b(\zeta)\right)=\underset{c,d}{\mathrm{argmin}}\,\Phi_{b,\zeta}\left(c,d\right). \tag{B.79}$$

To show Theorem 3.3.3, we proceed with the following four steps.
Step 1: *Show that*

$$\begin{aligned}\Phi_{b,\zeta}(c,d) &= \frac{c^\top}{\sqrt{n}}\sum_{i=1}^n\omega^{(i)}_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)+\frac{d^\top}{\sqrt{n}}\sum_{i=1}^n\omega^{(i)}_{\Theta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\\ &\quad+\frac{1}{2}c^\top\mathcal{I}_{\beta,\mathcal{S}_1;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)c+\frac{1}{2}d^\top\mathcal{I}_{\Theta,\mathcal{S}_2;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)d+o_p(1),\end{aligned} \tag{B.80}$$

where similar to the structures of (B.14) and (B.22), we define

$$\omega^{(i)}_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)=\int_0^\tau-\left\{W^{(i)}_b(\zeta)-\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right\}dN_i(t)$$

$$-\int_0^\tau\left[\frac{Y_i(t)\exp\left(\sum\limits_{r\in\mathcal{S}_1}W^{(i)}_{b,r}(\zeta)\beta_{0r}+\sum\limits_{(s,\nu)\in\mathcal{S}_2}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{0s\nu}\right)}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right.$$

$$\left.\times\left\{W^{(i)}_b(\zeta)-\frac{\mathcal{G}^{(1)}_{\beta;b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right\}\right]d\mathbf{N}(t)$$

and

$$\omega^{(i)}_{\Theta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)=\int_0^\tau\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}-\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right\}dN_i(t)$$

$$-\int_0^\tau\left[\frac{Y_i(t)\exp\left(\sum\limits_{r\in\mathcal{S}_1}W^{(i)}_{b,r}(\zeta)\beta_{0r}+\sum\limits_{(s,\nu)\in\mathcal{S}_2}W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\theta_{0s\nu}\right)}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right.$$

$$\left.\times\left\{\left(W^{(i)}_{b,s}(\zeta)W^{(i)}_{b,\nu}(\zeta)\right)_{s\neq\nu}-\frac{\mathcal{G}^{(1)}_{\Theta;b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}{\mathcal{G}_{b,\zeta}(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})}\right\}\right]d\mathbf{N}(t).$$

First, we write

$$\Phi_{b,\zeta}\left(c,d\right)=V_1+V_2+V_3,\tag{B.81}$$

where

$$V_1=\ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1}+\frac{c}{\sqrt{n}},\Theta_{0;\mathcal{S}_2}+\frac{d}{\sqrt{n}}\right)-\ell_{b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right);\tag{B.82a}$$

$$V_2=\lambda_{n1}\left\{\rho_1\left(\beta_{0;\mathcal{S}_1}+\frac{c}{\sqrt{n}}\right)-\rho_1\left(\beta_{0;\mathcal{S}_1}\right)\right\};\tag{B.82b}$$

$$V_3=\lambda_{n2}\left\{\rho_2\left(\Theta_{0;\mathcal{S}_2}+\frac{d}{\sqrt{n}}\right)-\rho_2\left(\Theta_{0;\mathcal{S}_2}\right)\right\}.\tag{B.82c}$$

We now examine V_1, V_2 and V_3 in the following three steps.
Step 1.1: *Show that*

$$\begin{aligned} V_1 &= \frac{c^\top}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\beta;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + \frac{d^\top}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\Theta;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ &\quad + \frac{1}{2} c^\top \mathcal{I}_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) c + \frac{1}{2} d^\top \mathcal{I}_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) d + o_p(1). \end{aligned} \tag{B.83}$$

By the second order Taylor series expansion, (B.82a) becomes

$$\begin{aligned} V_1 &= \frac{c^\top}{\sqrt{n}} U_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + \frac{1}{2!} \frac{c^\top}{\sqrt{n}} I_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \frac{c}{\sqrt{n}} \\ &\quad + \frac{d^\top}{\sqrt{n}} U_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + \frac{1}{2!} \frac{d^\top}{\sqrt{n}} I_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \frac{d}{\sqrt{n}}. \end{aligned} \tag{B.84}$$

By the similar derivations of (B.60a) and (B.60b), we have

$$\frac{1}{n} I_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathcal{I}_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1) \tag{B.85}$$

and

$$\frac{1}{n} I_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathcal{I}_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1). \tag{B.86}$$

On the other hand, by the derivations similar to (B.14) and (B.22), we have

$$\frac{1}{\sqrt{n}} U_{\beta,\mathcal{S}_1;b,\zeta} \left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\beta;b,\zeta} \left(\beta_{0;\mathcal{S}_2}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1) \tag{B.87}$$

and

$$\frac{1}{\sqrt{n}} U_{\Theta,\mathcal{S}_2;b,\zeta} \left(\beta_{0;\mathcal{S}_2}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\Theta;b,\zeta} \left(\beta_{0;\mathcal{S}_2}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1), \tag{B.88}$$

Therefore, combining (B.85), (B.86), (B.87) and (B.88) with (B.84) gives (B.83).

Step 1.2: *Show that*

$$V_2 = o_p(1). \tag{B.89}$$

Indeed,

$$\begin{aligned} V_2 &= \lambda_{n1}\left\{\rho_1\left(\beta_{0;\mathcal{S}_1}+\frac{c}{\sqrt{n}}\right)-\rho_1\left(\beta_{0;\mathcal{S}_1}\right)\right\} \\ &= \lambda_{n1}\sum_{r\in\mathcal{S}_1}\left(w_r\left|\beta_{r0}+\frac{c_r}{\sqrt{n}}\right|-w_r\left|\beta_{r0}\right|\right) \\ &= \lambda_{n1}\sum_{r\in\mathcal{S}_1}\left(w_r\text{sign}\left(\beta_{r0}\right)\frac{c_r}{\sqrt{n}}+o(1)\right) \\ &= \frac{\lambda_{n1}}{\sqrt{n}}\sum_{r\in\mathcal{S}_1}\left(\text{sign}\left(\beta_{r0}\right)w_r c_r+o(1)\right). \end{aligned}$$

By the derivations similar to (B.63) in Appendix B.3 yields (B.89) as $\frac{\lambda_{n1}}{\sqrt{n}}\to 0$.

Step 1.3: *Show that*

$$V_3 = o_p(1). \tag{B.90}$$

Finally, by the derivations similar to (B.82b), we have that as $\frac{\lambda_{n2}}{\sqrt{n}}\to 0$,

$$\lambda_{n2}\left\{\rho_2\left(\Theta_{0;\mathcal{S}_2}+\frac{d}{\sqrt{n}}\right)-\rho_2\left(\Theta_{0;\mathcal{S}_2}\right)\right\}=o_p(1),$$

and thus (B.90) holds.

Therefore, combining (B.83), (B.89) and (B.90) with (B.81) gives (B.80).

Step 2: *Let* $\Phi_\zeta(c,d)=\frac{1}{B}\sum_{b=1}^{B}\Phi_{b,\zeta}(c,d)$*, where* $\Phi_{b,\zeta}(c,d)$ *is given by (B.80). Now we examine the minimum of* $\Phi_\zeta(c,d)$*.*

Indeed, $\Phi_\zeta(c,d)$ can be expressed as

$$\begin{aligned} \Phi_\zeta(c,d) &= \frac{c^\top}{\sqrt{n}}\sum_{i=1}^{n}\omega^{(i)}_{\beta;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)+\frac{d^\top}{\sqrt{n}}\sum_{i=1}^{n}\omega^{(i)}_{\Theta;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ &\quad+\frac{1}{2}c^\top\mathcal{I}_{\beta,\mathcal{S}_1;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)c+\frac{1}{2}d^\top\mathcal{I}_{\Theta,\mathcal{S}_2;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)d+o_p(1), \end{aligned}$$

where

$$\begin{aligned}
\omega^{(i)}_{\beta;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) &= \frac{1}{B}\sum_{b=1}^{B}\omega^{(i)}_{\beta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right),\\
\omega^{(i)}_{\Theta;\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) &= \frac{1}{B}\sum_{b=1}^{B}\omega^{(i)}_{\Theta;b,\zeta}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right),\\
\mathcal{I}_{\beta,\mathcal{S}_1;\zeta}(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}) &= \frac{1}{B}\sum_{b=1}^{B}\mathcal{I}_{\beta,\mathcal{S}_1;b,\zeta}(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}),\quad\text{and}\\
\mathcal{I}_{\Theta,\mathcal{S}_2;\zeta}(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}) &= \frac{1}{B}\sum_{b=1}^{B}\mathcal{I}_{\Theta,\mathcal{S}_2;b,\zeta}(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}).
\end{aligned}$$

Recall that $\widehat{c}_b(\zeta)=\sqrt{n}\left(\widehat{\beta}_{b,\mathcal{S}_1}(\zeta)-\beta_{0;\mathcal{S}_1}\right)$ and $\widehat{d}_b(\zeta)=\sqrt{n}\left\{\mathrm{vec}(\widehat{\Theta}_{b,\mathcal{S}_2}(\zeta))-\mathrm{vec}(\Theta_{0;\mathcal{S}_2})\right\}$, which are defined in (B.78). Similar to the definition in (3.15), we define

$$\left(\widehat{c}(\zeta),\widehat{d}(\zeta)\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{c}_b(\zeta),\widehat{d}_b(\zeta)\right).$$

Therefore, according to (B.79), we have

$$\begin{aligned}
\left(\widehat{c}(\zeta),\widehat{d}(\zeta)\right) &= \underset{c,d}{\mathrm{argmin}}\left\{\frac{1}{B}\sum_{b=1}^{B}\Phi_{b,\zeta}\left(c,d\right)\right\}\\
&= \underset{c,d}{\mathrm{argmin}}\,\Phi_{\zeta}\left(c,d\right).
\end{aligned}$$

Step 3: *Show that as* $n\to\infty$,

$$\sqrt{n}\left(\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z})-\beta_{0;\mathcal{S}_1}(\mathcal{Z}),vec\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right)-vec\left(\Theta_{0;\mathcal{S}_2}(\mathcal{Z})\right)\right)\xrightarrow{d}\underset{c,d}{\mathrm{argmin}}\,\Phi_0(c,d),$$

where

$$\begin{aligned}
\Phi_0(c,d) &= \left(c^\top,d^\top\right)\begin{pmatrix}\mathcal{U}\\ \mathcal{V}\end{pmatrix}\\
&\quad+\frac{1}{2}\left(c^\top,d^\top\right)\begin{pmatrix}\mathcal{I}_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) & 0\\ 0 & \mathcal{I}_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}\begin{pmatrix}c\\ d\end{pmatrix}.
\end{aligned}\tag{B.91}$$

Let $\widehat{c}(\mathcal{Z}) = \sqrt{n}\left(\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z}) - \beta_{0;\mathcal{S}_1}(\mathcal{Z})\right)$ and $\widehat{d}(\mathcal{Z}) = \sqrt{n}\left(\mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right) - \mathrm{vec}\left(\Theta_{0;\mathcal{S}_2}(\mathcal{Z})\right)\right)$. Write

$$\left(\widehat{c}(\mathcal{Z}), \widehat{d}(\mathcal{Z})\right) = \mathrm{vec}\left\{\left(\widehat{c}(\zeta), \widehat{d}(\zeta)\right) : \zeta \in \mathcal{Z}\right\}. \tag{B.92}$$

We define

$$\omega^{(i)}_{\beta,\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathrm{vec}\left\{\omega^{(i)}_{\beta;\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) : \zeta \in \mathcal{Z}\right\},$$

$$\omega^{(i)}_{\Theta,\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathrm{vec}\left\{\omega^{(i)}_{\Theta;\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) : \zeta \in \mathcal{Z}\right\},$$

$$\mathcal{I}_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathrm{diag}\left\{\mathcal{I}_{\beta,\mathcal{S}_1;\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) : \zeta \in \mathcal{Z}\right\}, \tag{B.93}$$

and

$$\mathcal{I}_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \mathrm{diag}\left\{\mathcal{I}_{\Theta,\mathcal{S}_2;\zeta}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) : \zeta \in \mathcal{Z}\right\}. \tag{B.94}$$

Write

$$\frac{1}{\sqrt{n}} U_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\beta,\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1)$$

and

$$\frac{1}{\sqrt{n}} U_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega^{(i)}_{\Theta,\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + o_p(1).$$

Define

$$\begin{aligned} \Phi_{\mathcal{Z}}(c,d) &= \frac{c}{\sqrt{n}} U_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) + \frac{d}{\sqrt{n}} U_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ &\quad + \frac{1}{2} c^\top \mathcal{I}_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) c + \frac{1}{2} d^\top \mathcal{I}_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) d + o_p(1). \end{aligned} \tag{B.95}$$

By the derivations similar to Section 3 of Carroll et al. (1996), we have that as $n \to \infty$,

$$\frac{1}{\sqrt{n}} U_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \xrightarrow{d} \mathcal{U} \tag{B.96a}$$

$$\frac{1}{\sqrt{n}} U_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \xrightarrow{d} \mathcal{V}, \tag{B.96b}$$

where $\mathcal{U}$ is a random variable having the distribution $N(0, \Sigma_\beta)$, $\mathcal{V}$ is a random variable having the distribution $N(0, \Sigma_\Theta)$,

$$\Sigma_\beta = \operatorname{cov}\left\{\omega^{(i)}_{\beta;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)\right\}, \quad \text{and} \quad \Sigma_\Theta = \operatorname{cov}\left\{\omega^{(i)}_{\Theta;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)\right\}.$$

Hence, combining (B.93), (B.94), (B.96a), (B.96b), and (B.95) yields that as $n \to \infty$,

$$\begin{aligned}\Phi_{\mathcal{Z}}(c,d) &\xrightarrow{d} c^\top \mathcal{U} + d^\top \mathcal{V} + \frac{1}{2} c^\top \mathcal{I}_{\beta,\mathcal{S}_1;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) c + \frac{1}{2} d^\top \mathcal{I}_{\Theta,\mathcal{S}_2;\mathcal{Z}}\left(\beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) d \\ &= \Phi_0(c,d),\end{aligned}$$

where $\Phi_0(\cdot,\cdot)$ is given in (B.91).

On the other hand, by the argmin continuous mapping theorem (Kim and Pollard 1990; Huang et al. 2014), we have that as $n \to \infty$,

$$\operatorname*{argmin}_{c,d} \Phi_{\mathcal{Z}}(c,d) \xrightarrow{d} \operatorname*{argmin}_{c,d} \Phi_0(c,d), \tag{B.97}$$

and together with (B.97) and the fact that $\left(\widehat{c}_n(\mathcal{Z}), \widehat{d}_n(\mathcal{Z})\right) = \operatorname*{argmin}_{c,d} \Phi_{\mathcal{Z}}(c,d)$, we have that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z}) - \beta_{0,\mathcal{S}_1}(\mathcal{Z}), \operatorname{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right) - \operatorname{vec}\left(\Theta_{\mathcal{S}_2,0}(\mathcal{Z})\right)\right) \xrightarrow{d} \operatorname*{argmin}_{c,d} \Phi_0(c,d). \tag{B.98}$$

Step 4: *Establish the result in Theorem 3.3.3.*

Finally, we need to extrapolate the estimators from (B.98). Let $\mathcal{R}_\beta(\Gamma_1) = \widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z}) - \varphi_1(\mathcal{Z};\Gamma_1)$ and $\mathcal{R}_\Theta(\Gamma_2) = \operatorname{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right) - \varphi_2(\mathcal{Z};\Gamma_2)$. Applying the least squares method to (3.16), $\widehat{\Gamma}_1$ and $\widehat{\Gamma}_2$ can be obtained by solving

$$\varphi^\top_{\Gamma_1,1} \mathcal{R}_\beta(\Gamma_1) = 0 \tag{B.99}$$

and

$$\varphi^\top_{\Gamma_2,2} \mathcal{R}_\Theta(\Gamma_2) = 0, \tag{B.100}$$

respectively, where $\varphi_{\Gamma_j,j} = \frac{\partial}{\partial \Gamma_j} \varphi_j(\mathcal{Z};\Gamma_j)$ with $j = 1, 2$.

By (B.99), we have

$$\varphi^\top_{\Gamma_1,1}\left\{\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z}) - \beta_{\mathcal{S}_1}(\mathcal{Z})\right\} = \varphi^\top_{\Gamma_1,1}\left\{\varphi_1\left(\mathcal{Z};\widehat{\Gamma}_1\right) - \varphi_1(\mathcal{Z};\Gamma_1)\right\}. \tag{B.101}$$

Multiplying $\sqrt{n}$ on both sides of (B.101) gives

$$\varphi_{\Gamma_1,1}^{\top}\sqrt{n}\left\{\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z})-\beta_{\mathcal{S}_1}(\mathcal{Z})\right\}=\varphi_{\Gamma_1,1}^{\top}\sqrt{n}\left\{\varphi_1\left(\mathcal{Z};\widehat{\Gamma}_1\right)-\varphi_1\left(\mathcal{Z};\Gamma_1\right)\right\}. \tag{B.102}$$

For the right-hand-side of (B.102), applying the Mean Value Theorem, there exists Γ^* which lies on the "line segment" between $\widehat{\Gamma}_1$ and Γ_1, such that

$$\begin{aligned}
& \sqrt{n}\left\{\varphi_1\left(\mathcal{Z};\widehat{\Gamma}_1\right)-\varphi_1\left(\mathcal{Z};\Gamma_1\right)\right\} \\
= & \sqrt{n}\left\{\frac{\partial}{\partial\Gamma_1}\varphi_1(\mathcal{Z},\Gamma^*)\left(\widehat{\Gamma}_1-\Gamma_1\right)\right\} \\
= & \sqrt{n}\left\{\frac{\partial}{\partial\Gamma_1}\varphi_1(\mathcal{Z},\Gamma_1)\left(\widehat{\Gamma}_1-\Gamma_1\right)+o_p(1)\left(\widehat{\Gamma}_1-\Gamma_1\right)\right\} \\
= & \sqrt{n}\left\{\frac{\partial}{\partial\Gamma_1}\varphi_1(\mathcal{Z},\Gamma_1)\left(\widehat{\Gamma}_1-\Gamma_1\right)+o_p(n^{-1/2})\right\} \\
= & \varphi_{\Gamma_1,1}\sqrt{n}\left(\widehat{\Gamma}_1-\Gamma_1\right)+o_p(1),
\end{aligned} \tag{B.103}$$

where the third and fourth steps are due to that $\widehat{\Gamma}_1$ is a consistent estimator. Therefore, combining (B.102) and (B.103) gives

$$\varphi_{\Gamma_1,1}^{\top}\sqrt{n}\left\{\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z})-\beta_{\mathcal{S}_1}(\mathcal{Z})\right\}=\varphi_{\Gamma_1,1}^{\top}\varphi_{\Gamma_1,1}\sqrt{n}\left(\widehat{\Gamma}_1-\Gamma_1\right)+o_p(1). \tag{B.104}$$

Thus, we can derive

$$\sqrt{n}\left(\widehat{\Gamma}_1-\Gamma_1\right)=\left(\varphi_{\Gamma_1,1}^{\top}\varphi_{\Gamma_1,1}\right)^{-1}\varphi_{\Gamma_1,1}^{\top}\sqrt{n}\left\{\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z})-\beta_{\mathcal{S}_1}(\mathcal{Z})\right\}+o_p(1). \tag{B.105}$$

On the other hand, by (B.100), we first have

$$\varphi_{\Gamma_2,2}^{\top}\left\{\mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right)-\mathrm{vec}\left(\Theta_{\mathcal{S}_2}(\mathcal{Z})\right)\right\}=\varphi_{\Gamma_2,2}^{\top}\left\{\varphi_2\left(\mathcal{Z};\widehat{\Gamma}_2\right)-\varphi_2\left(\mathcal{Z};\Gamma_2\right)\right\}, \tag{B.106}$$

then similar to the derivations for (B.104), we obtain that

$$\varphi_{\Gamma_2,2}^{\top}\sqrt{n}\left\{\mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right)-\mathrm{vec}\left(\Theta_{\mathcal{S}_2}(\mathcal{Z})\right)\right\}=\left(\varphi_{\Gamma_2,2}^{\top}\varphi_{\Gamma_2,2}\right)\sqrt{n}\left(\widehat{\Gamma}_2-\Gamma_2\right)+o_p(1),$$

which gives

$$\begin{aligned}
& \sqrt{n}\left(\widehat{\Gamma}_2-\Gamma_2\right) \\
= & \left(\varphi_{\Gamma_2,2}^{\top}\varphi_{\Gamma_2,2}\right)^{-1}\varphi_{\Gamma_2,2}^{\top}\sqrt{n}\left\{\mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right)-\mathrm{vec}\left(\Theta_{\mathcal{S}_2}(\mathcal{Z})\right)\right\}+o_p(1)
\end{aligned} \tag{B.107}$$

by derivations similar to those for (B.105).

Recall that

$$\widehat{c}(\mathcal{Z}) = \sqrt{n}\left\{\widehat{\beta}_{\mathcal{S}_1}(\mathcal{Z}) - \beta_{\mathcal{S}_1}(\mathcal{Z})\right\}$$

and

$$\widehat{d}(\mathcal{Z}) = \sqrt{n}\left\{\mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}(\mathcal{Z})\right) - \mathrm{vec}\left(\Theta_{\mathcal{S}_2}(\mathcal{Z})\right)\right\}$$

given by (B.92), and let $\Gamma = \left(\Gamma_1^\top, \Gamma_2^\top\right)^\top$. By (B.105) and (B.107), we further have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) &= \begin{pmatrix} \left(\varphi_{\Gamma_1,1}^\top \varphi_{\Gamma_1,1}\right)^{-1} \varphi_{\Gamma_1,1}^\top \widehat{c}(\mathcal{Z}) \\ \left(\varphi_{\Gamma_2,2}^\top \varphi_{\Gamma_2,2}\right)^{-1} \varphi_{\Gamma_2,2}^\top \widehat{d}(\mathcal{Z}) \end{pmatrix} \\
&= \begin{pmatrix} \left(\varphi_{\Gamma_1,1}^\top \varphi_{\Gamma_1,1}\right)^{-1} & 0 \\ 0 & \left(\varphi_{\Gamma_2,2}^\top \varphi_{\Gamma_2,2}\right)^{-1} \end{pmatrix} \begin{pmatrix} \varphi_{\Gamma_1,1}^\top & 0 \\ 0 & \varphi_{\Gamma_2,2}^\top \end{pmatrix} \begin{pmatrix} \widehat{c}(\mathcal{Z}) \\ \widehat{d}(\mathcal{Z}) \end{pmatrix} \\
&\triangleq \left(\varphi_\Gamma'^\top \varphi_\Gamma'\right)^{-1} \varphi_\Gamma'^\top \begin{pmatrix} \widehat{c}(\mathcal{Z}) \\ \widehat{d}(\mathcal{Z}) \end{pmatrix}, \qquad \text{(B.108)}
\end{aligned}$$

where $\varphi_\Gamma' = \begin{pmatrix} \varphi_{\Gamma_1,1} & 0 \\ 0 & \varphi_{\Gamma_2,2} \end{pmatrix}$.

Therefore, combining (B.98) and (B.108) yields that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) \xrightarrow{d} \operatorname*{argmin}_{c,d} \left(\varphi_\Gamma'^\top \varphi_\Gamma'\right)^{-1} \varphi_\Gamma'^\top \Phi_0(c,d). \qquad \text{(B.109)}$$

Write $\begin{pmatrix} \widehat{\beta}_{\mathcal{S}_1} \\ \mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}\right) \end{pmatrix} = \begin{pmatrix} \varphi_1(-1;\widehat{\Gamma}_1) \\ \varphi_2(-1;\widehat{\Gamma}_2) \end{pmatrix} \triangleq \varphi(-1;\widehat{\Gamma})$. Then

$$\begin{aligned}
\begin{pmatrix} \widehat{c} \\ \widehat{d} \end{pmatrix} &\triangleq \sqrt{n} \begin{pmatrix} \widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1} \\ \mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}\right) - \mathrm{vec}\left(\Theta_{0;\mathcal{S}_2}\right) \end{pmatrix} \\
&= \sqrt{n}\left\{\varphi(-1;\widehat{\Gamma}) - \varphi(-1;\Gamma)\right\} \\
&= \varphi'(-1;\Gamma)\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) + o_p(1), \qquad \text{(B.110)}
\end{aligned}$$

where $\varphi'(-1;\Gamma) = \begin{pmatrix} \frac{\partial\varphi_1(-1;\Gamma_1)}{\partial\Gamma_1} & 0 \\ 0 & \frac{\partial\varphi_2(-1;\Gamma_2)}{\partial\Gamma_2} \end{pmatrix}$, and the third equality is due to the Mean Value Theorem, consistency of the estimator and the derivations similar to those for (B.103).

Therefore,combining (B.109) and (B.110) and applying the delta method, we have that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1}, \mathrm{vec}\left(\widehat{\Theta}_{\mathcal{S}_2}\right) - \mathrm{vec}\left(\Theta_{0;\mathcal{S}_2}\right)\right) \xrightarrow{d} \underset{c,d}{\mathrm{argmin}}\, \varphi'\left(-1;\Gamma\right)\left(\varphi_\Gamma'^{\top}\varphi_\Gamma'\right)^{-1}\varphi_\Gamma'^{\top}\Phi_0(c,d),$$

and the proof completes. □

B.6 Proof of Theorem 3.3.4

We write

$$\begin{aligned}\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;b,\zeta) - \Lambda_0(t)\right\} &= \sqrt{n}\left\{\int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n g\left(W_b^{(i)}(\zeta);\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right)Y_i(u)} - \int_0^t d\Lambda_0(u)\right\} \\ &= W_1(t;b,\zeta) + W_2(t;b,\zeta), \end{aligned} \tag{B.111}$$

where

$$\begin{aligned} W_1(t;b,\zeta) &= \sqrt{n}\left\{\int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n g(W_b^{(i)}(\zeta);\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2})Y_i(u)}\right. \\ &\quad \left. - \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n g(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})Y_i(u)}\right\} \end{aligned} \tag{B.112}$$

and

$$W_2(t;b,\zeta) = \sqrt{n}\left\{\int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n g(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})Y_i(u)} - \int_0^t d\Lambda_0(u)\right\}. \tag{B.113}$$

Step 1: *Examine (B.112) and show that*

$$W_1(t;b,\zeta) = \sqrt{n}\left(\begin{array}{c} F_{\beta;b,\zeta}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;b,\zeta}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \end{array}\right)^{\top}\left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1}-\beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2})-\mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array}\right).$$

Let $G_{b,\zeta}\left(u;\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right) = \sum_{i=1}^{n} g\left(W_b^{(i)}(\zeta);\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right) Y_i(u)$ where $g\left(W_b^{(i)}(\zeta);\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right)$ is defined in Section 3.2.3. Applying the Taylor series expansion for $G_{b,\zeta}\left(u;\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right)$ around $(\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2})$ yields

$$\begin{aligned} & \frac{1}{G_{b,\zeta}\left(u;\widehat{\beta}_{\mathcal{S}_1},\widehat{\Theta}_{\mathcal{S}_2}\right)} - \frac{1}{G_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)} \\ = & \frac{1}{\left\{G_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\right\}^2}\left(\begin{array}{c} G^{(1)}_{\beta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ G^{(1)}_{\Theta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \end{array}\right)^{\top}\left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1}-\beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2})-\mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array}\right) \\ & +o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Then (B.112) becomes

$$\begin{aligned} & W_1(t;b,\zeta) \\ = & \sqrt{n}\int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\left\{G_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\right\}^2}\left(\begin{array}{c} G^{(1)}_{\beta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ G^{(1)}_{\Theta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \end{array}\right)^{\top}\left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1}-\beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2})-\mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array}\right) \\ & +o_p(1) \\ = & \int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\left\{G_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\right\}^2}\left(\begin{array}{c} G^{(1)}_{\beta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ G^{(1)}_{\Theta;b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \end{array}\right)^{\top}\sqrt{n}\left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1}-\beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2})-\mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array}\right) \\ & +o_p(1). \end{aligned} \tag{B.114}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} dN_i(u) \xrightarrow{a.s.} dE\left\{N_i(u)\right\}$$

uniformly at u as $n \to \infty$. Therefore, as $n \to \infty$,

$$\int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\left\{G_{b,\zeta}\left(u; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)\right\}^2} \left(\begin{array}{c} G^{(1)}_{\beta;b,\zeta}\left(u; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ G^{(1)}_{\Theta;b,\zeta}\left(u; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \end{array} \right)^\top \xrightarrow{a.s.} \left(\begin{array}{c} F_{\beta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \end{array} \right)^\top, \tag{B.115}$$

where

$$F_{\beta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \int_0^t \frac{dE\left\{N_i(u)\right\} \mathcal{G}^{(1)}_{\beta;b,\zeta}\left(u, \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)}{\left\{\mathcal{G}_{b,\zeta}\left(u, \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)\right\}^2}$$

and

$$F_{\Theta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \int_0^t \frac{dE\left\{N_i(u)\right\} \mathcal{G}^{(1)}_{\Theta;b,\zeta}\left(u, \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)}{\left\{\mathcal{G}_{b,\zeta}\left(u, \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)\right\}^2}.$$

Therefore, combining (B.114) and (B.115) gives

$$W_1(t; b, \zeta) = \sqrt{n} \left(\begin{array}{c} F_{\beta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \end{array} \right)^\top \left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array} \right). \tag{B.116}$$

Taking average on (B.116) with respect to b yields

$$\begin{aligned} W_1(t; \zeta) &= \frac{1}{B} \sum_{b=1}^{B} W_1(t; b, \zeta) \\ &= \sqrt{n} \left(\begin{array}{c} F_{\beta;\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) \end{array} \right)^\top \left(\begin{array}{c} \widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0;\mathcal{S}_2}) \end{array} \right) + o_p(1) \\ &\triangleq \sqrt{n}\mathbf{F}(t; \zeta) + o_p(1), \end{aligned} \tag{B.117}$$

where

$$F_{\beta;\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{B} \sum_{b=1}^{B} F_{\beta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right)$$

and

$$F_{\Theta;\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right) = \frac{1}{B} \sum_{b=1}^{B} F_{\Theta;b,\zeta}\left(t; \beta_{0;\mathcal{S}_1}, \Theta_{0;\mathcal{S}_2}\right).$$

Step 2: *Examine (B.113).*

Since

$$\begin{aligned} W_2(t;b,\zeta) &= \sqrt{n}\int_0^t \left[\frac{\sum_{i=1}^n \left\{dN_i(u) - g\left(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)Y_i(u)d\Lambda_0(u)\right\}}{\sum_{i=1}^n g\left(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)Y_i(u)}\right] \\ &= \frac{1}{\sqrt{n}}\int_0^t \left[\frac{\sum_{i=1}^n \left\{dN_i(u) - g\left(W_b^{(i)}(\zeta);\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)Y_i(u)d\Lambda_0(u)\right\}}{\mathcal{G}_{b,\zeta}\left(u;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)}\right] + o_p(1) \\ &\triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{W}_i(t;b,\zeta) + o_p(1). \end{aligned} \tag{B.118}$$

Taking average of (B.118) with respect to b gives

$$W_2(t;\zeta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{W}_i(t;\zeta) + o_p(1), \tag{B.119}$$

where $\mathbf{W}_i(t;\zeta) = \frac{1}{B}\sum_{b=1}^B \mathbf{W}_i(t;b,\zeta)$.

Step 3: *Establish the result in Theorem 3.3.4.*

Combining (B.117) and (B.119) yields

$$\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;\zeta) - \Lambda_0(t)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\mathbf{F}(t;\zeta) + \mathbf{W}_i(t;\zeta)\right\} + o_p(1). \tag{B.120}$$

Suppose that $\varphi_\Lambda(\zeta;\Gamma_\Lambda)$ is a regression function, and Γ_Λ is the associated parameter. Let $\mathcal{R}_\Lambda(\Gamma_\Lambda) = \widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;\mathcal{Z}) - \varphi_\Lambda(\mathcal{Z};\Gamma_\Lambda)$ for a given time point t, and let $\widehat{\Gamma}_\Lambda$ denote the solution of

$$\varphi'^{\top}_{\Gamma,\Lambda}\mathcal{R}_\Lambda(\Gamma_\Lambda) = 0.$$

Similar to the derivations for (B.105), we have

$$\begin{aligned}
&\sqrt{n}\left\{\widehat{\Gamma}_\Lambda - \Gamma_\Lambda\right\} \\
= \;& \left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t;\mathcal{Z}) - \Lambda_0(t)\right\} \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\left\{\mathbf{F}(t;\mathcal{Z}) + \mathbf{W}_i(t;\mathcal{Z})\right\} + o_p(1). \qquad \text{(B.121)}
\end{aligned}$$

Finally, applying the delta method to (B.121) and taking $\zeta = -1$ as the extrapolation gives

$$\begin{aligned}
&\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t) - \Lambda_0(t)\right\} \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\left\{\mathbf{F}(t;\mathcal{Z}) + \mathbf{W}_i(t;\mathcal{Z})\right\} + o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\mathbf{F}(t;\mathcal{Z}) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\mathbf{W}_i(t;\mathcal{Z}) + o_p(1) \\
\triangleq \;& \mathbf{A}(t) + \mathbf{B}(t) + o_p(1). \qquad \text{(B.122)}
\end{aligned}$$

Noting that $\varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\mathbf{F}(t;\mathcal{Z})$ is free of index i, then by the definition of $\mathbf{F}(\cdot)$ in (B.117), $\mathbf{A}(t)$ can be re-written as

$$\begin{aligned}
\mathbf{A}(t) \;=\; & \sqrt{n}\varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1} \\
& \times\varphi_{\Gamma,\Lambda}'^{\top}\begin{pmatrix} F_{\beta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}^{\top}\begin{pmatrix} \widehat{\beta}_{\mathcal{S}_1} - \beta_{0;\mathcal{S}_1} \\ \mathrm{vec}(\widehat{\Theta}_{\mathcal{S}_2}) - \mathrm{vec}(\Theta_{0;\mathcal{S}_2})\end{pmatrix} + o_p(1).
\end{aligned}$$

Thus, when $n \to \infty$,

$$\begin{aligned}
\mathbf{A}(t) \;\xrightarrow{d}\; & \varphi_\Lambda'(-1;\Gamma_\Lambda)\left(\varphi_{\Gamma,\Lambda}'^{\top}\varphi_{\Gamma,\Lambda}'\right)^{-1}\varphi_{\Gamma,\Lambda}'^{\top}\begin{pmatrix} F_{\beta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}^{\top} \\
& \times \operatorname*{argmin}_{c,d}\varphi'\left(-1;\Gamma\right)\left(\varphi_\Gamma'^{\top}\varphi_\Gamma'\right)^{-1}\varphi_\Gamma'^{\top}\Phi_0(c,d). \qquad \text{(B.123)}
\end{aligned}$$

Let $\mathbf{W}_i(t) = \varphi'_\Lambda(-1;\Gamma_\Lambda)\left(\varphi'^\top_{\Gamma,\Lambda}\varphi'_{\Gamma,\Lambda}\right)^{-1}\varphi'^\top_{\Gamma,\Lambda}\mathbf{W}_i(t;\mathcal{Z})$. Since $\mathbf{W}_i(t;\zeta)$ are i.i.d. with mean zero due to Condition (C5), so by the Central Limit Theorem, when $n \to \infty$,

$$\begin{aligned}\mathbf{B}(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}_i(t) + o_p(1) \\ &\xrightarrow{d} \mathcal{W}(t), \end{aligned} \tag{B.124}$$

where $\mathcal{W}(t)$ is a Gaussian process with mean zero and covariance $E\{\mathbf{W}_i(t)\mathbf{W}_i(s)\}$. Therefore, combining (B.123) and (B.124) with (B.122) gives that as $n \to \infty$,

$$\begin{aligned}\sqrt{n}\left\{\widehat{\Lambda}_{\widehat{\mathcal{N}},0}(t) - \Lambda_0(t)\right\} \xrightarrow{d}\ & \mathcal{W}(t) + \varphi'_\Lambda(-1;\Gamma_\Lambda)\left(\varphi'^\top_{\Gamma,\Lambda}\varphi'_{\Gamma,\Lambda}\right)^{-1}\varphi'^\top_{\Gamma,\Lambda}\begin{pmatrix} F_{\beta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right) \\ F_{\Theta;\mathcal{Z}}\left(t;\beta_{0;\mathcal{S}_1},\Theta_{0;\mathcal{S}_2}\right)\end{pmatrix}^\top \\ & \times \underset{c,d}{\operatorname{argmin}}\,\varphi'\left(-1;\Gamma\right)\left(\varphi'^\top_\Gamma\varphi'_\Gamma\right)^{-1}\varphi'^\top_\Gamma\Phi_0(c,d).\end{aligned}$$

□

Appendix C

Proofs for the Results in Chapter 4

C.1 Regularity Conditions

(C1) $P\left(R_i(\tau)=1\right)>0$, where $R_i(t)=I\{Y_i>t\}$ and τ is an upper bound of survival times which is assumed to be finite.

(C2) Censoring time is non-informative. That is, the survival time and the censoring time are independent

(C3) The $\{I(Y_i\leq t), X_i^*\}$ are independent and identically distributed for $i=1,\cdots,n$.

(C4) The bandwidth h falls in the interval $H_{\kappa;n}=[h_l n^{-\kappa}, h_u n^{-\kappa}]$ for some constants h_l and h_u and $\kappa\in(1/(4q), 1/\max\{2d+2, d+4\})$.

(C5) $\mathcal{A}$ is nonsingular.

(C6) Both X and Y satisfy the subexponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0<s\leq 2s_0$,

$$\sup_p \max_{1\leq k\leq p} E\left\{\exp\left(s\|X_k\|_1^2\right)\right\}<\infty \quad \text{and} \quad E\left\{\exp\left(s\|Y\|_1^2\right)\right\}<\infty.$$

(C7) The minimum DC of active predictors satisfies

$$\min_{k\in\mathcal{I}} \omega_k^* \geq 2cn^{-\xi}$$

for some constants $c>0$ and $0\leq\xi<1/2$.

Conditions (C1) to (C3) are regular assumptions in survival analysis for the establishment of the asymptotic properties (e.g., Andersen and Gill 1982). Conditions (C4) is a constraint for bandwidth and is used to establish the $\sqrt{n}$-consistency of $\widehat{B}$ (Huang and Chiang 2017). Condition (C5) indicates that $\mathcal{A}$ is a positive definite matrix and is used to establish the asymptotic distribution of $\widehat{B}$. Conditions (C6) and (C7) come from the requirement for the feature selection (e.g., Li et al. 2012).

C.2 Technical Lemmas

Based on definitions in (4.22) and (4.23), we further define

$$\widetilde{\mathbb{F}}^{(j)}_{l,B,L}(y,u) = \widehat{\mathbb{F}}^{(j)}_{l,B,L}(y, B^\top u) - \mathbb{F}^{(j)}_{l,B,L}(y,u) \tag{C.1}$$

and

$$\widetilde{\mathbb{F}}^{(j)}_{l,B,\widehat{L}}(y,u) = \widehat{\mathbb{F}}^{(j)}_{l,B,\widehat{L}}(y, B^\top u) - \mathbb{F}^{(j)}_{l,B,L}(y,u) \tag{C.2}$$

for $l = 0, 1$ and $j = 0, 1, 2$. In particular, we let $\widetilde{\mathbb{F}}_{l,B,L}(y,u) = \widetilde{\mathbb{F}}^{(0)}_{l,B,L}(y,u)$, $\widehat{\mathbb{F}}_{l,B,L}(y,u) = \widehat{\mathbb{F}}^{(0)}_{l,B,L}(y,u)$, and $\mathbb{F}_{l,B,L}(y,u) = \mathbb{F}^{(0)}_{l,B,L}(y,u)$ for $l = 0, 1$.

Furthermore, we define

$$\widetilde{F}^{(j)}(y,u) = \widehat{F}^{(j)}(y, B^\top u) - F^{(j)}(y,u) \tag{C.3}$$

for $j = 0, 1$, where $\widehat{F}^{(j)}(y, B^\top u) = \nabla^j_{\mathrm{vec}(B)} \widehat{F}(y, B^\top u)$, and $F^{(j)}(y,u)$ with $j = 0$ and 1 are defined in (4.25) and (4.26), respectively. In particular, when $j = 0$, we have $\widehat{F}^{(0)}(y, B^\top u) = \widehat{F}(y, B^\top u)$ and $F^{(0)}(y,u) = F(y,u)$. As a result, let $\widetilde{F}(y,u) = \widetilde{F}^{(0)}(y,u)$ if $j = 0$. Noting that $\widehat{F}^{(j)}(y, B^\top u)$ involves the measurement error correction L, so here we add the subscript in (C.3) to emphasize the involvement of L. That is, if L is known, then we re-write (C.3) by

$$\widetilde{F}^{(j)}_{B,L}(y,u) = \widehat{F}^{(j)}_{B,L}(y, B^\top u) - F^{(j)}_{B,L}(y,u);$$

if L is unknown and is estimated by $\widehat{L}$, then we express (C.3) by

$$\widetilde{F}^{(j)}_{B,\widehat{L}}(y,u) = \widehat{F}^{(j)}_{B,\widehat{L}}(y, B^\top u) - F^{(j)}_{B,L}(y,u).$$

In the following two lemmas, we present the convergence rates of (C.1) and (C.3).

Lemma C.2.1 *Suppose that regularity conditions in Appendix C.1 holds. For $j = 0, 1, 2$ and $l = 0, 1$, if L is known, then*

$$\sup_{y,u,B} \left\| \widetilde{\mathbb{F}}^{(j)}_{l,B,L}(y,u) \right\| = O\left(h^q\right) + o\left(\frac{\log(n)}{\sqrt{nh^{j+d}}}\right) \tag{C.4}$$

almost surely (a.s.); if L is unknown and $\widehat{L}$ is the estimator of L, then

$$\sup_{y,u,B} \left\| \widetilde{\mathbb{F}}^{(j)}_{l,B,\widehat{L}}(y,u) \right\| = o\left(\frac{p\ \log(m)}{\sqrt{m}}\right) + O\left(h^q\right) + o\left(\frac{\log(n)}{\sqrt{nh^{j+d}}}\right)\ \ a.s.. \tag{C.5}$$

Proof:

We first show (C.4). Since $\{I(Y_i \leq y) : y \geq 0\}$, $\left\{\mathcal{K}(B^\top U_i - u) : B \in \mathbb{R}^{p\times d}\right\}$ and $\{(U_i - u)^{\otimes j} : j = 0, 1, 2\}$ are the *Vapnik-Červonenkis* (VC) classes by Giné and Guillou (2002, p.911) and Lemma 2.4 in Pakes and Pollard (1989). Besides, Lemma 2.12 in Pakes and Pollard (1989) implies that those three classes are Euclidean, and thus, Lemma 2.14 in Pakes and Pollard (1989) indicates that $\left\{\delta_i \left\{I(Y_i \leq y)\right\}^l \nabla^j_{\mathrm{vec}(B)} \mathcal{K}\left(B^\top \widehat{U}_i - u\right) : y, u, B\right\}$ is also a Euclidean. As a result, by Theorem II.37 in Pollard (1984) and derivations similar to Giné and Guillou (2002), we have

$$\widehat{\mathbb{F}}^{(j)}_{l,B,L}(y, B^\top u) - E\left\{\widehat{\mathbb{F}}^{(j)}_{l,B,L}(y, B^\top u)\right\} = o\left(\frac{\log(n)}{\sqrt{nh^{j+d}}}\right)\ \ a.s. \tag{C.6}$$

and

$$E\left\{\widehat{\mathbb{F}}^{(j)}_{l,B,L}(y, B^\top u)\right\} - \mathbb{F}^{(j)}_{l,B,L}(y, B^\top u) = O\left(h^q\right)\ \ a.s. \tag{C.7}$$

for $l = 0, 1$ and $j = 0, 1, 2$. Therefore, combining (C.6) and (C.7) gives (C.4).

We next show (C.5). Since $\widehat{L}$ is involved, then we consider

$$\begin{aligned}
& \widehat{\mathbb{F}}^{(j)}_{l,B,\widehat{L}}(y,u) - \widehat{\mathbb{F}}^{(j)}_{l,B,L}(y,u) \\
= & \frac{1}{n}\sum_{i=1}^n \left[\delta_i \left\{I(Y_i \leq y)\right\}^l \nabla^j_{\mathrm{vec}(B)} \left\{\mathcal{K}\left(B^\top \widehat{U}_i - u\right) - \mathcal{K}\left(B^\top U_i - u\right)\right\}\right] \\
= & \frac{1}{n}\sum_{i=1}^n \left[\delta_i \left\{I(Y_i \leq y)\right\}^l \nabla^j_{\mathrm{vec}(B)} \left\{\nabla^1_L \mathcal{K}\left(B^\top U_i - u\right)\right\}\right] \left(\widehat{L} - L\right).
\end{aligned}$$

Since $\widehat{L}$ is estimated by either repeated measurements of validation sample, by the similar derivations of Theorem 1 in Zhang et al. (2014), we have

$$\widehat{L} - L = o\left(\frac{p\ \log(m)}{\sqrt{m}}\right).$$

As the result, we have

$$\sup_{y,u,B}\left\|\widehat{\mathbb{F}}^{(j)}_{l,B,\widehat{L}}(y,u) - \widehat{\mathbb{F}}^{(j)}_{l,B,L}(y,u)\right\| = o\left(\frac{p\ \log(m)}{\sqrt{m}}\right),$$

and combining the result (C.4) gives the desired result of (C.5). □

Let $\zeta^{(0)}_{i,B_0}(y,u)$ be as defined in (4.24). In addition, define

$$\begin{aligned}\zeta^{(1)}_{i,B_0}(y,u) &= \sum_{j,l=0}^{1}\left\{\widetilde{\mathbb{F}}^{(j)}_{i,l,B,L}(y,u)\frac{(-1)^{1+l}\mathbb{F}_{1-l,B,L}(y,B^\top u)}{\mathbb{F}^{3-l}_{0,B,L}(y,B^\top u)}\right. \\ &\left.\times\left(\sum_{l'=0}^{1}(l+l'-2)\mathbb{F}^{l'}_{1-l',B,L}(u,B^\top u)\mathbb{F}^{l'}_{l',B,L}(y,B^\top u)\mathbb{F}^{1}_{l',B,L}(y,B^\top u)\right)^{1-j}\right\}.\end{aligned} \tag{C.8}$$

Lemma C.2.2 *Suppose that regularity conditions in Appendix C.1 holds. For $j = 0, 1$, if L is known, then*

$$\sup_{y,u,B}\left\|\widetilde{F}^{(j)}_{B,L}(y,u) - \frac{1}{n}\sum_{i=1}^{n}\zeta^{(j)}_{i,B}(y,u)\right\| = o_p\left(\frac{1}{\sqrt{n}}\right)\ a.s.; \tag{C.9}$$

if L is unknown and $\widehat{L}$ is the estimator, then

$$\sup_{y,u,B}\left\|\widetilde{F}^{(j)}_{B,\widehat{L}}(y,u) - \frac{1}{n}\sum_{i=1}^{n}\zeta^{(j)}_{i,B}(y,u)\right\| = o_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(\frac{1}{\sqrt{m}}\right)\ a.s.. \tag{C.10}$$

Proof:

For $j = 0$, by expressions (4.19) and (4.22), we observe that

$$\begin{aligned}
\widetilde{F}_{B,L}(y,u) &= \widehat{F}_{B,L}(y,B^\top u) - F_{B,L}(y,u) \\
&= \frac{\widehat{\mathbb{F}}_{1,B,L}(y,B^\top u)}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)} - \frac{\mathbb{F}_{1,B,L}(y,u)}{\mathbb{F}_{0,B,L}(y,u)} \\
&= \frac{\widehat{\mathbb{F}}_{1,B,L}(y,B^\top u)\mathbb{F}_{0,B,L}(y,u) - \widehat{\mathbb{F}}_{0,B,L}(y,u)\mathbb{F}_{1,B,L}(y,u)}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)\mathbb{F}_{0,B,L}(y,u)} \\
&= \frac{\mathbb{F}_{0,B,L}(y,u)\left\{\widehat{\mathbb{F}}_{1,B,L}(y,B^\top u) - \mathbb{F}_{1,B,L}(y,B^\top u)\right\}}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)\mathbb{F}_{0,B,L}(y,u)} \\
&\quad - \frac{\mathbb{F}_{1,B,L}(y,u)\left\{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u) - \mathbb{F}_{0,B,L}(y,B^\top u)\right\}}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)\mathbb{F}_{0,B,L}(y,u)} \\
&= \frac{\widetilde{\mathbb{F}}_{1,B,L}(y,B^\top u)}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)} - \frac{F(y,u)\widetilde{\mathbb{F}}_{0,B,L}(y,B^\top u)}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)}, \qquad \text{(C.11)}
\end{aligned}$$

where the fourth equality is due to adding and subtracting an additional term $\mathbb{F}_{0,B,L}(y,u)$ $\times \mathbb{F}_{1,B,L}(y,u)$, the fifth equality comes from (C.1) with $j = 0$ and $l = 0, 1$.

Moreover, applying the Taylor series expansion on $\left\{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)\right\}^{-1}$ gives

$$\frac{1}{\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)} = \frac{1}{\mathbb{F}_{0,B,L}(y,B^\top u)} - \frac{\widetilde{\mathbb{F}}_{0,B,L}(y,B^\top u)}{\mathbb{F}^2_{0,B,L}(y,B^\top u)} + \frac{2\left\{\widetilde{\mathbb{F}}_{0,B,L}(y,B^\top u)\right\}^2}{\widehat{\mathbb{F}}^{*3}_{0,B,L}(y,B^\top u)}, \qquad \text{(C.12)}$$

where $\widehat{\mathbb{F}}^{*}_{0,B,L}(y,B^\top u)$ is between $\widehat{\mathbb{F}}_{0,B,L}(y,B^\top u)$ and $\mathbb{F}_{0,B,L}(y,B^\top u)$. Combining (C.11) with (C.12) and applying Lemma C.2.1 give

$$\begin{aligned}
\widetilde{F}^{(0)}_{B,L}(y,u) &= \frac{1}{n}\sum_{i=1}^{n}\zeta^{(0)}_{i,B}(y,u) - \frac{\widetilde{\mathbb{F}}^{(0)}_{0,B,L}(y,u)\widetilde{\mathbb{F}}^{(0)}_{1,B,L}(y,u)}{\mathbb{F}^2_{0,B,L}(y,B^\top u)} \\
&\quad + \frac{2\widetilde{\mathbb{F}}_{1,B,L}(y,B^\top u)\left\{\widetilde{\mathbb{F}}^{(0)}_{0,B,L}(y,u)\right\}^2}{\widehat{\mathbb{F}}^{*3}_{0,B,L}(y,B^\top u)}, \qquad \text{(C.13)}
\end{aligned}$$

where $\zeta^{(0)}_{i,B}(y,u)$ is given by (4.24).

By Lemma C.2.1, we have

$$\sup_{y,u,B}\left\|\frac{\widetilde{\mathbb{F}}^{(0)}_{0,B,L}(y,u)\widetilde{\mathbb{F}}^{(0)}_{1,B,L}(y,u)}{\mathbb{F}^{2}_{0,B,L}(y,B^{\top}u)}-\frac{2\widetilde{\mathbb{F}}_{1,B,L}(y,B^{\top}u)\left\{\widetilde{\mathbb{F}}^{(0)}_{0,B,L}(y,u)\right\}^{2}}{\widehat{\mathbb{F}}^{*3}_{0,B,L}(y,B^{\top}u)}\right\|=o_p\left(\frac{1}{\sqrt{n}}\right)\ a.s.. \tag{C.14}$$

Therefore, combining with (C.13) and (C.14) gives (C.9) with $j=0$. Similar procedure gives (C.9) with $j=1$.

We next discuss the derivation of (C.10). Since

$$\begin{aligned}\widetilde{F}^{(j)}_{B,\widehat{L}}(y,u) &= \widehat{F}^{(j)}_{B,\widehat{L}}(y,B^{\top}u)-F^{(j)}_{B,L}(y,u)\\ &= \left(\widehat{F}^{(j)}_{B,\widehat{L}}(y,B^{\top}u)-\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)\right)+\left(\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)-F^{(j)}_{B,L}(y,u)\right)\\ &= \left(\widehat{F}^{(j)}_{B,\widehat{L}}(y,B^{\top}u)-\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)\right)+\widetilde{F}^{(j)}_{B,L}(y,u)\end{aligned} \tag{C.15}$$

for $j=0,1$. The second term of (C.15) is derived, so the remaining target is the first term of (C.15). By the Taylor series expansion on $\widehat{F}^{(j)}_{B,\widehat{L}}(y,u)$ with respect to L gives

$$\widehat{F}^{(j)}_{B,\widehat{L}}(y,B^{\top}u)-\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)=\left(\nabla^{1}_{L}\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)\right)\left(\widehat{L}-L\right)$$

for $j=0,1$. By the result in Lemma C.2.1, we have

$$\sup_{y,u,B}\left\|\widehat{F}^{(j)}_{B,\widehat{L}}(y,B^{\top}u)-\widehat{F}^{(j)}_{B,L}(y,B^{\top}u)\right\|=o_p\left(\frac{1}{\sqrt{m}}\right). \tag{C.16}$$

Consequently, combining (C.16) and (C.9) with (C.15) gives (C.10). □

Let

$$\begin{aligned}\sigma^{2}_{0} &= E\left\{\left\|I\left(Y_i\le y\right)-F(y,B^{\top}_{0}U_i)\right\|^{2}\right\}\\ b^{2}_{0}(B) &= E\left\{\left\|F(y,B^{\top}_{0}U_i)-F(y,B^{\top}U_i)\right\|^{2}\right\}\\ \mathcal{B}_{B}(y,u) &= \frac{\int v^{q}K(v)dv}{q!}(dh^{q})\sum_{l=0}^{1}\frac{\left\{-F\left(y,u\right)\right\}^{1-l}\nabla^{q}\mathbb{F}_{0,B,L}(y,u)}{\mathbb{F}_{0,B,L}\left(y,u\right)}\\ \mathcal{V}_{B}(y,u) &= \frac{\left(\int K(v)dv\right)^{d}F(y,u)\left\{1-F(y,u)\right\}}{nh^{d}f_{B^{\top}U}(u)}.\end{aligned}$$

In addition, define

$$AMISE_B(h) = E\left\{\mathcal{B}_B^2(y, B^\top u) + \mathcal{V}_B(y, B^\top u)\right\}$$

and

$$ECV(B,d,h) = \begin{cases} \sigma_0^2 + AMISE_B(h) & \text{if } \mathcal{S}_{T|U} \subseteq \mathcal{S}(B) \\ \sigma_0^2 + b_0^2(B) + AMISE_B(h) & \text{if } \mathcal{S}_{T|U} \nsubseteq \mathcal{S}(B). \end{cases}$$

In the next lemma, we examine the behavior of the CV value.

Lemma C.2.3 *Under regularity conditions in Appendix C.1, if* $\mathcal{S}_{T|U} \subseteq \mathcal{S}(B)$*, then*

$$\sup_{B,d,h} \frac{|CV(B,d,h) - ECV(B,d,h)|}{AMISE_B(h)} = o_p(1) \ a.s.; \tag{C.17}$$

if $\mathcal{S}_{T|U} \nsubseteq \mathcal{S}(B)$*, then*

$$\sup_{B,d,h} \frac{|CV(B,d,h) - ECV(B,d,h)|}{b_0(B) AMISE_B(h)} = O_p(1) \ a.s.. \tag{C.18}$$

Proof:

Let

$$\begin{aligned} \mathcal{E}_{1;i,Y_j} &= I\left(Y_i \le y\right) - F(Y_j, B_0^\top U_i), \\ \mathcal{E}_{2;i,Y_j} &= F(Y_j, B_0^\top U_i) - F(Y_j, B^\top U_i), \\ \mathcal{E}_{3;i,Y_j} &= F(Y_j, B^\top U_i) - \widehat{F}^{(-i)}(Y_j, B^\top U_i), \end{aligned}$$

and

$$\mathcal{E}_{4;i,Y_j} = \widehat{F}^{(-i)}(Y_j, B^\top U_i) - \widehat{F}^{(-i)}(Y_j, B^\top \widehat{U}_i).$$

Then the cross-validation criterion (4.20) can be decomposed by

$$\begin{aligned} CV(B,d,h) &= \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{1;i,Y_j}^2 + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{2;i,Y_j}^2 + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{3;i,Y_j}^2 + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{4;i,Y_j}^2 \\ &\quad + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{1;i,Y_j}\mathcal{E}_{2;i,Y_j} + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{1;i,Y_j}\mathcal{E}_{3;i,Y_j} + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{1;i,Y_j}\mathcal{E}_{4;i,Y_j} \\ &\quad + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{2;i,Y_j}\mathcal{E}_{3;i,Y_j} + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{2;i,Y_j}\mathcal{E}_{4;i,Y_j} + \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{E}_{3;i,Y_j}\mathcal{E}_{4;i,Y_j} \\ &\triangleq S_1 + S_2 + S_3 + S_4 + R_1 + R_2 + R_3 + R_4 + R_5 + R_6. \end{aligned} \tag{C.19}$$

To study the uniform consistency of $CV(B,d,h)$ and $ECV(B,d,h)$, we consider the following two scenarios.

Case 1: $\mathcal{S}_{T|U} \subseteq \mathcal{S}(B)$.

In this case, we have $F(y, B_0^\top u) = F(y, B^\top u)$ (e.g., Huang and Chiang 2017). Thus, we immediately have

$$S_2 = R_1 = R_4 = R_5 = 0.$$

Since S_1 is the form of U-statistic, then applying the convergence property of U-statistic (e.g., van der Vaart 1998, Chapter 12) gives that as $n \to \infty$,

$$S_1 \xrightarrow{p} \sigma_0^2. \tag{C.20}$$

By Lemmas C.2.1 and C.2.2, S_3 can be expressed as

$$\begin{aligned} S_3 &= \left\{ \frac{1}{n^2(n-1)^2} \sum_{i \neq j_1 \neq j_2} \sum_{k=1}^{n} \zeta_{j_1,B}\left(Y_k, B^\top U_i\right) \zeta_{j_2,B}\left(Y_k, B^\top U_i\right) \right. \\ &\quad \left. + \frac{1}{n^2(n-1)^2} \sum_{i \neq j} \sum_{k=1}^{n} \zeta_{j,B}^2\left(Y_k, B^\top U_i\right) \right\} \{1 + o_p(1)\} \\ &\triangleq \{K_1(B) + K_2(B)\} \{1 + o_p(1)\}. \end{aligned} \tag{C.21}$$

Since the class $\{\zeta_{j_1,B}(y,u)\}$ is Euclidean (Pakes and Pollard 1989, Theorems 2.13 and 2.14), then by the derivations similar to Theorem of Huang and Chiang (2017), we can show that

$$\sup_B \left| K_1(B) - E\left\{\mathcal{B}_B^2\left(Y_k, B^\top U_i\right)\right\} \right| = o_p(1) \ a.s. \tag{C.22}$$

and

$$\sup_B \left| K_2(B) - E\left\{\mathcal{V}_B\left(Y_k, B^\top U_i\right)\right\} \right| = o_p(1) \ a.s.. \tag{C.23}$$

As a result, by (C.22) and (C.23), we have

$$\sup_B |S_3 - AMISE_B(h)| = o_p\left(dh^{2q} + \frac{1}{nh^d}\right). \tag{C.24}$$

On the other hand, for the parameter corresponding to measurement error model, if L is known, then $S_4 = R_3 = R_6 = 0$. If L is unknown and $\widehat{L}$ is the corresponding estimator, then

$$\begin{aligned} S_4 &= \frac{1}{n}\sum_{i=1}^{n}\int \left\{\widehat{F}^{(-i)}\left(y, B^\top U_i\right) - \widehat{F}^{(-i)}\left(y, B^\top \widehat{U}_i\right)\right\}^2 dF_Y(y) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int \left\{\nabla_L^1 \widehat{F}^{(-i)}\left(y, B^\top U_i\right)\right\}^2 dF_Y(y)\left(\widehat{L} - L\right)^2. \end{aligned} \tag{C.25}$$

As shown in Li and Yin (2007), $\widehat{L}$ is the estimator of L with $\sqrt{m}$-rate. It means that $\widehat{L} - L = O_p\left(\frac{1}{\sqrt{m}}\right)$. As a result, applying Lemma C.2.1 gives $S_4 = o_p(1)$.

For R_3 and R_6, similar to the derivations, we have $\sup_B |R_3| = \sup_B |R_6| = o_p(1)$. Combining all results (C.20) - (C.25) with (C.19), we have

$$\begin{aligned} CV(B,d,h) &= \sigma_0^2 + S_3 + o_p(1) \\ &= \sigma_0^2 + S_3 - AMISE(h) + AMISE_B(h) + o_p(1) \\ &= ECV(B,d,h) + \{S_3 - AMISE_B(h)\} + o_p(1), \end{aligned}$$

where $ECV(B,d,h) = \sigma_0^2 + AMISE_B(h)$. Consequently, by (C.24) and similar derivation of Proposition 2 in Huang and Chiang (2017) that $AMISE_B(h) = O\left(dh^{2q} + \frac{1}{nh^d}\right)$, we have

$$\sup_{d,B,h}\left|\frac{CV(B,d,h) - ECV(B,d,h)}{AMISE_B(h)}\right| = o(1)\ a.s..$$

Case 2: $\mathcal{S}_{T|U} \not\subseteq \mathcal{S}(B)$.
In this case, the derivations of S_1, S_3, R_2, R_3, and R_6 can be determined in Case 1, but S_2, R_1, R_4, and R_5 do not equal zero. The main goal in this case is to discuss those remaining parts.

Noting that by Theorem 3.1 in Arcones and Giné (1993), we have that

$$\sup_B \left|S_2 - b_0^2(B)\right| = o(1)\ a.s.,$$

which is equivalent to

$$\sup_B \left|\frac{S_2}{b_0^2(B)} - 1\right| = o(1)\ a.s.. \tag{C.26}$$

Besides, by the CauchySchwarz inequality, we have $R_1^2 \leq S_1 S_2$. Then dividing $b_0(B)$ on both sides and applying (C.20) and (C.26) yield

$$\sup_B \frac{|R_1|}{b_0(B)} = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Furthermore, by the Cauchy-Schwarz inequality again, Lemma C.2.1 and (C.26), we have

$$\sup_B \frac{|R_4|}{b_0(B)} = O\left(h^q\right) + o_p\left(\frac{\log(n)}{\sqrt{nh^d}}\right)$$

and

$$\sup_B \frac{|R_5|}{b_0(B)} = o\left(\frac{p\ \log(m)}{\sqrt{m}}\right).$$

Consequently, combining those results with (C.19) yields

$$\sup_{B,d,h} \frac{|CV(B,d,h) - ECV(B,d,h)|}{b_0(B) AMISE(h)} = O_p(1)\ a.s.,$$

where $ECV(B,h) = \sigma_0^2 + b_0^2(B) + AMISE(h)$. □

C.3 Proofs of Proposition in Section 4.2

C.3.1 Proof of Proposition 4.2.2

We first observe that

$$\begin{aligned} E\left\{I\left(Y \leq t\right)|\Delta = 1, T, U\right\} &= P\left(Y \leq t|\Delta = 1, T, U\right) \\ &= \frac{P\left(Y \leq t, \Delta = 1|T, U\right)}{P\left(\Delta = 1|T, U\right)} \\ &= \frac{P\left(T \leq t, T \leq C|T, U\right)}{P\left(T \leq C|T, U\right)} \\ &= P\left(T \leq t|U\right). \end{aligned} \tag{C.27}$$

On the other hand, we can express $P(T \leq t|U)$ by

$$\begin{aligned} P(T \leq t|U) &= E\{I(T \leq t)|U\} \\ &= E\left\{ \frac{\Delta I(Y \leq t)}{P(T \leq C|U)} \middle| U \right\} \\ &= \frac{P(\Delta = 1, Y \leq t|U)}{P(T \leq C|U)} \\ &= P(\Delta = 1, Y \leq t|\Delta = 1, U) \\ &= E\{\Delta I(Y \leq t)|\Delta = 1, U\}. \end{aligned} \tag{C.28}$$

Consequently, using (C.27) and (C.28) gives

$$\begin{aligned} & E\{I(Y \leq t)|\Delta = 1, U\} \\ = \; & E[E\{I(Y \leq t)|\Delta = 1, T, U\}|\Delta = 1, U] \\ = \; & E[E\{\Delta I(Y \leq t)|\Delta = 1, U\}|\Delta = 1, U] \\ = \; & E\{\Delta I(Y \leq t)|\Delta = 1, U\} \\ = \; & P(T \leq t|U), \end{aligned} \tag{C.29}$$

which yields the desired result. □

C.4 Proofs of Theorems in Section 4.3

C.4.1 Proof of Theorem 4.3.1

Note that $\widehat{L}$ is the consistent estimator in the sense that $\widehat{L} = L + o_p(1)$, then we have $\widehat{U}_i = U_i + o_p(1)$. Hence, in the remaining proof, we focus on U_i. In addition, we separate this proof to two steps. In step 1, we discuss the consistency of $\widehat{B}$, and then discuss the asymptotic performances of $\widehat{h}$ and $\widehat{d}$ in Step 2.

Step 1: The consistency of $\widehat{B}$.

Let $\widehat{B}$ denote the the minimizer of $CV(B, d_0, h_0)$ with $d = d_0$ and $h = h_0$. Then we have

$$P\left\{ CV\left(\widehat{B}, d_0, h_0\right) < CV(B_0, d_0, h_0) \right\} = 1. \tag{C.30}$$

Let $DCV(B, d_0, h_0) = |CV(B, d_0, h_0) - ECV(B, d_0, h_0)|$. For every $\epsilon > 0$, we further have

$$\begin{aligned}
& \left\{CV\left(\widehat{B}, d_0, h_0\right) < CV\left(B_0, d_0, h_0\right)\right\} \\
= & \left\{b_0^2(\widehat{B}) < \epsilon, b_0^2(\widehat{B}) > \epsilon, CV\left(\widehat{B}, d_0, h_0\right) < CV\left(B_0, d_0, h_0\right)\right\} \\
\subseteq & \left\{b_0^2(\widehat{B}) < \epsilon\right\} \cup \left\{b_0^2(\widehat{B}) > \epsilon, DCV\left(\widehat{B}, d_0, h_0\right) + DCV\left(B_0, d_0, h_0\right)\right. \\
& \left. > ECV\left(\widehat{B} d_0, h_0\right) - ECV\left(B_0, d_0, h_0\right)\right\}. \qquad \text{(C.31)}
\end{aligned}$$

By (C.30) and (C.31), we have

$$\begin{aligned}
1 \leq & P\left\{b_0^2(\widehat{B}) < \epsilon\right\} \\
& + P\left\{b_0^2(\widehat{B}) > \epsilon, DCV\left(\widehat{B}, d_0, h_0\right) + DCV\left(B_0, d_0, h_0\right)\right. \\
& \left. > ECV\left(\widehat{B}, d_0, h_0\right) - ECV\left(B_0, d_0, h_0\right)\right\}. \qquad \text{(C.32)}
\end{aligned}$$

For the second term of (C.32), we further have

$$\begin{aligned}
& P\left\{b_0^2(\widehat{B}) > \epsilon, DCV\left(\widehat{B}, d_0, h_0\right) + DCV\left(B_0, d_0, h_0\right)\right. \\
& \left. > ECV\left(\widehat{B}, d_0, h_0\right) - ECV\left(B_0, d_0, h_0\right)\right\} \\
= & P\left\{b_0^2(\widehat{B}) > \epsilon, \frac{DCV\left(\widehat{B}, d_0, h_0\right)}{b_0(\widehat{B})} + \frac{DCV\left(B_0, d_0, h_0\right)}{b_0(\widehat{B})}\right. \\
& \left. > \frac{ECV\left(\widehat{B}, d_0, h0\right)}{b_0(\widehat{B})} - \frac{ECV\left(B_0, d_0, h_0\right)}{b_0(\widehat{B})}\right\} \\
\leq & P\left\{b_0^2(\widehat{B}) > \epsilon, \frac{DCV\left(\widehat{B}, d_0, h_0\right)}{b_0(\widehat{B})} + \frac{DCV\left(B_0, d_0, h_0\right)}{\sqrt{\epsilon}}\right. \\
& \left. > \frac{ECV\left(\widehat{B}, d_0, h_0\right)}{\sqrt{\epsilon}} - \frac{ECV\left(B_0, d_0, h_0\right)}{\sqrt{\epsilon}}\right\}, \qquad \text{(C.33)}
\end{aligned}$$

where the last step is due to $b_0^{-1}(\widehat{B}) < \epsilon^{-1/2}$.

Since $b_0^2(\widehat{B}) > \epsilon > 0$ implies $\mathcal{S}_{T|U} \not\subseteq \mathcal{S}(B)$, then $ECV(B,d,h) = \sigma_0^2 + b_0^2(B) + AMISE(h)$. Therefore, (C.33) becomes

$$\begin{aligned}
& P\left\{b_0^2(\widehat{B}) > \epsilon, DCV\left(\widehat{B}, d_0, h_0\right) + DCV\left(B_0, d_0, h_0\right)\right. \\
& \left. > ECV\left(\widehat{B}, d_0, h_0\right) - ECV\left(B_0, d_0, h_0\right)\right\} \\
\leq\ & P\left\{b_0^2(\widehat{B}) > \epsilon, O_p\left\{AMISE_{\widehat{B}}(h)\right\} + o_p\left\{AMISE_{B_0}(h)\right\}\right. \\
& \left. > \sqrt{\epsilon} + \frac{AMISE_{\widehat{B}}(h) - AMISE_{B_0}(h)}{\sqrt{\epsilon}}\right\} \\
\to\ & P\left\{b_0^2(\widehat{B}) > \epsilon, 0 > \sqrt{\epsilon}\right\} \\
=\ & 0
\end{aligned} \tag{C.34}$$

as $n \to \infty$, where the first step is due to (C.17) and (C.18), and the second step is due to both $O_p\left\{AMISE_{B_0}(h)\right\} \to 0$ and $o_p\left\{AMISE_{B_0}(h)\right\} \to 0$. As a result, by (C.32), we have that as $n \to \infty$,

$$P\left\{b_0^2(\widehat{B}) < \epsilon\right\} \to 1. \tag{C.35}$$

Therefore, by (C.35), we have that as $n \to \infty$,

$$\widehat{B} \xrightarrow{p} B_0.$$

Step 2: The asymptotic performance of $\left(\widehat{d}, \widehat{h}\right)$.

Let $\epsilon = \inf_{B:d<d_0} b_0^2(B)$ in (C.35), we can observe that

$$\begin{aligned}
& P\left\{b_0^2(\widehat{B}) < \inf_{B:d<d_0} b_0^2(B)\right\} \\
\leq\ & P\left\{b_0^2(\widehat{B}) < \inf_{B:d<d_0} b_0^2(B), d < d_0\right\} + P\left\{b_0^2(\widehat{B}) < \inf_{B:d<d_0} b_0^2(B), d \geq d_0\right\} \\
\leq\ & P\left\{d \geq d_0\right\} \\
\leq\ & 1.
\end{aligned}$$

Combining (C.35) gives that as $n \to \infty$,

$$P\left\{d \geq d_0\right\} \to 1. \tag{C.36}$$

Furthermore, define

$$\begin{aligned} W_1 &= \left\{ b_0^2(\widehat{B}) < \frac{\log(n)}{n}, \widehat{d} = d_0, \left| \frac{\widehat{h}}{h_0} - 1 \right| < \eta \right\} \\ W_2 &= \left\{ b_0^2(\widehat{B}) \geq \frac{\log(n)}{n} \right\} \\ W_3 &= \left\{ \widehat{d} < d_0 \right\} \\ W_4 &= \left\{ b_0^2(\widehat{B}) < \frac{\log(n)}{n}, \widehat{d} \geq d_0, \left| \frac{\widehat{h}}{h_0} - 1 \right| \geq \eta \right\} \\ W_5 &= \left\{ b_0^2(\widehat{B}) < \frac{\log(n)}{n}, \widehat{d} > d_0, \left| \frac{\widehat{h}}{h_0} - 1 \right| < \eta \right\} \end{aligned}$$

and

$$W_{CV} = \left\{ DCV\left(\widehat{B}, d_0, h_0\right) + DCV\left(B_0, d_0, h_0\right) > ECV\left(\widehat{B}, d_0, h_0\right) - ECV\left(B_0, d_0, h_0\right) \right\}.$$

Similar to the procedure and the decomposition of (C.30) and (C.31), we have

$$1 \leq P(W_1) + \sum_{k=2}^{5} P\left(W_k \cap W_{CV}\right). \tag{C.37}$$

For $k = 2$, applying (C.35) with $\epsilon = \frac{\log(n)}{n}$ gives that as $n \to \infty$,

$$0 < P\left(W_2 \cap W_{CV}\right) \leq P\left(W_2\right) \to 0. \tag{C.38}$$

For $k = 3$, applying (C.36) gives that as $n \to \infty$,

$$0 < P\left(W_3 \cap W_{CV}\right) \leq P\left(W_3\right) \to 0. \tag{C.39}$$

For $k = 4$, we have that as $n \to \infty$,

$$P\left(W_4 \cap W_{CV}\right) \to 0. \tag{C.40}$$

For $k = 5$, we have that as $n \to \infty$,

$$P\left(W_5 \cap W_{CV}\right) \to 0. \tag{C.41}$$

Therefore, combining (C.38)-(C.41) with (C.37) yields that as $n \to \infty$,

$$P\left\{ b_0^2(\widehat{B}) < \frac{\log(n)}{n}, \widehat{d} = d_0, \left| \frac{\widehat{h}}{h_0} - 1 \right| < \eta \right\} \to 1,$$

and we conclude that $\left(\widehat{d}, \widehat{h}\right)$ are consistent estimators. □

C.4.2 Proof of Theorem 4.3.2

Note that $\widehat{L}$ is the consistent estimator in the sense that $\widehat{L} = L + o_p(1)$, then we have $\widehat{U}_i = U_i + o_p(1)$. Hence, in the remaining proof, we focus on U_i.

By adding and subtracting an additional term $\widehat{F}\left(y, B_0^\top u\right)$, we have

$$\begin{aligned}\widehat{F}\left(y, \widehat{B}^\top u\right) - F(y, B_0^\top u) &= \left\{\widehat{F}\left(y, \widehat{B}^\top u\right) - \widehat{F}\left(y, B_0^\top u\right)\right\} + \left\{\widehat{F}\left(y, B_0^\top u\right) - F(y, B_0^\top u)\right\} \\ &\triangleq A_1 + A_2. \end{aligned} \tag{C.42}$$

For A_1, applying the first order Taylor series expansion at B_0 gives

$$\begin{aligned} A_1 &= \widehat{F}^{(1)}\left(y, B_0^\top u\right)\left(\widehat{B} - B_0\right) \\ &= \left\{\widehat{F}^{(1)}\left(y, B_0^\top u\right) - F^{(1)}\left(y, B_0^\top u\right)\right\}\left(\widehat{B} - B_0\right) + F^{(1)}\left(y, B_0^\top u\right)\left(\widehat{B} - B_0\right) \\ &= o_p(1), \end{aligned} \tag{C.43}$$

where the third equality is due to Lemma C.2.1 and Theorem 4.3.1. On the other hand, applying Lemma C.2.2 with $j = 0$ for A_2 gives

$$\sup_{y,u}\left|\widehat{F}\left(y, B_0^\top u\right) - F(y, B_0^\top u) - \frac{1}{n}\sum_{i=1}^{n}\zeta_{i,B_0}^{(0)}(y, u)\right| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{C.44}$$

Finally, combining (C.43) and (C.44) with (C.42) gives the desired result in Theorem 4.3.2. □

C.4.3 Proof of Theorem 4.3.3

Let $\widehat{B}$ denote the minimizer of $CV(B, d_0, h_0)$ with $d = d_0$ and $h = h_0$ and satisfy $\nabla^1_{\mathrm{vec}(B)} CV(\widehat{B}, d_0, h_{d0}) = 0$. Then by the first order Taylor series expansion, we have

$$\begin{aligned} 0 &= \nabla^1_{\mathrm{vec}(B)} CV(\widehat{B}, d_0, h_0) \\ &= \nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0) + \nabla^2_{\mathrm{vec}(B)} CV(B^*, d_0, h_0)\left\{\mathrm{vec}(\widehat{B}) - \mathrm{vec}(B_0)\right\}, \end{aligned}$$

where B^* is between $\widehat{B}$ and B_0. Equivalently, we have

$$\begin{aligned} &\sqrt{n}\left\{\mathrm{vec}(\widehat{B}) - \mathrm{vec}(B_0)\right\} \\ =\ &\left\{\nabla^2_{\mathrm{vec}(B)} CV(B^*, d_0, h_0)\right\}^{-1}\sqrt{n}\nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0). \end{aligned} \tag{C.45}$$

We first discuss the asymptotic result of $\nabla^2_{\mathrm{vec}(B)}CV(B^*, d_0, h_0)$. By the result in Theorem 4.3.1, we have $B^* \xrightarrow{p} B_0$ as $n \to \infty$. By simple calculations, $\nabla^2_{\mathrm{vec}(B)}CV(B_0, d_0, h_0)$ can be written as

$$\begin{aligned}\nabla^2_{\mathrm{vec}(B)}CV(B_0, d_0, h_0) &= \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\Big\{\nabla^1_{\mathrm{vec}(B)}\widehat{F}^{(-i)}\left(y, B_0^\top U_i\right)\Big\}^{\otimes 2} \\ &\quad - \nabla^2_{\mathrm{vec}(B)}\widehat{F}^{(-i)}\left(y, B_0^\top U_i\right)\Big\{I\left(Y_i \le y\right) - \widehat{F}^{(-i)}\left(y, B_0^\top U_i\right)\Big\}\Big] d\widehat{F}_Y(y).\end{aligned}$$

By Lemma C.2.1 and the Law of Large Numbers, we have that as $n \to \infty$,

$$\nabla^2_{\mathrm{vec}(B)}CV(B_0, d_0, h_0) \xrightarrow{p} \mathcal{A}, \tag{C.46}$$

where

$$\mathcal{A} = 2E\left(\int_0^\tau \Big[\big\{F^{(1)}\left(y, B_0^\top U_i\right)\big\}^{\otimes 2} - F^{(2)}\left(y, B_0^\top U_i\right)\big\{I\left(Y_i \le y\right) - F\left(y, B_0^\top U_i\right)\big\}\Big] dF_Y(y)\right).$$

On the other hand, if L is known, then $\nabla^1_{\mathrm{vec}(B)}CV(B_0, d_0, h_0)$ can be expressed as

$$\begin{aligned}&\nabla^1_{\mathrm{vec}(B)}CV(B_0, d_0, h_0) \\ =\ & \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{I\left(Y_i \le y\right) - \widehat{F}^{(-i)}(y, B_0^\top U_i)\Big\}\Big\{-\nabla^1_{\mathrm{vec}(B)}\widehat{F}^{(-i)}(y, B_0^\top U_i)\Big\} d\widehat{F}_Y(y) \\ =\ & \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\Big\{I\left(Y_i \le y\right) - F(y, B_0^\top u) + F(y, B_0^\top u) - \widehat{F}^{(-i)}(y, B_0^\top U_i)\Big\} \\ & \times \Big\{-\nabla^1_{\mathrm{vec}(B)}\widehat{F}^{(-i)}(y, B_0^\top U_i) + F^{(1)}(y, B_0^\top u) - F^{(1)}(y, B_0^\top u)\Big\}\Big] d\widehat{F}_Y(y) \\ =\ & -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \mathcal{E}_{1;iy}F^{(1)}(y, U_i)d\widehat{F}_Y(y) + \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \mathcal{E}_{3;iy}F^{(1)}(y, U_i)d\widehat{F}_Y(y) \\ & -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \mathcal{E}_{1;iy}\Big\{\nabla^1_{\mathrm{vec}(B)}\widehat{F}^{(-i)}(y, B_0^\top U_i) - F^{(1)}(y, U_i)\Big\} d\widehat{F}_Y(y),\end{aligned}$$

where the second equality comes from adding and subtracting additional terms $F(y, B_0^\top u)$

and $F^{(1)}(y, B_0^\top u)$, and the last step is due to

$$\begin{aligned}
& \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\widehat{F}^{(-i)}(y, B_0^\top U_i) - F(y, B_0^\top u)\right\}\left\{\nabla^1_{\mathrm{vec}(B)}\widehat{F}^{(-i)}(y, B_0^\top U_i) - F^{(1)}(y, U_i)\right\} d\widehat{F}_Y(y) \\
= \; & \frac{1}{n}\nabla^1_{\mathrm{vec}(B)}\sum_{i=1}^{n}\int_0^\tau \left\{\widehat{F}^{(-i)}(y, B_0^\top U_i) - F(y, B_0^\top u)\right\}^2 d\widehat{F}_Y(y) \\
= \; & 0.
\end{aligned}$$

By the result in Lemma C.2.2, $\nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0)$ can be further written as

$$\begin{aligned}
& \nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0) \\
= \; & -\frac{2}{n^2}\sum_{i\neq j}\mathcal{E}_{1;iy}F^{(1)}(Y_j, U_i) + \frac{2}{n^2(n-1)}\sum_{i\neq j\neq k} F^{(1)}(Y_j, U_i)\zeta^{(0)}_{j,B}(Y_k, U_i) \\
& -\frac{2}{n^2(n-1)}\sum_{i\neq j\neq k}\mathcal{E}_{1;iy}\zeta^{(1)}_{j,B}(Y_k, U_i) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
\triangleq \; & T_1 + T_2 + T_3 + o_p\left(\frac{1}{\sqrt{n}}\right). \qquad \text{(C.47)}
\end{aligned}$$

For T_1, applying the convergence property of U-statistic in Hoeffding (1948), we have that as $n \to \infty$,

$$\sqrt{n}T_1 \xrightarrow{d} N\left(0, \mathcal{B}\right), \qquad \text{(C.48)}$$

where $\mathcal{B} = E\left\{U^{\otimes 2}(B_0)\right\}$ with

$$U(B_0) = \int_0^\tau \left\{I\left(Y_i \leq y\right) - F(y, B_0^\top U_i)\right\} F^{(1)}(y, U_i) dF_Y(y). \qquad \text{(C.49)}$$

Since T_2 and T_3 contain $\zeta^{(0)}_{i,B_0}(y, u)$ and $\zeta^{(1)}_{i,B_0}(y, u)$ in (4.24) and (C.8), respectively, and involve $\widetilde{\mathbb{F}}^{(j)}_{i,l,B_0,L}(y, u)$ with $j = 0, 1$. Then by Lemma C.2.1, we can show that

$$\sqrt{n}T_k \xrightarrow{p} 0 \text{ for } k = 2, 3. \qquad \text{(C.50)}$$

Consequently, combining (C.48) and (C.50) with (C.47), we have that as $n \to \infty$,

$$\nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0) \xrightarrow{d} N\left(0, \mathcal{B}\right). \qquad \text{(C.51)}$$

Finally, combining (C.46) and (C.51) with (C.45) gives that as $n \to \infty$,

$$\sqrt{n}\left\{\text{vec}(\widehat{B}) - \text{vec}(B_0)\right\} \xrightarrow{d} N\left(0, \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right).$$

If L is unknown and $\widehat{L}$ is the estimator, then $\nabla^1_{\text{vec}(B)} CV(B_0, d_0, h_0)$ can be written as

$$\sqrt{n}\nabla^1_{\text{vec}(B)} CV(B_0, d_0, h_0) = \sqrt{n}T_1 + \sqrt{n}T_2 + \sqrt{n}T_2 + \sqrt{n}T_4 + o_p(1), \tag{C.52}$$

where $\sqrt{n}T_k$ with $k = 1, 2, 3$ have been derived in (C.48) and (C.50), and T_4 is

$$\begin{aligned} T_4 &= \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \mathcal{E}_{4;iy} F^{(1)}(y, B_0^\top U_i) d\widehat{F}_Y(y) \\ &= \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\widehat{F}^{(-i)}(y, B_0^\top U_i) - \widehat{F}^{(-i)}(y, B_0^\top \widehat{U}_i)\right\} F^{(1)}(y, B_0^\top U_i) d\widehat{F}_Y(y) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\nabla^1_{\text{vec}(B_0)}\widehat{F}^{(-i)}(y, B_0^\top U_i)\right\} F^{(1)}(y, B_0^\top U_i) d\widehat{F}_Y(y)\left(\widehat{L} - L\right). \end{aligned} \tag{C.53}$$

If $\widehat{L}$ is determined by repeated measurements, then

$$\begin{aligned} \widehat{L} - L &= \left(I - \widehat{\Sigma}_\epsilon \widehat{\Sigma}^{-1}_{X^*}\right) - \left(I - \Sigma_\epsilon \Sigma^{-1}_{X^*}\right) \\ &= -\left(\widehat{\Sigma}_\epsilon \widehat{\Sigma}^{-1}_{X^*} - \Sigma_\epsilon \Sigma^{-1}_{X^*}\right) \\ &= -\Sigma^{-1}_{X^*}\left(\widehat{\Sigma}_\epsilon - \Sigma_\epsilon\right) \\ &= -\Sigma^{-1}_{X^*}\frac{1}{2m}\sum_{i=1}^{m}\left\{(x^*_{i1} - x^*_{i2})(x^*_{i1} - x^*_{i2})^\top - 2\Sigma_\epsilon\right\} + o_p(1), \end{aligned}$$

and if $\widehat{L}$ is obtained by validation data, then

$$\begin{aligned} \widehat{L} - L &= \widehat{\text{cov}(X_i, X_i^*)}\widehat{\Sigma}^{-1}_{X^*} - \text{cov}(X_i, X_i^*)\Sigma^{-1}_{X^*} \\ &= -\Sigma^{-1}_{X^*}\left\{\widehat{\text{cov}(X_i, X_i^*)} - \text{cov}(X_i, X_i^*)\right\} \\ &= -\Sigma^{-1}_{X^*}\frac{1}{m}\sum_{i=1}^{m}\left\{(x_i - \mu_X)(x_i^* - \mu_{X^*})^\top - \Sigma_{XX^*}\right\} + o_p(1), \end{aligned}$$

where $\Sigma_{XX^*} = \text{cov}(X_i, X_i^*)$.

Applying the result in Lemma C.2.1 and combining (C.53) with the expression of $\widehat{L} - L$ gives

$$\sqrt{n}T_4 = \frac{\sqrt{n}}{m}\sum_{i=1}^{m}\mathcal{T}(B_0)\Phi_i + o_p(1), \tag{C.54}$$

where

$$\mathcal{T}(B_0) = E\left[\int_0^{\tau}\left\{F^{(1)}(y, B_0^{\top}U_i)\right\}^{\otimes 2} dF_Y(y)\right]\Sigma_{X^*}$$

and

$$\Phi_i = \begin{cases} (x_{i1}^* - x_{i2}^*)(x_{i1}^* - x_{i2}^*)^{\top} - 2\Sigma_{\epsilon}, & \text{based on repeated measurements;} \\ (x_i - \mu_X)(x_i^* - \mu_{X^*})^{\top} - \Sigma_{XX^*}, & \text{based on validation data.} \end{cases}$$

Therefore, combining results (C.48), (C.50), and (C.54) with (C.53) gives that as $n \to \infty$,

$$\nabla^1_{\mathrm{vec}(B)} CV(B_0, d_0, h_0) \xrightarrow{d} N(0, \mathcal{B}_L), \tag{C.55}$$

where $\mathcal{B}_L = E\left[\{U(B_0) + \mathcal{T}(B_0)\Phi_i\}^{\otimes 2}\right]$. As a result, by the Slusky Theorem on (C.46) and (C.55), we have that as $n \to \infty$,

$$\sqrt{n}\left\{\mathrm{vec}(\widehat{B}) - \mathrm{vec}(B_0)\right\} \xrightarrow{d} N\left(0, \mathcal{A}_L^{-1}\mathcal{B}_L\mathcal{A}_L^{-1}\right),$$

where $\mathcal{A}_L = \mathcal{A}$. □

C.5 Proofs of Theorems in Section 4.4

C.5.1 Proof of Theorem 4.4.1

We first consider $\mathrm{dcov}(Y^*, X_k)$ and $\mathrm{dcov}^*(Y^*, X_k^*)$. Note that the former formulation is based on the true covariates X, while the latter formulation is based on the surrogate covariates X^*.

Since the error term ϵ follows normal distribution $N(0, \Sigma_{\epsilon})$, then the characteristic function of kth entry of ϵ is given by

$$E\{\exp(\mathbf{i}s\epsilon_k)\} = \exp\left(-\frac{1}{2}s^2\sigma_{\epsilon,kk}\right) \tag{C.56}$$

for $k = 1, \cdots, p$. By the direct computation, we have

$$\begin{aligned}
\phi_{X_k^*}(s) &= E\left\{\exp\left(\mathbf{i}sX_k^*\right)\right\} \exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right) \\
&= E\left\{\exp\left(\mathbf{i}sX_k\right)\right\} E\left\{\exp\left(\mathbf{i}s\epsilon_k\right)\right\} \exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right) \\
&= E\left\{\exp\left(\mathbf{i}sX\right)\right\}, \qquad \text{(C.57)}
\end{aligned}$$

where the second equality is due to the independence of X and ϵ, and the last equality is due to (C.56).

In addition, we can also derive

$$\begin{aligned}
\phi_{Y^*,X_k^*}(r,s) &= E\left\{\exp\left(\mathbf{i}rY^* + \mathbf{i}sX_k^*\right)\right\} \exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right) \\
&= E\left\{\exp\left(\mathbf{i}rY^* + \mathbf{i}sX_k\right)\right\} E\left\{\exp\left(\mathbf{i}s\epsilon_k\right)\right\} \exp\left(\frac{1}{2}s^2\sigma_{\epsilon,kk}\right) \\
&= E\left\{\exp\left(\mathbf{i}rY^* + \mathbf{i}sX_k\right)\right\}, \qquad \text{(C.58)}
\end{aligned}$$

where the second equality is due to the independence of ϵ and X, Y, and the last equality again comes from (C.56). As a result, combining (C.57) and (C.58) with $\mathrm{dcov}^*\left(Y^*, X_k^*\right)$ gives the same expression of $\mathrm{dcov}\left(Y^*, X_k\right)$.

The equivalence of $\mathrm{dcov}^*\left(X_k^*, X_k^*\right)$ and $\mathrm{dcov}\left(X_k, X_k\right)$ holds by the similar derivations. Therefore, we conclude that $\mathrm{dcorr}\left(Y^*, X_k\right)$ and $\mathrm{dcorr}^*\left(Y^*, X_k^*\right)$ are equivalent in the sense that $\mathrm{dcorr}\left(Y^*, X_k\right) > 0$ if and only if $\mathrm{dcorr}^*\left(Y^*, X_k^*\right) > 0$. Consequently, the same active features can be determined for X^* and X. □

C.5.2 Proof of Theorem 4.4.2

Under Condition (C7) in Appendix C.1 and the similar derivations of Theorem 1 in Li et al. (2012) with replacing X by X^* gives

$$P\left(\max_{1\le k\le p} |\widehat{\omega}_k^* - \omega_k^*| \ge cn^{-\zeta}\right) \le O\left\{\exp\left(-c_1 n^{(1-2c_2)/3}\right)\right\}$$

for some positive constants c_1 and c_2. Let $\mathcal{E} = \left\{ \max_{k \in \mathcal{I}} |\widehat{\omega}_k^* - \omega_k^*| \leq cn^{-\zeta} \right\}$, and we have $\mathcal{E} \subseteq \left\{ \mathcal{I} \subseteq \widehat{\mathcal{I}} \right\}$. As a result, we can obtain

$$\begin{aligned} P\left(\mathcal{I} \subseteq \widehat{\mathcal{I}}\right) &\geq P(\mathcal{E}) \\ &= 1 - P(\mathcal{E}^c) \\ &\geq 1 - |\mathcal{I}| \, P\left(\max_{1 \leq k \leq p} |\widehat{\omega}_k^* - \omega_k^*| \geq cn^{-\zeta} \right) \\ &\geq 1 - |\mathcal{I}| \, O\left\{ \exp\left(-c_1 n^{(1-2c_2)/3} \right) \right\}. \end{aligned} \tag{C.59}$$

By (C.59), when $n \to \infty$, we have

$$P\left(\mathcal{I} \subseteq \widehat{\mathcal{I}}\right) \to 1,$$

which completes the proof. □

Appendix D

Proofs for the Results in Chapter 5

D.1 Regularity Conditions

(C1) Θ is a compact set, and the true parameter value β_0 is an interior point of Θ.

(C2) $\int_0^\tau \lambda_0(t)dt < \infty$, where τ is the finite maximum support of the failure time.

(C3) The $\{N_i(t), Y_i(t), Z_i, X_i\}$ are independent and identically distributed for $i = 1, \cdots, n$.

(C4) The covariates Z_i and X_i are bounded.

(C5) Conditional on $\widetilde{V}_i$, $\left(\widetilde{T}_i, \widetilde{V}_i\right)$ are independent of $\widetilde{A}_i$.

(C6) Censoring time C_i is non-informative. That is, the failure time T_i and the censoring time C_i are independent, given the covariates $\{Z_i, X_i\}$.

(C7) Define

$$\kappa_P = E\left(\int_0^\tau \left[V_i^{*\top}\beta + \frac{1}{2}\beta_x^\top \Sigma_\epsilon \beta_x - \log\left\{E\left(\exp(V_i^{*\top}\beta) I(A_i \le u \le Y_i)\right)\right\}\right] dN_i(u)\right)$$

and assume that β_0 is the unique maximizer of κ_P. Define

$$\begin{aligned}\kappa \;=\; & E\left(\int_0^\tau \left[V_i^{*\top}\beta + \frac{1}{2}\beta_x^\top \Sigma_\epsilon \beta_x - \log\left\{E\left(\exp(V_i^{*\top}\beta) I(A_i \le u \le Y_i)\right)\right\}\right] dN_i(u)\right. \\ & + \left[\log\{dH(A_i)\} - \Lambda_0(A_i)\exp\left(\widehat{V}_i^\top \beta\right)\right. \\ & \left.\left. - \log\left\{\int_0^\tau \exp\left\{-\Lambda_0(u)\exp(\widetilde{X}_{\mathrm{RC},i}^\top \beta_x + Z_i^\top \beta_z)\right\} dH(u)\right\}\right]\right),\end{aligned}$$

and assume that β_0 is the unique maximizer of κ.

(C8) Both $E\left(-\frac{\partial^2 \widehat{\ell}_P^*}{\partial\beta\partial\beta^\top}\right)$ and $E\left(-\frac{\partial^2 \widehat{\ell}_M^*}{\partial\beta\partial\beta^\top}\right)$ are positive definite matrices.

Condition (C1) is a basic condition that is used to derive the maximizer of the target function. (C2) to (C6) are standard conditions for survival analysis, which allow us to obtain the sum of i.i.d. random variables and hence to derive the asymptotic properties of the estimators. Condition (C7) is used to establish the consistency of the estimators $\widehat{\beta}$ and $\widetilde{\beta}$, respectively, given in Theorems 5.2.2 and 5.3.1. The requirement of positive definite matrices in Condition (C8) is standard which ensures asymptotic covariance matrices of $\widehat{\ell}_P^*$ and $\widehat{\ell}_M^*$ meaningful.

D.2 Preliminary Results

In this Appendix, we present the lemmas that are useful for proving the theorems.

Lemma D.2.1 *Let*

$$L_P^* = \prod_{i=1}^{n} \{m(\beta_x)\}^{\delta_i} \left[\frac{\exp\left(v_i^{*\top}\beta\right)}{\sum_{j=1}^{n} \exp\left(v_j^{*\top}\beta\right) I(a_j \le y_i \le y_j)} \right]^{\delta_i}$$

and

$$\begin{aligned} L_R^* &= \prod_{i=1}^{n} \left[\lambda_0(y_i) \sum_{j=1}^{n} \exp\left(v_j^{*\top}\beta\right) \{m(\beta_x)\}^{-1} I(a_j \le y_i \le y_j) \right]^{\delta_i} \\ &\times \exp\left[-\int \lambda_0(u) \sum_{j=1}^{n} \exp\left(v_j^{*\top}\beta\right) \{m(\beta_x)\}^{-1} I(a_j \le u \le y_j) du \right]. \end{aligned}$$

Then

(1) $L_C^* = L_P^* \times L_R^*$;

(2) L_R^* *is ancillary which does not convey the information of* β.

Proof:

Let $L_C^* = \exp(\ell_C^*)$ where ℓ_C^* is given by (5.9), and let

$$\mathcal{U}_i = \left[\sum_{j=1}^{n} \exp\left(v_j^{*\top}\beta\right)\{m(\beta_x)\}^{-1} I(a_j \le y_i \le y_j)\right]^{\delta_i}.$$

Then L_C^* can be written as

$$\begin{aligned} L_C^* &= \prod_{i=1}^{n} \frac{\left\{\lambda_0(y_i)\exp\left(v_i^{*\top}\beta\right)\right\}^{\delta_i} \exp\left[-\Lambda_0(y_i)\exp\left(v_i^{*\top}\beta\right)\{m(\beta_x)\}^{-1}\right]}{\exp\left[-\Lambda_0(a_i)\exp\left(v_i^{*\top}\beta\right)\{m(\beta_x)\}^{-1}\right]} \\ &= \prod_{i=1}^{n} \frac{\left\{\lambda_0(y_i)\exp\left(v_i^{*\top}\beta\right)\right\}^{\delta_i}}{\mathcal{U}_i} \times \prod_{i=1}^{n} \frac{\mathcal{U}_i \exp\left[-\Lambda_0(y_i)\exp\left(v_i^{*\top}\beta\right)\{m(\beta_x)\}^{-1}\right]}{\exp\left[-\Lambda_0(a_i)\exp\left(v_i^{*\top}\beta\right)\{m(\beta_x)\}^{-1}\right]} \\ &= L_P^* \times L_R^*. \end{aligned}$$

Analogous to the derivations of Wang et al. (1993), we can show that L_R^* is ancillary which does not convey the information of β, and hence, it is sufficient to obtain the estimator of β by maximizing L_P^*, or equivalently, deriving an estimator of β from $\log(L_C^*)$ is equivalent to deriving an estimator from using $\log(L_P^*)$ alone. □

Let

$$\widehat{\ell}_P^* = \sum_{i=1}^{n}\left[\delta_i\left\{v_i^{*\top}\beta\right\} + \frac{1}{2}\delta_i\beta_x^\top\Sigma_\epsilon\beta_x - \delta_i\log\left\{\sum_{j=1}^{n}\exp(v_j^{*\top}\beta)I(a_j \le y_i \le y_j)\right\}\right] \quad \text{(D.1)}$$

and $\widehat{L}_P^* = \exp\left(\widehat{\ell}_P^*\right)$. Let $\widehat{L}_C^* = \exp\left(\widehat{\ell}_C^*\right)$ where $\widehat{\ell}_C^*$ is given (5.12). By Lemma D.2.1, inference based on $\widehat{\ell}_C^* = \log\widehat{L}_C^*$ is equivalent to that based on $\widehat{\ell}_P^* = \log(\widehat{L}_P^*)$, given by

$$\widehat{\ell}_P^* = \sum_{i=1}^{n}\int_0^\tau\left[v_i^{*\top}\beta + \frac{1}{2}\beta_x^\top\Sigma_\epsilon\beta_x - \log\left\{\sum_{j=1}^{n}\exp(v_j^{*\top}\beta)I(a_j \le u \le y_j)\right\}\right]dN_i(u);$$

this expression was also derived by Lawless (2003, p.351).

Hence, in the maximization of (5.22), using $\widehat{\ell}_C^* + \widehat{\ell}_M^*$ to perform inference about β is equivalent to using $\widehat{\ell}^* = \widehat{\ell}_P^* + \widehat{\ell}_M^*$. Corresponding to $\widehat{\ell}_P^*$ and $\widehat{\ell}_M^*$, we let

$$\widetilde{\ell}_P^* = \sum_{i=1}^{n}\int_0^\tau\left[v_i^{*\top}\beta + \frac{1}{2}\beta_x^\top\Sigma_\epsilon\beta_x - \log\left\{E\left(\exp(V_i^{*\top}\beta)I(A_i \le u \le Y_i)\right)\right\}\right]dN_i(u)$$

and

$$\begin{aligned}\widetilde{\ell}_M^* &= \sum_{i=1}^{n} \Big[\log\{dH(a_i)\} - \Lambda_0(a_i)\exp\left(\widehat{v}_i^\top\beta\right) \\ &\quad - \log\left\{\int_0^\tau \exp\left\{-\Lambda_o(u)\exp(\widetilde{x}_{\mathrm{RC},i}^\top\beta_x + z_i^\top\beta_z)\right\}dH(u)\right\}\Big],\end{aligned}$$

where $\widetilde{x}_{\mathrm{RC},i}$ is defined in (5.16). Define $\widetilde{\ell}^* = \widetilde{\ell}_P^* + \widetilde{\ell}_M^*$.

Lemma D.2.2 *Under regularity conditions in Appendix D.1,*

$$\sup_{\beta\in\Theta, t\in[0,\tau]} \left|\widehat{\Lambda}_0(t) - \Lambda_0(t)\right| \xrightarrow{a.s.} 0 \text{ as } n\to\infty.$$

Proof:

By (5.14) and the definition of $\Lambda_0(t)$, we need only to show that

$$\sup_{\beta\in\Theta, t\in[0,\tau]} \left| \int_0^t \frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)}{\{m(\beta_x)\}^{-1} S^{(0)}(u,\beta)} - \int_0^t \frac{dP(\Delta_i=1, Y_i\le u)}{\{m(\beta_x)\}^{-1}\mathcal{S}^{(0)}(u,\beta)} \right| \xrightarrow{a.s.} 0 \text{ as } n\to\infty. \quad \text{(D.2)}$$

Since two sets of indicator functions $\{I(A_i \le t \le Y_i) : t\in[0,\tau]\}$ and $\{I(Y_i\le t) : t\in[0,\tau]\}$ are Glivanko-Cantelli classes (van der Vaart and Wellner 1996, Example 2.4.2; van der Vaart 1998, Example 19.6), so are $\{V_i^{*\otimes k}\exp\left(V_i^{*\top}\beta\right)\{m(\beta_x)\}^{-1} I(A_i\le t\le Y_i) : \beta\in\Theta, t\in[0,\tau]\}$ and $\{\Delta_i I(Y_i\le t) : t\in[0,\tau]\}$. Hence, we have that as $n\to\infty$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \Delta_i I(Y_i\le t) &\xrightarrow{a.s.} E\{\Delta_i I(Y_i\le t)\} \\ &= P(\Delta_i=1, Y_i\le t) \quad \text{(D.3)}\end{aligned}$$

and

$$S^{(k)}(u,\beta) \xrightarrow{a.s.} \mathcal{S}^{(k)}(u,\beta) \quad \text{(D.4)}$$

for $k=0,1,2$ and for all $\beta\in\Theta$ and $t\in[0,\tau]$. Combining (D.3) and (D.4) gives that as $n\to\infty$,

$$\frac{\frac{1}{n}\sum_{i=1}^n dN_i(t)}{\{m(\beta_x)\}^{-1} S^{(0)}(t,\beta)} \xrightarrow{a.s.} \frac{dP(\Delta_i=1, Y_i\le t)}{\{m(\beta_x)\}^{-1}\mathcal{S}^{(0)}(t,\beta)}$$

for all $\beta\in\Theta$ and $t\in[0,\tau]$. Therefore, taking integration gives (D.2). □

Lemma D.2.3 *Under regularity conditions in Appendix D.1,*

$$\sup_{\beta\in\Theta,t\in[0,\tau]}\left|\frac{1}{n}\widehat{\ell}^{*}-\frac{1}{n}\widetilde{\ell}^{*}\right|\xrightarrow{a.s.}0 \text{ as } n\to\infty.$$

Proof:

Claim 1: $\sup_{\beta\in\Theta}\left|\frac{1}{n}\widehat{\ell}_P^{*}-\frac{1}{n}\widetilde{\ell}_P^{*}\right|\xrightarrow{a.s.}$ as $n\to\infty$.

Since $\left\{\exp(V_j^{*\top}\beta)I(A_j\le u\le Y_j):\beta\in\Theta,u\in[0,\tau]\right\}$ is a Glivenko-Cantelli class (e.g., van der Vaart and Wellner 1996, Example 2.4.2), and $\log(\cdot)$ is a continuous and monotone function. Hence, we have that as $n\to\infty$,

$$\log\left\{S^{(0)}(u,\beta)\right\}\xrightarrow{a.s.}\log\left\{\mathcal{S}^{(0)}(u,\beta)\right\}$$

uniformly for all $\beta\in\Theta$ and $t\in[0,\tau]$ (Huang et al. 2012). Hence, we conclude that $n\to\infty$,

$$\sup_{\beta\in\Theta}\left|\frac{1}{n}\widehat{\ell}_P^{*}-\frac{1}{n}\widetilde{\ell}_P^{*}\right|\xrightarrow{a.s.}0.$$

Claim 2: $\sup_{\beta\in\Theta}\left|\frac{1}{n}\widehat{\ell}_M^{*}-\frac{1}{n}\widetilde{\ell}_M^{*}\right|\xrightarrow{a.s.}$ as $n\to\infty$.

Since $\widehat{\mu}_{W^*}=\mu_{W^*}+o_p(1)$ and $\widehat{\Sigma}_{W^*}=\Sigma_{W^*}+o_p(1)$, we obtain that $\widehat{X}_i=\widetilde{X}_{\mathrm{RC},i}+o_p(1)$ by the Law of Large Numbers, where $\widetilde{X}_{\mathrm{RC},i}=E\left(X_i|W_i^*\right)$ is defined by (5.16). Then by Lemma D.2.2, we conclude that as $n\to\infty$,

$$\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widehat{X}_i^{\top}\beta_x+Z_i^{\top}\beta_z\right)\right\}\xrightarrow{a.s.}\exp\left\{-\Lambda_0(u)\exp\left(\widetilde{X}_{\mathrm{RC},i}^{\top}\beta_x+Z_i^{\top}\beta_z\right)\right\}. \quad \text{(D.5)}$$

In addition, by the similar result in Lemma 4.2 of Wang (1991), $\widehat{H}(a)$ is strongly consistent estimate of $H(a)$ for each a. Then we have that as $n\to\infty$,

$$\begin{aligned}&\sum_{i=1}^{n}\log\int_0^{\tau}\exp\left\{-\widehat{\Lambda}_0(u)\exp(\widehat{X}_i^{\top}\beta_x+Z_i^{\top}\beta_z)\right\}d\widehat{H}(u)\\ \xrightarrow{a.s.}\ &\sum_{i=1}^{n}\log\int_0^{\tau}\exp\left\{-\Lambda_0(u)\exp(\widetilde{X}_{\mathrm{RC},i}^{\top}\beta_x+Z_i^{\top}\beta_z)\right\}dH(u). \quad \text{(D.6)}\end{aligned}$$

Using the similar derivations in Huang et al. (2012), we have that as $n\to\infty$,

$$\sup_{\beta\in\Theta}n^{-1}\sum_{i=1}^{n}\left\{\widehat{\Lambda}_0(A_i)-\Lambda_0(A_i)\right\}\exp(\widehat{V}_i^{\top}\beta)\xrightarrow{a.s.}0. \quad \text{(D.7)}$$

Hence, combining (D.5), (D.6) and (D.7) gives

$$\sup_{\beta\in\Theta}\left|\frac{1}{n}\widehat{\ell}^*_M - \frac{1}{n}\widetilde{\ell}^*_M\right| \xrightarrow{a.s.} 0 \quad \text{as} \quad n\to\infty.$$

Therefore, combining Claim 1 and Claim 2 yields the result of Lemma D.2.3. □

We also present Theorem 5.7 of van der Vaart (1998) here as the following lemma which will be used in subsequent proof.

Lemma D.2.4 *Let $M_n(\cdot)$ be random functions and let $M(\cdot)$ be a real-valued function of θ. Let θ_0 be the true value of θ. Suppose that for any $\epsilon > 0$,*

$$\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \xrightarrow{p} 0;$$
$$\sup_{\theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0);$$

where $d(\theta,\theta_0) = \|\theta-\theta_0\|$ is the Euclidean distance between θ and θ_0. Then any sequence of estimators $\widehat{\theta}_n$ with $M_n(\widehat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$ converges in probability to θ_0.

D.3 Proofs of the Theorems in Section 5.2

D.3.1 Proof of Theorem 5.2.1

The uniformly consistency of $\widehat{\Lambda}_0(t)$ comes from Lemma D.2.2. □

D.3.2 Proof of Theorem 5.2.2

Proof of Theorem 5.2.2 (1):

By Conditions (C3) and (C4), $\widetilde{\ell}^*_P$ is the sum of i.i.d. random functions. Then by (C7) and the Law of Large Numbers, we have that as $n\to\infty$,

$$\frac{1}{n}\widetilde{\ell}^*_P \xrightarrow{p} \kappa_P$$

for every β. By Claim 1 in Lemma D.2.3, we have that as $n\to\infty$,

$$\sup_{\beta\in\Theta,t\in[0,\tau]}\left|n^{-1}\widehat{\ell}^*_P - \kappa_P\right| \xrightarrow{a.s.} 0.$$

Therefore, by Lemma D.2.4, we have that as $n \to \infty$,

$$\widehat{\beta} \xrightarrow{p} \beta_0. \tag{D.8}$$

Proof of Theorem 5.2.2 (2):

Since $\widehat{\ell}_P^* = \sum_{i=1}^{n} \int_0^\tau \left[v_i^{*\top}\beta + \frac{1}{2}\beta_x^\top \Sigma_\epsilon \beta_x - \log\left\{S^{(0)}(u,\beta)\right\}\right] dN_i(u)$, taking the derivative of $\widehat{\ell}_P^*$ with respect of β gives that

$$U_P(\beta) \triangleq \frac{\partial \widehat{\ell}_P^*}{\partial \beta} = \sum_{i=1}^{n} \int_0^\tau \left\{ v_i^* + \begin{pmatrix} \Sigma_\epsilon \beta_x \\ \mathbf{0}_q \end{pmatrix} - \frac{S^{(1)}(u,\beta)}{S^{(0)}(u,\beta)} \right\} dN_i(u). \tag{D.9}$$

Since $\widehat{\beta}$ is the estimator satisfying $U_P(\widehat{\beta}) = 0$ and $\widehat{\beta}$ is the consistent estimator of β by (D.8), to show the asymptotic distribution of $\widehat{\beta}$, we consider the Taylor series expansion of $U_P(\widehat{\beta})$ around the true parameter β_0:

$$0 = U_P(\widehat{\beta}) = U_P(\beta_0) + \frac{\partial U_P(\beta_0)}{\partial \beta}(\widehat{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{D.10}$$

yielding that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \;\; = \;\; -\left(\frac{1}{n}\frac{\partial U_P(\beta_0)}{\partial \beta}\right)^{-1} \times \frac{1}{\sqrt{n}} U_P(\beta_0) + o_p(1). \tag{D.11}$$

To work out the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta_0\right)$, it suffices to determine the asymptotic behavior of $\frac{\partial U_P(\beta_0)}{\partial \beta}$ and $U_P(\cdot)$. To this end, we proceed with the following two steps.

Step 1: To examine the convergence of $\frac{\partial U_P(\beta_0)}{\partial \beta}$, we first note that $\{N_i(t) : t \in [0,\tau]\}$ is a Glivenko-Cantelli class (van der Vaart and Wellner 1996, Example 2.4.2), which gives that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} dN_i(t) \xrightarrow{a.s.} dE\left\{N_i(t)\right\}$$

uniformly (van der Vaart 1998, Theorem 19.1). Then by (D.9) and the Uniform Law of Large Numbers, we have that as $n \to \infty$,

$$\begin{aligned} & \frac{-1}{n}\frac{\partial U_P(\beta_0)}{\partial \beta} \\ = \quad & \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\left\{\frac{S^{(2)}(u,\beta_0)}{S^{(0)}(u,\beta_0)} - \left(\frac{S^{(1)}(u,\beta_0)}{S^{(0)}(u,\beta_0)}\right)^{\otimes 2}\right\} - \begin{pmatrix} \Sigma_\epsilon & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q} \end{pmatrix}\right] dN_i(u) \\ \xrightarrow{p} \quad & \mathcal{A}_P, \end{aligned} \tag{D.12}$$

where $\mathcal{A}_P$ is given by (5.15).

Step 2: To determine the asymptotic distribution of $U_P(\beta_0)$, we sort out the leading term of $U_P(\beta_0)$ which can be expressed as a sum of i.i.d. random variables. By the similar derivations for Theorem 2.1 of Lin and Wei (1989), we express

$$\frac{1}{\sqrt{n}}U_P(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) + o_p(1), \tag{D.13}$$

where

$$\Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) \tag{D.14}$$

$$\begin{aligned} = \quad & \int_0^\tau \left\{V_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} + \begin{pmatrix} \Sigma_\epsilon\beta_{x0} \\ \mathbf{0}_q \end{pmatrix}\right\} dN_i(u) \\ & - \int_0^\tau \frac{\exp\left(V_i^{*\top}\beta_0\right) I(A_i \le u \le Y_i)}{\mathcal{S}^{(0)}(u,\beta_0)}\left(V_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)}\right) dE\left\{N_i(u)\right\}. \end{aligned} \tag{D.15}$$

Since the $\Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right)$ are i.i.d. random functions, applying the Central Limit Theorem yields that as $n \to \infty$,

$$\frac{1}{\sqrt{n}}U_P(\beta_0) \xrightarrow{d} N\left(0, \mathcal{B}_P\right), \tag{D.16}$$

where $\mathcal{B}_P = E\left(\Phi_i^{\otimes 2}\right)$ and $\Phi_i = \Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right)$.

Finally, combining (D.12) and (D.16) with (D.11) and applying the Slutsky's Theorem, we conclude that as $n \to \infty$,

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N\left(0, \mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1}\right).$$

□

D.4 Proofs of the Theorems in Section 5.3

The derivations in this appendix are in principle analogous to those of Appendix D.3. However, the technical details are a lot more complex than those of Appendix D.3, because no infinite dimensional parameters are involved with the key estimating function $U_P(\beta)$ in Appendix D.3 while such parameters are contained in the estimating function considered here.

D.4.1 Proof of Theorem 5.3.1

Proof of Theorem 5.3.1 (1):
The proof is the same as that of Theorem 5.2.2 (2) except that $\widehat{\ell}_P^*$ and κ_P are replaced by $\widehat{\ell}^*$ and κ, respectively.

Proof of Theorem 5.3.1 (2):
To find the asymptotic distribution of $\widetilde{\beta}$, we note that $\widetilde{\beta}$ solves $U(\beta) = 0$, where

$$U(\beta) = \frac{\partial \widehat{\ell}^*}{\partial \beta} = \frac{\partial \widehat{\ell}_P^*}{\partial \beta} + \frac{\partial \widehat{\ell}_M^*}{\partial \beta}. \tag{D.17}$$

Considering the Taylor series expansion of $U(\widetilde{\beta})$ around β_0 gives that

$$0 = U(\widetilde{\beta}) = U(\beta_0) + \frac{\partial U(\beta_0)}{\partial \beta}(\widetilde{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{D.18}$$

or equivalently,

$$\sqrt{n}(\widetilde{\beta} - \beta_0) = -\left(\frac{1}{n}\frac{\partial U(\beta_0)}{\partial \beta}\right)^{-1}\left(\frac{1}{\sqrt{n}}U(\beta_0)\right) + o_p(1).$$

Analogous to the examination of (D.11) in Appendix D.3.2, we proceed with the following two steps, separately examining $\frac{\partial U(\beta_0)}{\partial \beta}$ and $U(\beta_0)$. By (D.17), we note that the main difficulty here is caused by the involvement of the term $U_M(\beta) = \frac{\partial \widehat{\ell}_M^*}{\partial \beta}$, while $\frac{\partial \widehat{\ell}_P^*}{\partial \beta}$ is examined in Appendix D.3.2.

Step 1: To show the convergence of $\frac{\partial U(\beta_0)}{\partial \beta}$, we first define

$$\widehat{\mu}(\widehat{x}_i, z_i) = \int_0^\tau \exp\left\{-\widehat{\Lambda}_0(u)\exp(\widehat{x}_i^\top \beta_{x0} + z_i^\top \beta_{z0})\right\} d\widehat{H}(u), \tag{D.19}$$

and we let $\widehat{\mu}_i$ denote $\widehat{\mu}(\widehat{x}_i, z_i)$ for ease of notation. By Theorem 5.2.1 and (D.5), we conclude that $\widehat{\mu}_i \xrightarrow{p} \mu_i$ as $n \to \infty$, where

$$\mu_i = \mu(\widetilde{x}_{\mathrm{RC},i}, z_i) = \int_0^\tau \exp\left\{-\Lambda_0(u)\exp(\widetilde{x}_{\mathrm{RC},i}^\top \beta_{x0} + z_i^\top \beta_{z0})\right\} dH(u).$$

Noting that

$$\begin{aligned}\frac{\partial U_M(\beta_0)}{\partial \beta} &= -\sum_{i=1}^n \frac{\partial^2}{\partial\beta\partial\beta^\top}\widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_i^\top\beta_{x0} + z_i^\top\beta_{z0}\right) \\ &\quad -\sum_{i=1}^n \left\{\frac{1}{\widehat{\mu}_i}\frac{\partial^2\widehat{\mu}_i}{\partial\beta\partial\beta^\top} - \frac{1}{\widehat{\mu}_i^2}\left(\frac{\partial\widehat{\mu}_i}{\partial\beta}\right)^{\otimes 2}\right\},\end{aligned}$$

we obtain, by Theorem 5.2.2 and the Law of Large Numbers, that as $n \to \infty$,

$$\frac{-1}{n}\frac{\partial U(\beta_0)}{\partial\beta} \xrightarrow{p} \mathcal{A}, \tag{D.20}$$

where $\mathcal{A} = \mathcal{A}_P + \mathcal{A}_M$, and $\mathcal{A}_P$ and $\mathcal{A}_M$ are given by (5.15) and (5.26), respectively.

Step 2: We now derive the asymptotic distribution of $\frac{1}{\sqrt{n}}U(\beta_0)$. Since a sum of i.i.d. random variables of $\frac{\partial\widehat{\ell}_P^*}{\partial\beta}$ is established in (D.13) of Appendix D.3.2, it remains to examine $U_M(\beta_0) = \frac{\partial\widehat{\ell}_M^*}{\partial\beta}$ by (D.17). To this end, we make an important comment. Different from the partial likelihood score function $U_P(\beta)$ which involves the parameter β only, $U_M(\cdot)$ involves not only the parameter β but also the infinite dimensional parameter $\Lambda_0(\cdot)$. The goal here is to sort out the key term in $U_M(\beta_0)$ which can be expressed as a sum of i.i.d. random functions.

Define

$$\begin{aligned}\widetilde{U}_M(\beta_0) &= -\sum_{i=1}^n \frac{\partial}{\partial\beta}\Lambda_0(a_i)\exp\left(\widehat{v}_i^\top\beta_0\right) - \sum_{i=1}^n \frac{1}{\mu_i}\frac{\partial\mu_i}{\partial\beta} \\ &\triangleq -\sum_{i=1}^n U_{M,i},\end{aligned}$$

and write the difference between $U_M(\beta_0)$ and $\widetilde{U}_M(\beta_0)$ as

$$\frac{1}{\sqrt{n}}\left\{U_M(\beta_0) - \widetilde{U}_M(\beta_0)\right\} = U_1 + U_2, \tag{D.21}$$

where

$$U_1 = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left\{ \widehat{\Lambda}_0(a_i) - \Lambda_0(a_i) \right\} \exp\left(\widehat{v}_i^\top \beta_0 \right) \tag{D.22}$$

and

$$U_2 = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{\widehat{\mu}_i} \frac{\partial \widehat{\mu}_i}{\partial \beta} - \frac{1}{\mu_i} \frac{\partial \mu_i}{\partial \beta} \right\}. \tag{D.23}$$

We first examine U_1. Recall that $\mathcal{N}(t) = P\left(\Delta_i = 1, Y_i \leq t\right)$ and let $d\bar{N}(t) = \frac{1}{n} \sum_{i=1}^{n} dN_i(t)$. Then

$$\begin{aligned}
& U_1 \\
= & -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\partial}{\partial \beta} \left\{ \widehat{\Lambda}_0(a_j) - \Lambda_0(a_j) \right\} \exp\left(\widehat{v}_j^\top \beta_0 \right) \\
= & -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{d\bar{N}(u)}{S^{(0)}(u, \beta_0)} - \frac{d\mathcal{N}(u)}{\mathcal{S}^{(0)}(u, \beta_0)} \right\} m(\beta_{x0}) \exp\left(\widehat{v}_j^\top \beta_0 \right) I(u \leq a_j \leq \tau) \\
= & -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{d\bar{N}(u) - d\mathcal{N}(u)}{\mathcal{S}^{(0)}(u, \beta_0)} + \frac{d\mathcal{N}(u) \mathcal{S}^{(0)}(u, \beta_0) - d\bar{N}(u) S^{(0)}(u, \beta_0)}{\mathcal{S}^{(0)}(u, \beta_0) S^{(0)}(u, \beta_0)} \right\} \\
& \times m(\beta_{x0}) \exp\left(\widehat{v}_j^\top \beta_0 \right) I(u \leq a_j \leq \tau) \\
= & -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left[\frac{d\bar{N}(u) - d\mathcal{N}(u)}{\mathcal{S}^{(0)}(u, \beta_0)} + \frac{d\mathcal{N}(u)}{\left\{ \mathcal{S}^{(0)}(u, \beta_0) \right\}^2} \left\{ \mathcal{S}^{(0)}(u, \beta_0) - S^{(0)}(u, \beta_0) \right\} \right] \\
& \times m(\beta_{x0}) \exp\left(\widehat{v}_j^\top \beta_0 \right) I(u \leq a_j \leq \tau) + o_p(1) \\
= & -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left[\frac{d\bar{N}(u)}{\mathcal{S}^{(0)}(u, \beta_0)} - \frac{d\mathcal{N}(u) S^{(0)}(u, \beta_0)}{\left\{ \mathcal{S}^{(0)}(u, \beta_0) \right\}^2} \right] m(\beta_{x0}) \exp\left(\widehat{v}_j^\top \beta_0 \right) \\
& \times I(u \leq a_j \leq \tau) + o_p(1).
\end{aligned} \tag{D.24}$$

In addition, since $\frac{1}{n} \sum_{j=1}^{n} \exp\left(\widehat{V}_j^\top \beta_0 \right) I(u \leq A_j \leq \tau)$ is an average of i.i.d. random variables due to Conditions (C3), (C4), and (C5), we have that by the Law of Large

Numbers, as $n \to \infty$,

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} \exp\left(\widehat{V}_j^\top \beta_0\right) I(u \le A_j \le \tau) \\ \xrightarrow{p} \; & E\left\{\exp\left(\widehat{V}_j^\top \beta_0\right) I(u \le A_j \le \tau)\right\} \\ = \; & \int_{-\infty}^{\infty}\int_{0}^{\tau} \left\{\exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau)\right\} dG(a, \widehat{v}), \end{aligned}$$

or we write (e.g., Jiang 2010, p.61)

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n} \exp\left(\widehat{V}_j^\top \beta_0\right) I(u \le A_j \le \tau) \; = \; & \int_{-\infty}^{\infty}\int_{0}^{\tau} \left\{\exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau)\right\} dG(a, \widehat{v}) \\ & + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{D.25}$$

Therefore, combining (D.24) and (D.25) gives

$$\begin{aligned} U_1 \; = \; & -\sqrt{n}\frac{\partial}{\partial \beta}\left[\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{d\bar{N}(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u)S^{(0)}(u,\beta_0)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2}\right\} m(\beta_{x0}) \right. \\ & \left. \times \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau) dG(a, \widehat{v})\right] + o_p(1) \\ = \; & -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial \beta}\left[\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u)\exp\left(v_i^{*\top}\beta_0\right) I(a_i \le u \le y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2}\right\}\right. \\ & \left. \times m(\beta_{x0}) \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau)\right] dG(a, \widehat{v}) + o_p(1) \\ \triangleq \; & -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_1\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right). \end{aligned} \tag{D.26}$$

Next, we examine U_2. By analogy with the derivations of (D.25), (D.23) can be re-written as

$$\begin{aligned} U_2 \; = \; & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{1}{\widehat{\mu}_j}\frac{\partial \widehat{\mu}_j}{\partial \beta} - \frac{1}{\mu_j}\frac{\partial \mu_j}{\partial \beta}\right) \\ = \; & \sqrt{n}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{\widehat{\mu}_j}\frac{\partial \widehat{\mu}_j}{\partial \beta} - \frac{1}{\mu_j}\frac{\partial \mu_j}{\partial \beta}\right) \\ = \; & \sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\left(\frac{1}{\widehat{\mu}}\frac{\partial \widehat{\mu}}{\partial \beta} - \frac{1}{\mu}\frac{\partial \mu}{\partial \beta}\right) dG(a, \widehat{v}) + o_p(1), \end{aligned} \tag{D.27}$$

where $\widehat{\mu} = \widehat{\mu}(\widehat{x}, z)$ and $\mu = \mu(\widetilde{x}_{\rm RC}, z)$.

We now express $\sqrt{n}(\widehat{\mu} - \mu)$ as a sum of i.i.d. random functions. Since

$$\begin{aligned}
\sqrt{n}(\widehat{\mu} - \mu) &= \sqrt{n}\left[\int_0^\tau \exp\left\{-\widehat{\Lambda}_0(u)\exp(\widehat{x}^\top\beta_{x0} + z^\top\beta_{z0})\right\}d\widehat{H}(u)\right. \\
&\quad \left. - \int_0^\tau \exp\left\{-\Lambda_0(u)\exp(\widetilde{x}_{\rm RC}^\top\beta_{x0} + z^\top\beta_{z0})\right\}dH(u)\right] \\
&= \sqrt{n}\int_0^\tau \left[\exp\left\{-\widehat{\Lambda}_0(u)\exp(\widetilde{x}_{\rm RC}^\top\beta_{x0} + z^\top\beta_{z0})\right\}\right. \\
&\quad \left. - \exp\left\{-\Lambda_0(u)\exp(\widetilde{x}_{\rm RC}^\top\beta_{x0} + z^\top\beta_{z0})\right\}\right] dH(u) + o_p(1), \quad \text{(D.28)}
\end{aligned}$$

where the second equality is due to $\widehat{X}_j = \widetilde{X}_{{\rm RC},j} + o_p(1)$ and $\widehat{H}(u) = H(u) + o_p(1)$ (e.g., Wang 1991).

Next, we examine the integrand of (D.28) with $\widetilde{x}_{{\rm RC},j}$ and z_j replaced by the corresponding random variables. Applying the Taylor series expansion to

$$\exp\left\{-\widehat{\Lambda}_0(u)\exp(\widetilde{X}_{{\rm RC},j}^\top\beta_{x0} + Z_j^\top\beta_{z0})\right\}$$

with respect to $\Lambda_0(\cdot)$, we obtain that

$$\begin{aligned}
&\exp\left\{-\widehat{\Lambda}_0(u)\exp(\widetilde{X}_{{\rm RC},j}^\top\beta_{x0} + Z_j^\top\beta_{z0})\right\} - \exp\left\{-\Lambda_0(u)\exp(\widetilde{X}_{{\rm RC},j}^\top\beta_{x0} + Z_j^\top\beta_{z0})\right\} \\
= \; & -\exp\left\{-\Lambda_0(u)\exp(\widetilde{X}_{{\rm RC},j}^\top\beta_{x0} + Z_j^\top\beta_{z0})\right\}\left\{\widehat{\Lambda}_0(u) - \Lambda_0(u)\right\} \\
& \times \exp\left(\widetilde{X}_{{\rm RC},j}^\top\beta_{x0} + Z_j^\top\beta_{z0}\right) + o_p\left(\frac{1}{\sqrt{n}}\right). \quad \text{(D.29)}
\end{aligned}$$

By the similar derivation in (D.24), we have

$$\begin{aligned}
&\left\{\widehat{\Lambda}_0(\tau) - \Lambda_0(\tau)\right\} \\
= \; & \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u)\exp\left(w_i^{*\top}\beta_{x0} + z_i^\top\beta_{z0}\right)I(A_i \le u \le Y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2}\right\} m(\beta_{x0}) \\
& + o_p(1). \quad \text{(D.30)}
\end{aligned}$$

Combining (D.29) and (D.30) with (D.28) gives

$$\sqrt{n}\left(\widehat{\mu} - \mu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(\beta_0|\widetilde{x}_{\rm RC}, z) + o_p\left(1\right), \quad \text{(D.31)}$$

where $\psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}}, z)$ is given by (5.24), and $S(\xi|\widetilde{x}_{\mathrm{RC}}, z) = \exp\{-\Lambda_0(\xi)\exp(\widetilde{x}_{\mathrm{RC}}^\top\beta_{x0} + z^\top\beta_{z0})\}$.

Therefore, combining (D.31) and (D.27) yields

$$\begin{aligned}
U_2 &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\frac{1}{\widehat{\mu}_j}\frac{\partial\widehat{\mu}_j}{\partial\beta} - \frac{1}{\mu_j}\frac{\partial\mu_j}{\partial\beta}\right) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left(\frac{1}{\widehat{\mu}}\frac{\partial\widehat{\mu}}{\partial\beta} - \frac{1}{\mu}\frac{\partial\widehat{\mu}}{\partial\beta} + \frac{1}{\mu}\frac{\partial\widehat{\mu}}{\partial\beta} - \frac{1}{\mu}\frac{\partial\mu}{\partial\beta}\right)dG(a,\widehat{v}) + o_p(1) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{\partial\widehat{\mu}}{\partial\beta}\left(\frac{1}{\widehat{\mu}} - \frac{1}{\mu}\right) + \frac{1}{\mu}\left(\frac{\partial\widehat{\mu}}{\partial\beta} - \frac{\partial\mu}{\partial\beta}\right)\right\}dG(a,\widehat{v}) + o_p(1) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu}\left(\frac{\partial\widehat{\mu}}{\partial\beta} - \frac{\partial\mu}{\partial\beta}\right) - \frac{\partial\widehat{\mu}}{\partial\beta}\left(\frac{\widehat{\mu}-\mu}{\widehat{\mu}\mu}\right)\right\}dG(a,\widehat{v}) + o_p(1) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu}\left(\frac{\partial\widehat{\mu}}{\partial\beta} - \frac{\partial\mu}{\partial\beta}\right) - \frac{\partial\mu}{\partial\beta}\left(\frac{\widehat{\mu}-\mu}{\mu^2}\right)\right\}dG(a,\widehat{v}) + o_p(1) \\
&= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu}\frac{\partial}{\partial\beta}(\widehat{\mu}-\mu) - \frac{\partial\mu}{\partial\beta}\frac{1}{\mu^2}(\widehat{\mu}-\mu)\right\}dG(a,\widehat{v}) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu}\frac{\partial}{\partial\beta}\psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}}, z) - \frac{\partial\mu}{\partial\beta}\frac{1}{\mu^2}\psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}}, z)\right\}\right]dG(a,\widehat{v}) \\
&\quad + o_p(1) \\
&\triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_2\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right). \qquad \text{(D.32)}
\end{aligned}$$

Finally, combining (D.13), (D.21), (D.26) and (D.32) gives that

$$\frac{1}{\sqrt{n}}U(\beta_0) = \frac{1}{\sqrt{n}}\left\{U_P(\beta_0) + U_M(\beta_0)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) + o_p(1),$$

where

$$\begin{aligned}
\Psi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) &= \Phi\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) - \Psi_1\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) \\
&\quad + \Psi_2\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, Y_i, A_i\right) - U_{M,i},
\end{aligned}$$

shown as in (5.25).

By the Central Limit Theorem, we have that as $n \to \infty$,

$$\frac{1}{\sqrt{n}} U(\beta_0) \xrightarrow{d} N\left(0, \mathcal{B}\right), \tag{D.33}$$

where $\mathcal{B} = E\left\{\Psi^{\otimes 2}\left(X_i^*, \widetilde{X}_{\mathrm{RC},i}, Z_i, A_i, Y_i\right)\right\}$. Therefore, using (D.20) and (D.33) and applying the Slutsky's Theorem yields that as $n \to \infty$,

$$\sqrt{n}(\widetilde{\beta} - \beta) \xrightarrow{d} N\left(0, \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right).$$

□

D.4.2 Proof of Theorem 5.3.2

By Theorems 5.2.2 and 5.3.1, to prove that $\widetilde{\beta}$ is more efficient than $\widehat{\beta}$, it suffices to show that for any non-zero column vector t,

$$t^\top \left(\mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right) t > 0.$$

Condition (C8) gives that $\mathcal{A}_M = \mathcal{A} - \mathcal{A}_P$ is positive definite, which yields that $\mathcal{A}_P^{-1} - \mathcal{A}^{-1}$ is positive definite.

Let

$$\begin{aligned} a &= \int_0^\tau \left\{ v_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} + \begin{pmatrix} \Sigma_\epsilon \beta_{x0} \\ \mathbf{0}_q \end{pmatrix} \right\} dN_i(u) \\ &\quad - \int_0^\tau \frac{\exp\left(V_i^{*\top}\beta_0\right) I(A_i \le u \le Y_i)}{\mathcal{S}^{(0)}(u,\beta_0)} \left\{ V_i^* - \frac{\mathcal{S}^{(1)}(u,\beta_0)}{\mathcal{S}^{(0)}(u,\beta_0)} \right\} dE\left\{N_i(u)\right\} \end{aligned}$$

and

$$\begin{aligned} b &= \left\{ \int_{-\infty}^{\infty} \int_0^\tau \frac{\partial}{\partial \beta} \left[\frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)} - \frac{d\mathcal{N}(u) \exp\left(V_i^{*\top}\beta_0\right) I(A_i \le u \le Y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2} \right] m(\beta_{x0}) \right. \\ &\quad \left. \times \exp\left(\widehat{v}^\top \beta_0\right) I(u \le a \le \tau) dG(a,\widehat{v}) \right\} \\ &\quad - \left[\int_{-\infty}^{\infty} \int_0^\tau \left\{ \frac{1}{\mu} \frac{\partial}{\partial \beta} \psi_i(\beta_0 | \widetilde{x}_{\mathrm{RC}}, z) - \frac{\partial \widehat{\mu}}{\partial \beta} \frac{1}{\mu^2} \psi_i(\beta_0 | \widetilde{x}_{\mathrm{RC}}, z) \right\} dG(a,\widehat{v}) \right] \\ &\quad + \frac{\partial}{\partial \beta} \Lambda_0(A_i) \exp\left(\widehat{V}_i^\top \beta_0\right) + \frac{1}{\mu_i} \frac{\partial}{\partial \beta} \mu_i. \end{aligned}$$

Define $\mathcal{B}_P = E\left(aa^\top\right)$, $\mathcal{B} = E\left\{(a-b)(a-b)^\top\right\}$ and $\mathcal{B}_M = E\left(ab^\top + ba^\top - bb^\top\right)$. Then it is immediate that $\mathcal{B} = \mathcal{B}_P - \mathcal{B}_M$. Representing the asymptotic covariance matrix related to $U_P(\cdot)$ in (D.9), $\mathcal{B}_P$ is a positive definite matrix. Since $\mathcal{B}$ is the asymptotic covariance matrix related to the function $U(\cdot)$ in (D.17), $\mathcal{B}$ is a positive definite matrix. Hence, for any vector t, $t^\top \mathcal{B} t = t^\top (\mathcal{B}_P - \mathcal{B}_M) t > 0$, or equivalently, $t^\top \mathcal{B}_P t - t^\top \mathcal{B}_M t > 0$.

Finally, for any $t \neq 0$,

$$\begin{aligned}
& t^\top \left\{\mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right\} t \\
= \; & t^\top \left\{\mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1} - \mathcal{A}^{-1}(\mathcal{B}_P - \mathcal{B}_M)\mathcal{A}^{-1}\right\} t \\
= \; & t^\top \left\{\mathcal{A}_P^{-1}\mathcal{B}_P\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\mathcal{B}_P\mathcal{A}^{-1} + \mathcal{A}^{-1}\mathcal{B}_M\mathcal{A}^{-1}\right\} t \\
\geqslant \; & t^\top \left\{\left(\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\right)^\top \mathcal{B}_P \left(\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\right) + \mathcal{A}^{-1}\mathcal{B}_M\mathcal{A}^{-1}\right\} t \\
> \; & 0,
\end{aligned}$$

where the last inequality comes from the fact that $\left(\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\right)^\top \mathcal{B}_P \left(\mathcal{A}_P^{-1} - \mathcal{A}^{-1}\right)$ is a positive definite matrix. Hence, the conclusion follows. □

D.5 Proofs of the Theorems in Section 5.4

The proofs in this appendix are more complicated than those derivations of Appendices D.3 and D.4, because the parameters in the measurement error model have to be estimated from validation data and the induced variability must be incorporated when establishing asymptotic results.

D.5.1 Proof of Theorem 5.4.1

Proof of Theorem 5.4.1 (1):

Since $\widehat{\gamma}$ is a consistent estimator of γ, consistency of $\widehat{\beta}_{val}$ can be established following the proof of Theorem 5.2.2 (1) in Appendix D.3.

Proof of Theorem 5.4.1 (2):

To derive the asymptotic distribution of $\widehat{\beta}_{val}$, we begin with examining the estimator $\widehat{\gamma}$. By the Taylor series expansion of (5.28) with respect to γ, we obtain that

$$\sqrt{n}\,(\widehat{\gamma} - \gamma) = \frac{\sqrt{n}}{m} \sum_{i \in \mathcal{V}} (X_i^* - X_i) + o_p(1). \tag{D.34}$$

Next, we define

$$\widehat{\ell}^*_{P,val} = \sum_{i=1}^{n} \int_0^{\tau} \left[\widetilde{v}_i^{*\top} \beta + \frac{1}{2} \beta_x^{\top} \widehat{\Sigma}_{\epsilon} \beta_x - \log \left\{ \sum_{j=1}^{n} \exp(\widetilde{v}_j^{*\top} \beta) I(a_j \leq u \leq y_j) \right\} \right] dN_i(u) \quad \text{(D.35)}$$

and $\widetilde{v}_i^* = \left((x_i^* - \widehat{\gamma})^{\top}, z_i^{\top} \right)^{\top}$. Similar to Lemma D.2.1, we can show that $\widehat{\ell}^*_{C,val} = \widehat{\ell}^*_{P,val} + \widehat{\ell}^*_{R,val}$ and that $\widehat{\ell}^*_{R,val}$ is ancillary, and thus inference about β based on $\widehat{\ell}^*_{C,val}$ is equivalent to that based on $\widehat{\ell}^*_{P,val}$.

Let $U_{P,val}(\beta) = \frac{\partial \widehat{\ell}^*_{P,val}}{\partial \beta}$. Since $\widehat{\beta}_{val}$ solves $U_{P,val}(\beta) = 0$, then by the Taylor series expansion of $U_{P,val}(\beta)$ around β_0, we have that

$$\sqrt{n} \left(\widehat{\beta}_{val} - \beta_0 \right) = - \left\{ \frac{1}{n} \frac{\partial}{\partial \beta} U_{P,val}(\beta_0) \right\}^{-1} \times \frac{1}{\sqrt{n}} U_{P,val}(\beta_0) + o_p(1). \quad \text{(D.36)}$$

Analogous to the derivation of (D.9) in Appendix D.3.2, we proceed with the following two steps by examining $\frac{\partial}{\partial \beta} U_{P,val}(\beta_0)$ and $U_{P,val}(\beta_0)$, respectively. The main difference here is the involvement of estimators in measurement error model.

Step 1: By the consistency of $\widehat{\gamma}$, we have $\widehat{\gamma} = \gamma + o_p(1)$. By the similar derivations of (D.12), we have that as $n \to \infty$,

$$\frac{-1}{n} \frac{\partial}{\partial \beta} U_{P,val}(\beta_0) \xrightarrow{p} \mathcal{A}_{P,val}, \quad \text{(D.37)}$$

where $\mathcal{A}_{P,val}$ is determined by (5.37).

Step 2: By (D.35) and that $U_{P,val}(\beta) = \frac{\partial \widehat{\ell}^*_{P,val}}{\partial \beta}$, we have that

$$\frac{1}{\sqrt{n}} U_{P,val}(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^{\tau} \left\{ \begin{pmatrix} x_i^* - \widehat{\gamma} \\ z_i \end{pmatrix} + \widehat{\Sigma}_{\epsilon} \beta_{x0} - \frac{\widehat{S}^{(1)}(u; \beta_0)}{\widehat{S}^{(0)}(u; \beta_0)} \right\} dN_i(u), \quad \text{(D.38)}$$

where

$$\widehat{S}^{(k)}(u; \beta_0) = \frac{1}{n} \sum_{i \in \mathcal{M}} \begin{pmatrix} x_i^* - \widehat{\gamma} \\ z_i \end{pmatrix}^{\otimes k} \exp \left\{ \begin{pmatrix} x_i^* - \widehat{\gamma} \\ z_i \end{pmatrix}^{\top} \begin{pmatrix} \beta_{x0} \\ \beta_{z0} \end{pmatrix} I(a_i \leq u \leq y_i) \right\}$$

for $k = 0, 1$.

Since (D.38) involves the estimators $\widehat{\gamma}$ and $\widehat{\Sigma}_\epsilon$, so by adding and subtracting γ and Σ_ϵ, (D.38) can be re-written as

$$\begin{aligned}
&\frac{1}{\sqrt{n}} U_{P,val}\left(\beta_0\right) \\
= &\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau \left[- \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} + \left(\widehat{\Sigma}_\epsilon - \Sigma_\epsilon \right) \beta_{x0} - \left\{ \frac{\widehat{S}^{(1)}\left(u;\beta_0\right)}{\widehat{S}^{(0)}\left(u;\beta_0\right)} - \frac{S^{(1)}\left(u;\beta_0\right)}{S^{(0)}\left(u;\beta_0\right)} \right\} \right] dN_i(u) \\
&+ \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau \left\{ \begin{pmatrix} x_i^* - \gamma \\ z_i \end{pmatrix} + \Sigma_\epsilon \beta_{x0} - \frac{S^{(1)}\left(u;\beta_0\right)}{S^{(0)}\left(u;\beta_0\right)} \right\} dN_i(u) \\
= &\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau \left[- \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} + \left(\widehat{\Sigma}_\epsilon - \Sigma_\epsilon \right) \beta_{x0} - \left\{ \frac{\widehat{S}^{(1)}\left(u;\beta_0\right)}{\widehat{S}^{(0)}\left(u;\beta_0\right)} - \frac{S^{(1)}\left(u;\beta_0\right)}{S^{(0)}\left(u;\beta_0\right)} \right\} \right] dN_i(u) \\
&+ \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \Phi\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right) + o_p(1) \\
= &\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau \left[- \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} + \left(\widehat{\Sigma}_\epsilon - \Sigma_\epsilon \right) \beta_{x0} - \left\{ \frac{\widehat{S}^{(1)}\left(u;\beta_0\right)}{\widehat{S}^{(0)}\left(u;\beta_0\right)} - \frac{S^{(1)}\left(u;\beta_0\right)}{S^{(0)}\left(u;\beta_0\right)} \right\} \right] dN_i(u) \\
&+ \frac{\sqrt{1+\rho}}{\sqrt{m+n}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \zeta_i \Phi\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right) + o_p(1), \qquad \text{(D.39)}
\end{aligned}$$

where $\Phi\left(x_i^*, \widetilde{x}_{\mathrm{RC},i}, z_i, y_i, a_i\right)$ is given by (D.14), and ζ_i is a indicator that $\zeta_i = 1$ if $i \in \mathcal{M}$ and $\zeta_i = 0$ if $i \in \mathcal{V}$.

We now examine the integral in (D.39), which is done by evaluating each term in the integrand separately.

Since

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau - \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} dN_i(u) \\
= &-\sqrt{n} \times \frac{1}{n} \sum_{i \in \mathcal{M}} \int_0^\tau dN_i(u) \times \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} \\
= &-E\left\{N_i(\tau)\right\} \times \sqrt{n} \begin{pmatrix} \widehat{\gamma} - \gamma \\ 0 \end{pmatrix} + o_p(1), \qquad \text{(D.40)}
\end{aligned}$$

by (D.34), we re-write (D.40) as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\int_0^\tau -\begin{pmatrix}\widehat{\gamma}-\gamma\\ 0\end{pmatrix} dN_i(u)\\
=\;& -\frac{\sqrt{n}}{m}\sum_{i\in\mathcal{V}} E\left\{N_i(\tau)\right\}\begin{pmatrix}X_i^*-X_i\\ 0\end{pmatrix}+o_p(1)\\
=\;& -\frac{\sqrt{1+\rho}}{\rho}\frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}(1-\zeta_i)E\left\{N_i(\tau)\right\}\begin{pmatrix}X_i^*-X_i\\ 0\end{pmatrix}+o_p(1). \qquad \text{(D.41)}
\end{aligned}$$

We next derive $\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\int_0^\tau\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}dN_i(u)$ as follows:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\int_0^\tau\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}dN_i(u)\\
=\;&\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}N_i(\tau)\\
=\;&\sqrt{n}\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}\left[\frac{1}{n}\sum_{i\in\mathcal{M}}N_i(\tau)-E\left\{N_i(\tau)\right\}\right]\\
&+\sqrt{n}\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}E\left\{N_i(\tau)\right\}\\
=\;&\sqrt{n}E\left\{N_i(\tau)\right\}\left\{\frac{1}{m-1}\sum_{i\in\mathcal{V}}\left(X_i^*-X_i-\widehat{\gamma}\right)\left(X_i^*-X_i-\widehat{\gamma}\right)^\top-\Sigma_\epsilon\right\}\beta_{x0}+o_p(1)\\
=\;&\sqrt{n}E\left\{N_i(\tau)\right\}\left\{\frac{1}{m-1}\sum_{i\in\mathcal{V}}\left(X_i^*-X_i-\gamma\right)\left(X_i^*-X_i-\gamma\right)^\top-\Sigma_\epsilon\right\}\beta_{x0}+o_p(1)\\
=\;&\frac{E\left\{N_i(\tau)\right\}}{m-1}\sqrt{n}\sum_{i\in\mathcal{V}}\left\{\epsilon_i\epsilon_i^\top-(m-1)\Sigma_\epsilon\right\}\beta_{x0}+o_p(1)\\
=\;&\frac{mE\left\{N_i(\tau)\right\}}{m-1}\frac{\sqrt{1+\rho}}{\rho}\frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}(1-\zeta_i)\left\{\epsilon_i\epsilon_i^\top-(m-1)\Sigma_\epsilon\right\}\beta_{x0}\\
&+o_p(1), \qquad \text{(D.42)}
\end{aligned}$$

where the third equality is due to $\sqrt{n}\left(\widehat{\Sigma}_\epsilon-\Sigma_\epsilon\right)\beta_{x0}\left[\frac{1}{n}\sum_{i\in\mathcal{M}}N_i(\tau)-E\left\{N_i(\tau)\right\}\right]=o_p(1)$, and the fourth equality is due to the consistency of $\widehat{\gamma}$.

Finally,

$$
\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \int_0^\tau \left\{ \frac{\widehat{S}^{(1)}(u;\beta_0)}{\widehat{S}^{(0)}(u;\beta_0)} - \frac{S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} dN_i(u) \\
= & \sqrt{n} \left\{ \frac{\widehat{S}^{(1)}(u;\beta_0)}{\widehat{S}^{(0)}(u;\beta_0)} - \frac{S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} \left[\frac{1}{n} \sum_{i \in \mathcal{M}} N_i(\tau) - E\{N_i(\tau)\} \right] \\
& + \sqrt{n} \left\{ \frac{\widehat{S}^{(1)}(u;\beta_0)}{\widehat{S}^{(0)}(u;\beta_0)} - \frac{S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} E\{N_i(\tau)\} \\
= & \sqrt{n} \left\{ \frac{\widehat{S}^{(1)}(u;\beta_0)}{\widehat{S}^{(0)}(u;\beta_0)} - \frac{S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} E\{N_i(\tau)\} + o_p(1) \\
= & E\{N_i(\tau)\} \sqrt{n} \left\{ \frac{\widehat{S}^{(1)}(u;\beta_0) - S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} + o_p(1) \\
= & \frac{E\{N_i(\tau)\}}{S^{(0)}(u;\beta_0)} \frac{\partial S^{(1)}(u;\beta_0)}{\partial \gamma} \sqrt{n}(\widehat{\gamma} - \gamma) + o_p(1) \\
= & \frac{E\{N_i(\tau)\}}{\mathcal{S}^{(0)}(u;\beta_0)} \frac{\partial \mathcal{S}^{(1)}(u;\beta_0)}{\partial \gamma} \sqrt{n}(\widehat{\gamma} - \gamma) + o_p(1) \\
= & \frac{E\{N_i(\tau)\}}{\mathcal{S}^{(0)}(u;\beta_0)} \frac{\partial \mathcal{S}^{(1)}(u;\beta_0)}{\partial \gamma} \frac{\sqrt{n}}{m} \sum_{i \in \mathcal{V}} (X_i^* - X_i) + o_p(1) \\
= & \frac{E\{N_i(\tau)\}}{\mathcal{S}^{(0)}(u;\beta_0)} \frac{\partial \mathcal{S}^{(1)}(u;\beta_0)}{\partial \gamma} \frac{\sqrt{1+\rho}}{\rho} \frac{1}{\sqrt{n+m}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} (1 - \zeta_i)(X_i^* - X_i) + o_p(1), \quad \text{(D.43)}
\end{aligned}
$$

where the second equality is due to $\sqrt{n}\left\{ \frac{\widehat{S}^{(1)}(u;\beta_0)}{\widehat{S}^{(0)}(u;\beta_0)} - \frac{S^{(1)}(u;\beta_0)}{S^{(0)}(u;\beta_0)} \right\} \left[\frac{1}{n} \sum_{i \in \mathcal{M}} N_i(\tau) - E\{N_i(\tau)\} \right] = o_p(1)$, the third equality is due to the consistency of $\widehat{\gamma}$, and the fourth equality comes from applying the Mean Value Theorem to $S^{(1)}(u;\beta_0)$ with respect to γ.

As a consequence, we combine (D.39), (D.41), (D.42), and (D.43) and obtain that

$$
\frac{1}{\sqrt{n}} U_{P,val}(\beta_0) = \frac{1}{\sqrt{m+n}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \mathcal{B}_{val1,i} + o_p(1), \quad \text{(D.44)}
$$

where $\mathcal{B}_{val1,i}$ is given by (5.35).

By the Central Limit Theorem, we conclude that as $n \to \infty$,

$$
\frac{1}{\sqrt{n}} U_{P,val}(\beta_0) \xrightarrow{d} N(0, \mathcal{B}_{P,val}), \quad \text{(D.45)}
$$

where $\mathcal{B}_{P,val} = E\left\{\left(\mathcal{B}_{val1,i}\right)^{\otimes 2}\right\}$.

Finally, combining (D.36), (D.37) and (D.45) and applying the Slutsky's Theorem, we have that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \mathcal{A}_{P,val}^{-1}\mathcal{B}_{P,val}\mathcal{A}_{P,val}^{-1}\right).$$

□

D.5.2 Proof of Theorem 5.4.2

Proof of Theorem 5.4.2 (1):
This can be done by following the proof of Theorem 5.3.1 in Appendix D.4.
Proof of Theorem 5.4.2 (2) :

Let $U_{P,val}(\beta) = \frac{\partial \widehat{\ell}^*_{P,val}}{\partial \beta}$, $U_{M,val}(\beta) = \frac{\partial \widehat{\ell}^*_{M,val}}{\partial \beta}$, and $U_{val}(\beta) = U_{P,val}(\beta) + U_{M,val}(\beta)$. Since $\widetilde{\beta}_{val}$ solves $U_{val}(\beta) = 0$, then by the Taylor series expansion of $U_{val}(\beta)$ around β_0, we have

$$\sqrt{n}\left(\widetilde{\beta}_{val} - \beta_0\right) = -\left\{\frac{1}{n}\frac{\partial}{\partial \beta}U_{val}\left(\beta_0\right)\right\}^{-1} \times \frac{1}{\sqrt{n}}U_{val}\left(\beta_0\right) + o_p(1). \tag{D.46}$$

Analogous to the proof of Theorem 5.4.1 (2) in Appendix D.5.1, we now examine the asymptotic behaviours of $\frac{\partial}{\partial \beta}U_{val}\left(\beta_0\right)$ and $U_{val}\left(\beta_0\right)$.

First, by the derivations of (D.20) and (D.37), we have that as $n \to \infty$,

$$\frac{-1}{n}\frac{\partial}{\partial \beta}U_{val}\left(\beta_0\right) \xrightarrow{p} \mathcal{A}_{val}, \tag{D.47}$$

where $\mathcal{A}_{val}$ is given by (5.39). So the remaining part is to examine $U_{val}(\beta_0)$. The additional difficulty here is to deal with $U_{M,val}(\beta_0)$ that is involved in $U_{val}(\beta_0)$ since the derivation of $U_{P,val}(\beta)$ is done in Appendix D.5.1.

To study $U_{M,val}(\beta)$, similar to the idea of (D.21), we define

$$\begin{aligned}
\widetilde{U}_{M,val}(\beta_0) &= -\sum_{i\in\mathcal{M}} \frac{\partial}{\partial \beta}\Lambda_0(A_i)\exp\left(\widehat{V}_i^\top \beta_{x0}\right) - \sum_{i\in\mathcal{M}} \frac{1}{\mu_{val,i}}\frac{\partial \mu_{val,i}}{\partial \beta} \\
&= -\sum_{i\in\mathcal{M}} \widetilde{U}_{M,val,i}
\end{aligned} \tag{D.48}$$

and write

$$\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\left\{U_{M,val}(\beta_0)-\widetilde{U}_{M,val}(\beta_0)\right\}=U_{val,1}(\beta_0)+U_{val,2}(\beta_0), \tag{D.49}$$

where

$$\begin{aligned}U_{val,1}(\beta_0) &= \frac{-1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\widehat{\Lambda}_0(A_i)\exp\left\{\widehat{X}_i^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\}\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\Lambda_0(A_i)\exp\left\{\widetilde{X}_{\mathrm{RC},i}^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\},\end{aligned}$$

$$U_{val,2}(\beta_0)=-\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\left\{\frac{1}{\widehat{\mu}_{val,i}}\frac{\partial\widehat{\mu}_{val,i}}{\partial\beta}-\frac{1}{\mu_{val,i}}\frac{\partial\mu_{val,i}}{\partial\beta}\right\},$$

$$\widehat{\mu}_{val,i}=\widehat{\mu}(\widehat{x}_{val,i},z_i)=\int_0^\tau\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widehat{x}_{val,i}^\top\beta_{x0}+z_i^\top\beta_{z0}\right)\right\}d\widehat{H}_{val}(u),$$

and $\mu_{val,i}=\mu(\widetilde{x}_{\mathrm{RC},i},z_i)$.

Now we carry out the following steps to examine each term of (D.49).
Step 1: We first analyze $U_{val,1}(\beta_0)$.

By the similar derivations of (D.21) and (D.22), we express

$$
\begin{aligned}
U_{val,1}(\beta_0) &= \frac{-1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \widehat{\Lambda}_0(A_i) \exp\left\{ \widehat{X}_i^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \Lambda_0(A_i) \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} \\
&= \frac{-1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\bar{N}(u)}{\widehat{S}^{(0)}(u; \beta_0)} m(\beta_{x0}) \exp\left\{ \widehat{X}_i^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\mathcal{N}(u)}{\mathcal{S}^{(0)}(u; \beta_0)} m(\beta_{x0}) \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \\
&= -\left[\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\bar{N}(u)}{\widehat{S}^{(0)}(u; \beta_0)} \exp\left\{ \widehat{X}_i^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \right. \\
&\quad \left. - \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\bar{N}(u)}{S^{(0)}(u; \beta_0)} \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \right] m(\beta_{x0}) \\
&\quad - \left[\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\bar{N}(u)}{S^{(0)}(u; \beta_0)} \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \right. \\
&\quad \left. - \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \frac{\partial}{\partial \beta} \int_0^\tau \frac{d\mathcal{N}(u)}{\mathcal{S}^{(0)}(u; \beta_0)} \exp\left\{ \widetilde{X}_{\mathrm{RC},i}^\top \beta_{x0} + Z_i^\top \beta_{z0} \right\} I(u \le a_i \le \tau) \right] m(\beta_{x0}) \\
&\triangleq T_1 + T_2. \qquad \text{(D.50)}
\end{aligned}
$$

For T_1 in (D.50), we have

$$\begin{aligned}
T_1 &= \frac{-1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\int_0^\tau d\bar{N}(u)\left\{\frac{1}{\widehat{S}^{(0)}(u;\beta_0)}-\frac{1}{S^{(0)}(u;\beta_0)}\right\}m(\beta_{x0}) \\
&\quad\times\exp\left\{\widetilde{X}_{\mathrm{RC},i}^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\}I(u\le A_i\le\tau)+o_p(1) \\
&= \frac{-1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\int_0^\tau d\bar{N}(u)\frac{-1}{\left\{S^{(0)}(u;\beta_0)\right\}^2}\frac{\partial S^{(0)}(u;\beta_0)}{\partial\gamma}\left(\widehat{\gamma}-\gamma\right)m(\beta_{x0}) \\
&\quad\times\exp\left\{\widetilde{X}_{\mathrm{RC},i}^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\}I(u\le A_i\le\tau)+o_p(1) \\
&= \frac{1}{n}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\int_0^\tau d\mathcal{N}(u)\frac{1}{\left\{\mathcal{S}^{(0)}(u;\beta_0)\right\}^2}\frac{\partial\mathcal{S}^{(0)}(u;\beta_0)}{\partial\gamma}m(\beta_{x0})\exp\left\{\widetilde{X}_{\mathrm{RC},i}^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\} \\
&\quad\times I(u\le A_i\le\tau)\sqrt{n}\left(\widehat{\gamma}-\gamma\right)+o_p(1) \\
&= E\left[\frac{\partial}{\partial\beta}\int_0^\tau d\mathcal{N}(u)\frac{1}{\left\{\mathcal{S}^{(0)}(u;\beta_0)\right\}^2}\frac{\partial\mathcal{S}^{(0)}(u;\beta_0)}{\partial\gamma}m(\beta_{x0})\exp\left\{\widetilde{X}_{\mathrm{RC},i}^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\}\right. \\
&\quad\left.\times I(u\le A_i\le\tau)\right]\frac{\sqrt{n}}{m}\sum_{i\in\mathcal{V}}\left(X_i^*-X_i\right)+o_p(1) \\
&\triangleq \mathcal{E}_{val,1}\frac{1}{\sqrt{m+n}}\frac{\sqrt{1+\rho}}{\rho}\sum_{i\in\mathcal{M}\cup\mathcal{V}}(1-\zeta_i)(X_i^*-X_i)+o_p(1), \qquad \text{(D.51)}
\end{aligned}$$

where the first equality comes from using the Mean Value Theorem on $S^{(0)}(u;\beta)$ with respect to γ, the third equality is by (D.34), and the Law of Large Numbers.

T_2 in (D.50) is exactly the form in (D.24). Therefore, we directly have

$$\begin{aligned}
T_2 &= -\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\frac{\partial}{\partial\beta}\left[\int_{-\infty}^{\infty}\int_0^\tau\left\{\frac{dN_i(u)}{\mathcal{S}^{(0)}(u,\beta_0)}-\frac{d\mathcal{N}(u)\exp\left(V_i^{*\top}\beta_0\right)I(A_i\le u\le Y_i)}{\left\{\mathcal{S}^{(0)}(u,\beta_0)\right\}^2}\right\}m(\beta_{x0})\right. \\
&\quad\left.\times\exp\left(\widehat{v}^\top\beta_0\right)I(u\le a\le\tau)\right]dG(a,\widehat{v})+o_p(1) \\
&\triangleq -\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\Psi_{M1}\left(X_i^*,\widetilde{X}_{\mathrm{RC},i},Z_i,Y_i,A_i\right)+o_p(1) \\
&= \frac{-1}{\sqrt{n+m}}\sqrt{1+\rho}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\zeta_i\Psi_{M1}\left(X_i^*,\widetilde{X}_{\mathrm{RC},i},Z_i,Y_i,A_i\right)+o_p(1). \qquad \text{(D.52)}
\end{aligned}$$

Therefore, combining (D.50), (D.51), and (D.52) gives

$$U_{val,1}(\beta_0) = \frac{1}{\sqrt{n+m}} \sum_{i \in \mathcal{M} \cup \mathcal{V}} \mathcal{B}_{val2,i} + o_p(1), \tag{D.53}$$

where $\mathcal{B}_{val2,i}$ is given by (5.36).

Step 2: We next examine $U_{val,2}(\beta_0)$.

Similar to the derivations in (D.27), we have

$$\begin{aligned} U_{val,2}(\beta_0) &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}} \left(\frac{1}{\widehat{\mu}_{val,i}} \frac{\partial \widehat{\mu}_{val,i}}{\partial \beta} - \frac{1}{\mu_{val,i}} \frac{\partial \mu_{val,i}}{\partial \beta} \right) \\ &= \sqrt{n} \int_{-\infty}^{\infty} \int_0^{\tau} \left(\frac{1}{\widehat{\mu}_{val}} \frac{\partial \widehat{\mu}_{val}}{\partial \beta} - \frac{1}{\mu_{val}} \frac{\partial \mu_{val}}{\partial \beta} \right) dG(a, \widehat{v}) + o_p(1). \end{aligned} \tag{D.54}$$

Similar to the derivations of Theorem 5.3.1 (2) in Appendix D.4, we first derive $\sqrt{n}(\widehat{\mu}_{val} - \mu_{val})$, where $\widehat{\mu}_{val} = \widehat{\mu}(\widehat{x}_{val}, z)$ and $\mu_{val} = \mu(\widetilde{x}_{\mathrm{RC}}, z)$.

Note that

$$\begin{aligned} \sqrt{n}(\widehat{\mu}_{val} - \mu_{val}) &= \sqrt{n} \int_0^{\tau} \Big[\exp\left\{ -\widehat{\Lambda}_0(u) \exp(\widetilde{x}_{\mathrm{RC}}^{\top} \beta_{x0} + z^{\top} \beta_{z0}) \right\} \\ &\quad - \exp\left\{ -\Lambda_0(u) \exp(\widetilde{x}_{\mathrm{RC}}^{\top} \beta_{x0} + z^{\top} \beta_{z0}) \right\} \Big] \, dH(u) + o_p(1). \end{aligned} \tag{D.55}$$

On the other hand, the difference $\widehat{\Lambda}_0(\tau) - \Lambda_0(\tau)$ can be expressed as

$$\begin{aligned} & \widehat{\Lambda}_0(\tau) - \Lambda_0(\tau) \\ = & \int_0^{\tau} \left\{ \frac{d\bar{N}(t)}{\widehat{S}^{(0)}(t; \beta_0)} - \frac{d\mathcal{N}(t)}{\mathcal{S}^{(0)}(t; \beta_0)} \right\} m(\beta_{x0}) \\ = & \int_0^{\tau} \left\{ \frac{d\bar{N}(t)}{\widehat{S}^{(0)}(t; \beta_0)} - \frac{d\bar{N}(t)}{S^{(0)}(t; \beta_0)} \right\} m(\beta_{x0}) + \int_0^{\tau} \left\{ \frac{d\bar{N}(t)}{S^{(0)}(t; \beta_0)} - \frac{d\mathcal{N}(t)}{\mathcal{S}^{(0)}(t; \beta_0)} \right\} m(\beta_{x0}) \\ \triangleq & A + B. \end{aligned} \tag{D.56}$$

By the Mean Value Theorem on $S^{(0)}(t; \beta)$ with respect to γ in A, we have

$$\begin{aligned} A &= \int_0^{\tau} \frac{-d\bar{N}(t)}{\{S^{(0)}(t; \beta_0)\}^2} \left\{ \frac{\partial S^{(0)}(t; \beta_0)}{\partial \gamma} \right\} m(\beta_{x0}) \, (\widehat{\gamma} - \gamma) \\ &= \int_0^{\tau} \frac{-d\mathcal{N}(t)}{\{\mathcal{S}^{(0)}(t; \beta_0)\}^2} \left\{ \frac{\partial \mathcal{S}^{(0)}(t; \beta_0)}{\partial \gamma} \right\} m(\beta_{x0}) \, (\widehat{\gamma} - \gamma) + o_p(1) \\ &\triangleq \mathbf{A} \, (\widehat{\gamma} - \gamma) + o_p(1). \end{aligned} \tag{D.57}$$

Moreover, B in (D.56) is equal to (D.30), so we obtain that

$$\begin{aligned} B &= \frac{1}{n}\sum_{i\in\mathcal{M}}\int_0^\tau \left[\frac{dN_i(t)}{\mathcal{S}^{(0)}(t;\beta_0)} - \frac{d\mathcal{N}(t)\exp\left\{(X_i^*-\gamma)^\top\beta_{x0}+Z_i^\top\beta_{z0}\right\}I(A_i\le u\le Y_i)}{\left\{\mathcal{S}^{(0)}(t;\beta_0)\right\}^2}\right] \\ &\quad \times m(\beta_{x0}) \\ &\triangleq \frac{1}{n}\sum_{i\in\mathcal{M}}\mathbf{B}_i + o_p(1). \end{aligned} \tag{D.58}$$

Therefore, combining (D.57) and (D.58) with (D.56) yields

$$\widehat{\Lambda}_0(\tau)-\Lambda_0(\tau) = \left\{\mathbf{A}\left(\widehat{\gamma}-\gamma\right)+\frac{1}{n}\sum_{i\in\mathcal{M}}\mathbf{B}_i\right\}+o_p(1). \tag{D.59}$$

Applying the Taylor series expansion to $\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widetilde{X}_{\mathrm{RC}}^\top\beta_{x0}+Z^\top\beta_{z0}\right)\right\}$ with respect to $\Lambda_0(\cdot)$ gives

$$\begin{aligned} &\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widetilde{X}_{\mathrm{RC}}^\top\beta_{x0}+Z^\top\beta_{z0}\right)\right\}-\exp\left\{-\Lambda_0(u)\exp\left(\widetilde{X}_{\mathrm{RC}}^\top\beta_{x0}+Z^\top\beta_{z0}\right)\right\} \\ =\ & -\exp\left\{-\Lambda_0(u)\exp\left(\widetilde{X}_{\mathrm{RC}}^\top\beta_{x0}+Z^\top\beta_{z0}\right)\right\}\left\{\widehat{\Lambda}_0(u)-\Lambda_0(u)\right\} \\ & \times\exp\left(\widetilde{X}_{\mathrm{RC}}^\top\beta_{x0}+Z^\top\beta_{z0}\right)+o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{D.60}$$

Therefore, combining (D.59) and (D.60) with (D.55) yields

$$\begin{aligned} \sqrt{n}\left(\widehat{\mu}_{val}-\mu_{val}\right) &= \sqrt{n}\int_0^\tau S(\nu|\widetilde{x}_{\mathrm{RC}},z)\left\{\mathbf{A}\left(\widehat{\gamma}-\gamma\right)+\frac{1}{n}\sum_{i\in\mathcal{M}}\mathbf{B}_i\right\}\exp\left(\widetilde{x}_{\mathrm{RC}}^\top\beta_{x0}+z^\top\beta_{z0}\right) \\ &\quad \times dH(\nu)+o_p(1) \\ &= \frac{\sqrt{n}}{m}\sum_{i\in\mathcal{V}}\left(X_i^*-X_i\right)\int_0^\tau S(\nu|\widetilde{x}_{\mathrm{RC}},z)\mathbf{A}\exp\left(\widetilde{x}_{\mathrm{RC}}^\top\beta_{x0}+z^\top\beta_{z0}\right)dH(\nu) \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{M}}\psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}},z)+o_p(1) \\ &= \frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\left[\frac{\sqrt{1+\rho}}{\rho}\left(X_i^*-X_i\right)\int_0^\tau\left\{S(\nu|\widetilde{x}_{\mathrm{RC}},z)\mathbf{A}\right.\right. \\ &\quad \left.\left.\times\exp\left(\widetilde{x}_{\mathrm{RC}}^\top\beta_{x0}+z^\top\beta_{z0}\right)dH(\nu)\right\}+\sqrt{1+\rho}\psi_i(\beta_0|\widetilde{x}_{\mathrm{RC}},z)\right]+o_p(1) \\ &\triangleq \frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\varphi_{val,i}+o_p(1). \end{aligned} \tag{D.61}$$

Similar to the derivations for (D.32), combining (D.54) and (D.61) gives

$$\begin{aligned} U_{val,2}(\beta_0) &= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left(\frac{1}{\widehat{\mu}_{val}}\frac{\partial\widehat{\mu}_{val}}{\partial\beta}-\frac{1}{\mu_{val}}\frac{\partial\widehat{\mu}_{val}}{\partial\beta}+\frac{1}{\mu_{val}}\frac{\partial\widehat{\mu}_{val}}{\partial\beta}-\frac{1}{\mu_{val}}\frac{\partial\mu_{val}}{\partial\beta}\right)dG(a,\widehat{v}) \\ &\quad +o_p(1) \\ &= \sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu_{val}}\frac{\partial}{\partial\beta}\left(\widehat{\mu}_{val}-\mu_{val}\right)-\frac{\partial\mu_{val}}{\partial\beta}\frac{1}{\mu_{val}^2}\left(\widehat{\mu}_{val}-\mu_{val}\right)\right\}dG(a,\widehat{v}) \\ &\quad +o_p(1) \\ &= \frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\left[\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{1}{\mu_{val}}\frac{\partial}{\partial\beta}\varphi_{val,i}-\frac{\partial\mu_{val}}{\partial\beta}\frac{1}{\mu_{val}^2}\varphi_{val,i}\right\}\right]dG(a,\widehat{v}) \\ &\quad +o_p(1) \\ &\triangleq \frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\mathcal{B}_{val3,i}+o_p(1). \end{aligned} \tag{D.62}$$

To summarize, combining (D.44), (D.48), (D.49), (D.53) and (D.62) yields

$$\frac{1}{\sqrt{n}}U_{val}(\beta_0)=\frac{1}{\sqrt{n+m}}\sum_{i\in\mathcal{M}\cup\mathcal{V}}\left(\mathcal{B}_{val1,i}+\mathcal{B}_{val2,i}+\mathcal{B}_{val3,i}+\sqrt{1+\rho}\zeta_i\widetilde{U}_{M,val,i}\right)+o_p(1).$$

Finally, by the Central Limit Theorem, we have that as $n\to\infty$,

$$\frac{1}{\sqrt{n}}U_{val}(\beta_0)\xrightarrow{d}N\left(0,\mathcal{B}_{val}\right), \tag{D.63}$$

where $\mathcal{B}_{val}$ is given by (5.38). Therefore, combining (D.47) and (D.63) with (D.46) and applying the Slutsky's Theorem give that as $n\to\infty$,

$$\sqrt{n}\left(\widetilde{\beta}_{val}-\beta_0\right)\xrightarrow{d}N\left(0,\mathcal{A}_{val}^{-1}\mathcal{B}_{val}\mathcal{A}_{val}^{-1}\right).$$

□

D.5.3 Proof of Theorem 5.4.3

This is done by the similar derivations in Appendix D.4.2. □

D.6 Proofs of the Theorems in Section 5.6

D.6.1 Proof of Theorem 5.6.1

Proof of Theorem 5.6.1 (1):
The proof of this theorem is similar to the proof of Theorem 5.3.1 (1) except for the inference of function $dH(a)$.
Proof of Theorem 5.6.1 (2):
The proof of this theorem can be done by modifying the proof of Theorem 5.3.1 (2) and that of Huang et al. (2012) who developed asymptotic normality under the length-biased sampling setting. □

Appendix E

Proofs for the Results in Chapter 6

E.1 Regularity Conditions

(C1) Θ is a compact set, and the true parameter value β_0 is an interior point of Θ.

(C2) $\int_0^\tau \lambda_0(t)dt < \infty$, where τ is the finite maximum support of the failure time.

(C3) The $\{N_i(t), Y_i(t), Z_i, X_i^*\}$ are independent and identically distributed for $i = 1, \cdots, n$.

(C4) The covariates Z and X^* are bounded.

(C5) Conditional on $\widetilde{V}$, $\left(\widetilde{T}, C, \widetilde{V}\right)$ are independent of $\widetilde{A}$.

(C6) Censoring time is non-informative. That is, the failure time and the censoring time are independent, given the covariate.

Condition (C1) is the basic condition that is used to the derivation of the maximizer from the target function. (C2) to (C6) are standard conditions for survival analysis, which allows us to obtain the sum of i.i.d. random variables and hence to derive the asymptotic properties of the estimators.

E.2 Proofs for the Results in Section 6.3

E.2.1 Proof of Lemma 6.3.1

For any given candidate model S, we have that

$$\begin{aligned}
\Sigma_{X_S^*} &= E\left\{\left(X_S^* - \mu_{X_S^*}\right)\left(X_S^* - \mu_{X_S^*}\right)^\top\right\} \\
&= E\left\{\left(\pi_S X^* - \pi_S \mu_{X^*}\right)\left(\pi_S X^* - \pi_S \mu_{X^*}\right)^\top\right\} \\
&= \pi_S E\left\{\left(X^* - \mu_{X^*}\right)\left(X^* - \mu_{X^*}\right)^\top\right\}\pi_S^\top \\
&= \pi_S \Sigma_{X^*} \pi_S^\top,
\end{aligned}$$

which yields that $I_{|S|\times|S|} = \Sigma_{X_S^*} \cdot \Sigma_{X_S^*}^{-1} = \pi_S \Sigma_{X^*} \pi_S^\top \cdot \Sigma_{X_S^*}^{-1}$, provided $\Sigma_{X_S^*}^{-1}$ exists. Multiplying π_S on both sides gives

$$\pi_S \Sigma_{X^*} \pi_S^\top \cdot \Sigma_{X_S^*}^{-1} \pi_S = \pi_S$$

or $\pi_S \left(\Sigma_{X^*} \pi_S^\top \cdot \Sigma_{X_S^*}^{-1} \pi_S - I_{|S|\times|S|}\right) = 0$, which implies that

$$\Sigma_{X^*} \pi_S^\top \cdot \Sigma_{X_S^*}^{-1} \pi_S = I_{|S|\times|S|},$$

or equivalently,

$$\pi_S^\top \cdot \Sigma_{X_S^*}^{-1} \pi_S = \Sigma_{X^*}^{-1},$$

and this proof is completed. □

E.2.2 Proof of Lemma 6.3.2

Proof of (a):
First, for any candidate model S, we denote

$$G_S^{(1)}(u, \beta_x, \beta_z) = \frac{1}{n}\sum_{i=1}^{n} \Pi_S \begin{pmatrix} x_i^* \\ z_i \end{pmatrix} Y_i(u) \exp\left\{\left((\pi_S x_i^*)^\top, z_i^\top\right)\begin{pmatrix} \beta_x \\ \beta_z \end{pmatrix}\right\}. \tag{E.1}$$

Let

$$G^{(1)}\left(u,\beta_x,\beta_z\right)=\left(\begin{array}{c} G_x^{(1)}(u,\beta_x,\beta_z) \\ G_z^{(1)}(u,\beta_x,\beta_z) \end{array}\right), \tag{E.2}$$

where

$$\left(\begin{array}{c} G_x^{(1)}(u,\beta_x,\beta_z) \\ G_z^{(1)}(u,\beta_x,\beta_z) \end{array}\right) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right) Y_i(u)\exp\left\{\left(x_i^{*\top},z_i^{\top}\right)\left(\begin{array}{c} \beta_x \\ \beta_z \end{array}\right)\right\},$$

and

$$G^{(2)}\left(u,\beta_x,\beta_z\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right)^{\otimes 2} Y_i(u)\exp\left\{\left(x_i^{*\top},z_i^{\top}\right)\left(\begin{array}{c} \beta_x \\ \beta_z \end{array}\right)\right\}, \tag{E.3}$$

where $a^{\otimes 2}=aa^{\top}$ for any vector a.

Then setting $(\beta_x,\beta_z)=(0,\beta_z)$ gives

$$\begin{aligned} G_S^{(1)}\left(u,0,\beta_z\right) &= \frac{1}{n}\sum_{i=1}^{n}\Pi_S\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right) Y_i(u)\exp\left\{\left((\pi_S x_i^*)^{\top},z_i^{\top}\right)\left(\begin{array}{c} 0 \\ \beta_z \end{array}\right)\right\} \\ &= \frac{1}{n}\Pi_S\sum_{i=1}^{n}\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right) Y_i(u)\exp\left\{\left(x_i^{*\top},z_i^{\top}\right)\left(\begin{array}{c} 0 \\ \beta_z \end{array}\right)\right\} \\ &= \Pi_S G^{(1)}\left(u,0,\beta_z\right). \end{aligned}$$

Similarly, from (6.10) and (6.20), one has

$$\begin{aligned} G_S^{(0)}\left(u,0,\beta_z\right) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(u)\exp\left\{(\pi_S x_i^*)^{\top}0+z_i^{\top}\beta_z\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(u)\exp\left(z_i^{\top}\beta_z\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(u)\exp\left(x_i^{*\top}0+z_i^{\top}\beta_z\right) \\ &= G^{(0)}\left(u,0,\beta_z\right). \end{aligned}$$

Therefore, for any β_z and $j=0,1$, we have

$$G_S^{(j)}\left(u,0,\beta_z\right)=\Pi_S^{\otimes j}G^{(j)}\left(u,0,\beta_z\right), \tag{E.4}$$

where $A^{\otimes 0} = I_{p\times p}$ and $A^{\otimes 1} = A$ for any matrix A.

Consequently, direct calculations show that

$$\begin{aligned} U_{P,S}\left(\beta_{x,S},\beta_{z,S}\right) &= \frac{\partial}{\partial\beta_S}\ell^*_{P,S}\left(\beta_S\right) \\ &= \sum_{i=1}^{n}\int_0^\tau\left\{\Pi_S\begin{pmatrix} x_i^* \\ z_i\end{pmatrix} + \begin{pmatrix} \pi_S\Sigma_\epsilon\pi_S^\top\beta_{x,S} \\ 0\end{pmatrix} - \frac{G_S^{(1)}\left(u,\beta_{x,S},\beta_{z,S}\right)}{G_S^{(0)}\left(u,\beta_{x,S},\beta_{z,S}\right)}\right\}dN_i\left(u\right), \end{aligned}$$

and

$$\begin{aligned} U_P\left(\beta_x,\beta_z\right) &= \frac{\partial}{\partial\beta}\ell^*_P\left(\beta\right) \\ &= \sum_{i=1}^{n}\int_0^\tau\left\{\begin{pmatrix} x_i^* \\ z_i\end{pmatrix} + \begin{pmatrix} \Sigma_\epsilon\beta_x \\ 0\end{pmatrix} - \frac{G^{(1)}\left(u,\beta_x,\beta_z\right)}{G^{(0)}\left(u,\beta_x,\beta_z\right)}\right\}dN_i\left(u\right). \end{aligned} \quad \text{(E.5)}$$

Thus, plugging in $\left(\beta_x^\top,\beta_z^\top\right)^\top = \left(0^\top,\beta_{z0}^\top\right)^\top$ and $\left(\beta_{x,S}^\top,\beta_{z,S}^\top\right)^\top = \left(0^\top,\beta_{z0}^\top\right)^\top$ to $U_{P,S}\left(\beta_{x,S},\beta_{z,S}\right)$ and $U_P\left(\beta_x,\beta_z\right)$, respectively, gives

$$U_{P,S}\left(0,\beta_{z0}\right) = \Pi_S U_P\left(0,\beta_{z0}\right).$$

Proof of (b):
We first show the relationship between $\widehat{x}_{i,S}$ and $\widehat{x}_i$, the quantities defined by (6.14) and (6.21) in Section 6.2.2. Applying $X_S^* = \pi_S X^*$ to (6.14), we have

$$\begin{aligned} \widehat{x}_{i,S} &= \pi_S\widehat{\mu}_{X^*} + \left(I_{|S|\times|S|} - \pi_S\Sigma_\epsilon\pi_S^\top\widehat{\Sigma}_{X_S^*}^{-1}\right)\left(x_{i,S}^* - \widehat{\mu}_{X_S^*}\right) \\ &= \pi_S\widehat{\mu}_{X^*} + \left(I_{|S|\times|S|} - \pi_S\Sigma_\epsilon\pi_S^\top\widehat{\Sigma}_{X_S^*}^{-1}\right)\pi_S\left(x_i^* - \widehat{\mu}_{X^*}\right) \\ &= \pi_S\widehat{\mu}_{X^*} + \left(\pi_S - \pi_S\Sigma_\epsilon\pi_S^\top\widehat{\Sigma}_{X_S^*}^{-1}\pi_S\right)\left(x_i^* - \widehat{\mu}_{X^*}\right) \\ &= \pi_S\left\{\widehat{\mu}_{X^*} + \left(I_{p\times p} - \Sigma_\epsilon\pi_S^\top\widehat{\Sigma}_{X_S^*}^{-1}\pi_S\right)\left(x_i^* - \widehat{\mu}_{X^*}\right)\right\} \\ &= \pi_S\left\{\widehat{\mu}_{X^*} + \left(I_{p\times p} - \Sigma_\epsilon\widehat{\Sigma}_{X^*}^{-1}\right)\left(x_i^* - \widehat{\mu}_{X^*}\right)\right\} \\ &= \pi_S\left\{\widehat{\mu}_{X^*} + \left(\widehat{\Sigma}_{X^*} - \Sigma_\epsilon\right)^\top\widehat{\Sigma}_{X^*}^{-1}\left(x_i^* - \widehat{\mu}_{X^*}\right)\right\} \\ &= \pi_S\widehat{x}_i, \end{aligned}$$

where the second identity is due to $\widehat{\mu}_{X_S^*} = \pi_S \widehat{\mu}_{X^*}$, and the third last step is due to Lemma 6.3.1.

To prove $U_{M,S}(0, \beta_{z0}) = \Pi_S U_M(0, \beta_{z0})$, we first examine the partial derivative of $\widehat{\ell}^*_{M,S}$. Note that we can express $\widehat{\ell}^*_{M,S} = \widehat{\ell}^*_{M1,S} - \widehat{\ell}^*_{M2,S}$, where

$$\widehat{\ell}^*_{M1,S} = \sum_{i=1}^{n} \left[\log \left\{ d\widehat{H}_S(a_i) \right\} - \widehat{\Lambda}_{0,S}(a_i) \exp\left(\widehat{x}_{i,S}^\top \beta_x + z_i^\top \beta_z \right) \right],$$

and

$$\widehat{\ell}^*_{M2,S} = \sum_{i=1}^{n} \log \int_0^\tau \exp\left\{ -\widehat{\Lambda}_{0,S}(u) \exp\left(\widehat{x}_{i,S}^\top \beta_x + z_i{}^\top \beta_z \right) \right\} d\widehat{H}_S(u).$$

Let $\mathbf{A} = \begin{pmatrix} \pi_S \Sigma_\epsilon \pi_S^\top & \mathbf{0}_{|S| \times q} \\ \mathbf{0}_{q \times |S|} & \mathbf{0}_{q \times q} \end{pmatrix}$ and $\beta_S = \begin{pmatrix} \beta_{x,S} \\ \beta_{z,S} \end{pmatrix}$. Then direct calculations give us

$$\begin{aligned}
& U_{M1,S}(\beta_{x,S}, \beta_{z,S}) \\
= \; & \frac{\partial}{\partial \beta_S} \widehat{\ell}^*_{M1,S}(\beta_S) \\
= \; & \frac{\partial}{\partial \beta_S} \left(\sum_{i=1}^{n} \left[\log \left\{ d\widehat{H}_S(a_i) \right\} - \widehat{\Lambda}_{0,S}(a_i) \exp\left(\widehat{x}_{i,S}^\top \beta_{x,S} + z_i^\top \beta_{z,S} \right) \right] \right) \\
= \; & -\sum_{i=1}^{n} \frac{\partial}{\partial \beta_S} \left\{ \widehat{\Lambda}_{0,S}(a_i) \exp\left(\widehat{x}_{i,S}^\top \beta_{x,S} + z_i^\top \beta_{z,S} \right) \right\} \\
= \; & -\sum_{i=1}^{n} \frac{\partial}{\partial \beta_S} \left\{ \int_0^{a_i} \frac{\frac{1}{n} \sum_{j=1}^{n} dN_j(u)}{m_S^{-1}(\beta_{x,S}) G_S^{(0)}(u, \beta_{x,S}, \beta_{z,S})} \exp\left(\widehat{x}_{i,S}^\top \beta_{x,S} + z_i^\top \beta_{z,S} \right) \right\}
\end{aligned}$$

where the fourth equality is due to the estimator (6.9). Note that

$$\begin{aligned}
\frac{\partial m_S^{-1}(\beta_{x,S})}{\partial \beta_S} &= \frac{\partial}{\partial \beta_S} \exp\left(\frac{-1}{2} \beta_{x,S}^\top \pi_S \Sigma_\epsilon \pi_S^\top \beta_{x,S} \right) \\
&= \frac{\partial}{\partial \beta_S} \exp\left(\frac{-1}{2} \beta_S^\top \mathbf{A} \beta_S \right) \\
&= -\mathbf{A} \beta_S m_S^{-1}(\beta_{x,S}),
\end{aligned}$$

where $m_S(\beta_{x,S})$ is defined by $\exp\left(\frac{1}{2} \beta_{x,S}^\top \pi_S \Sigma_\epsilon \pi_S^\top \beta_{x,S} \right)$ in Section 6.2.1.

Then plugging in $\beta_{x,S}=0$ and $\beta_{z,S}=\beta_{z0}$ to $U_{M1,S}\left(\beta_{x,S},\beta_{z,S}\right)$ gives

$$\begin{aligned} U_{M1,S}\left(0,\beta_{z0}\right) &= -\sum_{i=1}^{n}\left[\int_0^{a_i}\frac{\frac{1}{n}\sum_{j=1}^{n}dN_j\left(u\right)\Pi_S G^{(1)}\left(u,0,\beta_{z0}\right)}{\left\{G^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\exp\left\{z_i^\top\beta_{z0}\right\}\right. \\ &\quad \left.+\widehat{\Lambda}_0\left(a_i\right)\Pi_S\begin{pmatrix}\widehat{x}_i\\ z_i\end{pmatrix}\exp\left\{z_i^\top\beta_{z0}\right\}\right] \\ &= \Pi_S U_{M1}\left(u,0,\beta_{z0}\right), \end{aligned}$$

where the last step is due to (E.4) and that $m_S^{-1}(0)=1$.

Similarly, we examine $U_{M2,S}\left(\beta_{x,S},\beta_{z,S}\right)=\frac{\partial}{\partial\beta}\widehat{\ell}^*_{M2,S}$ and plug in $\beta_{x,S}=0$ and $\beta_{z,S}=\beta_{z0}$ to $U_{M2,S}$ and apply Lemma 6.3.1, yielding

$$\begin{aligned} U_{M2,S}\left(0,\beta_{z0}\right) &= \frac{1}{\int_0^\tau\exp\left\{-\widehat{\Lambda}_0\left(u\right)\exp\left(z_i^\top\beta_{z0}\right)\right\}d\widehat{H}_S(u)} \\ &\quad\times\int_0^\tau\exp\left\{-\widehat{\Lambda}_0\left(u\right)\exp\left(z_i^\top\beta_{z0}\right)\right\}\left\{\Pi_S\left(\frac{\partial}{\partial\beta}\widehat{\Lambda}_0\left(u\right)\right)\exp\left(z_i^\top\beta_{z0}\right)\right. \\ &\quad\left.+\widehat{\Lambda}_0\left(u\right)\Pi_S\begin{pmatrix}\widehat{x}_i\\ z_i\end{pmatrix}\exp\left(z_i^\top\beta_{z0}\right)\right\}d\widehat{H}_S(u) \\ &= \Pi_S\frac{\partial}{\partial\beta}\log\int_0^\tau\exp\left\{-\widehat{\Lambda}_0\left(u\right)\exp\left(\widehat{x}_i^\top\beta_x+z_i^\top\beta_z\right)\right\}d\widehat{H}_S(u)\Big|_{\beta_x=0,\beta_z=\beta_{z0}} \\ &= \Pi_S U_{M2}\left(u,0,\beta_{z0}\right). \end{aligned}$$

Thus, we complete the proof. □

E.2.3 Proof of Theorem 6.3.1

Proof of (a): The proof consists of the following two steps.
Step 1:

Let $U_P(\beta_x,\beta_z)=\frac{\partial\ell^*_P}{\partial\beta}$, $U_M(\beta_x,\beta_z)=\frac{\partial\widehat{\ell}^*_M}{\partial\beta}$ and $U(\beta_x,\beta_z)=U_P(\beta_x,\beta_z)+U_M(\beta_x,\beta_z)$, where $\beta=\left(\beta_x^\top,\beta_z^\top\right)^\top$, and ℓ^*_P and $\widehat{\ell}^*_M$ are given by (6.23) and (6.18), respectively. Applying the Taylor expansion of $U\left(\widehat{\beta}_x,\widehat{\beta}_z\right)$ and $U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)$ around $\left(\beta_x^\top,\beta_z^\top\right)^\top=\left(0,\beta_{z0}^\top\right)^\top$,

respectively, gives

$$0 = U\left(\widehat{\beta}_x, \widehat{\beta}_z\right) = U\left(0, \beta_{z0}\right) + \frac{\partial U\left(0, \beta_{z0}\right)}{\partial \beta^\top} \begin{pmatrix} \widehat{\beta}_x - 0 \\ \widehat{\beta}_z - \beta_{z0} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right) \tag{E.6}$$

and

$$U\left(\frac{\eta}{\sqrt{n}}, \beta_{z0}\right) = U\left(0, \beta_{z0}\right) + \frac{\partial U\left(0, \beta_{z0}\right)}{\partial \beta^\top} \begin{pmatrix} \frac{\eta}{\sqrt{n}} \\ 0 \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{E.7}$$

Combining (E.6) and (E.7) gives

$$0 = U\left(\frac{\eta}{\sqrt{n}}, \beta_{z0}\right) + \frac{\partial U\left(0, \beta_{z0}\right)}{\partial \beta^\top} \begin{pmatrix} \widehat{\beta}_x \\ \widehat{\beta}_z - \beta_{z0} \end{pmatrix} - \frac{\partial U\left(0, \beta_{z0}\right)}{\partial \beta^\top} \begin{pmatrix} \frac{\eta}{\sqrt{n}} \\ 0 \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{E.8}$$

and re-scaling (E.8) yields

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_x \\ \widehat{\beta}_z - \beta_{z0} \end{pmatrix} = \left\{ \frac{-1}{n} \frac{\partial U\left(0, \beta_{z0}\right)}{\partial \beta^\top} \right\}^{-1} \frac{1}{\sqrt{n}} U\left(\frac{\eta}{\sqrt{n}}, \beta_{z0}\right) + \begin{pmatrix} \eta \\ 0 \end{pmatrix} + o_p\left(1\right). \tag{E.9}$$

Let

$$\widehat{\zeta}_i^*\left(\beta_x, \beta_z\right) = \int_0^\tau \exp\left\{ -\widehat{\Lambda}_0\left(u\right) \exp\left(\widehat{x}_i^\top \beta_x + z_i^\top \beta_z\right) \right\} d\widehat{H}\left(u\right). \tag{E.10}$$

Since $\widehat{\mu}_{X^*} = \mu_X + o_p(1)$ and $\widehat{\Sigma}_{X^*} = \Sigma_{X^*} + o_p(1)$, we obtain that $\widehat{X}_i = \widetilde{X}_{\mathrm{RC},i} + o_p(1)$ by the Law of Large Numbers, where

$$\begin{aligned} \widetilde{X}_{\mathrm{RC},i} &= E\left(X_i | X_i^*\right) \\ &= \mu_X + \left(\Sigma_{X^*} - \Sigma_\epsilon\right)^\top \Sigma_{X^*}^{-1}\left(X_i^* - \mu_{X^*}\right). \end{aligned}$$

Since the indicator functions

$$\{I(A \le t \le Y) : t \in [0, \tau]\} \quad \text{and} \quad \{I(Y \le t) : t \in [0, \tau]\}$$

are Glivanko-Cantelli classes (van der Vaart and Wellner 1996, Example 2.4.2), by Uniformly Strong Law of Large Numbers, we have that as $n \to \infty$,

$$G^{(k)}(u, \beta_x, \beta_z) \xrightarrow{a.s.} \mathcal{G}^{(k)}(u, \beta_x, \beta_z)$$

uniformly at u, where

$$\mathcal{G}^{(k)}(u,\beta_x,\beta_z)=E\left\{\begin{pmatrix} X^* \\ Z \end{pmatrix}^{\otimes k}\exp(X^{*\top}\beta_x+Z^\top\beta_z)I(A\le u\le Y)\right\} \tag{E.11}$$

for $k=0,1,2$. By the similar proof of Theorem 5.2.1, we have that as $n\to\infty$,

$$\sup_{\beta\in\Theta,\ t\in[0,\tau]}|\widehat{\Lambda}_0(t)-\Lambda_0^*(t)|\xrightarrow{a.s.}0, \tag{E.12}$$

where

$$\Lambda_0^*(t)=\int_0^t\frac{dP\left(\Delta=1,Y\le u\right)}{m^{-1}\left(\beta_{x0}\right)\mathcal{G}^{(0)}(u,\beta_{x0},\beta_{z0})}. \tag{E.13}$$

In addition, by the similar derivations of Lemma 4.2 in Wang (1991), we have that as $n\to\infty$,

$$\widehat{H}(u)\xrightarrow{a.s.}H(u) \tag{E.14}$$

uniformly. Combining (E.12) and (E.14) yields that as $n\to\infty$,

$$\widehat{\zeta}_i^*\left(\beta_x,\beta_z\right)\xrightarrow{a.s.}\zeta_i^*\left(\beta_x,\beta_z\right),$$

where

$$\zeta_i^*\left(\beta_x,\beta_z\right)=\int_0^\tau\exp\left\{-\Lambda_0^*\left(u\right)\exp\left(\widetilde{x}_{\mathrm{RC},i}^\top\beta_x+z_i^\top\beta_z\right)\right\}dH\left(u\right).$$

Noting that by (E.5) and $U_M=\frac{\partial\widehat{\ell}_M^*}{\partial\beta}$ together with (6.18), we obtain that

$$\begin{aligned}
&\frac{-1}{n}\frac{\partial U\left(0,\beta_{z0}\right)}{\partial\beta}\\
=\ &\frac{-1}{n}\left(\frac{\partial U_P\left(0,\beta_{z0}\right)}{\partial\beta}+\frac{\partial U_M\left(0,\beta_{z0}\right)}{\partial\beta}\right)\\
=\ &\frac{-1}{n}\frac{\partial}{\partial\beta}\sum_{i=1}^n\int_0^\tau\left\{\begin{pmatrix} x_i^* \\ z_i \end{pmatrix}+\begin{pmatrix} \Sigma_\epsilon\beta_x \\ 0 \end{pmatrix}-\frac{G^{(1)}\left(u,\beta_x,\beta_z\right)}{G^{(0)}\left(u,\beta_x,\beta_z\right)}\right\}dN_i\left(u\right)\Bigg|_{(\beta_x,\beta_z)=(0,\beta_{z0})}\\
&+\frac{1}{n}\frac{\partial}{\partial\beta}\sum_{i=1}^n\frac{\partial}{\partial\beta}\left\{\widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_i^\top\beta_x+z_i^\top\beta_z\right)\right\}\Bigg|_{(\beta_x,\beta_z)=(0,\beta_{z0})}\\
&+\frac{1}{n}\frac{\partial}{\partial\beta}\sum_{i=1}^n\frac{\frac{\partial}{\partial\beta}\left[\int_0^\tau\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widehat{x}_i^\top\beta_x+z_i^\top\beta_z\right)\right\}d\widehat{H}(u)\right]}{\int_0^\tau\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(\widehat{x}_i^\top\beta_x+z_i^\top\beta_z\right)\right\}d\widehat{H}(u)}\Bigg|_{(\beta_x,\beta_z)=(0,\beta_{z0}).}
\end{aligned} \tag{E.15}$$

Then exchanging the order of differentiation and summation and plugging (E.10) to (E.15) with (β_x, β_z) evaluated at $(0, \beta_{z0})$, yields

$$\begin{aligned}
&\frac{-1}{n}\frac{\partial U\left(0,\beta_{z0}\right)}{\partial\beta} \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\left\{\frac{G^{(2)}(u,0,\beta_{z0})}{G^{(0)}(u,0,\beta_{z0})} - \left(\frac{G^{(1)}(u,0,\beta_{z0})}{G^{(0)}(u,0,\beta_{z0})}\right)^{\otimes 2}\right\} - \begin{pmatrix} \Sigma_\epsilon & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q}\end{pmatrix}\right] dN_i(u) \\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta^\top}\left\{\widehat{\Lambda}_0(a_i)\exp\left(\widehat{x}_i^\top 0 + z_i^\top\beta_{z0}\right)\right\} \\
&+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{\widehat{\zeta}_i^*}\frac{\partial^2\widehat{\zeta}_i^*}{\partial\beta\partial\beta^\top} - \frac{1}{(\widehat{\zeta}_i^*)^2}\left(\frac{\partial\widehat{\zeta}_i^*}{\partial\beta}\right)^{\otimes 2}\right\},
\end{aligned}$$

where $G^{(k+1)}(u,\beta_x,\beta_z) = \frac{\partial}{\partial\beta}G^{(k)}(u,\beta_x,\beta_z)$ for $k = 0, 1$, and $\widehat{\zeta}_i^* = \widehat{\zeta}_i^*\left(0,\beta_{z0}\right)$.

We conclude that by the Law of Large Numbers, as $n \to \infty$,

$$\frac{-1}{n}\frac{\partial U\left(0,\beta_{z0}\right)}{\partial\beta} \quad \xrightarrow{p} \quad \mathcal{A}, \tag{E.16}$$

where

$$\begin{aligned}
\mathcal{A} \;=\; & \int_0^\tau \left[\left\{\frac{\mathcal{G}^{(2)}(u,0,\beta_{z0})}{\mathcal{G}^{(0)}(u,0,\beta_{z0})} - \left(\frac{\mathcal{G}^{(1)}(u,0,\beta_{z0})}{\mathcal{G}^{(0)}(u,0,\beta_{z0})}\right)^{\otimes 2}\right\} - \begin{pmatrix} \Sigma_\epsilon & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q}\end{pmatrix}\right] dE\left\{N_i(u)\right\} \\
&+E\left\{\frac{\partial^2}{\partial\beta\partial\beta^\top}\Lambda_0^*(A)\exp\left(\widetilde{X}_{\mathrm{RC}}^\top 0 + Z^\top\beta_{z0}\right) + (\zeta^*)^{-2}\left(\zeta^*\frac{\partial(\zeta^*)^2}{\partial^2\beta} - \left(\frac{\partial\zeta^*}{\partial\beta}\right)^{\otimes 2}\right)\right\}
\end{aligned}$$

with $\mathcal{G}^{(k)}(\cdot)$ is given by (E.11) for $k = 0, 1, 2$, and

$$\zeta^* = \int_0^\tau \exp\left\{-\Lambda_0^*(u)\exp\left(\widetilde{x}_{\mathrm{RC}}^\top 0 + z^\top\beta_{z0}\right)\right\} dH(u).$$

Step 2:

Since $U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)$ contains the sample size n, it cannot be directly expressed as a sum of i.i.d. random functions. We now want to re-express it in order to derive a sum of i.i.d. random functions. Since $\exp\left(x^{*\top}\frac{\eta}{\sqrt{n}}\right) = \frac{x^{*\top}\eta}{\sqrt{n}} + O_p(1)$, then by (6.20) and (E.2),

$$G^{(j)}\left(u,\frac{\eta}{\sqrt{n}},\beta_z\right) = \frac{1}{\sqrt{n}}\widetilde{G}^{(j)}(u,\eta,\beta_z) + O_p(1) \text{ for } j = 0, 1, \tag{E.17}$$

where

$$\widetilde{G}^{(j)}\left(u,\eta,\beta_z\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right)^{\otimes j} Y_i(u)\exp\left(z_i^\top\beta_z\right)x_i^{*\top} \tag{E.18}$$

which is a sum of i.i.d. random variables for $j=0,1$.

Combining (E.17) and $U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)$ gives

$$\frac{1}{\sqrt{n}}U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)=\frac{1}{\sqrt{n}}\widetilde{U}\left(\eta,\beta_{z0}\right)+o_p(1), \tag{E.19}$$

where

$$\widetilde{U}\left(\eta,\beta_{z0}\right)=\widetilde{U}_P\left(\eta,\beta_{z0}\right)+\widetilde{U}_M\left(\eta,\beta_{z0}\right), \tag{E.20}$$

$$\widetilde{U}_P\left(\eta,\beta_{z0}\right)=\sum_{i=1}^{n}\int_0^\tau\left\{\left(\begin{array}{c} x_i^* \\ z_i \end{array}\right)+\left(\begin{array}{c} \Sigma_\epsilon\eta \\ 0 \end{array}\right)-\frac{\widetilde{G}^{(1)}\left(u,\eta,\beta_z\right)}{\widetilde{G}^{(0)}\left(u,\eta,\beta_z\right)}\right\}dN_i\left(u\right)$$

and

$$\begin{aligned}\widetilde{U}_M\left(\eta,\beta_{z0}\right) &= -\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\left\{\int_0^{a_i}\frac{\frac{1}{n}\sum_{j=1}^{n}dN_j(u)}{(\eta^\top\Sigma_\epsilon\eta)^{-1}\widetilde{G}^{(0)}\left(u,\eta,\beta_{z0}\right)}\exp\left(z_i^\top\beta_z\right)\widehat{x}_i^\top\eta\right\} \\ &\quad -\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\log\int_0^\tau\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(z_i^\top\beta_{z0}\right)\widehat{x}_i^\top\eta\right\}d\widehat{H}(u).\end{aligned} \tag{E.21}$$

(E.19) suggests that to study the asymptotic behavior of $\frac{1}{\sqrt{n}}U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)$, it suffices to study $\frac{1}{\sqrt{n}}\widetilde{U}\left(\eta,\beta_{z0}\right)$ by expressing it as a sum of i.i.d. random functions. To this end, by (E.20), we separately examine $\widetilde{U}_P\left(\eta,\beta_{z0}\right)$ and $\widetilde{U}_M\left(\eta,\beta_{z0}\right)$. First, using the arguments similar to the derivations in Theorem 2.1 of Lin and Wei (1989), we derive

$$\frac{1}{\sqrt{n}}\widetilde{U}_P\left(\eta,\beta_{z0}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{1i}+o_p(1), \tag{E.22}$$

where

$$\begin{aligned}\Psi_{1i} &= \int_0^\tau \left\{ \begin{pmatrix} x_i^* \\ z_i \end{pmatrix} - \frac{\widetilde{\mathcal{G}}^{(1)}(u,\eta,\beta_{z0})}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})} + \begin{pmatrix} \Sigma_\epsilon \eta \\ \mathbf{0}_q \end{pmatrix} \right\} dN_i(u) \\ &\quad - \int_0^\tau \frac{\exp\left(z_i^\top \beta_z\right) x_i^{*\top} \eta I(A_i \le u \le Y_i)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})} \left\{ \begin{pmatrix} x_i^* \\ z_i \end{pmatrix} - \frac{\widetilde{\mathcal{G}}^{(1)}(u,\eta,\beta_{z0})}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})} \right\} dE\left\{N_i(u)\right\}\end{aligned} \tag{E.23}$$

with

$$\widetilde{\mathcal{G}}^{(j)}(u,\eta,\beta_{z0}) = E\left\{ \begin{pmatrix} X^* \\ Z \end{pmatrix}^{\otimes j} Y(u) \exp\left(Z^\top \beta_z\right) X^{*\top} \eta \right\} \text{ for } j = 0, 1.$$

Next, we examine $\widetilde{U}_M(\eta, \beta_{z0})$. Let

$$\zeta(\widetilde{x}_{\mathrm{RC}}, z) = \int_0^\tau \exp\left\{-\Lambda_0^*(u)\exp(z^\top \beta_{z0})\widetilde{x}_{\mathrm{RC}}^\top \eta\right\} dH(u), \tag{E.24}$$

$$\widetilde{X}_{\mathrm{RC}} = \mu_X + \left(\Sigma_{X^*} - \Sigma_\epsilon\right)^\top \Sigma_{X^*}^{-1}\left(X^* - \mu_{X^*}\right)$$

and

$$\widehat{\zeta}(\widehat{x}, z) = \int_0^\tau \exp\left\{-\widehat{\Lambda}_0(u)\exp(z^\top \beta_{z0})\widehat{x}^\top \eta\right\} d\widehat{H}(u). \tag{E.25}$$

To derive a sum of i.i.d. random functions and study the asymptotic behavior of $\widetilde{U}_M(\eta, \beta_{z0})$, we further define

$$\widetilde{U}_M^*(\eta, \beta_{z0}) = -\sum_{i=1}^n \frac{\partial}{\partial \beta} \Lambda_0^*(a_i) \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta - \sum_{i=1}^n \frac{1}{\zeta(\widetilde{x}_{\mathrm{RC},i}, z_i)} \frac{\partial \zeta(\widetilde{x}_{\mathrm{RC},i}, z_i)}{\partial \beta}. \tag{E.26}$$

Then by (E.21) and (E.26), the difference between $\widetilde{U}_M$ and $\widetilde{U}_M^*$ can be written as

$$\frac{1}{\sqrt{n}}\left\{\widetilde{U}_M - \widetilde{U}_M^*\right\} = \widetilde{U}_1 + \widetilde{U}_2, \tag{E.27}$$

where

$$\widetilde{U}_1 = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} \left\{\widehat{\Lambda}_0(a_i) - \Lambda_0^*(a_i)\right\} \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta$$

and

$$\widetilde{U}_2 = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{\widehat{\zeta}\left(\widehat{x}_i, z_i\right)} \frac{\partial \widehat{\zeta}\left(\widehat{x}_i, z_i\right)}{\partial \beta} - \frac{1}{\zeta\left(\widetilde{x}_{\mathrm{RC},i}, z_i\right)} \frac{\partial \zeta\left(\widetilde{x}_{\mathrm{RC},i}, z_i\right)}{\partial \beta} \right\}. \tag{E.28}$$

To study the asymptotic behaviour of (E.27), we examine $\widetilde{U}_1$ and $\widetilde{U}_2$ individually. First, let $\mathcal{N}(t) = P\left(\Delta = 1, Y \leq t\right)$ and $d\bar{N}(t) = \frac{1}{n} \sum_{i=1}^{n} dN_i(t)$. Then by (6.19) and (E.13),

$$\begin{aligned}
\widetilde{U}_1 &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left\{ \widehat{\Lambda}_0(a_i) - \Lambda_0^*(a_i) \right\} \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{d\bar{N}(u)}{\widetilde{G}^{(0)}(u, \eta, \beta_{z0})} - \frac{d\mathcal{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u, \eta, \beta_{z0})} \right\} \left(\eta^\top \Sigma_\epsilon \eta\right) \exp\left(z_i^\top \beta_{z0}\right) \\
&\quad \times \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a_i \leq \tau) \\
&= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \frac{\partial}{\partial \beta} \left\{ \frac{d\bar{N}(u) - d\mathcal{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u, \eta, \beta_{z0})} + \frac{d\mathcal{N}(u)\widetilde{\mathcal{G}}^{(0)}(u, \eta, \beta_{z0}) - d\bar{N}(u)\widetilde{G}^{(0)}(u, \eta, \beta_{z0})}{\widetilde{\mathcal{G}}^{(0)}(u, \eta, \beta_{z0})\widetilde{G}^{(0)}(u, \eta, \beta_{z0})} \right\} \\
&\quad \times \left(\eta^\top \Sigma_\epsilon \eta\right) \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a_i \leq \tau).
\end{aligned} \tag{E.29}$$

Since $\frac{1}{n} \sum_{i=1}^{n} \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a_i \leq \tau)$ is an average of i.i.d. random variables due to Conditions (C3), (C4) and (C5). Then by the Law of Large Numbers, we have that as $n \to \infty$,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a_i \leq \tau) \\
\xrightarrow{p} \; & E\left\{ \exp\left(Z^\top \beta_{z0}\right) \widetilde{X}_{\mathrm{RC}}^\top \eta I(u \leq A \leq \tau) \right\} \\
= \; & \int_{-\infty}^{\infty} \int_0^\tau \exp\left(z^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC}}^\top \eta I(u \leq a \leq \tau) dQ(a, \widehat{v}),
\end{aligned}$$

i.e.,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \exp\left(z_i^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a_i \leq \tau) \\
= \; & \int_{-\infty}^{\infty} \int_0^\tau \exp\left(z^\top \beta_{z0}\right) \widetilde{x}_{\mathrm{RC},i}^\top \eta I(u \leq a \leq \tau) dQ(a, \widehat{v}) + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{E.30}$$

(e.g., Jiang 2010, p.61), where $Q(a, \widehat{v})$ is the joint density function of $(A, \widehat{V})$ with $\widehat{V} = \left(\widetilde{X}_{\mathrm{RC}}, Z\right)$.

Then plugging (E.30) into (E.29) gives

$$\begin{aligned}
&\widetilde{U}_1 \\
= \; & -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\frac{\partial}{\partial\beta}\left\{\frac{d\bar{N}(u)-d\mathcal{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}+\frac{d\mathcal{N}(u)\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})-d\bar{N}(u)\widetilde{G}^{(0)}(u,\eta,\beta_{z0})}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\widetilde{G}^{(0)}(u,\eta,\beta_{z0})}\right\} \\
& \times\left(\eta^{\top}\Sigma_{\epsilon}\eta\right)\exp\left(z^{\top}\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^{\top}\eta I(u\le a\le\tau)dQ(a,\widehat{v})+o_p(1), \qquad \text{(E.31)}
\end{aligned}$$

where the order term is determined by $\sqrt{n}\times o_p(1)\times O_p\left(\frac{1}{\sqrt{n}}\right)=o_p(1)$.

Furthermore, noting that by the Uniformly Strong Law of Large Numbers,

$$d\bar{N}(t)\xrightarrow{a.s.}d\mathcal{N}(t)$$

and

$$\widetilde{G}^{(0)}(u,\eta,\beta_{z0})\xrightarrow{a.s.}\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})$$

uniformly as $n\to\infty$. That is,

$$d\bar{N}(t)=d\mathcal{N}(t)+o_p(1) \qquad \text{(E.32)}$$

and

$$\widetilde{G}^{(0)}(u,\eta,\beta_{z0})=\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})+o_p(1). \qquad \text{(E.33)}$$

Then we obtain that

$$\begin{aligned}
\widetilde{U}_1 \; = \; & -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\frac{\partial}{\partial\beta}\left\{\frac{d\bar{N}(u)-d\mathcal{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}+\frac{d\mathcal{N}(u)\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})-d\bar{N}(u)\widetilde{G}^{(0)}(u,\eta,\beta_{z0})}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\widetilde{G}^{(0)}(u,\eta,\beta_{z0})}\right\} \\
& \times\left(\eta^{\top}\Sigma_{\epsilon}\eta\right)\exp\left(z^{\top}\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^{\top}\eta I(u\le a\le\tau)dQ(a,\widehat{v})+o_p(1) \\
= \; & -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\frac{\partial}{\partial\beta}\left[\frac{d\bar{N}(u)-d\mathcal{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}+\frac{d\mathcal{N}(u)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right.\right. \\
& \left.\left.-\widetilde{G}^{(0)}(u,\eta,\beta_{z0})\right\}\right]\times\left(\eta^{\top}\Sigma_{\epsilon}\eta\right)\exp\left(z^{\top}\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^{\top}\eta I(u\le a\le\tau)dQ(a,\widehat{v})+o_p(1),
\end{aligned}$$

where we apply (E.32) and (E.33) to the numerator and denominator of the second term, respectively. Then by definition of $d\bar{N}(t)$ and (E.18), we obtain that

$$\begin{aligned}
\widetilde{U}_1 &= -\sqrt{n}\int_{-\infty}^{\infty}\int_0^{\tau}\frac{\partial}{\partial\beta}\left[\frac{d\bar{N}(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\widetilde{G}^{(0)}(u,\eta,\beta_{z0})}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right]\left(\eta^\top\Sigma_\epsilon\eta\right)\\
&\quad\times\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta I(u\le a\le\tau)dQ(a,\widehat{v})+o_p(1)\\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\left[\int_{-\infty}^{\infty}\int_0^{\tau}\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\exp\left(z_i^\top\beta_{z0}\right)x_i^{*\top}\eta I(a_i\le u\le y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\right.\\
&\quad\left.\times\left(\eta^\top\Sigma_\epsilon\eta\right)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta I(u\le a\le\tau)\right]dQ(a,\widehat{v})+o_p(1). \qquad \text{(E.34)}
\end{aligned}$$

We next examine $\widetilde{U}_2$. To do so, we first derive the asymptotic result of

$$\sqrt{n}\left\{\widehat{\zeta}\left(\widehat{x},z\right)-\zeta\left(\widetilde{x}_{\mathrm{RC}},z\right)\right\}.$$

Since $\widehat{\mu}_{X^*}=\mu_X+o_p(1)$ and $\widehat{\Sigma}_{X^*}=\Sigma_{X^*}+o_p(1)$, we obtain that $\widehat{X}_i=\widetilde{X}_{\mathrm{RC},i}+o_p(1)$ by the Law of Large Numbers. Hence,

$$\begin{aligned}
&-\widehat{\Lambda}_0(u)\exp(z^\top\beta_{z0})\widehat{x}^\top\eta+\Lambda_0^*(u)\exp(z^\top\beta_{z0})\widetilde{x}_{\mathrm{RC}}^\top\eta\\
&= -\left\{\widehat{\Lambda}_0(u)-\Lambda_0^*(u)\right\}\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta+o_p(1)\\
&= \frac{-1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\exp\left(z_i^\top\beta_{z0}\right)x_i^{*\top}\eta I(a_i\le u\le y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\\
&\quad\times\left(\eta^\top\Sigma_\epsilon\eta\right)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta+o_p(1), \qquad \text{(E.35)}
\end{aligned}$$

where the last equality is due to the expression of $\widehat{\Lambda}_0(u)-\Lambda_0^*(u)$ in (E.29).

Applying the Taylor expansion to $\exp\left\{-\widehat{\Lambda}_0(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\}$ with respect to $\Lambda_0(\cdot)$

yields

$$
\begin{aligned}
& \exp\left\{-\widehat{\Lambda}_0(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\}-\exp\left\{-\Lambda_0^*(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\} \\
= & -\exp\left\{-\Lambda_0^*(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\}\left\{\widehat{\Lambda}_0(u)-\Lambda_0^*(u)\right\}\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta+o_p\left(\frac{1}{\sqrt{n}}\right) \\
= & \exp\left\{-\Lambda_0^*(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\} \\
& \times\frac{1}{n}\sum_{i=1}^n\int_0^\tau\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\exp\left(z_i^\top\beta_{z0}\right)x_i^{*\top}\eta I(a_i\le u\le y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\left(\eta^\top\Sigma_\epsilon\eta\right) \\
& \times\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta+o_p\left(\frac{1}{\sqrt{n}}\right) \\
= & S(u,\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z) \\
& \times\frac{1}{n}\sum_{i=1}^n\int_0^\tau\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\exp\left(z_i^\top\beta_{z0}\right)x_i^{*\top}\eta I(a_i\le u\le y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\left(\eta^\top\Sigma_\epsilon\eta\right) \\
& \times\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta+o_p\left(\frac{1}{\sqrt{n}}\right), \qquad \text{(E.36)}
\end{aligned}
$$

where the second equality is due to (E.35), and

$$
S(u,\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)=\exp\left\{-\Lambda_0^*(u)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta\right\}.
$$

Finally, using (E.24) and (E.25) in combination with (E.14) and (E.36), we obtain that

$$
\sqrt{n}\left\{\widehat{\zeta}(\widehat{x},z)-\zeta(\widetilde{x}_{\mathrm{RC}},z)\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)+o_p\left(1\right), \qquad \text{(E.37)}
$$

where

$$
\begin{aligned}
& \psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z) \\
= & \int_0^\tau\int_0^\tau h(\xi)S(\xi,\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}\right. \\
& \left.-\frac{d\mathcal{N}(u)\exp\left(z_i^\top\beta_{z0}\right)x_i^{*\top}\eta I(a_i\le u\le y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\times\left(\eta^\top\Sigma_\epsilon\eta\right)\exp\left(z^\top\beta_{z0}\right)\widetilde{x}_{\mathrm{RC}}^\top\eta d\xi.
\end{aligned}
$$

Then by (E.28) and similar to the derivations of (E.31), we obtain that

$$\begin{aligned}
\widetilde{U}_2 &= \frac{-1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\frac{1}{\widehat{\zeta}(\widehat{x}_j,z_j)}\frac{\partial\widehat{\zeta}(\widehat{x}_j,z_j)}{\partial\beta}-\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC},j},z_j)}\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC},j},z_j)}{\partial\beta}\right\} \\
&= -\sqrt{n}\times\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{\widehat{\zeta}(\widehat{x}_j,z_j)}\frac{\partial\widehat{\zeta}(\widehat{x}_j,z_j)}{\partial\beta}-\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC},j},z_j)}\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC},j},z_j)}{\partial\beta}\right\} \\
&= -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{1}{\widehat{\zeta}(\widehat{x},z)}\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}-\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC}},z)}{\partial\beta}\right\}dQ(a,\widehat{v})+o_p(1). \quad \text{(E.38)}
\end{aligned}$$

To sort out a sum of i.i.d. random functions from (E.38), we add and subtract the term $\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}$ and then regroup the differences, yielding

$$\begin{aligned}
\widetilde{U}_2 &= -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{1}{\widehat{\zeta}(\widehat{x},z)}\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}-\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}+\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}\right. \\
&\quad \left.-\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC}},z)}{\partial\beta}\right\}dQ(a,\widehat{v})+o_p(1) \\
&= -\sqrt{n}\int_{-\infty}^{\infty}\int_{0}^{\tau}\left[\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\left\{\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}-\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC}},z)}{\partial\beta}\right\}\right. \\
&\quad \left.-\frac{\partial\widehat{\zeta}(\widehat{x},z)}{\partial\beta}\left\{\frac{\widehat{\zeta}(\widehat{x},z)-\zeta(\widetilde{x}_{\mathrm{RC}},z)}{\widehat{\zeta}(\widehat{x},z)\zeta(\widetilde{x}_{\mathrm{RC}},z)}\right\}\right]dQ(a,\widehat{v})+o_p(1) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{1}{\zeta(\widetilde{x}_{\mathrm{RC}},z)}\frac{\partial}{\partial\beta}\psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)\right.\right. \\
&\quad \left.\left.-\frac{\partial\zeta(\widetilde{x}_{\mathrm{RC}},z)}{\partial\beta}\frac{1}{\zeta^2(\widetilde{x}_{\mathrm{RC}},z)}\psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)\right\}\right]dQ(a,\widehat{v})+o_p(1), \quad \text{(E.39)}
\end{aligned}$$

where the second equality is due to $\widehat{\zeta}\left(\widehat{x},z\right)=\zeta\left(\widetilde{x}_{\mathrm{RC}},z\right)+o_p(1)$, and the last step is due to (E.37).

Combining (E.27), (E.34) and (E.39) gives

$$\frac{1}{\sqrt{n}}\widetilde{U}_M\left(\eta,\beta_{z0}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{2i}+o_p(1), \quad \text{(E.40)}$$

where

$$\begin{aligned}\Psi_{2i} &= -\left[\int_{-\infty}^{\infty}\int_{0}^{\tau}\frac{\partial}{\partial\beta}\left\{\frac{dN_i(u)}{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})}-\frac{d\mathcal{N}(u)\exp\left(z_i^{\top}\beta_z\right)x_i^{*\top}\eta I(a_i\leq u\leq y_i)}{\left\{\widetilde{\mathcal{G}}^{(0)}(u,\eta,\beta_{z0})\right\}^2}\right\}\right.\\ &\quad\left.\times\left(\eta^{\top}\Sigma_{\epsilon}\eta\right)\exp\left(z^{\top}\beta_z\right)\widetilde{x}_{\mathrm{RC}}^{\top}\eta I(u\leq a\leq\tau)\right]dQ(a,\widehat{v})\\ &\quad-\left[\int_{-\infty}^{\infty}\int_{0}^{\tau}\left\{\frac{1}{\zeta\left(\widetilde{x}_{\mathrm{RC}},z\right)}\frac{\partial}{\partial\beta}\psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)\right.\right.\\ &\quad\left.\left.-\frac{\partial\zeta\left(\widetilde{x}_{\mathrm{RC}},z\right)}{\partial\beta}\frac{1}{\zeta^2\left(\widetilde{x}_{\mathrm{RC}},z\right)}\psi_i(\eta,\beta_{z0}|\widetilde{x}_{\mathrm{RC}},z)\right\}dQ(a,\widehat{v})\right]\\ &\quad-\frac{\partial}{\partial\beta}\Lambda_0^*(a_i)\exp\left(z_i^{\top}\beta_z\right)\widetilde{x}_{\mathrm{RC},i}^{\top}\eta-\frac{1}{\zeta\left(\widetilde{x}_{\mathrm{RC},i},z_i\right)}\frac{\partial}{\partial\beta}\zeta\left(\widetilde{x}_{\mathrm{RC},i},z_i\right). \qquad \text{(E.41)}\end{aligned}$$

Therefore, using (E.19), (E.20), (E.22) and (E.40) and applying the Central Limit Theorem, we obtain that as $n\to\infty$,

$$\frac{1}{\sqrt{n}}U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)\xrightarrow{d}N\left(0,\mathcal{B}\right), \qquad \text{(E.42)}$$

where $\mathcal{B}=E(\Psi_i^{\otimes 2})$ with $\Psi_i=\Psi_{1i}+\Psi_{2i}$, and Ψ_{1i} and Ψ_{2i} are given by (E.23) and (E.41), respectively.

Finally, applying Slutsky's Theorem in combination with (E.9), (E.16) and (E.42), we obtain that as $n\to\infty$,

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_x\\ \widehat{\beta}_x-\beta_{z0}\end{pmatrix}\xrightarrow{d}N\left(\begin{pmatrix}\eta\\ 0\end{pmatrix},\mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1}\right).$$

Proof of (b):

We first re-scale (E.7), which gives

$$\frac{1}{\sqrt{n}}U\left(0,\beta_{z0}\right)=\frac{1}{\sqrt{n}}U\left(\frac{\eta}{\sqrt{n}},\beta_{z0}\right)-\frac{1}{n}\frac{\partial U\left(0,\beta_{z0}\right)}{\partial\beta^{\top}}\begin{pmatrix}\eta\\ 0\end{pmatrix}+o_p\left(1\right). \qquad \text{(E.43)}$$

Combining (E.16), (E.42) and (E.43) and applying Slutsky's Theorem, we obtain that as $n\to\infty$,

$$\frac{1}{\sqrt{n}}U\left(0,\beta_{z0}\right)\xrightarrow{d}N\left(\mathcal{A}\begin{pmatrix}\eta\\ 0\end{pmatrix},\mathcal{B}\right). \qquad \text{(E.44)}$$

Now we consider any candidate model S. Applying the Taylor expansion to $U_S(\widehat{\beta}_{x,S}, \widehat{\beta}_{z,S})$ around $(0, \beta_{z0})$ gives

$$0 = U_S\left(\widehat{\beta}_{x,S}, \widehat{\beta}_{z,S}\right) = U_S\left(0, \beta_{z0}\right) + \frac{\partial U_S\left(0, \beta_{z0}\right)}{\partial \beta^\top} \begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

yielding that

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} &= -\left(\frac{1}{n}\frac{\partial U_S\left(0, \beta_{z0}\right)}{\partial \beta^\top}\right)^{-1} \frac{1}{\sqrt{n}} U_S\left(0, \beta_{z0}\right) + o_p\left(1\right) \\ &= -\left(\frac{1}{n}\frac{\partial U_S\left(0, \beta_{z0}\right)}{\partial \beta^\top}\right)^{-1} \frac{1}{\sqrt{n}} \Pi_S U\left(0, \beta_{z0}\right) + o_p\left(1\right) \\ &\xrightarrow{d} \mathcal{A}_S^{-1} \Pi_S N\left(\mathcal{A} \begin{pmatrix} \eta \\ 0 \end{pmatrix}, \mathcal{B}\right) \text{ as } n \to \infty, \end{aligned}$$

where the second identity is from Lemma 6.3.2 and the third step is due to (E.44). Thus,

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} \xrightarrow{d} N\left(\mathcal{A}_S^{-1} \Pi_S \mathcal{A} \begin{pmatrix} \eta \\ 0 \end{pmatrix}, \mathcal{A}_S^{-1} \mathcal{B}_S \mathcal{A}_S^{-1}\right) \text{ as } n \to \infty,$$

where $\mathcal{B}_S = \Pi_S \mathcal{B} \Pi_S^\top$. □

E.2.4 Proof of Theorem 6.3.2

The proof consists of the following three steps.

Step 1:

For a given candidate model S, by (6.9), we have

$$\sqrt{n}\left\{\widehat{\Lambda}_{0,S}(t) - \Lambda_0(t)\right\} = \sqrt{n}\left\{\int_0^t \frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)}{m_S^{-1}(\widehat{\beta}_{x,S}) G_S^{(0)}\left(u, \widehat{\beta}_{x,S}, \widehat{\beta}_{z,S}\right)} - \Lambda_0(t)\right\} = A + B,$$

where

$$A = \sqrt{n}\left\{\int_0^t \frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)}{m_S^{-1}(\widehat{\beta}_{x,S}) G_S^{(0)}\left(u, \widehat{\beta}_{x,S}, \widehat{\beta}_{z,S}\right)} - \int_0^t \frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)}{m^{-1}(\frac{\eta}{\sqrt{n}}) G^{(0)}\left(u, \frac{\eta}{\sqrt{n}}, \beta_{z0}\right)}\right\} \quad \text{(E.45)}$$

and

$$B = \sqrt{n}\left\{ \int_0^t \frac{\frac{1}{n}\sum_{i=1}^{n} dN_i(u)}{m^{-1}(\frac{\eta}{\sqrt{n}})G^{(0)}\left(u, \frac{\eta}{\sqrt{n}}, \beta_{z0}\right)} - \Lambda_0(t) \right\}.$$

Step 2:

We first examine A. Applying the Taylor expansion to $\frac{m_S(\widehat{\beta}_{x,S})}{G_S^{(0)}\left(u,\widehat{\beta}_{x,S},\widehat{\beta}_{z,S}\right)}$ and $\frac{m(\frac{\eta}{\sqrt{n}})}{G^{(0)}\left(u,\frac{\eta}{\sqrt{n}},\beta_{z0}\right)}$, respectively, around $(0, \beta_{z0})$, we have

$$\begin{aligned} & \frac{m_S(\widehat{\beta}_{x,S})}{G_S^{(0)}\left(u, \widehat{\beta}_{x,S}, \widehat{\beta}_{z,S}\right)} \\ = \; & \frac{1}{G_S^{(0)}\left(u,0,\beta_{z0}\right)} - \frac{1}{\left\{G_S^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2} \begin{pmatrix} G_{x,S}^{(1)}\left(u,0,\beta_{z0}\right) \\ G_{z,S}^{(1)}\left(u,0,\beta_{z0}\right) \end{pmatrix}^{\top} \begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} \\ & + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{E.46}$$

and

$$\frac{m(\frac{\eta}{\sqrt{n}})}{G^{(0)}\left(u, \frac{\eta}{\sqrt{n}}, \beta_{z0}\right)} = \frac{1}{G^{(0)}\left(u,0,\beta_{z0}\right)} - \frac{G_x^{(1)}\left(u,0,\beta_{z0}\right)}{\left\{G^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\frac{\eta}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{E.47}$$

where

$$G_{x,S}^{(1)}(u, \beta_{x,S}, \beta_{z,S}) = \frac{\partial}{\partial \beta_{x,S}} G_S^{(0)}(u, \beta_{x,S}, \beta_{z,S})$$

and

$$G_{z,S}^{(1)}(u, \beta_{x,S}, \beta_{z,S}) = \frac{\partial}{\partial \beta_{z,S}} G_S^{(0)}(u, \beta_{x,S}, \beta_{z,S}).$$

Since (E.4) with $j = 0$ gives $G_S^{(0)}\left(u,0,\beta_{z0}\right) = G^{(0)}\left(u,0,\beta_{z0}\right)$, so we combine $(E.46)$ and

$(E.47)$ and obtain that

$$\begin{aligned}
&\frac{m_S(\widehat{\beta}_{x,S})}{G_S^{(0)}\left(u,\widehat{\beta}_{x,S},\widehat{\beta}_{z,S}\right)}-\frac{m(\frac{\eta}{\sqrt{n}})}{G^{(0)}\left(u,\frac{\eta}{\sqrt{n}},\beta_{z0}\right)}\\
= &\frac{-1}{\left\{G_S^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\begin{pmatrix} G_{x,S}^{(1)}\left(u,0,\beta_{z0}\right)\\ G_{z,S}^{(1)}\left(u,0,\beta_{z0}\right)\end{pmatrix}^{\top}\begin{pmatrix}\widehat{\beta}_{x,S}\\ \widehat{\beta}_{z,S}-\beta_{z0}\end{pmatrix}\\
&+\frac{G_x^{(1)}\left(u,0,\beta_{z0}\right)}{\left\{G^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\frac{\eta}{\sqrt{n}}+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Hence, applying (E.45) gives that

$$\begin{aligned}
A = &-\int_0^t\frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)}{\left\{G_S^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\begin{pmatrix} G_{x,S}^{(1)}\left(u,0,\beta_{z0}\right)\\ G_{z,S}^{(1)}\left(u,0,\beta_{z0}\right)\end{pmatrix}^{\top}\sqrt{n}\begin{pmatrix}\widehat{\beta}_{x,S}\\ \widehat{\beta}_{z,S}-\beta_{z0}\end{pmatrix}\\
&+\int_0^t\frac{\frac{1}{n}\sum_{i=1}^n dN_i(u)G_x^{(1)}\left(u,0,\beta_{z0}\right)}{\left\{G^{(0)}\left(u,0,\beta_{z0}\right)\right\}^2}\eta+o_p(1). \qquad (E.48)
\end{aligned}$$

Now we examine the terms in (E.48) separately. Since $\{Y_i(t):t\in[0,\tau]\}$ and $\{N_i(t):t\in[0,\tau]\}$ are Glivenko-Cantelli class (van der Vaart and Wellner 1996, Theorems 2.4.1 and 2.7.5), then we have as that $n\to\infty$,

$$\frac{1}{n}\sum_{i=1}^n dN_i(t)\xrightarrow{a.s.}E\left\{dN_i(t)\right\},$$

and

$$G_S^{(k)}\left(t,0,\beta_{z0}\right)\xrightarrow{a.s.}\mathcal{G}_S^{(k)}\left(t,0,\beta_{z0}\right)\text{ for }k=0,1$$

uniformly at t, where $G_S^{(0)}(\cdot)$ and $G_S^{(1)}(\cdot)$ are given by (6.10) and (E.1), respectively, and

$$\mathcal{G}_S^{(k)}(u,\beta_x,\beta_z)=E\left\{\begin{pmatrix}\pi_S X^*\\ Z\end{pmatrix}^{\otimes k}\exp((\pi_S X^*)^{\top}\beta_x+Z^{\top}\beta_z)I(A\le u\le Y)\right\}.$$

Therefore, as $n \to \infty$,

$$\int_0^t \frac{\frac{1}{n}\sum_{i=1}^{n} dN_i(u)}{\left\{G_S^{(0)}(u,0,\beta_{z0})\right\}^2} \begin{pmatrix} G_{x,S}^{(1)}(u,0,\beta_{z0}) \\ G_{z,S}^{(1)}(u,0,\beta_{z0}) \end{pmatrix}^\top \xrightarrow{a.s.} \begin{pmatrix} F_{x,S}(t) \\ F_z(t) \end{pmatrix}^\top \tag{E.49}$$

uniformly at t, where

$$F_{x,S}(t) = \int_0^t \frac{E(dN_i(u))\mathcal{G}_{x,S}^{(1)}(u,0,\beta_{z0})}{\left\{\mathcal{G}_S^{(0)}(u,0,\beta_{z0})\right\}^2}$$

and

$$F_z(t) = \int_0^t \frac{E(dN_i(u))\mathcal{G}_z^{(1)}(u,0,\beta_{z0})}{\left\{\mathcal{G}^{(0)}(u,0,\beta_{z0})\right\}^2}.$$

Regarding the term $\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix}$ in (E.48), we apply Theorem 6.3.1 (b) and let W_S be a random vector whose distribution is the same as the limiting distribution of $\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix}$, i.e.,

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix} \xrightarrow{d} W_S \text{ as } n \to \infty. \tag{E.50}$$

Then applying Slutsky's Theorem to (E.48) in combination with (E.49) and (E.50), we have that as $n \to \infty$,

$$A \xrightarrow{d} -\begin{pmatrix} F_{x,S}(t) \\ F_z(t) \end{pmatrix}^\top W_S + F_x(t)^\top \eta. \tag{E.51}$$

Step 3:

Finally, we examine the asymptotic behavior of B. Noting that $\exp\left(\frac{\eta^\top \Sigma_\epsilon \eta}{n}\right) = \frac{\eta^\top \Sigma_\epsilon \eta}{n} + O(1)$ and $\frac{1}{n} = o\left(\frac{1}{\sqrt{n}}\right)$, by the arguments similar to Appendix A.4 of Lin et al. (2000), we

have

$$
\begin{aligned}
B &= \sqrt{n}\int_0^t \left\{ \frac{m(\frac{\eta}{\sqrt{n}})\frac{1}{n}\sum_{i=1}^n dN_i(u)}{G^{(0)}\left(u, \frac{\eta}{\sqrt{n}}, \beta_{z0}\right)} - d\Lambda_0(u) \right\} \\
&= \sqrt{n}\int_0^t \left\{ \frac{\frac{\eta^\top \Sigma_\epsilon \eta}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n dN_i(u)}{\frac{1}{\sqrt{n}}\widetilde{G}^{(0)}\left(u, \eta, \beta_{z0}\right)} - d\Lambda_0(u) \right\} + o_p(1) \\
&= \sqrt{n}\int_0^t \left\{ \frac{\eta^\top \Sigma_\epsilon \eta \frac{1}{n}\sum_{i=1}^n dN_i(u)}{\widetilde{G}^{(0)}\left(u, \eta, \beta_{z0}\right)} - d\Lambda_0(u) \right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\int_0^t \left\{ \frac{\sum_{i=1}^n \eta^\top \Sigma_\epsilon \eta dN_i(u) - \sum_{i=1}^n Y_i(u)\exp\left(z_i^\top \beta_{z0}\right)(x_i^{*\top}\eta) d\Lambda_0(t)}{\widetilde{G}^{(0)}\left(u, \eta, \beta_{z0}\right)} \right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\int_0^t \left[\frac{\sum_{i=1}^n \left\{ \eta^\top \Sigma_\epsilon \eta dN_i(u) - Y_i(u)\exp\left(z_i^\top \beta_{z0}\right)(x_i^{*\top}\eta) d\Lambda_0(u) \right\}}{\widetilde{\mathcal{G}}^{(0)}\left(u, \eta, \beta_{z0}\right)} \right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \left\{ \frac{\eta^\top \Sigma_\epsilon \eta dN_i(u) - Y_i(u)\exp\left(z_i^\top \beta_{z0}\right)(x_i^{*\top}\eta) d\Lambda_0(u)}{\widetilde{\mathcal{G}}^{(0)}\left(u, \eta, \beta_{z0}\right)} \right\} + o_p(1) \\
&\triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi_i(t) + o_p(1),
\end{aligned}
$$

where the second equality is due to (E.17) and $m(\frac{\eta}{\sqrt{n}}) = \frac{\eta^\top \Sigma_\epsilon \eta}{\sqrt{n}} + O(1)$, the fourth equality is due to (E.18), and the fifth equality is due to (E.33).

By (E.32) and (E.33), $E\left(\frac{1}{n}\sum_{i=1}^n \Phi_i(t)\right) = 0$. Then by Condition (C3), the $\Phi_i(t)$ are i.i.d. with mean zero, and hence, by the Central Limit Theorem, we conclude that

$$
B \xrightarrow{d} \mathcal{V}(t) \text{ as } n \to \infty, \tag{E.52}
$$

where $\mathcal{V}(t)$ is the Gaussian process with mean zero and covariance function $E\{\Phi_i(t)\Phi_i(s)\}$.

Finally, combining (E.51) and (E.52) gives that as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\Lambda}_{0,S}(t) - \Lambda_0(t)\right) \xrightarrow{d} \mathcal{V}(t) - \begin{pmatrix} F_{x,S}(t) \\ F_z(t) \end{pmatrix}^\top W_S + F_x(t)^\top \eta,$$

which completes the proof. □

E.2.5 Proof of Theorem 6.3.3

For ease of exposition, we simply write $\frac{\partial\mu(0,\beta_{z0})}{\partial\beta_{x,S}}$ and $\frac{\partial\mu(0,\beta_{z0})}{\partial\beta_{z,S}}$ as $\frac{\partial\mu}{\partial\beta_{x,S}}$ and $\frac{\partial\mu}{\partial\beta_{z,S}}$, respectively.
Proof of (a):
The proof consists of the following two steps.
Step 1:

Let $\begin{pmatrix} J \\ M \end{pmatrix}$ be a random vector whose distribution is $N(0, \mathcal{B})$, where J is a $p \times 1$ random vector and M is a $q \times 1$ random vector. Define $\begin{pmatrix} J_S \\ M \end{pmatrix} = \Pi_S \begin{pmatrix} J \\ M \end{pmatrix}$. Then $\begin{pmatrix} J_S \\ M \end{pmatrix}$ is a random vector whose distribution is $N(0, \mathcal{B}_S)$. Let

$$W_S = \begin{pmatrix} C_S \\ D_S \end{pmatrix} = \mathcal{A}_S^{-1}\left\{\Pi_S \mathcal{A}\begin{pmatrix} \eta \\ 0 \end{pmatrix} + \begin{pmatrix} J_S \\ M \end{pmatrix}\right\} \tag{E.53}$$

be a random vector whose distribution is the asymptotic distribution of $\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,S} \\ \widehat{\beta}_{z,S} - \beta_{z0} \end{pmatrix}$, where C_S and D_S, respectively, have the distribution identical to the asymptotic distributions of $\sqrt{n}\widehat{\beta}_{x,S}$ and $\sqrt{n}(\widehat{\beta}_{z,S} - \beta_{z0})$.

Furthermore, we express $\mathcal{A}$ as

$$\mathcal{A} = \begin{pmatrix} A_{xx} & A_{xz} \\ A_{zx} & A_{zz} \end{pmatrix}$$

by making the block matrices A_{xx}, A_{xz}, A_{zx} and A_{zz} be of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$, respectively. Similarly, the inverse matrix of $\mathcal{A}$, $\mathcal{A}_S$ and $\mathcal{A}_S^{-1}$ are expressed as $\mathcal{A}^{-1} = \begin{pmatrix} A^{xx} & A^{xz} \\ A^{zx} & A^{zz} \end{pmatrix}$, $\begin{pmatrix} A_{xxS} & A_{xzS} \\ A_{zxS} & A_{zzS} \end{pmatrix}$ and $\begin{pmatrix} A^{xxS} & A^{xzS} \\ A^{zxS} & A^{zzS} \end{pmatrix}$, respectively.

Consequently, by $(E.53)$, we write

$$C_S = \left(A^{xxS}\pi_S A_{xx} + A^{xzS}A_{zx}\right)\eta + A^{xxS}J_S + A^{xzS}M \tag{E.54}$$

and

$$D_S = \left(A^{zxS}\pi_S A_{xx} + A^{zzS}A_{zx}\right)\eta + A^{zxS}\pi_S J + A^{zzS}M.$$

To continue the proof, we need the following lemma.

Lemma E.2.1 *Under regularity conditions in Appendix E.1, we have*

$$\mathcal{A}_S = \Pi_S \mathcal{A} \Pi_S^\top,$$

where $\mathcal{A}$ and $\mathcal{A}_S$ are the asymptotic covariances matrices in Theorem 6.3.1.

By Lemma E.2.1, we have

$$\begin{aligned} A^{xxS} &= \left(A_{xxS} - A_{xzS}A_{zzS}^{-1}A_{zxS}\right)^{-1} \\ &= \left(\pi_S A_{xx}\pi_S^\top - \pi_S A_{xz}A_{zz}^{-1}A_{zx}\pi_S^\top\right)^{-1} \\ &= \left\{\pi_S\left(A_{xx} - A_{xz}A_{zz}^{-1}A_{zx}\right)\pi_S^\top\right\}^{-1} \\ &= \left\{\pi_S (A^{xx})^{-1}\pi_S^\top\right\}^{-1} \end{aligned} \tag{E.55}$$

and

$$\begin{aligned} A^{xzS} &= -A^{xxS}A_{xzS}A_{zzS}^{-1} \\ &= -A^{xxS}A_{xzS}A_{zzS}^{-1}. \end{aligned} \tag{E.56}$$

Then combining (E.54), (E.55) and (E.56) gives

$$\begin{aligned} C_S &= \left(A^{xxS}\pi_S A_{xx} - A^{xxS}A_{xzS}A_{zzS}^{-1}A_{zx}\right)\eta + A^{xxS}J_S - A^{xxS}A_{xzS}A_{zzS}^{-1}M \\ &= A^{xxS}\pi_S\left(A_{xx} - A_{xz}A_{zzS}^{-1}A_{zx}\right)\eta + A^{xxS}J_S - A^{xxS}A_{xzS}A_{zzS}^{-1}M \\ &= A^{xxS}\pi_S(A^{xx})^{-1}\eta + A^{xxS}\pi_S J - A^{xxS}\pi_S A_{xz}A_{zz}^{-1}M \\ &= A^{xxS}\pi_S(A^{xx})^{-1}\left(\eta + A^{xx}J - A^{xx}A_{xz}A_{zz}^{-1}M\right) \\ &\triangleq A^{xxS}\pi_S(A^{xx})^{-1}\left(\eta + \mathcal{W}\right), \end{aligned} \tag{E.57}$$

where the third equality is due to $J_S = \pi_S J$ and $A^{xx} = \left(A_{xx} - A_{xz}A_{zz}^{-1}A_{zx}\right)^{-1}$,

$$\mathcal{W} = A^{xx}J - A^{xx}A_{xz}A_{zz}^{-1}M, \tag{E.58}$$

and $(A^{xx})^{-1}$ stands for the inverse of matrix A^{xx}.

Similarly,

$$\begin{aligned} A^{zzS} &= \left\{A_{zz} - A_{zx}\pi_S^\top \left(\pi_S A_{xx}\pi_S^\top\right)^{-1}\pi_S A_{xz}\right\}^{-1} \\ &= \left(A_{zz} - A_{zx}A_{xx}^{-1}A_{xz}\right)^{-1} \\ &= A^{zz}, \end{aligned} \tag{E.59}$$

and

$$\begin{aligned} A^{zxS} &= -A^{zzS}A_{zxS}A_{xxS}^{-1} \\ &= -A^{zz}A_{zx}\pi_S^\top A_{xxS}^{-1}. \end{aligned} \tag{E.60}$$

Thus, using (E.59) and (E.60), and direct calculations give

$$\begin{aligned} D_S &= A_{zz}^{-1}A_{zx}\left\{I_{p\times p} - \pi_S^\top A^{xxS}\pi_S\left(A^{xx}\right)^{-1}\right\}\eta + A_{zz}^{-1}M - A_{zz}^{-1}A_{zx}\pi_S^\top\left(\pi_S A_{xx}\pi_S^\top\right)^{-1}\pi_S J \\ &\quad + A_{zz}^{-1}A_{zx}\pi_S^\top\left(A_{xxS}\right)^{-1}\pi_S A_{xz}A_{zz}^{-1}M \\ &= A_{zz}^{-1}A_{zx}\left[I_{p\times p} - (A^{xx})^{1/2}(A^{xx})^{-1/2}\pi_S^\top\left\{\pi_S(A^{xx})^{-1}\pi_S^\top\right\}^{-1}\pi_S(A^{xx})^{-1/2}(A^{xx})^{-1/2}\right]\eta \\ &\quad + A_{zz}^{-1}M - A_{zz}^{-1}A_{zx}\pi_S^\top\left\{\pi_S(A^{xx})^{-1}\pi_S^\top\right\}^{-1}\pi_S\left(J - A_{xz}A_{zz}^{-1}M\right) \\ &= A_{zz}^{-1}A_{zx}\left[I_{p\times p} - (A^{xx})^{1/2}(A^{xx})^{-1/2}\pi_S^\top\left\{\pi_S(A^{xx})^{-1}\pi_S^\top\right\}^{-1}\pi_S(A^{xx})^{-1/2}(A^{xx})^{-1/2}\right]\eta \\ &\quad + A_{zz}^{-1}M - \Big[A_{zz}^{-1}A_{zx}(A^{xx})^{1/2}(A^{xx})^{-1/2}\pi_S^\top\left\{\pi_S(A^{xx})^{-1}\pi_S^\top\right\}^{-1} \\ &\quad \times\pi_S(A^{xx})^{-1/2}(A^{xx})^{-1/2}\left\{A^{xx}J - A^{xx}A_{xz}A_{zz}^{-1}M\right\}\Big] \\ &\triangleq A_{zz}^{-1}A_{zx}\left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\eta + A_{zz}^{-1}M \\ &\quad - A_{zz}^{-1}A_{zx}(A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\mathcal{W}, \end{aligned} \tag{E.61}$$

where the second equality is due to (E.55), and

$$\mathbb{H}_S = (A^{xx})^{-1/2}\pi_S^\top\left\{\pi_S(A^{xx})^{-1}\pi_S^\top\right\}^{-1}\pi_S(A^{xx})^{-1/2}.$$

Step 2:

Applying the Taylor expansion to $\widehat{\mu}_S$ and μ_{true} around $(0, \beta_{z0})$, respectively, gives

$$\widehat{\mu}_S - \mu\left(0, \beta_{z0}\right) = \left(\frac{\partial\mu}{\partial\beta_{x,S}}\right)^\top\widehat{\beta}_{x,S} + \left(\frac{\partial\mu}{\partial\beta_{z,S}}\right)^\top\left(\widehat{\beta}_{z,S} - \beta_{z0}\right) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{E.62}$$

and

$$\mu_{true} - \mu\left(0, \beta_{z0}\right) = \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \frac{\eta}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{E.63}$$

Combining $(E.62)$ and $(E.63)$ gives

$$\begin{aligned}
& \sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right) \\
= \quad & \left(\frac{\partial \mu}{\partial \beta_{x,S}}\right)^\top \sqrt{n}\widehat{\beta}_{x,S} + \left(\frac{\partial \mu}{\partial \beta_{z,S}}\right)^\top \sqrt{n}(\widehat{\beta}_{z,S} - \beta_{z0}) - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta + o_p\left(1\right) \\
\xrightarrow{d} \quad & \left(\frac{\partial \mu}{\partial \beta_{x,S}}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \beta_{z,S}}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta \text{ as } n \to \infty.
\end{aligned} \tag{E.64}$$

Then plugging in expressions (E.57) and (E.61) to $(E.64)$ and applying $\frac{\partial \mu}{\partial \beta_{x,S}} = \pi_S \frac{\partial \mu}{\partial \beta_x}$ and $\frac{\partial \mu}{\partial \beta_{z,S}} = \frac{\partial \mu}{\partial \beta_z}$ yield that as $n \to \infty$,

$$\begin{aligned}
& \sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right) \\
\xrightarrow{d} \quad & \left(\frac{\partial \mu}{\partial \beta_{x,S}}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \beta_{z,S}}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta \\
= \quad & \left(\pi_S \frac{\partial \mu}{\partial \beta_x}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta \\
= \quad & \left(\pi_S \frac{\partial \mu}{\partial \beta_x}\right)^\top A^{xxS} \pi_S (A^{xx})^{-1}\left(\eta + \mathcal{W}\right) \\
& + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} A_{zx}\left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\right\}\eta + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M \\
& - \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} A_{zx} (A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\mathcal{W} - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta \\
= \quad & \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \left\{(A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\right\}\left(\eta + \mathcal{W}\right) \\
& + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} A_{zx}\left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\right\}\eta + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M \\
& - \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} A_{zx} (A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\mathcal{W} - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta.
\end{aligned} \tag{E.65}$$

where (E.57) and (E.61) are used in the third equality, and the last equality is obtained by using the formulation of $\mathbb{H}_S$. Then combining common terms together, the last expression of (E.65) can be re-written as

$$\begin{aligned}
& \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\eta \\
& + \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1}A_{zx}\left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\eta \\
& + \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \left\{(A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\mathcal{W} \\
& - \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1}A_{zx}\left\{(A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\mathcal{W} \\
= \; & \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M \\
& + \left(\frac{\partial \mu}{\partial \beta_x} - A_{zx}^\top A_{zz}^{-1}\frac{\partial \mu}{\partial \beta_z}\right)^\top \left[\left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\eta\right. \\
& \left. - \left\{I_{p\times p} - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\right\}\mathcal{W}\right] \\
\triangleq \; & \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M + \omega^\top\left\{\eta - (A^{xx})^{1/2}\mathbb{H}_S(A^{xx})^{-1/2}\mathcal{U}\right\},
\end{aligned}$$

where $\omega = \frac{\partial \mu}{\partial \beta_x} - A_{zx}^\top A_{zz}^{-1}\frac{\partial \mu}{\partial \beta_z}$ and $\mathcal{U} = \eta + \mathcal{W}$, which completes the proof.

Proof of (b):

The proof for (b) is similar to that of (a); the only difference is to include $\Lambda_0(\cdot)$ and its estimator in the derivations, and we view $\Lambda_0(\cdot)$ as a parameter in the same way as we do for β_x and β_z in (a). In this case, the Taylor expansion becomes

$$\begin{aligned}
\widehat{\mu}_S - \mu_{true} \; = \; & \left(\frac{\partial \mu}{\partial \beta_{x,S}}\right)^\top \widehat{\beta}_{x,S} + \left(\frac{\partial \mu}{\partial \beta_{z,S}}\right)^\top \left(\widehat{\beta}_{z,S} - \beta_{z0}\right) \\
& + \frac{\partial \mu}{\partial \Lambda_0}\left(\widehat{\Lambda}_{0,S} - \Lambda_0\right) - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \frac{\eta}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{E.66}$$

Multiplying $\sqrt{n}$ on both sides and plugging in the results of Theorems 6.3.1 and 6.3.2

to (E.66) give that as $n \to \infty$,

$$\begin{aligned}
& \sqrt{n}\,(\widehat{\mu}_S - \mu_{true}) \\
\xrightarrow{d} \;& \left(\frac{\partial \mu}{\partial \beta_{x,S}}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \beta_{z,S}}\right)^\top D_S \\
& + \frac{\partial \mu}{\partial \Lambda_0}\left[\mathcal{V}(t) - \{F_{x,S}(t)\}^\top C_S - \{F_z(t)\}^\top D_S + \{F_x(t)\}^\top \eta\right] - \left(\frac{\partial \mu}{\partial \beta_x}\right)^\top \eta \\
= \;& \frac{\partial \mu}{\partial \Lambda_0}\mathcal{V}(t) + \left\{\frac{\partial \mu}{\partial \beta_{x,S}} - \frac{\partial \mu}{\partial \Lambda_0}F_{x,S}(t)\right\}^\top C_S + \left\{\frac{\partial \mu}{\partial \beta_z} - \frac{\partial \mu}{\partial \Lambda_0}F_z(t)\right\}^\top D_S \\
& - \left\{\frac{\partial \mu}{\partial \beta_x} + \frac{\partial \mu}{\partial \Lambda_0}F_x(t)\right\}^\top \eta.
\end{aligned}$$

Similar to the proof of (a), we derive that as $n \to \infty$,

$$\begin{aligned}
\sqrt{n}\,(\widehat{\mu}_S - \mu_{true}) \;\xrightarrow{d}\; & \frac{\partial \mu}{\partial \Lambda_0}\mathcal{V}(t) + \left\{\frac{\partial \mu}{\partial \beta_z} - \frac{\partial \mu}{\partial \Lambda_0}F_z(t)\right\}^\top A_{zz}^{-1} M \\
& + (\omega + \kappa)^\top \left\{\eta - (A^{xx})^{1/2}\mathbb{H}_S (A^{xx})^{-1/2}\mathcal{U}\right\},
\end{aligned}$$

where $\omega = \frac{\partial \mu}{\partial \beta_x} - A_{zx}^\top A_{zz}^{-1}\frac{\partial \mu}{\partial \beta_z}$ and $\kappa = \frac{\partial \mu}{\partial \Lambda_0}F_x(t) - A_{zx}^\top A_{zz}^{-1}\frac{\partial \mu}{\partial \Lambda_0}F_z(t)$, which completes the proof.

□

E.2.6 Proof of Theorem 6.3.4

Proof of (a):

Recall that $\widehat{\mu}_{ave} = \sum_{S \in \mathcal{S}} w(S|\widehat{\eta})\widehat{\mu}_S$ with weights $w(S|\widehat{\eta})$ satisfying conditions in Section 6.3.4. Since $\widehat{\eta} = \sqrt{n}\widehat{\beta}_x$, then by Theorem 6.3.1 (a), we have that as $n \to \infty$,

$$\begin{aligned}
\widehat{\eta} &= \sqrt{n}\widehat{\beta}_{x,full} \\
&= (I_{p\times p}, 0)\sqrt{n}\begin{pmatrix} \widehat{\beta}_{x,full} \\ \widehat{\beta}_{z,full} - \beta_{z0} \end{pmatrix} \\
&\xrightarrow{d} (I_{p\times p}, 0)\left\{\begin{pmatrix} \eta \\ 0 \end{pmatrix} + \mathcal{A}^{-1}\begin{pmatrix} J \\ M \end{pmatrix}\right\} \\
&= \eta + \mathcal{W} \\
&= \mathcal{U}. \qquad \text{(E.67)}
\end{aligned}$$

Therefore, let $w(S|\mathcal{U})$ denote the weight to which $w(S|\widehat{\eta})$ converges.

Then by the result of Theorem 6.3.3 (a) and (E.67), we have that as $n \to \infty$,

$$\begin{aligned}
\sqrt{n}\left(\widehat{\mu}_{ave} - \mu_{true}\right) &= \sum_{S \in \mathcal{S}} w(S|\widehat{\eta}) \left\{\sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right)\right\} \\
&\xrightarrow{d} \sum_{S \in \mathcal{S}} w(S|\mathcal{U}) \left[\left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M + \omega^\top \left\{\mathcal{U} - (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\}\right] \\
&= \left(\frac{\partial \mu}{\partial \beta_z}\right)^\top A_{zz}^{-1} M + \omega^\top \left\{\mathcal{U} - \sum_{S \in \mathcal{S}} w(S|\mathcal{U}) (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\}.
\end{aligned}$$

Proof of (b):

Similar to (a), using $\widehat{\mu}_{ave} = \sum_{S \in \mathcal{S}} w(S|\widehat{\eta}) \widehat{\mu}_S$ with $\sum_{S \in \mathcal{S}} w(S|\widehat{\eta}) = 1$ and applying Theorem 6.3.3 (b) give that as $n \to \infty$,

$$\begin{aligned}
\sqrt{n}\left(\widehat{\mu}_{ave} - \mu_{true}\right) &= \sum_{S \in \mathcal{S}} w(S|\widehat{\eta}) \left\{\sqrt{n}\left(\widehat{\mu}_S - \mu_{true}\right)\right\} \\
&\xrightarrow{d} \sum_{S \in \mathcal{S}} w(S|\mathcal{U}) \left[\frac{\partial \mu}{\partial \Lambda_0} \mathcal{V}(t) + \left\{\frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t)\right\}^\top A_{zz}^{-1} M \right. \\
&\qquad \left. + (\omega + \kappa)^\top \left\{\mathcal{U} - (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\}\right] \\
&= \frac{\partial \mu}{\partial \Lambda_0} \mathcal{V}(t) + \left\{\frac{\partial \mu}{\partial \beta_z} + \frac{\partial \mu}{\partial \Lambda_0} F_z(t)\right\}^\top A_{zz}^{-1} M \\
&\qquad + (\omega + \kappa)^\top \left\{\mathcal{U} - \sum_{S \in \mathcal{S}} w(S|\mathcal{U}) (A^{xx})^{1/2} \mathbb{H}_S (A^{xx})^{-1/2} \mathcal{U}\right\}.
\end{aligned}$$

Therefore, the proof of Theorem 6.3.4 is completed. □

E.2.7 Proof of Lemma 6.3.3

By an argument similar to (E.56), we have $A^{xz} = -A^{xx} A_{xz} A_{zz}^{-1}$, or equivalently,

$$(A^{xx})^{-1} A^{xz} = -A_{xz} A_{zz}^{-1}.$$

We re-write $(E.58)$ as

$$\begin{aligned}\mathcal{W} &= A^{xx}J - A^{xx}A_{xz}A_{zz}^{-1}M \\ &= A^{xx}\left(J - A_{xz}A_{zz}^{-1}M\right) \\ &= A^{xx}\left\{J + (A^{xx})^{-1}A^{xz}M\right\}. \end{aligned} \tag{E.68}$$

Write $\mathcal{B}$ as the block matrix $\begin{pmatrix} B_{xx} & B_{xz} \\ B_{zx} & B_{zz} \end{pmatrix}$ where B_{xx}, B_{xz}, B_{zx} and B_{zz} of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$, respectively. Noting that $\mathcal{A}^{-1}$ is a symmetric matrix, then $(A^{xx})^\top = A^{xx}$, $(A^{xz})^\top = A^{zx}$, $(A^{zx})^\top = A^{xz}$ and $(A^{zz})^\top = A^{zz}$. From $(E.68)$, the variance of $\mathcal{W}$ can be expressed as

$$\begin{aligned}\text{var}\left(\mathcal{W}\right) &= A^{xx}\text{var}\left\{J + (A^{xx})^{-1}A^{xz}M\right\}A^{xx} \\ &= A^{xx}\text{var}\left(J\right)A^{xx} + A^{xx}(A^{xx})^{-1}A^{xz}\text{var}\left(M\right)A^{zx}(A^{xx})^{-1}A^{xx} \\ &\quad + A^{xx}\text{cov}\left\{J, (A^{xx})^{-1}A^{xz}M\right\}A^{xx} + A^{xx}\text{cov}\left\{(A^{xx})^{-1}A^{xz}M, J\right\}A^{xx} \\ &= A^{xx}B_{xx}A^{xx} + A^{xz}B_{zz}A^{zx} \\ &\quad + A^{xx}B_{xz}A^{zx}(A^{xx})^{-1}A^{xx} + A^{xx}(A^{xx})^{-1}A^{xz}B_{zx}A^{xx} \\ &= A^{xx}B_{xx}A^{xx} + A^{xz}B_{zz}A^{zx} + A^{xz}B_{zx}A^{xx} + A^{xx}B_{xz}A^{zx}. \end{aligned}$$

On the other hand, directly calculations give

$$\begin{aligned}\mathcal{A}^{-1}\mathcal{B}\mathcal{A}^{-1} &= \begin{pmatrix} A^{xx} & A^{xz} \\ A^{zx} & A^{zz} \end{pmatrix}\begin{pmatrix} B_{xx} & B_{xz} \\ B_{zx} & B_{zz} \end{pmatrix}\begin{pmatrix} A^{xx} & A^{xz} \\ A^{zx} & A^{zz} \end{pmatrix} \\ &= \begin{pmatrix} A^{xx}B_{xx} + A^{xz}B_{zx} & A^{xx}B_{xz} + A^{xz}B_{zz} \\ A^{zx} & A^{zz} \end{pmatrix}\begin{pmatrix} A^{xx} & A^{xz} \\ A^{zx} & A^{zz} \end{pmatrix} \end{aligned}$$

leading to the upper left block matrix $\sigma_{xx} = A^{xx}B_{xx}A^{xx} + A^{xz}B_{zx}A^{xx} + A^{xx}B_{xz}A^{zx} + A^{xz}B_{zz}A^{zx}$, which is $\text{var}\left(\mathcal{W}\right)$; the proof is then completed. □