RECENT ADVANCES IN STATISTICAL MODELS: TOPICS IN MODEL SELECTION AND SEMI-PARAMETRIC INFERENCE

BY WENQIAN QIAO

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Minge Xie

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

Recent advances in statistical models: Topics in model selection and semi-parametric inference

by WENQIAN QIAO

Dissertation Director: Minge Xie

This dissertation consists of three chapters. It develops new methodologies to address two specific problems of recent statistical research:

- How to incorporate hierarchical structure in high dimensional regression model selection.
- How to achieve semi-parametric efficiency in the presence of missing data.

For the first problem, we provide a new approach to explicitly incorporate a given hierarchical structure among the predictors into high dimensional regression model selection. The proposed estimation approach has a *hierarchical grouping property* so that a pair of variables that are "close" in the hierarchy will be more likely grouped in the estimated model than those that are "far away". We also prove that the proposed method can consistently select the true model. These properties are demonstrated numerically in simulation and a real data analysis on peripheral-blood mononuclear cell (PBMC) study.

For the second problem, two frameworks are considered: *generalized partially linear model (GPLM)* and *causal inference of observational study*. Specifically, under the GPLM framework, we consider a broad range of missing patterns which subsume most publications on the same topic. We use the concept of least favorable curve and extend the generalized profile likelihood approach [Severini and Wong (1992)] to estimate the parametric component of the model, and prove that the proposed estimator is consistent and semi-parametrically efficient. Also, under the causal inference framework, we propose to estimate the mean treatment effect with non-randomized treatment exposures in the presence of missing data. An appealing aspect of this development is that we incorporate the post-baseline covariates which are often excluded from causal effect inference due to their inherent confounding effect with treatment. We derive the semiparametric efficiency bound for regular asymptotically linear (RAL) estimators and propose an estimator which achieves this bound. Moreover, we prove that the proposed estimator is robust against four types of model mis-specifications. The performance of the proposed approaches are illustrated numerically through simulations and real data analysis on group testing dataset from Nebraska Infertility Prevention Project and burden of illness dataset from Duke University Medical Center.

Acknowledgements

First of all, I would like to express my gratitude to my advisor, Professor Minge Xie for his guidance and continuous support during my years of research. He has introduced me wide variety of research topics and give me the opportunity to work on many interesting projects. He also has spent much of his valuable time listening and understanding my questions and providing me great ideas and encourages all the time.

My thanks also go to the Department of Statistics and Biostatistics of Rutgers University for providing me support and a great learning and research environment.

Finally, I would like to thank my wife Lu Xin and my parents Ming Qiao, Zhiping Guan for their understanding and unwavering support throughout my studies.

Dedication

To my wife Lu Xin and my parents Ming Qiao, Zhiping Guan.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

The recent advances in computational power have greatly facilitated the collection of massive data. However, the increasing magnitude of the dataset poses many challenges including: 1) redundancy; 2) missingness.

Redundancy refers to the fact that, even though more data are collected, only a small fraction of them contain the information we need to draw statistical conclusions. The redundant or nuisance information does not help to improve the inference. Contrarily, it behaves as additional noise which reduces the accuracy and predictability of the statistical conclusion. To tackle this problem, we need to identify the informative data and get rid of redundant ones. In the ordinary linear regression framework, such feature selection problem has been extensively discussed and penalized regression has played a significant role to provide consistent model selection approaches. One commonly occurring scenario when the number of covariates exceeds the number of observation is that the covariates tend to be highly correlated, and it is shown that the correlation should be taken into account in the model selection procedure to produce a stable and consistent result. Several important methods have been established for this particular application, including Elastic Net (Enet) method [Zou and Hastie (2005)], the OSCAR (octagonal shrinkage and clustering algorithm for regression) approach [Bondell and Reich (2008)], the Mnet method [Huang et al (2010)], and the SLS method [Huang et al (2010)].

It is also common that the covariates in some studies are hierarchically associated or correlated. It is prudent to utilize this type of information in our analysis and model selection. To date, no methods in the existing literatures are able to fully utilize the hierarchy information in the model selection problem. In this dissertation, we propose a novel approach to incorporate the hierarchical structures of covariates into the variable selection problem. We illustrate our development using a Enet-type of penalty, which is one of the most widely used methods that encourage sparsity and grouping simultaneously. It also consistently selects the true model under certain condition; c.f [Jia and Yu (2010)].

In order to clearly introduce our scoring scheme without other complications, we start with a simplified scenario in which each terminal node of the hierarchical tree represents only one predictor. In this setting, each predictor will be assigned a score which is derived by the hierarchical structure. This scoring scheme is used to quantify the hierarchical information among the predictors. Then, we integrate the hierarchical score into the Enet penalty function. It can be shown that the resulting procedure not only performs model selection and estimation simultaneously, but also enjoys a desired feature, called "hierarchical grouping property", which can be loosely stated as follows:

- Two highly correlated variables in the same hierarchical cluster will have the same chance to be added or dropped from the model.

- Two highly correlated variables which are "close" to each other in the hierarchical tree will more likely be included or dropped together in the estimated model than those which are "far away".

This hierarchical group property will be formally defined in Section 2.2.1.

Often, the terminal nodes of a hierarchical tree contain multiple variables; see, e.g., [Breiman et al. (1984)]. We generalize the idea to the scenario in which each terminal node can contain potentially multiple predictors. The hierarchical

grouping property parallel to the simplified case is established and model selection consistency in both between and within group level is proved.

As opposed to the redundancy problem described above, when the dataset grows in dimension, the completeness of the dataset is sometimes compromised. The data entries on certain dimensions could be missing or coarsened. Traditional methodologies relying on the complete dataset won't work in the presence of missing data. New techniques have been developed to handle this problem especially under the parametric setting. In this dissertation, we focus on the semi-parametric setting with missing data under two frameworks: generalized partially linear model (GPLM) and causal inference of observational study.

Under the GPLM framework, there are very few existing discussions in the presence of missing data, except in some missing at random (MAR) cases; c.f., [Wang, Linton and Härdle (2004)] and [Wang, Rotnitzky and Lin (2010)]. The MAR patterns in [Wang, Linton and Härdle (2004)] and [Wang, Rotnitzky and Lin (2010)] are very simple and limited, and they only include the case where each outcome is independently observed or missing according to certain simple independent MAR probability models. We consider in this dissertation a very general missing structure similar to the one studied in [Wu (1983)]. In particular, we assume that there is a *many to one mapping* from the unobservable complete data to the observed data. We use the concept of least favorable curve and extend the generalized profile likelihood approach ([Severini and Wong (1992)]) to estimate the parametric component. It is shown that the estimator is consistent and semi-parametric efficient under some regularity conditions. We also develop a computing algorithm, which runs iteratively between fitting parametric part by a semiparametric estimating equation and fitting nonparametric part by smoothing techniques. Because of missing data, EM algorithm ([Dempster, Laird and Rubin (1977)]) is used to carry out the estimations, and has once again been proven to be an effective tool for a wide range

of different types of missing data. This work is a continuation of the development in [Li (2009)] in which the partially linear model(PLM), a special case of our framework, is discussed. The regularity conditions for the consistency and semi-parametric efficiency are made less stringent and the numerical algorithms are re-designed due to the more general likelihood structure in GPLM.

Under the casual inference framework, we consider the observational study where the treatment exposures are not randomized but associated with demographic and physiologic characteristics of the study subjects. Thus, inference for the treatment effect should account for the potential imbalance in baseline covariates between treatment and control groups. In addition, we assume the response could be missing and the missing mechanism depends on treatment assignment, baseline variables, and post-baseline variables. The introduction of post-baseline variables is novel because, conventionally, such variables are often excluded from causal effect inference, due to their inherent confounding effect with the treatment. To tease out the confounding effect of post-baseline variables on the missing mechanism, and on the treatment effect of interest, we structure our problem using the potential outcome setup. Given the reasonable ignorability assumptions, we derive the semiparametric efficiency bound of RAL estimators for treatment effect, and propose a consistent and asymptotically efficient estimator inspired by the efficiency bound. This estimator is in the family of augmented inverse probability-weighted estimators. It improves upon existing methods by incorporating post-baseline variables in missing data mechanism and potential outcome model. Furthermore, we show that this estimator is quadruple paired robust, in the sense that it is consistent as long as one of the following four pairs of models are correctly specified: $\{M1, M2\}, \{M1, M3\}, \{M2, M4\}, \{M3, M4\}$, where $M1$ is the missing mechanism; $M2$ is the treatment assignment mechanism; $M3$ is the conditional distribution of potential responses given baseline variables; $M4$ is the conditional distribution of potential responses given baseline and post-baseline

variables. This property is an extension of the well-known double-robustness in missing data and causal inference models [Bang and Rubins (2005)]. This work is a continuation of the development in [Shentu (2006)]

The rest of the dissertation is organized as follows. In Chapter 2, the work on model selection under the hierarchical structure is summarized. It begins with a simplified scenario where each terminal node represents one predictor to illustrate the idea behind the proposed method. Then the generalized case is considered and parallel theoretical results are developed. Numerical simulation and a real data analysis on SLE dataset are carried out to test the performance of the estimator. Chapter 3 focuses on the semi-parametric efficient estimation under GPLM with missing data. We prove that the estimator is consistent and efficient. Also, an iterative algorithm is developed to carry out the estimation procedure. The proposed approach is benchmarked against existing estimators in the simulation study and an analysis on Nebraska IPP dataset. In Chapter 4, we study the missing data problem under causal inference framework. The semi-parametric efficiency bound and an estimator which is consistent and asymptotic efficient is derived. Moreover, we prove that the proposed estimator is robust against four types of model mis-specification. The theoretical results are illustrated by simulation studies and a real data analysis on the burden of illness data. Some concluding remarks are given in Chapter 5. We relegate technical proofs to the Appendix.

Chapter 2

Model Selection under Hierarchical Structure using Elastic Net

2.1 Background and Motivation

Consider the ordinary linear regression model with n observations and p predictors:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{2.1.1}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ deterministic design matrix, $\beta = (\beta_1, \cdots, \beta_p)^T$ is the corresponding regression coefficients and ε is the vector of independent random errors. In the case when p is large (comparable to or much larger than n) and β is sparse (in the sense that many of its entries are 0), penalized regression has played a significant role to provide consistent model selection approaches. General methodologies include, for example, Lasso [Tibshirani (1996)], SCAD [Fan and Li (2001), Fan and Lv (2011)], SIS (Sure independence screening) and two-scale method [Fan and Lv (2008)], MCP [Zhang (2010)] and many others.

It is common that the covariates in some studies are hierarchically associated. For instance, most organizational management structures are naturally hierarchical. In evolution biology, evolutionary tree is a branching diagram of the evolutionary relationships among various biological species based upon similarities and differences in their physical and genetic characteristics. In cell biology and genetic studies, genes are often characterized into groups, many times with

hierarchical layers, based on their biological characteristics or genetic functions. To incorporate the hierarchical structures of covariates into the variable selection problem, we illustrate our development using a Enet-type of penalty, which is defined as

$$\hat{\beta} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|D_1\beta\|_1 + \lambda_2 \beta^T D_2 \beta, \tag{2.1.2}$$

where D_1, D_2 are p by p matrix containing hierarchical information of the covariates. When $D_1 = D_2 = I_p$, (2.1.2) is the well-known vanilla Enet

$$\hat{\beta} = \text{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T \beta. \tag{2.1.3}$$

Our motivation comes from the peripheral-blood mononuclear cell (PBMC) study. The objective of the study is to eliminate the trivial genes and identify the important genes which can be used to predict the Systemic Lupus Erythematosus disease-activity index (SLEDAI) among 4779 potential candidates. Given such a great number of predictors, however, we have only 47 individual samples. According to [Chaussabel et al. (2008)], those 4779 genes are distributed among 28 modules (groups). The transcripts within each modules are highly correlated. On top of these 28 modules, we can also obtain a hierarchical structure with each terminal node representing a single module; see Figure 2.1. The challenge is how to take advantage of the hierarchical structure in the model selection procedure which selects variables at both group and individual level.

Note that a hierarchical structure can often be represented by a tree graph; see, for examples, Figures 2.1 and 2.2. Before we end this section, we introduce some terminologies which will be used throughout this chapter. We refer to the top node of the tree as the *root* which represents the entire collection of the predictors. The nodes at the bottom of tree are the *terminal nodes*. The nodes (splits) in between are the *internal nodes* which, viewed from bottom-up, show how individual predictors are grouped and how the groups are merged into supergroups. The concept *depth* is defined for each node as its position in the

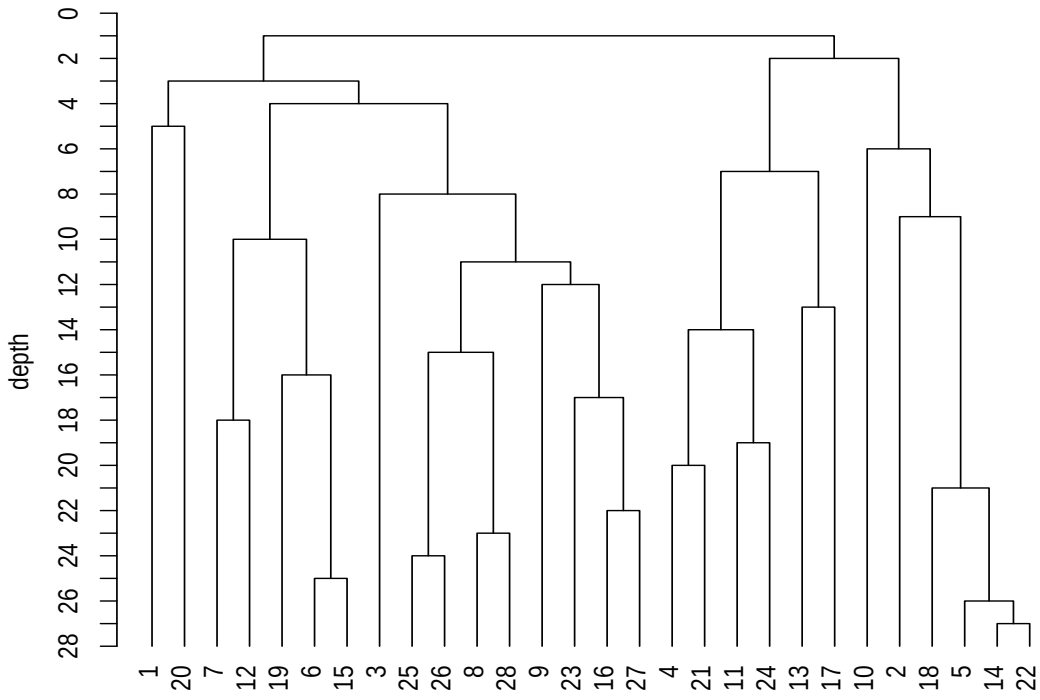


Figure 2.1: Hierarchical structure for SLE dataset. Each of the 28 terminal nodes contains a group (module) of genes.

sequence of splitting when viewed top-down. The root node, which represents the first split, has depth 1. Then the node representing the second split has depth 2, etc. See the vertical coordinates in Figures 2.1 and 2.2. We assume in this chapter that the entire hierarchical structure is known and the depth values are given for each given tree. The depth values can be obtained from different ways depending on the practice. For instance, the classical divisive hierarchical clustering algorithms recursively divide one of the existing clusters into two sub-clusters at each iteration; Thus, the iteration order can be treated as the depth. In the evolutionary tree, each splitting of lineages can be arranged in chronological order; In this case the depth values can be derived from the time

of origination of each new species. Also, to simplify our presentations and following [Zou and Hastie (2005)] and [Bondell and Reich (2008)], we assume that each predictor has been standardized before the data analysis throughout this chapter:

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = n. \tag{2.1.4}$$

The rest of the chapter is organized as follows. In Section 2.2, we define the "hierarchical grouping property" and construct the "hierarchical score" along with the corresponding "hierarchical Elastic Net" estimator for the simplified scenario where each terminal node represents only one predictor. We also prove that the proposed estimator has the "hierarchical grouping property" and can provide consistent result in model selection. Section 2.3 extends the results to the case where each terminal node corresponds to a group of predictors. Computational algorithm is illustrated in Section 2.4. Numerical studies including simulation and real data analysis are carried out in Section 2.5.

2.2 Hierarchical variable selection when a terminal node contains only a single predictor

We study in this section a simple scenario in which each terminal node of the hierarchical tree represents only one predictor. In this case, to account for the hierarchical structure in the model selection procedure, we focus on a special case of (2.1.2) where D_1 and D_2 are diagonal matrices. We design a *hierarchical score* s_j for each predictor $\mathbf{x}_j$, $j = 1, \cdots, p$. Here, s_j's are positive numbers and determined by the given hierarchy. We treat these "hierarchical scores" as a set of weights on the penalty terms in the Enet (2.1.3). As a result, we propose the following "hierarchical Enet (HEnet)" estimator:

$$\hat{\beta} = \text{argmin}_\beta \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|S^{-1}\beta\|_1 + \lambda_2 \beta^T S^{-1} \beta. \tag{2.2.1}$$

where λ_1, λ_2 are the tuning parameters and $S = \text{diag}\{s_1, \cdots, s_p\}$. Here, $s_i > 0$ and we also assume that $\max_{1 \le i \le p} s_i$ is bounded above. When all the s_j's are equal to 1, (2.2.1) reduces to the conventional Enet approach (2.1.3) which has been extensively studied in the literature.

We first introduce in this section the definition of hierarchical grouping property and an approach to construct hierarchical scores s_j. We then study the properties of the proposed estimator (2.2.1). These developments will be modified and extended to accommodate the general scenario in which each terminal node contains a group of predictors in Section 2.3.

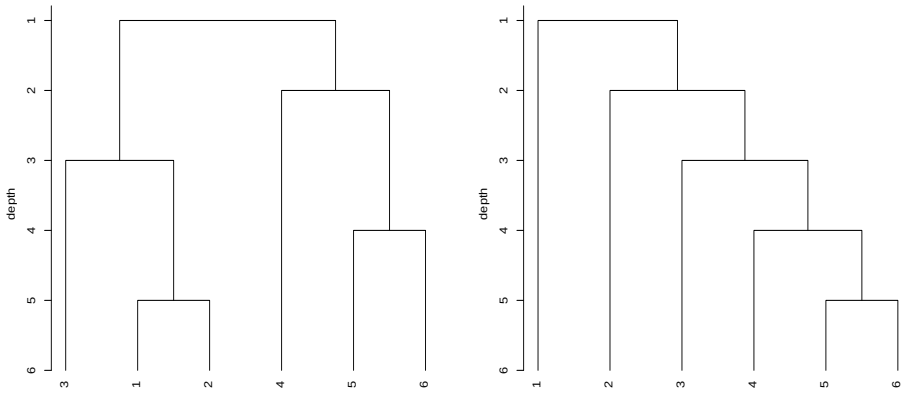
2.2.1 Hierarchical grouping property

The conventional Enet estimator (2.1.3) enjoys the *grouping property* in the sense that the highly correlated variables tend to be selected or dropped together in the estimated model. In the presence of hierarchical structure, the new procedure (2.2.1) also performs model selection and variable grouping simultaneously. The grouping property in this new proposal is in compliance with the hierarchy of the given tree. We refer it as *hierarchical grouping property* and it is a generalization of the grouping property in [Zou and Hastie (2005)]. To formally describe this property, we first provide the following definition of *ancestors*, which is used to help determine the closeness of a pair of predictor variables in a given tree.

Definition 2.2.1 (**Ancestors**)**.** For a predictor variable $\mathbf{x}$ in a given hierarchy structure, we define *ancestors* of this single variable $\mathbf{x}$ as the set of all the nodes on the branch of $\mathbf{x}$. We also define the *ancestors* of a set of predictors $\{\mathbf{x}_i, i \in I\}$ as the overlapping ancestors of $\mathbf{x}_i$ over $i \in I$.

For examples, the ancestors of predictor $\mathbf{x}_1$ (corresponding to the terminal node 1) in Figure 2.2(a) below are the nodes with depths $\{1, 3, 5\}$. The ancestors of predictor $\mathbf{x}_3$ (corresponding to the terminal node 3) are the nodes with depths

$\{1, 3\}$. The ancestors of the set of predictors $\{\mathbf{x}_1, \mathbf{x}_3\}$ are the nodes with depths $\{1, 3\} = \{1, 3, 5\} \cap \{1, 3\}$. Similarly, in Figure 2.2(b), the ancestors of predictor $\mathbf{x}_1$, predictor $\mathbf{x}_3$ and predictor set $\{\mathbf{x}_1, \mathbf{x}_3\}$ are the nodes with depths $\{1\}$, $\{1, 2, 3\}$ and $\{1\} = \{1\} \cap \{1, 2, 3\}$, respectively.



(a) Example1(a) (b) Example1(b)

Figure 2.2: Hierarchical structures for the the simulated data in Example 1 of Section 4. Each of the six terminal nodes in (a) and (b) has only one predictor variable; i.e., $\mathbf{x}_1, ..., \mathbf{x}_6$, respectively.

We formally state the *hierarchical grouping property* as follows:

Definition 2.2.2 (**Hierarchical grouping property**)**.** Let us call predictor variables are *grouped*, if they are selected or dropped together by the model selection procedure. The *hierarchical grouping property* then refers to:

P1. For two predictors sharing the same ancestors, they are likely grouped together if they are highly correlated.

P2. For any three predictors, say $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$, with same pairwise correlation, if $\mathbf{x}_i$ and $\mathbf{x}_j$ share more ancestors than $\mathbf{x}_i$ and $\mathbf{x}_k$ (i.e. $\mathbf{x}_j$ is closer to $\mathbf{x}_i$ than $\mathbf{x}_k$ on the hierarchy), then $\mathbf{x}_i$ and $\mathbf{x}_j$ are more likely grouped than $\mathbf{x}_i$ and $\mathbf{x}_k$.

The P1 property is similar to the grouping property discussed in [Zou and Hastie (2005)] and [Bondell and Reich (2008)]. The P2 property is in compliance with hierarchical structure: a pair of variables that are "close" in the hierarchy will be more likely grouped in the estimated model than those that are "far away" if both pairs have the same correlation.

To achieve the goal of hierarchical grouping property, the proposed hierarchical scores (i.e., s_i's) used in our proposed approach (2.2.1) need to satisfy certain conditions. First, let us study how the grouping property in Enet is realized. By take derivatives on (2.1.3) of Enet with respect to β_j and β_k, respectively, we have

$$-2\mathbf{x}_j'\{\mathbf{y}-\mathbf{X}\hat{\beta}\}+\lambda_1\text{sgn}\{\hat{\beta}_j\}+2\lambda_2\hat{\beta}_j=0$$
$$-2\mathbf{x}_k'\{\mathbf{y}-\mathbf{X}\hat{\beta}\}+\lambda_1\text{sgn}\{\hat{\beta}_k\}+2\lambda_2\hat{\beta}_k=0$$

assuming $\hat{\beta}_j\hat{\beta}_k\neq 0$. Under the condition $\hat{\beta}_j\hat{\beta}_k>0$, subtracting above two equations and combining with the fact that $\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2\leq\|\mathbf{y}\|_2$, we have

$$|\hat{\beta}_j-\hat{\beta}_k|\leq\frac{\|\mathbf{y}\|_2}{\lambda_2}\|\mathbf{x}_j-\mathbf{x}_k\|_2=\frac{\|\mathbf{y}\|_2}{\lambda_2}\sqrt{2n(1-\phi_{jk})} \tag{2.2.2}$$

where $\phi_{jk}=\text{cor}(\mathbf{x}_j,\mathbf{x}_k)$. The inequality (2.2.2) indicates that as $\phi_{jk}\to 1$, $|\hat{\beta}_j-\hat{\beta}_k|\to 0$; thus, the grouping property in Enet is realized.

For the HEnet (2.2.1), by taking derivative on the penalized function, we have

$$-2\mathbf{x}_j'\{\mathbf{y}-\mathbf{X}\hat{\beta}\}+\lambda_1\text{sgn}\{\hat{\beta}_j\}/s_j+2\lambda_2\hat{\beta}_j/s_j=0$$

Or, equivalently,

$$-2s_j\mathbf{x}_j'\{\mathbf{y}-\mathbf{X}\hat{\beta}\}+\lambda_1\text{sgn}\{\hat{\beta}_j\}+2\lambda_2\hat{\beta}_j=0.$$

As a result, (2.2.2) becomes

$$|\hat{\beta}_j-\hat{\beta}_k|\leq\frac{\|\mathbf{y}\|_2}{\lambda_2}\|s_j\mathbf{x}_j-s_k\mathbf{x}_k\|_2\leq\frac{\|\mathbf{y}\|_2 s^{(K)}}{\lambda_2}\sqrt{2n(1-\varphi_{jk}\phi_{jk})} \tag{2.2.3}$$

where $s^{(K)}=\max_{1\leq k\leq K}s_k$ and

$$\varphi_{jk}=\frac{2s_js_k}{s_j^2+s_k^2}=1-\frac{(s_j-s_k)^2}{s_j^2+s_k^2}. \tag{2.2.4}$$

Based on (2.2.3) and (2.2.4), the hierarchical grouping property P1 and P2 can be restated respectively in terms of the following two (sufficient) conditions on the hierarchical scores s_i's:

C1. For any given pair of predictors $\mathbf{x}_j$ and $\mathbf{x}_k$, $s_j = s_k$ if and only if they have exactly the same ancestors.

C2. For any given predictors $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$, if the ancestors of $\mathbf{x}_i$ and $\mathbf{x}_k$ are a subset of the ancestors of $\mathbf{x}_i$ and $\mathbf{x}_j$, then $\min(s_i/s_j, s_j/s_i) > \min(s_i/s_k, s_k/s_i)$.

Note that, when $s_j = s_k$, we have $\varphi_{jk} = 1$ and thus $|\hat{\beta}_j - \hat{\beta}_k| \to 0$ as $\phi_{jk} \to 1$ by (2.2.3) and (2.2.4). So, by C1, we have P1 in that, if a pair of predictors have the same ancestors, they are likely grouped together if they are also highly correlated. In addition, when $\min(s_i/s_j, s_j/s_i) > \min(s_i/s_k, s_k/s_i)$, we have $1 \geq \varphi_{ij} > \varphi_{ik} \geq 0$ by (2.2.4), and it follows that $1-\varphi_{ij}\phi_{ij} < 1-\varphi_{ik}\phi_{ik}$ if $\phi_{ij} = \phi_{ik} > 0$. In this case, based on (2.2.3), $\mathbf{x}_i$ and $\mathbf{x}_j$ are more likely grouped than $\mathbf{x}_i$ and $\mathbf{x}_k$. Thus, C2 implies P2.

We construct in the next subsection a set of hierarchical scores s_i's that satisfies conditions C1 and C2.

2.2.2 Construction of hierarchical scores

For any given terminal node $\mathbf{x}_i$, we denote by P_i the set that contains the depth values of the ancestors of $\mathbf{x}_i$. For example, the predictor $\mathbf{x}_1$ in Figure 2.2(a) has $P_1 = \{1, 3, 5\}$. Similarly, the predictor $\mathbf{x}_3$ has $P_3 = \{1, 3\}$. Also, we define the binary vector $\mathbf{v}_i \in \mathbb{R}^{p-1}$ such that for $l = 1, \cdots, p-1$,

$$\mathbf{v}_i(l) = \begin{cases} 1 & \text{if } l \in P_i, \\ 0 & \text{otherwise.} \end{cases}$$

For instance, corresponding to $P_1 = \{1, 3, 5\}$ of predictor $\mathbf{x}_1$ in Figure 2.2 (a), the binary vector $\mathbf{v}_1 = (1, 0, 1, 0, 1)$. Similarly, corresponding to $P_3 = \{1, 3\}$ of predictor $\mathbf{x}_3$, the binary vector $\mathbf{v}_3 = (1, 0, 1, 0, 0)$.

The hierarchical score s_i for the node or predictor $\mathbf{x}_i$ is then defined as

$$s_i = \left\{(\tau^{-1}, \tau^{-2}, \cdots, \tau^{-(p-1)}) \cdot \mathbf{v}_i\right\}^{\alpha} = \left\{ \sum_{l=1}^{p-1} \tau^{-l} \mathbf{v}_i(l) \right\}^{\alpha}. \tag{2.2.5}$$

Here, τ and α are two positive constants which will be further explained below. This set of s_i's are bounded above by $\max_{1\le i\le p} s_i \le (\sum_{l=1}^{p-1} \tau^{-l})^{\alpha} = \{(\tau^{-1} - \tau^{-p})/(1-\tau^{-1})\}^{\alpha}$ when $\tau > 1$. To ensure unique representations of terminal nodes with different ancestors, we assume that all the internal nodes have different depths in a given hierarchy structure.

The following Theorem states that s_i's satisfy condition C1 and C2 for a wide range of τ and α. The proof of the theorem is given in the appendix.

Theorem 2.2.1. *When $\tau \ge 3$ and $\alpha > 0$, the s_i defined in (2.2.5) satisfies conditions C1 and C2.*

In Theorem 2.2.1, $\tau \ge 3$ guarantees that the inequality in condition C2 holds. The other parameter α is a tuning parameter that controls the "degree" on how much a hierarchy can impact on the estimation. In particular,

- When $\alpha \to 0$, all $s_i \equiv 1$ and all $\varphi_{ij} \equiv 1$. In this case, the hierarchy is not taken into account.

- When $\alpha \to \infty$, only predictors sharing same ancestors have $\varphi \approx 1$ and otherwise $\varphi \approx 0$. So only predictors with the same ancestors are considered for grouping and the hierarchy structure is strictly enforced.

From (2.2.4), $0 \le \varphi_{ij} \le 1$. So, φ_{ij} can be treated as a shrinkage coefficient on the correlation ϕ_{ij} based on (2.2.3). Furthermore, $|s_i - s_j|$ is dominated by τ^{-b} under (2.2.5), where $b = \min\{k : \mathbf{v}_i(k) \ne \mathbf{v}_j(k)\}$ is the location of first difference between two sequence $\mathbf{v}_i$ and $\mathbf{v}_j$. By (2.2.4), $1 - \varphi_{ij}$ can be viewed as a standardized measure of the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ on the hierarchical tree. So the pair of predictors that are "close" in the hierarchy will have small $1 - \varphi_{ij}$ and vice versa.

2.2.3 Theoretical results

The HEnet estimator (2.2.1) can be re-expressed as

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \|\mathbf{y} - \sum_{j=1}^{p} \beta_j \mathbf{x}_j\|_2^2 + \lambda_1 \sum_{j=1}^{p} \frac{|\beta_j| + \delta\beta_j^2}{s_j} \right\}, \tag{2.2.6}$$

where $\delta = \lambda_2/\lambda_1$ is a new tuning parameter replacing λ_2. We first formally state the result derived in Section 2.2.1 in the following Theorem.

Theorem 2.2.2 (**Hierarchical grouping property**). *Let $\hat{\beta}$ be the estimator in (2.2.6). Suppose $\hat{\beta}_i\hat{\beta}_j > 0$, then*

$$|\hat{\beta}_i - \hat{\beta}_j| \leq \frac{\|\mathbf{y}\|_2}{\lambda_1\delta}\|s_i\mathbf{x}_i - s_j\mathbf{x}_j\|_2 \leq \frac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\sqrt{2(1-\varphi_{ij}\phi_{ij})}$$

where $s^{(K)} = \max_{1\leq k\leq} s_k$, $\varphi_{ij} = 2s_is_j/(s_i^2+s_j^2)$ and $\phi_{ij} = cor(\mathbf{x}_i, \mathbf{x}_j)$.

Theorem 2.2.2 entails the hierarchical grouping property for the proposed HEnet estimator. For instance, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated, i.e. $\phi \approx 1$ (if $\phi \approx -1$ then consider $-\mathbf{x}_j$), Theorem 2.2.2 indicates that $|\hat{\beta}_i - \hat{\beta}_j|$ is bounded by $C(\sqrt{1-\varphi_{ij}})$ with $C = \sqrt{2n}\|\mathbf{y}\|_2 s^{(K)}/(\lambda_1\delta)$. When $\lambda_1\delta = O(\sqrt{n}\|\mathbf{y}\|_2)$, C is bounded asymptotically. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are also "close" in the hierarchical structure with $\varphi_{ij} \approx 1$, we have $|\hat{\beta}_i - \hat{\beta}_j| \approx 0$ by Theorem 2.2.2 and thus $\mathbf{x}_i$ and $\mathbf{x}_j$ are grouped together (in the same sense as [Zou and Hastie (2005)] and [Bondell and Reich (2008)]). When $\varphi_{ij} < 1$, the addition of the φ_{ij} term channels the information of the hierarchical structure into the grouping process.

Now we move on to model selection consistency. Without loss of generality, we assume that the first q true parameters $\beta_j^0 \neq 0$ for $j \in \{1, \cdots, q\}$ and the rest $p-q$ true parameters $\beta_j^0 = 0$ for $j \in \{q+1, \cdots, p\}$. Write $\beta_{(1)}^0 = (\beta_1^0, \cdots, \beta_q^0)'$ and $\beta_{(2)}^0 = (\beta_{q+1}^0, \cdots, \beta_p^0)' = (0, \cdots, 0)'$. Let $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$ be the first q and last $p-q$ columns of the design matrix $\mathbf{X}$, respectively, and $\Sigma_{ij} = \mathbf{X}_{(i)}'\mathbf{X}_{(j)}/n$, for $i, j \in \{1, 2\}$.

Model selection consistency of the conventional Enet estimator has been discussed in details in [Jia and Yu (2010)]. According to [Jia and Yu (2010)], a simple Elastic Irrepresentable Condition (EIC) is necessary for the model selection consistency of the Enet.

EIC. There exists a positive constant $\eta < 1$ such that

$$\left\|\Sigma_{21}(\Sigma_{11} + \frac{\lambda_1\delta}{n}I)^{-1}(\text{sgn}(\beta^0_{(1)}) + 2\delta\beta^0_{(1)})\right\|_\infty \leq 1 - \eta. \tag{2.2.7}$$

For the HEnet estimator (2.2.6), we generalize the EIC to incorporate the hierarchical information. We define the Hierarchical Elastic Irrepresentable Condition (HEIC) as follows.

HEIC. There exists a positive constant $\eta < 1$ such that

$$\left\|S_{(2)}\Sigma_{21}(S_{(1)}\Sigma_{11} + \frac{\lambda_1\delta}{n}I)^{-1}(\text{sgn}(\beta^0_{(1)}) + 2\delta\beta^0_{(1)})\right\|_\infty \leq 1 - \eta, \tag{2.2.8}$$

where $S_{(1)} = \text{diag}(s_1, \cdots, s_q)$ and $S_{(2)} = \text{diag}(s_{q+1}, \cdots, s_p)$.
Note that the HEIC condition is exactly the EIC condition, when all the scores $s_i \equiv 1, S_{(1)} = I_q$ and $S_{(2)} = I_{p-q}$. We show in the following theorem that HEIC is necessary for the model selection consistency of the HEnet estimator. The proof of this theorem is almost identical to that of [Jia and Yu (2010)] and is thus omitted here.

Theorem 2.2.3 (**Model selection consistency**). *Suppose $\varepsilon \sim N(0, \sigma^2 I)$. Assume HEIC (2.2.8) holds. Then, the globally optimal HEnet estimator satisfies:*

$$P(sgn(\hat{\beta}_{(1)}) = sgn(\beta^0_{(1)}), \hat{\beta}_{(2)} = 0) \to 1$$

provided the tuning parameter λ_1 and δ are chosen such that
(a) $\frac{\lambda_1^2}{n\log(p-q)(\max_{q+1\leq i\leq p} s_i)^2} \to \infty$
(b) $\frac{1}{\rho}\left\{3\sqrt{\frac{\sigma^2\log q}{nC_{min}}} + \frac{\lambda_1}{2n}\|(S_{(1)}\Sigma_{11} + \frac{\lambda_1\delta}{n}I)^{-1}sign(\beta^0_{(1)})\|_\infty\right\} \to 0$, *where* $\rho = \min|(\Sigma_{11} + \frac{\lambda_1\delta}{n}S^{-1}_{(1)})^{-1}(\Sigma_{11}\beta^0_{(1)})|$, *and* $C_{\min} = \Lambda_{\min}(\Sigma_{11}) + \frac{\lambda_1\delta}{n}(\max_{1\leq i\leq q} s_i)^{-1}$. *Here* $\Lambda_{\min}(\cdot)$ *denotes the minimal eigenvalue.*

Note that, for any fixed α in (2.2.5), all the s_i's are bounded. Thus, incorporating the hierarchical scores into the Enet penalty does not introduce additional complexity to the original problem. Specifically for the classical setting when p and q are fixed, Σ_{11} converges to a non-negative definite matrix. Thus, HEIC implies model selection consistency of the HEnet estimator, if the tuning parameter is chosen such that

$$\lambda_1/\sqrt{n} \to \infty, \quad \lambda_1/n \to 0 \quad \text{and} \quad \lambda_1\delta = O(n).$$

This conclusion matches the one in [Jia and Yu (2010)] and also many other papers on Lasso approaches.

2.3 Hierarchical variable selection when a terminal node contains multiple predictors

To prevent overfitting problems yet still capture the important structure, many researchers have suggested to "prune" a large hierarchical trees; c.f.
[Breiman et al. (1984)]. The terminal nodes on a pruned tree will often represent multiple predictors. Indeed, the terminal nodes in most hierarchical clusters contain more than one member, whether it is by pruning, by their natural structure or by pther meaning. In the example of the SLE dataset described in Section 2.1, all the 4779 predictors are divided into 28 modules (terminal nodes) withing each module having more than one gene. The natural grouping structure by terminal nodes introduces another layer of complication to the model selection problem. In this situation, we need two levels of model selection: between the groups of predictors in different terminal nodes and within the group of predictors in each terminal node. Specifically, if we define "important variable" as the variable with non-zero true coefficient and "important group" as the group with at least one "important variable", we would like to identify the important groups (terminal

nodes) and the important variables (in each important terminal node), while taking into account that these groups (terminal nodes) have a certain hierarchical structure.

We extend the developments in Section 2.2 to deal with this more general but also more challenging case. The scoring scheme we proposed in Section 2.2.2 is applied to the hierarchy on the group structure. Consequently, every group will be assigned a score. These scores play an important role in our proposed group hierarchical Enet estimator.

2.3.1 Group Hierarchical Enet

In the linear model (2.1.1), suppose there are p predictors and they are contained in K terminal nodes of a known hierarchical tree. Let $G_k, k = 1, \cdots, K$ be the K non-overlapping subsets of the indices $(1, \cdots, p)$, corresponding to the K terminal nodes. Denote by p_k the number of predictor variables in the k^{th} subset and, thus, $\sum_{k=1}^{K} p_k = p$. To simplify the notations and without loss of generality, let the $p \times 1$ vector of regression coefficient β be arranged as $\beta_{kj}, j = 1, \cdots, p_k; k = 1, \cdots, K$. To introduce a coefficient for the group (terminal node) G_k and following [Wang et al. (2009)], we reparameterize the β's as:

$$\beta_{kj} = \gamma_k \theta_{kj}, k = 1, \cdots, K, j = 1, \cdots, p_k \tag{2.3.1}$$

where the parameter $\gamma_k \geq 0$ is a group level coefficient for G_k; the $\theta_{kj}, j = 1, \cdots, p_k$ reflect different coefficients within G_k. Since we introduced one new parameter for each group, we pose one constraint for each group as the identifiability condition

$$\| \sum_{j=1}^{p_k} \theta_{kj} \mathbf{x}_{kj} \|_2^2 = n \quad \text{for} \quad k \in \{k : \gamma_k > 0\}. \tag{2.3.2}$$

Explicit expression of γ_k and θ_{kj} can be derived from (2.3.1) and (2.3.2). Specifically, when there is at least one nonzero β_{kj} in kth group,

$$\gamma_k = \frac{\| \sum_{j=1}^{p_k} \beta_{kj}\mathbf{x}_{kj}\|_2}{\sqrt{n}} \quad \text{and} \quad \theta_{kj} = \frac{\beta_{kj}}{\gamma_k} = \frac{\beta_{kj}\sqrt{n}}{\| \sum_{j=1}^{p_k} \beta_{kj}\mathbf{x}_{kj}\|_2}. \tag{2.3.3}$$

Otherwise, in the case with $\beta_{kj} = 0$ for all $j = 1, \cdots, p_k$, we set $\gamma_k = 0$ and $\theta_{kj} = 0$ for $j = 1, \cdots, p_k$.

Following the idea of the HEnet discussed in Section 2.2, we propose the following constrained penalized likelihood estimator

$$\begin{aligned}(\hat{\gamma}, \hat{\theta}) = \text{argmin}_{\gamma,\theta} \|\mathbf{y} - \sum_{k=1}^{K}\sum_{j=1}^{p_k} \gamma_k\theta_{kj}\mathbf{x}_{kj}\|_2^2 \quad &+ \quad \lambda_1 \sum_{k=1}^{K} \frac{P_1(\gamma_k)}{s_k} \\ &+ \quad \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p_k} \frac{P_2(\gamma_k|\theta_{kj}|)}{s_k}\end{aligned} \tag{2.3.4}$$

subject to the constraint in (2.3.2). We refer to this estimator as the *group hierarchical Enet (GHEnet) estimator*. Here, The s_k represents "hierarchical score" for the k^{th} group (terminal node). Also, P_1 is Enet penalty encouraging sparse group selection and P_2 is a class of penalty function estimating within group coefficient θ. For instance, we can choose P_2 to be Bridge penalty [Frank and Friedman (1993)], Lasso [Tibshirani (1996)], SCAD [Fan and Li (2001)] or MCP [Zhang (2010)].

For illustration purpose, we let P_1 be Enet penalty and P_2 be the Lasso penalty. In this case, we re-write (2.3.4) and GHEnet as:

$$\begin{aligned}(\hat{\gamma}, \hat{\theta}) = \text{argmin}_{\gamma,\theta} \|\mathbf{y} - \sum_{k=1}^{K}\sum_{j=1}^{p_k} \gamma_k\theta_{kj}\mathbf{x}_{kj}\|_2^2 \quad &+ \quad \lambda_1 \{\sum_{k=1}^{K} \frac{\gamma_k + \delta\gamma_k^2}{s_k}\} \\ &+ \quad \lambda_2 \sum_{k=1}^{K} \frac{\gamma_k\|\theta_k\|_1}{s_k}\end{aligned} \tag{2.3.5}$$

subject to the constrain in (2.3.2). Here, λ_1, λ_2 and δ are the tuning parameters, and $\theta_k = (\theta_{k1}, \ldots, \theta_{kp_k})^T$.

The proposed GHEnet estimator (2.3.5) is an extension of (2.2.6). Specifically, if there is only one predictor within each group, i.e. $p_k = 1$ for all $k = 1, \cdots, K$,

we go back to the simplified case discussed in Section 2.2. In particular, the reparameterization (2.3.1) becomes

$$\beta_{k1} = |\beta_{k1}|\text{sign}(\beta_{k1}) \text{ with } \gamma_k = |\beta_{k1}| \text{ and } \theta_{k1} = \text{sign}(\beta_{k1}) \text{ for } k = 1, \cdots, K.$$

Plugging it into (2.3.5), we get back to (2.2.6).

2.3.2 Theoretical results

2.3.2.1 Hierarchical grouping property

Denote by $\tilde{\mathbf{x}}_k = \sum_{l=1}^{p_k} \hat{\theta}_{kl}\mathbf{x}_{kl}$, for $k = 1, \cdots, K$. The definition (2.3.5) and the constraint (2.3.2) imply that $\|\tilde{\mathbf{x}}_k\|_2^2 = n$ for $\{k : \hat{\gamma}_i > 0\}$. We interpret $\tilde{\mathbf{x}}_k$ as the "overall predictor vector" that represents group G_k. By treating $\tilde{\mathbf{x}}_k$'s as $\mathbf{x}_i$'s in (2.2.6), we provide in the next theorem a "group" version of the hierarchical grouping property. The proof of Theorem 2.3.1 is given in the appendix.

Theorem 2.3.1. *Let $(\hat{\theta}, \hat{\gamma})$ be the estimator in (2.3.5). Suppose $\hat{\gamma}_i\hat{\gamma}_j > 0$, then*

$$\begin{aligned} |\hat{\gamma}_i - \hat{\gamma}_i| &\leq \frac{\|\mathbf{y}\|_2}{\lambda_1\delta}\|s_i\tilde{\mathbf{x}}_i - s_j\tilde{\mathbf{x}}_j\|_2 + \frac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_i\|_1 - \|\hat{\theta}_j\|_1\Big| \\ &\leq \frac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\sqrt{2(1-\varphi_{ij}\phi_{ij})} + \frac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_i\|_1 - \|\hat{\theta}_j\|_1\Big|, \end{aligned}$$

where $s^{(K)} = \max s_k$, $\varphi_{ij} = 2s_is_j/(s_i^2 + s_j^2)$, and $\phi_{ij} = cor(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ is the correlation between group G_i and G_j. Furthermore, if $\lambda_2 \max_{1\leq k\leq K}\sqrt{p_k} = o(n)$, we have

$$\begin{aligned} &\tfrac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\sqrt{2(1-\varphi_{ij}\phi_{ij})} + \tfrac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_i\|_1 - \|\hat{\theta}_j\|_1\Big| \\ &\qquad = \tfrac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\left\{\sqrt{2(1-\varphi_{ij}\phi_{ij})} + o_p(1)\right\}. \end{aligned}$$

and

$$|\hat{\gamma}_i - \hat{\gamma}_i| \leq \frac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\left\{\sqrt{2(1-\varphi_{ij}\phi_{ij})} + o_p(1)\right\}.$$

In Theorem 2.3.1, the group correlation $\phi_{ij} = \text{cor}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ equals the weighted average of the correlation of the individual variables within group i and group j.

Comparing with Theorem 2.2.2, the additional term comes from the derivative of the Lasso penalty on individual level coefficient. This additional term for individual is dominated by the original term for grouping when $\lambda_2 \max_{1\le k\le K}\sqrt{p_k} = o(n)$. Thus, based on the theorem and asymptotically, we have a similar statement of the hierarchical grouping property as the the simplified cased discussed in Section 2.2. In particular, if group i and group j are highly correlated (i.e., $\phi_{ij} \approx 1$) and the corresponding terminal nodes are close in the hierarchical tree (i.e., $\varphi_{ij} \approx 1$), then we have $|\hat{\gamma}_i - \hat{\gamma}_j| \approx 0$. Using the terminology in Definition 2.2.2, group i and group j are *grouped*.

In the simplified special case with only one variable for each terminal node, the results in Theorem 2.3.1 reduce to the results in Theorem 2.2.2, thus Theorem 2.3.1 is a generalization of Theorem 2.2.2.

2.3.2.2 Model selection consistency

To derive the property of model selection consistency, we express (2.3.5) in terms of β. By plugging (2.3.3) into (2.3.5), we have:

$$P_n(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda_1}{\sqrt{n}} \sum_{k=1}^{K} \frac{\|\mathbf{X}_k\beta_k\|_2 + \delta\|\mathbf{X}_k\beta_k\|_2^2/\sqrt{n}}{s_k} + \lambda_2 \sum_{k=1}^{K} \frac{\|\beta_k\|_1}{s_k} \quad (2.3.6)$$

where $\mathbf{X}_k = (\mathbf{x}_{k1}, \cdots, \mathbf{x}_{kp_k})$ and $\beta_k = (\beta_{k1}, \cdots, \beta_{kp_k})^T$.

Write the true regression coefficients as $\beta_{kj}^0 = \gamma_k^0\theta_{kj}^0$ for $k = 1, \cdots, K, j = 1, \cdots, p_k$. Without lose of generality, assume $\gamma_k^0 > 0$ for $k = 1, \cdots, r$ and $\gamma_k^0 = 0$ for $k = r+1, \cdots, K$. Define $A = \{(k,j) : \beta_{kj}^0 \neq 0\}$, $B = \{(k,j) : \gamma_k^0 > 0, \beta_{kj}^0 = 0\}$, $A_i = \{(i,j) : \beta_{ij}^0 \neq 0\}$ and $B_i = \{(i,j) : \beta_{ij}^0 = 0\}$ for $i = 1, \cdots, r$. So A represents all the important variables and B is the set of non-important variables within important groups.

As a generalization of the HEIC condition, we propose a GHEIC condition which is necessary for the GHEnet to perform consistent model selection. In the GHEIC condition below, we denote by $S_A = \mathrm{diag}(d_{ij}|\{i,j\} \in A, d_{ij} = s_i)$,

$S_B = \text{diag}(d_{ij}|\{i,j\} \in B, d_{ij} = s_i)$, $\mathbf{X}_A = (\mathbf{x}_{kj}, (k,j) \in A)$, $\mathbf{X}_B = (\mathbf{x}_{kj}, (k,j) \in B)$, $\Sigma_{AA} = \mathbf{X}_A'\mathbf{X}_A/n$, $\Sigma_{BA} = \mathbf{X}_B'\mathbf{X}_A/n$, $\bar{\Sigma} = \text{diag}(\mathbf{X}_{A_k}'\mathbf{X}_{A_k}/n)_{1\le k\le r}$ and $\beta_A = (\beta_{kj}, (k,j) \in A)'$.

GHEIC. There exists a positive constant $\eta < 1$ such that

$$\left\| S_B\Sigma_{BA}(S_A\Sigma_{AA} + \frac{\lambda_1\delta}{n}\bar{\Sigma})^{-1}(\text{sgn}(\beta_A^0) + \frac{2\delta\lambda_1}{\lambda_2}\bar{\Sigma}\beta_A^0) \right\|_\infty \le 1 - \eta \qquad (2.3.7)$$

Note that when there is only one variable for each group, GHEIC is same as HEIC. Also we only require $\mathbf{X}_A'\mathbf{X}_A/n + \lambda_1\delta S_A^{-1}\bar{\Sigma}/n$ to be invertible instead of $\mathbf{X}_A'\mathbf{X}_A/n$. This condition is weaker, since $\mathbf{X}_A'\mathbf{X}_A/n + \lambda_1\delta S_A^{-1}\bar{\Sigma}/n$ is less likely to be singular than $\mathbf{X}_A'\mathbf{X}_A/n$ in a high dimensional setting. The following theorem states that GHEIC is necessary for the variable selection consistency of GHEnet estimator. The proof of Theorem 2.3.2 is given in the appendix.

Theorem 2.3.2 (**Model selection consistency**). *Suppose $\varepsilon \sim N(0, \sigma^2 I)$. Assume GHEIC (2.3.7) holds. Then, the GHEnet estimator $\hat{\beta}$ which maximizes (2.3.6) satisfies:*

1) *$\hat{\beta}$ is a minimizer of (2.3.6) on the subspace $\{\beta : \beta_{(A\cup B)^c} = 0\}$ such that $P(sgn(\hat{\beta}_A) = sgn(\beta_A^0), \hat{\beta}_B = 0) \to 1$.*

2) *$P_n(\tilde{\beta}) \ge P_n(\hat{\beta})$ for any $\tilde{\beta} \in \{\beta : \beta_{A\cup B} = \hat{\beta}_{A\cup B}\}$*

provided that the following conditions hold:

(a) $\|\tilde{\Sigma}^{-1}\|_\infty \le C_1$ for some positive constant C_1, where $\tilde{\Sigma} = \frac{1}{n}\mathbf{X}_A'\mathbf{X}_A + \frac{\lambda_1\delta}{n}S_A^{-1}\bar{\Sigma}$.

(b) $\|\mathbf{X}_B'\mathbf{X}_A\|_\infty/n \le C_3$ for some positive constant C_3.

(c) $\|\mathbf{X}_k'\mathbf{X}_A\|_\infty/n \le C_2, k = r+1, \cdots, K$ for some positive constant C_2.

(d) $p = o(\sqrt{n}\beta_* e^{\frac{n\beta_*^2}{2\sigma^2}})$, where $\beta_* = \min\{|\beta_A^0|\}$.

(e) The tuning parameters $\lambda_1, \lambda_2, \delta$ are chosen such that

- $\frac{\lambda_2}{2\sqrt{n}\max s_k} \ge C_2\sqrt{n}\beta_*$
- $C_1^{-1}(1-\alpha)\beta_*\sqrt{n} - C_1^{-1}\zeta\lambda_1/\sqrt{n} - \frac{\lambda_1+\lambda_2}{2\sqrt{n}\min s_k} = O(\sqrt{n}\beta_*) > 0$ *where* $\zeta = \|(\frac{1}{n}S_A\mathbf{X}_A'\mathbf{X}_A + \frac{\lambda_1\delta}{n}\bar{\Sigma})^{-1}(\delta\bar{\Sigma}\beta_A^0)\|_\infty$

- $\frac{\eta\lambda_2-\lambda_1}{2\sqrt{n}\max s_k} - \frac{\lambda_1 C_1 C_3}{2\sqrt{n}\min s_k} - \frac{\lambda_1\delta}{\sqrt{n}\max s_k} C_3(\beta_* + \|\beta_A^0\|_\infty) = O(\sqrt{n}\beta_*) > 0$

The first three conditions (a)-(c) in this theorem impose some regularity constraints on the design matrix $\mathbf{X}$. Specifically, conditions (b) and (c) constrain the growth rate of the absolute sum of covariances between the noise variable and all true variables. Condition (d) constrains the growth rate of the total number of parameter p. Similar conditions can be found in [Fan and Lv (2009)]. The last condition gives us a guideline of choosing tuning parameter $\lambda_1, \lambda_2, \delta$. In particular, the following choice satisfies both condition (e) and the condition $\lambda_2 \max_k \sqrt{p_k} = o(n)$ in Theorem 2.3.1:

$$\lambda_1 = O(n\beta_*), \lambda_2 = O(n\beta_*), \lambda_2\delta = O(n)$$

given $\|\beta_A^0\|_\infty = O(\beta_*)$ and $\beta_* = o(1/\max_k \sqrt{p_k})$. The condition $\|\beta_A^0\|_\infty = O(\beta_*)$ aligns with GHEIC condition which implies $\frac{\delta\lambda_1}{\lambda_2}\|\beta_A^0\|_\infty = O(1)$ or equivalently $\|\beta_A^0\|_\infty = O(\frac{\lambda_2}{\delta\lambda_1}) = O(\beta_*)$. Furthermore, under the classical setting when p are fixed, if $\beta_* = O(n^\nu)$ with $-0.5 < \nu < 0$, the rate of λ_2, δ matches that in [Jia and Yu (2010)].

2.4 Computation algorithm

In this section, we discuss the computational issues. A good feature of our proposed method is that, due to the fact that the hierarchical scores are bounded and positive, the hierarchical estimation procedure can inherit the computational algorithm from the corresponding conventional one. Specifically, the HEnet estimator in (2.2.6) can be solved by the Enet algorithm [Zou and Hastie (2005)] with only slight modifications. The computation details are given in the following algorithm, the proof of which is just algebra and omitted here.

Given data set $(\mathbf{y}, \mathbf{X})$ and hierarchical score $(s_1, \cdots, s_p)$, denote $S = \text{diag}(s_1, \cdot, s_p)$, $S^{1/2} = \text{diag}(s_1^{1/2}, \cdot, s_p^{1/2})$.

1. Define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X}S \\ \sqrt{\lambda_1 \delta} S^{1/2} \end{pmatrix}, \quad \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}.$$

2. Solve the lasso problem for all λ_1,

$$\hat{\beta}^* = \text{argmax}_\beta \|\mathbf{y}^* - \mathbf{X}^*\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|.$$

3. Output $\hat{\beta}_j = \hat{\beta}_j^* s_j, j = 1, \cdots, p$

For the estimator (2.3.5), the computational algorithm is similar to the one in [Wang et al. (2009)]. Specifically, the algorithm is

1. Obtain an initial value $\gamma_k^{(0)}$ for each γ_k; for example, $\gamma_k^{(0)} = 1$. Let $m = 1$.

2. At the mth iteration, let $\tilde{\mathbf{x}}_{kj} = \gamma_k^{(m-1)} \mathbf{x}_{kj}$ and estimate $\theta_{kj}^{(m)}$ by

$$\theta_{kj}^{(m)} = \text{argmin}_\theta \|\mathbf{y} - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \theta_{kj} \tilde{\mathbf{x}}_{kj}\|_2^2 + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{p_k} \frac{\gamma_k^{(m-1)} |\theta_{kj}|}{s_k}$$

   subject to $\| \sum_{l=1}^{p_k} \theta_{kj} \mathbf{x}_{kj} \|_2^2 = n \quad \text{for} \quad \{k : \gamma_k^{(m-1)} > 0\}$

3. Let $\tilde{\mathbf{x}}_k = \sum_{j=1}^{p_k} \theta_{kj}^{(m)} \mathbf{x}_{kj}$ and estimate $\gamma_k^{(m)}$ by

$$\gamma_k^{(m)} = \text{argmin}_\gamma \|\mathbf{y} - \sum_{k=1}^{K} \gamma_k \tilde{\mathbf{x}}_k\|_2^2 + \lambda_1 \{ \sum_{k=1}^{K} \frac{\gamma_k (1 + \|\theta_k^{(m)}\|_1 \lambda_2 / \lambda_1)}{s_k} + \delta \sum_{k=1}^{K} \frac{\gamma_k^2}{s_k} \}$$

4. Set $m = m + 1$ Repeat Step 2 and Step 3 until convergence.

Both Step 2 and Step 3 can be solved using the constrained quadratic programming algorithm.

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ can be chosen via different criterion. The most commonly used ones in the penalizing criteria include

- $\text{AIC}(\boldsymbol{\lambda}) = \log(\|\mathbf{y} - \mathbf{X}\hat{\beta}(\boldsymbol{\lambda})\|_2^2 / n) + 2\|\hat{\beta}(\boldsymbol{\lambda})\|_0 / n$

- $\text{BIC}(\boldsymbol{\lambda}) = \log(\|\mathbf{y} - \mathbf{X}\hat{\beta}(\boldsymbol{\lambda})\|_2^2/n) + \log(n)\|\hat{\beta}(\boldsymbol{\lambda})\|_0/n$

where $\|\hat{\beta}\|_0$ is the number of non-zero $\hat{\beta}$'s. Generally, AIC will select a big model and BIC tends to select less variables. In the our numerical studies in the next section, the tuning parameter is chosen from BIC criteria.

The parameter α in (2.2.5) can also be viewed as an additional tuning parameter in the proposed approach which allows researchers to determine the "degree" of impact of hierarchical structure on model selection. The choice of α is based on empirical study which will be discussed and illustrated using concrete examples in the following section of numerical studies.

2.5 Numerical studies

2.5.1 Simulation Studies

In this section, two simulation studies are carried out to evaluate the performance of the proposed method. In Example 1, we consider small scale dataset to test the HEnet estimator proposed in Section 2.2 where each terminal node contains only one predictor. In Example 2, we investigate the performance of GHEnet estimator proposed in Section 2.3 by considering a large dataset where each terminal node contains multiple predictors.

2.5.1.1 Example 1

The data consists of $n = 20$ samples with $p = 6$ variables. The true parameter $\beta = (0, 0, 0, 3, 3, 3)$ and $\sigma = 1$. Thus the true model is

$$\mathbf{y} = \mathbf{x}_4\beta_4 + \mathbf{x}_5\beta_5 + \mathbf{x}_6\beta_6 + \varepsilon.$$

We consider two scenarios with different hierarchical and covariance structure:

Scenario 1. The variables are assumed to have a balanced hierarchical structure as Figure 2.2(a). Each row vector of the design matrix $\mathbf{X}$ is generated from

the standard normal distribution with covariance matrix $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.9^{1\{i \neq j\}}$ for $(i, j) \in \{1, 2, 3\}$ and $(i, j) \in \{4, 5, 6\}$. For any other pair of (i, j), $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0$.

Scenario 2. The variables are assumed to have an unbalanced hierarchical structure as Figure 2.2(b). Each row vector of the design matrix $\mathbf{X}$ is generated from the standard normal distribution with covariance matrix $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = 0.9^{1\{i \neq j\}}$.

We benchmark our proposed estimator against three existing estimators: Enet (which is a special HEnet with $\alpha = 0$), Lasso and OSCAR. For the proposed approach, we have a tuning parameter α to control how much impact the hierarchy will have on the model selection results. Note that $\varphi_{i,j} = \frac{2s_i s_j}{s_i^2 + s_j^2}$, which is a decreasing function of α. Also, $\varphi_{i,j} = 1$ for $\alpha = 0$ and $\lim_{\alpha \to \infty} \varphi_{i,j} = 0$ for any pair of $(\mathbf{x}_i, \mathbf{x}_j)$ with different ancestors. We plot $\varphi_{i,j}$ against α for all pairs (i, j) in Figure 2.3. For each plot, the upper shade corresponds to the 75% or higher quantiles of all $\varphi_{i,j}$'s and the lower shade is the 50% or lower quantiles of all $\varphi_{i,j}$'s. The vertical red lines represent the different α's we pick in the simulation. For scenario 1, we first pick $\alpha = 15$ which represents the largest range of $\varphi_{i,j}$ meanwhile keeping the smallest φ away from 0. Then we pick $\alpha = 8$ to test the impact of different ranges of φ on the model selection. Similarly, we choose $\alpha \in \{5, 10\}$ for Scenario 2. We also include $\alpha = 0.1$ case in both Scenario to illustrate the idea that almost no hierarchy is contributed to the model selection result for small α.

The numerical results are summarized in Table 2.1. Denote by P_{ij}^S the frequency of occasions on which $\mathbf{x}_i$ and $\mathbf{x}_j$ being selected together and P_{ij}^D the frequency of occasions on which $\mathbf{x}_i$ and $\mathbf{x}_j$ being dropped from the model together. We can see that for Scenario 1, $P_{45}^S, P_{46}^S, P_{56}^S$ are very close under Enet, Lasso, OSCAR and HEnet with $\alpha = 0.1$ because pairwise correlations are the same within important variables and (almost) no hierarchy structure is considered. When α increases

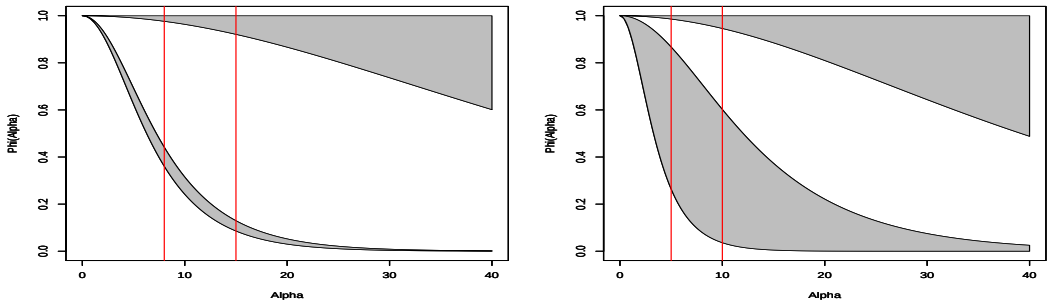
to 8 and 15, the hierarchy impacts on the model selection result. Specifically, in Figure 2.2(a) and among the pairs of the three important variables $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, the pair $\{\mathbf{x}_5, \mathbf{x}_6\}$ is closer than other two pairs. Such structure leads to higher value of P^S_{56} than P^S_{45} and P^S_{46}, i.e. $\{\mathbf{x}_5, \mathbf{x}_6\}$ is more likely selected together than other two pairs of important variables. Parallel results can be found in $P^D_{12}, P^D_{13}, P^D_{23}$ with $\alpha \in \{8, 15\}$ for the pairs of non-important variables: $\{\mathbf{x}_1, \mathbf{x}_2\}$ is more likely dropped together than the other two pairs because they are closer on the hierarchy.

Similar results can be found for Scenario 2. The $P^S_{45}, P^S_{46}, P^S_{56}$ are very close under Enet, Lasso, OSCAR and HEnet with $\alpha = 0.1$. The pair $\{\mathbf{x}_5, \mathbf{x}_6\}$ are closer than other two pairs of important variable on the hierarchical tree from Figure 2.2(b). Thus P^S_{56} is higher than P^S_{45} and P^S_{46} with $\alpha = 5$ or 10. For the non-important variable, $\{\mathbf{x}_2, \mathbf{x}_3\}$ is closer than other two pairs thus is more likely dropped together. Overall, the hierarchical grouping property of our proposed estimator is well illustrated by the above results.

We also record the variable sensitivity and specificity to gauge the variable selection performance. Sensitivity is defined as the proportion of the selected variables that are the important variables. Specificity is defined as the proportion of the excluded variables that are unimportant variables. From Table 2.1, there is no significant advantage of HEnet in Scenario 1. HEnet with $\alpha = 5, 10$ outperform others in Scenario 2.

2.5.1.2 Example 2

The data consists on $n = 50$ samples with $p = 4000$ variables which mimic the SLE dataset. The variables are distributed among 20 groups $(G_1, \cdots, G_{20})$ with 200 variables in each group. The true parameters are $\beta_{1,j} = \beta_{2,j} = \beta_{3,j} = 2$, for



(a) Example1(a) (b) Example1(b)

Figure 2.3: Plot of φ as a function of α. The left panel is for Scenario 1 and right panel is for Scenario 2.

$j = 1, \cdots, 10$ and all other $\beta_{k,j} = 0$. Also, $\sigma = 1.5$. Thus the true model is

$$\mathbf{y} = \sum_{i=1}^{10} \mathbf{x}_{1,i}\beta_{1,i} + \sum_{i=1}^{10} \mathbf{x}_{2,i}\beta_{2,i} + \sum_{i=1}^{10} \mathbf{x}_{3,i}\beta_{3,i} + \varepsilon.$$

Each row vector of the design matrix $\mathbf{X}$ is generated from the standard normal distribution with covariance matrix $\mathrm{Cov}(\mathbf{x}_{ki}, \mathbf{x}_{k'j}) = 0.7^{1\{i\neq j\}} \cdot 0.9^{1\{k\neq k'\}}$, $k, k' = 1, \cdots, 20$, $i, j = 1, \cdots, 200$. We assume that a hierarchical structure on top of the 20 groups is as Figure 2.4.

We still denote by P_{ij}^S and P_{ij}^D the frequencies of of occasions on which G_i and G_j are selected or dropped together, respectively. Here, we define a group being selected if and only if at least one of the variables in this group is selected. To demonstrate the hierarchical grouping property of our proposed GHEnet estimator, we calculate P_{ij}^S and P_{ij}^D for the important group pairs $\{i, j\} = \{1, 2\}, \{1, 3\}, \{2, 3\}$ and the pairs among the non-important groups $\{4, \cdots, 20\}$. We benchmark our proposed estimator with $\alpha \in \{0, 0.1, 10, 20\}$ against Group Lasso. The choice of α is based on a similar empirical study as Example 1.

Note that, the GEnet is GHEnet with $\alpha = 0$, i.e. no hierarchical structure is considered. From Table 2.2, $P_{12}^S, P_{13}^S, P_{23}^S$ are very close under GEnet, GHEnet with $\alpha = 0.1$ and Group Lasso because between-group correlations are the same within important groups and no hierarchical structure is considered. When α

Table 2.1: Example1: Frequency of variables being grouped, variable specificity and sensitivity

Scenario 1								
	Important Pairs			Non-important Pairs				
Method	P^S_{45}/P^D_{45}	P^S_{46}/P^D_{46}	P^S_{56}/P^D_{56}	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Spec.	Sens.
Enet	.741/.006	.737/.014	.724/.009	.071/.719	.067/.717	.061/.709	.823	.866
HEnet($\alpha = .1$)	.739/.007	.735/.014	.726/.008	.071/.721	.067/.718	.061/.712	.824	.867
HEnet($\alpha = 8$)	.672/.015	.685/.011	.793/.002	.051/.829	.050/.815	.048/.809	.882	.849
HEnet($\alpha = 15$)	.654/.007	.667/.011	.828/.001	.047/.843	.046/.827	.045/.820	.892	.860
Lasso	.745/.006	.746/.013	.733/.009	.078/.680	.075/.677	.066/.670	.800	.870
OSCAR	.744/.006	.744/.013	.730/.008	.076/.681	.074/.679	.066/.672	.801	.870
Scenario 2								
	Important Pairs			Non-important Pairs				
Method	P^S_{45}/P^D_{45}	P^S_{46}/P^D_{46}	P^S_{56}/P^D_{56}	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Spec.	Sens.
Enet	.595/.039	.615/.040	.613/.034	.124/.398	.118/.396	.133/.392	.640	.784
HEnet($\alpha = .1$)	.612/.034	.629/.038	.626/.033	.096/.410	.097/.409	.134/.398	.665	.794
HEnet($\alpha = 5$)	.625/.015	.650/.020	.718/.012	.007/.713	.008/.708	.057/.725	.872	.837
HEnet($\alpha = 10$)	.661/.015	.688/.021	.730/.011	.005/.740	.005/.748	.073/.768	.890	.826
Lasso	.591/.039	.612/.042	.609/.039	.122/.401	.117/.399	.131/.394	.642	.782
OSCAR	.568/.047	.592/.048	.582/.051	.125/.409	.119/.397	.131/.393	.640	.766

increases to 10 or 20, the fact that $\{G_1, G_3\}$ is closer than the other two pairs ($\{G_1, G_2\}$ and $\{G_2, G_3\}$) on Figure 2.4 leads to higher value of P^S_{13} than those of P^S_{12} and P^S_{23}, indicating $\{G_1, G_3\}$ is more likely selected together than other two pairs of important groups. Parallel results can be found in non-important group pairs P^S_{ij} and P^D_{ij} with $\{i, j\} \in \{4, \cdots, 20\}$ which is omitted in Table 2.2. These results demonstrate the proposed GHEnet estimator has indeed the hierarchical grouping property.

Besides, the sensitivity and specificity at both group level and individual level are also summarized in Table 2.2 to assess the model selection performance. GHEnet outperforms Group Lasso in specificity since Group Lasso tends to select a larger model and can not perform within group selection. The sensitivity of GHEnet with $\alpha \in \{5, 10\}$ is low because the important group G_2 is less likely selected.

In conclusion, the above simulation studies have demonstrated the hierarchical grouping property and model selection consistency of our proposed estimator.

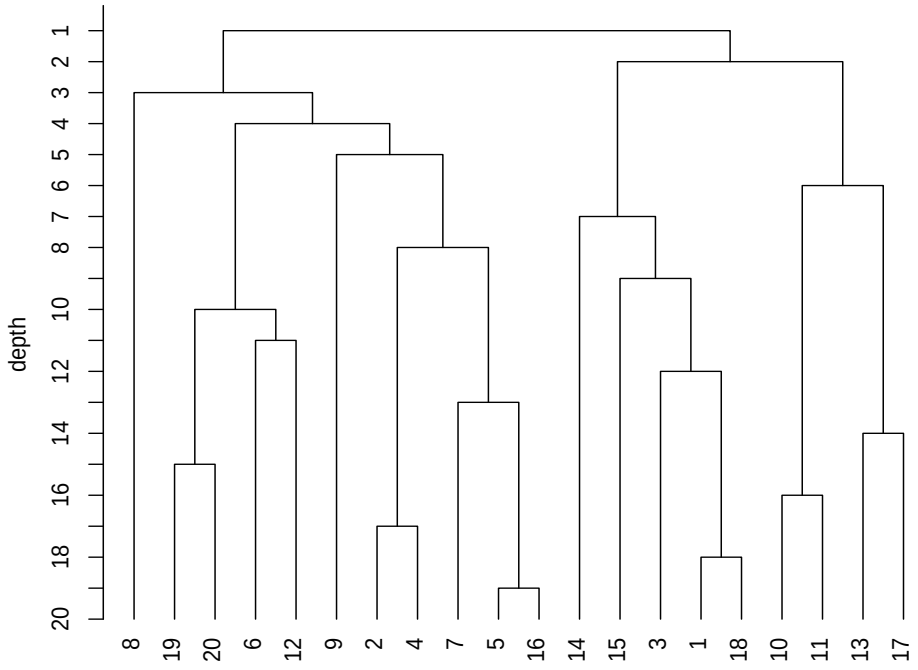


Figure 2.4: Hierarchical structure for Example2.

2.5.2 Real Data Analysis

We consider in this subsection the real data analysis problem on how to perform model selection at between and within group level while taking advantage of the hierarchical information in the Systemic Lupus Erythematosus (SLE) dataset in PBMC study [Chaussabel et al. (2008)].

In the blood genomic studies reported in [Chaussabel et al. (2008)], PBMC samples are obtained from $n = 47$ individual with SLE condition. Transcriptional profiles were generated with Affymetrix U133A and U133B GeneChips (> 44000 probe sets). The gene intensity signal is assessed and normalized using Microarray Suite, Version 5.0 for each probe set. Then logarithmic transformation is performed on the gene intensity level. Among these 44000+ transcripts, 4779 of

Table 2.2: Example2: Frequency of groups being grouped, specificity and sensitivity at group and variable level.

Method	Important Pairs			Specificity		Sensitivity	
	P^S_{12}/P^D_{12}	P^S_{13}/P^D_{13}	P^S_{23}/P^D_{23}	Grp	Var	Grp	Var
GEnet	.889/.003	.886/.003	.882/.009	.760	.661	.935	.864
GHEnet($\alpha = .1$)	.889/.002	.921/.003	.871/.003	.767	.667	.947	.876
GHEnet($\alpha = 10$)	.315/.003	.977/.000	.314/.005	.774	.673	.768	.703
GHEnet($\alpha = 20$)	.075/.006	.986/.000	.075/.007	.798	.694	.687	.616
Group Lasso	.914/.001	.900/.003	.894/.009	.690	.591	.950	.950

Note: GEnet corresponds to $\alpha = 0$ in (2.3.5), i.e. no hierarchical information is considered.

them considered "present" are selected as the input of the module-construction algorithm. Totally 28 modules (groups) are formed. Within each module, the transcripts are coordinately expressed, i.e. highly correlated and usually have similar functions. After a hierarchical clustering algorithm is applied with "complete" linkage on the correlation structure among these 28 modules, the hierarchy tree structure is shown in Figure 2.1. On average, each terminal node represents about 170 transcripts. The goal is to identify the modules and the transcripts within these modules which are potentially related to the individual's disease index: SLE disease-activity index (SLEDAI).

One approach is to use regression analysis which falls into the framework we discussed in Section 2.3. We can treat the SLEDAI as the response variable and all the transcripts as the predictors. Also, the interaction between different modules can be captured by hierarchical clustering. We can deploy the GHEnet procedure to perform the gene selection at both module and individual level. Since unlike the simulation studies we don't know which modules or transcripts are truly important in this dataset, we are not able to calculate P^S_{ij}, P^D_{ij}, sensitivity and specificity for our model selection. To examine the impact of hierarchical scores, we perform a sensitivity analysis by trying different α values.

We summarize in Table 2.3 Part I the selected modules and number of genes

for $\alpha \in \{0, 0.1, 5, 10, 25\}$. Table 2.3 Part II lists the functionality of these modules. The choice of α is based on an empirical study similar to that produces Figure 3 of Example 1. From Table 2.3 Part I, we can see how the identified modules evolve as α increases. The module identification results are the close for $\alpha = 0$ and $\alpha = 0.1$. For large α, the hierarchical structure starts impacting the module selection results as expected. For $\alpha \geq 5$, module #7, #20, #12 are dropped sequentially.

Table 2.3: Part I: Sensitivity Analysis: identified Modules under different $\alpha's$

	Modules	# of Genes
$\alpha = 0$	#7, #10, #12, #20	37
$\alpha = 0.1$	#10, #12, #20	37
$\alpha = 5$	#10, #12	32
$\alpha = 10$	#10	24
$\alpha = 25$	#10	37

Part II: Functionality of selected modules

Module	Functionality
#7	MHC/Ribosomal proteins. Almost exclusively formed by genes encoding MHC class I molecules (HLA-A,B,C,G,E)+Beta 2-microglobulin (B2M) or ribosomal proteins (RPLs,RPSs).
#10	Neutrophils. This set includes genes encoding innate molecules that are found in neutrophil granules (lactotransferrin: LTF, defensin: DEAF1, bacterial permeability increasing protein: BPI, cathelicidin antimicrobial protein: CAMP.).
#12	Ribosomal proteins. Includes genes encoding ribosomal proteins (RPLs, RPSs), Eukaryotic Translation Elongation Factor-family members (EEFs), and Nucleolar proteins (NPM1, NOAL2, NAP1L1).
#20	Interferon-inducible. This set includes interferon-inducible genes: antiviral molecules (OAS1/2/3/L, GBP1, G1P2, EIF2AK2/PKR, MX1, PML), chemokines (CXCL10/IP-10), signaling molecules (STAT1, STAt2, IRF7, ISGF3G).

Chapter 3

Semiparametrically Efficient Estimation and EM algorithm for Generalized Partially Linear Model with Missing Data

3.1 Background

Semiparametric models, which incorporate both the parametric and nonparametric components, have been studied extensively in statistics and econometrics. Most literature discusses estimation methods in complete data cases and, occasionally, some very simple missing at random (MAR) cases. There is little discussion on efficient semiparametric inference in the presence of more general types of missing data. In this chapter we study a general type of missing pattern and prove that the asymptotic covariance of the parameter estimator from a generalized profile likelihood approach can achieve the semiparametric efficiency bound under some mild conditions. We also propose an estimation algorithm to estimate both the parametric and nonparametric components.

Our developments are illustrated using a generalized partially linear regression model,

$$h(EY_i) = \eta_i \text{ with } \eta_i = \mathbf{W}_i^T\boldsymbol{\beta} + g(V_i), \quad \text{for } i = 1, 2, \cdots, N, \tag{3.1.1}$$

although the developments can be extended, with only mild modifications, to other types of semiparametric regression models that include varying coefficient models, single-index models, among others. Here, in Model (3.1.1), h is a known link function, Y_i is the response variable with known distribution (with realization

y_i), $\mathbf{W}_i$ is a $q \times 1$ vector of covariate variable (with realization $\mathbf{w}_i$), V_i is a scalar covariate variable (with realization v_i), $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown parameter and $g(\cdot)$ is an unknown smooth function. For simplicity, we assume that the parameter space $\mathcal{B}$ of $\boldsymbol{\beta}$ is a compact subset of $\mathbb{R}^q$, the domain $\mathcal{H}$ of $g(\cdot)$ is a compact subset of $\mathbb{R}$ and the support $\mathcal{V}$ of $g(\cdot)$ is a compact subset of $\mathbb{R}$.

In the case of complete data without any missing values, the generalized partially linear regression model (3.1.1) and its various extensions have been extensively studied in the literature; see, e.g., [Ruppert, Wang and Carroll (2003)] and the references therein. Often the main interest or objective is to estimate the finite dimensional parametric component $\boldsymbol{\beta}$, while the nonparametric component is considered as the infinite dimensional nuisance parameter. [Severini and Wong (1992)] proposed an estimation method that maximizes the generalized profile likelihood under the conditionally parametric model and proved that the estimation method leads to an asymptotically efficient estimator of the parameter of interest, where the efficiency refers to the usual asymptotic semiparametric efficiency as described in [Newey (1990)]. [Ahmad, Leelahanon and Li (2005)] proposed a general series method to estimate semiparametric partially linear varying coefficient model and showed that the estimator of the finite dimensional parameters is semiparametrically efficient when the error is conditionally homoscedastic. [Ma, Chiou and Wang (2006)] proposed a family of consistent estimators and showed that the optimal semiparametric efficiency bound can be reached by a semiparametric kernel estimator in this family. Some other publications include [Carroll et al. (1997)],[Cai, Fan and Li (2000)], [Fan and Huang (2005)], [Boente, He and Zhou (2006)], [Xie, Simpson and Carroll (2008)],
[Lam and Fan (2008)], among others.

Although missing data are common in many practices (c.f.,

[Little and Rubin (2002)]), there are few discussions on nonparametric or semiparametric regression in the presence of missing data. We consider in this chapter a very general missing structure similar to the one studied in [Wu (1983)]. In particular, suppose that $\mathscr{Z}$ is a sample space for the observed samples and $\mathscr{Y}$ is a sample space for the complete samples. Instead of observing the complete data $\mathbf{y} = (y_1, \cdots, y_N)^T$ in $\mathscr{Y}$, we observe the incomplete data $\mathbf{z} = \mathbf{z}(\mathbf{y}) = (z_1(\mathbf{y}), \cdots, z_n(\mathbf{y}))^T$ in $\mathscr{Z}$, where N is the number of complete responses and $n \leq N$ is the number of observed responses. Here, we assume that each element in $\mathscr{Z}$ is a many-to-one mapping from $\mathscr{Y}$, and the elements $z_1, \cdots, z_n$ of $\mathscr{Z}$ can potentially be correlated. Denote by $l\big(\boldsymbol{\beta}, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},g}\big(\mathbf{z}, \mathbf{w}, \mathbf{v}\big)$ the log-likelihood function of the observed data, where $\mathbf{w} = ({\mathbf{w}_1}^T, \cdots, \mathbf{w}_N^T)^T$ and $\mathbf{v} = (v_1, \cdots, v_N)^T$. The developments in this chapter only require that the observed data likelihood function satisfy the following condition (I),

Condition I. For a fixed but arbitrary $\boldsymbol{\beta} \in \mathcal{B}$ and any function $g(\cdot)$,

$$n^{-1}\Big\{ l\big(\boldsymbol{\beta}, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - E_0\big[l\big(\boldsymbol{\beta}, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big] \Big\} \to_p 0,$$

along with some mild conditions specified in Section 3.2.1. Note that, this assumption covers a wide range of missing patterns. In particular, it subsumes the cases of one-to-one mapping on independent responses, including those MAR assumptions considered in [Wang, Linton and Härdle (2004)] and [Wang, Rotnitzky and Lin (2010)]. In addition, it also covers many many-to-one mappings. For example, in the annual household income report [U.S. Census Bureau (2010)], it is common that only the median income of each region is recorded instead of each household due to privacy considerations. Similarly, in education, reported average scores at class- or grade-level are often available instead of scores at individual-level. Another example of the many-to-one mapping is the class of group testing samples (see, [Dorfman (1943)] and a large volume of papers followed) , where the many-to-one mapping is from individual-level responses (not observed complete responses) to

the observed group-level testing responses (observed responses). Furthermore, in some of the many-to-one mappings, different observed responses may depend on one or more common complete responses, which is often the case in multi-level group testing samples and also among many other situations. In this type of cases, the observed data $\mathbf{z}$ are correlated. Depending on the situation, the dependent structure in observed responses $\mathbf{z}$ could be either a block dependent series, or a Markov chain, or an α-mixing sequence, or some other patterns. In some of these cases, the observed likelihood function can be very complicated or even unavailable.

The rest of the chapter is organized as follows. In Section 3.2, we present our estimation method and the large sample properties of the estimator, including both consistency and efficiency. Estimation algorithm using EM algorithms is given in Section 3.3. In Section 3.4, we establish the connection between our proposed estimator and the estimator proposed by [Ahmad, Leelahanon and Li (2005)]. In Section 3.5, we evaluate the finite sample performance of proposed algorithm by two simulation studies and an analyze of real data from the Nebraska Infertility Prevention Project.

3.2 Generalized Profile Likelihood Approach

3.2.1 Generalized Partially Linear Model and Generalized Profile Likelihood Function

In the generalized partially linear model (3.1.1), we assume that the response y_i and covariates $\mathbf{w}_i$ and v_i are n independently and identically distributed replicates, i.e., $(y_i, \mathbf{w}_i, v_i)$ is a sample of $(Y_i, \mathbf{W}_i, V_i) \sim (Y, \mathbf{W}, V)$. Following [Severini and Wong (1992)], the conventional (complete) log-likelihood function

for the generalized partially linear model (3.1.1) is

$$l\big(\boldsymbol{\beta}, g(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},g}\big(\mathbf{y}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},g}\big(\mathbf{y}|\mathbf{w}, \mathbf{v}\big) + \log f(\mathbf{w}, \mathbf{v}), \quad (3.2.1)$$

where $f(\mathbf{w}, \mathbf{v})$ is the marginal density function of $(\mathbf{w}, \mathbf{v})$. But we are interested in the case where some information on responses y_i, $i = 1, \ldots, N$, is missing, and we only observe z_i, $i = 1, \cdots, n$, which could be a one-to-one or many-to-one mapping from the complete responses.

Since $\mathbf{y} = (y_1, \ldots, y_n)$ is not completely observed, we need to consider the log-likelihood function for the observed data $(\mathbf{z}, \mathbf{w}, \mathbf{v})$,

$$l\big(\boldsymbol{\beta}, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},g}\big(\mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},g}\big(\mathbf{z}|\mathbf{w}, \mathbf{v}\big) + \log f(\mathbf{w}, \mathbf{v}). \quad (3.2.2)$$

In addition to Condition (I), we impose a Lipschitz condition on the observed likelihood function:

CONDITION C. For any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and any $g_1(\cdot)$, $g_2(\cdot)$, there exist A_n and B_n such that in probability

$$\frac{1}{n}\left|l\big(\boldsymbol{\beta}_1, g_1(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\boldsymbol{\beta}_2, g_2(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\right| \leq A_n \left|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right| + B_n \left\|g_1 - g_2\right\|,$$

where A_n and B_n are bounded by constants A and B in probability, respectively.

Similar to [Severini and Wong (1992)], we denote by $g_{\boldsymbol{\beta}}(v)$ the *least favorable curve* of the nonparametric function $g(v)$, which is defined as

$$g_{\boldsymbol{\beta}}(v) = \text{argmax}_{\eta} E_0\left\{l\big(\boldsymbol{\beta}, \eta; \mathbf{z}, \mathbf{w}, \mathbf{v}\big)|\mathbf{v} = v \cdot \overrightarrow{1}\right\}$$

for each $\boldsymbol{\beta}$ and v. Here, $\overrightarrow{1}$ is a size N vector with each entry equals to 1, E_0 is the expectation taken under the true parameters $\boldsymbol{\beta}_0$ and $g_0(v)$. Also, let $\hat{g}_{\boldsymbol{\beta}}$ be an estimator of $g_{\boldsymbol{\beta}}$. Then, the function

$$l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f_{\boldsymbol{\beta},\hat{g}_{\boldsymbol{\beta}}}\big(\mathbf{z}, \mathbf{w}, \mathbf{v}\big)$$

is called a *generalized profile likelihood* for $\boldsymbol{\beta}$. See, [Severini and Wong (1992)] for detailed discussions on the concept of least favorable curve and the least favorable

direction. In our proposed generalized profile approach, the estimator of $\boldsymbol{\beta}$ is obtained by maximizing the observed generalized profile likelihood function,

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big). \tag{3.2.3}$$

Finally, we write

$$l_0(\boldsymbol{\beta}) = \lim_{n\to\infty} \frac{1}{n} E_0 \left\{ l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \right\}, \tag{3.2.4}$$

where we assume that the limit exists. Also, as in [Severini and Wong (1992)], we require that $l_0(\boldsymbol{\beta})$ is maximized at true parameter $\boldsymbol{\beta}_0$. We prove next in Section 3.2.2 and 3.2.3 that $\hat{\boldsymbol{\beta}}$ is a consistent and semiparametric efficient estimator of true $\boldsymbol{\beta}_0$, under some regularity conditions.

3.2.2 Consistency

3.2.2.1 Some Regularity Conditions

Following [Severini and Wong (1992)], we develop our theoretical results for a general class of estimators of the nuisance nonparametric function $\hat{g}_{\boldsymbol{\beta}}(\cdot)$. In particular, we impose the following NP conditions for $\hat{g}_{\boldsymbol{\beta}}(\cdot)$.

Condition NP. Denote by $\|\cdot\|$ the infinity norm of functions. We assume that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \big\|\hat{g}_{\boldsymbol{\beta}} - g_{\boldsymbol{\beta}}\big\| = o_p(1)$. We also assume that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\frac{\partial \hat{g}_{\boldsymbol{\beta}}(v)}{\partial \boldsymbol{\beta}}\|$ and $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\frac{\partial g_{\boldsymbol{\beta}}(v)}{\partial \boldsymbol{\beta}}\|$ are finite.

3.2.2.2 Theoretical Results

Theorem 3.2.1 states that the $\hat{\boldsymbol{\beta}}$ defined in (3.2.3) is a consistent estimator of true parameter $\boldsymbol{\beta}_0$, and the plug-in estimator $\hat{g}_{\hat{\boldsymbol{\beta}}}(\cdot)$ converges to the least favorable curve $g_{\boldsymbol{\beta}_0}(\cdot)$. The proof is given in Appendix.

Theorem 3.2.1 (**Consistency**). *Under the Conditions I, C and NP, the estimator of the parametric component $\hat{\boldsymbol{\beta}}$ is consistent:*

$$\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}_0, \quad as \quad n \to \infty.$$

Furthermore, if we plug the estimator $\hat{\boldsymbol{\beta}}$ *in the estimator of the nonparametric component* $\hat{g}_{\boldsymbol{\beta}}(\cdot)$*, the plug-in estimator* $\hat{g}_{\hat{\boldsymbol{\beta}}}(\cdot)$ *converges to the least favorable curve* $g_{\boldsymbol{\beta}_0}(\cdot)$*:*

$$\sup_{v\in\mathcal{V}}\left|\hat{g}_{\hat{\boldsymbol{\beta}}}(v)-g_{\boldsymbol{\beta}_0}(v)\right|=o_p(1)\quad \textit{and}\quad \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{g}_{\hat{\boldsymbol{\beta}}}(v_i)-g_{\boldsymbol{\beta}_0}(v_i)\right\}^2=o_p(1).$$

3.2.3 Asymptotic Normality

Because of the observed responses $\mathbf{z}=(z_1,\,\cdots,z_n)$ could potentially be correlated, we need some additional conditions on the likelihood function so that the Central Limit Theorem for dependent variables can be applied. In particular, for notation convenience and without loss of generality, we re-write the joint probability density function of observed data $\mathbf{z}=(z_1,\cdots,z_n)$ as

$$\begin{aligned} f_{\boldsymbol{\beta},g}(z_1,\cdots,z_n) &= f_{\boldsymbol{\beta},g}(z_1)f_{\boldsymbol{\beta},g}(z_2|z_1)\cdots f_{\boldsymbol{\beta},g}(z_n|z_{n-1},\cdots,z_1)\\ &= \prod_{k=1}^{n} f_k(\boldsymbol{\beta},g)=e^{\sum_{k=1}^{n}\log f_k(\boldsymbol{\beta},g)}, \end{aligned}$$

where $f_k(\boldsymbol{\beta},g)=f_{\boldsymbol{\beta},g}(z_k|z_{k-1},\cdots,z_1)$. Also, we denote by $l_k(\boldsymbol{\beta},g)=\log f_k(\boldsymbol{\beta},g)$ and $l(\boldsymbol{\beta},g)=\log f_{\boldsymbol{\beta},g}(\mathbf{z})=\sum_{k=1}^{n} l_k(\boldsymbol{\beta},g)$. Here, for notation simplicity, the order of the f_k sequence corresponds to the original indexes of the z_k responses. But it does not have to be the case and our developments work for any set of permuted indexes as long as its corresponding sequence follows the condition specified in Section 3.2.3.1 next.

3.2.3.1 Additional Conditions

To obtain the asymptotic normality for the estimators $\hat{\boldsymbol{\beta}}$, additional regularity conditions C_1-C_5 below are imposed on the likelihood function. These conditions are similar to the efficiency and normality conditions imposed in [Bhat (1974)] for parametric models with dependent variables such that the central limit theorem for martingales holds.

CONDITION C_1. Assume that the support of $f_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)$ is independent of $\boldsymbol{\beta}$. Differentiation with respect to $\boldsymbol{\beta}$ can be carried under the integral sign with respect to z_k up to second order for $l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)$, i.e., for $k = 1, \cdots, n$,

$$E\frac{\partial}{\partial \boldsymbol{\beta}} l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big) = 0,$$

$$E\left\{\frac{\partial l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}}\right\}\left\{\frac{\partial l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}}\right\}^T = -E\left\{\frac{\partial^2 l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right\}$$

CONDITION C_2. The third order derivatives $\{\partial^3/\partial\boldsymbol{\beta}_i\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_l\} l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)$ are bounded in probability uniformly for all $i, j, l = 1, \cdots, q$ and in k.

CONDITION C_3. As $n \to \infty$,

$$\frac{1}{n}\sum_{k=1}^{n} i_k(\boldsymbol{\beta}_0) \triangleq \frac{1}{n}\sum_{k=1}^{n} E\left\{\frac{\partial l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}}\right\}\left\{\frac{\partial l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}}\right\}^T \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \longrightarrow i_{\boldsymbol{\beta}_0},$$

where $i_{\boldsymbol{\beta}_0}$ is a positive definite matrix.

CONDITION C_4. There exits some $\delta > 0$ such that, as $n \to \infty$,

$$\frac{1}{n^{1+\delta/2}}\sum_{k=1}^{n} E\Big|\frac{\partial l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}_i}\Big|^{2+\delta} \longrightarrow 0,$$

for all $i = 1, \cdots, q$.

CONDITION C_5. As $n \to \infty$,

$$\frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}\left\{\frac{\partial^2 l_k\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}\big)}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j}\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} < \infty,$$

for all $i = 1, \cdots, q$ and $i = 1, \cdots, q$.

Condition C_6 and C_7 below are requirements on the estimator of least favorable curve. They are the same as those shown in Lemma 2 and Lemma 3 of [Severini and Wong (1992)].

CONDITION C_6.

$$\frac{1}{\sqrt{n}}\left\{\frac{\partial l(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} - \frac{\partial l(\boldsymbol{\beta}, g_{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right\}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = o_p(1).$$

CONDITION C_7.

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{n}\Big\| \frac{\partial^2 l(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - \frac{\partial^2 l(\boldsymbol{\beta}, g_{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \Big\| = o_p(1).$$

3.2.3.2 Theoretical Results

Theorem 3.2.2 establishes that $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with asymptotic variance equal to the semiparametric efficiency bound, $i_{\boldsymbol{\beta}}^{-1}$. The proof is given in Appendix.

Theorem 3.2.2 (**Semiparametric Efficiency**). *Under Conditions I, C and NP and Condition $C_1 - C_7$ above, we have, as $n \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_{\mathscr{D}} N(0, i_{\boldsymbol{\beta}}^{-1}),$$

where $i_{\boldsymbol{\beta}}^{-1}$ is the semiparametric efficiency bound. Also,

$$\hat{i}_{\boldsymbol{\beta}} = -\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

can be used to consistently estimate the information bound $i_{\boldsymbol{\beta}}$ with

$$\hat{i}_{\boldsymbol{\beta}} \to_p i_{\boldsymbol{\beta}} \quad as \quad n \to \infty.$$

3.3 Estimation Algorithm

3.3.1 An Iterative Algorithm

The observed log-likelihood $l(\boldsymbol{\beta}, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v})$ in (3.2.2) depends on the missing structure and, in general, it can be very complicated or even unavailable. Direct maximization of the observed log-likelihood function may be unpractical, except in some very special cases. However, the complete log-likelihood function $l(\boldsymbol{\beta}, g(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v})$ in (3.2.1) often has a much simpler form, and the EM algorithm can be used to deal with missing data in this case. We propose an iterative approach to estimate the parameters $\boldsymbol{\beta}$ and the nonparametric component $g(\cdot)$ in

model (3.1.1) with missing data. The key idea of the approach is to use a backfitting algorithm in combination with the generalized profile likelihood method. In this approach, we first fix the parametric component and estimate the nonparametric component using an EM algorithm and a spline smoothing method. Note that, this estimator of the nonparametric component depends on the value at which the parameter $\boldsymbol{\beta}$ is held fixed, thus the estimator of the nonparametric part can be considered as a function of the parameter $\boldsymbol{\beta}$. This estimator of nonparametric component is then used to create a generalized profile likelihood of parameter $\boldsymbol{\beta}$ using the observed log-likelihood function. The estimator of parameter $\boldsymbol{\beta}$ can be obtained by using an EM algorithm for a set of semiparametric estimating equations obtained from the generalized profile likelihood and the least favorable curve. We iterate between the estimation of the parameters and the estimation of the nonparametric component until the algorithm converges. In particular, the proposed estimating approach iterates between the following two modules:

- Module I (Estimating parametric component): Fix the old estimator of nonparametric function and its first derivative with respect to $\boldsymbol{\beta}$, say $\hat{g}^{old}(\cdot)$ and $\hat{g}'^{old}(\cdot)$. Then update the estimate of parametric part $\hat{\boldsymbol{\beta}}^{new}$ using the weighted iterative least square (WILS) and EM algorithm.

- Module II (Estimating nonparametric component): Fix the old estimators of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}^{old}$. Update the estimator of nonparametric function $\hat{g}(\cdot)$ and its first derivative to $\boldsymbol{\beta}$, say $\hat{g}'(\cdot)$, iteratively until converge to get $\hat{g}^{new}(\cdot)$ and $\hat{g}'^{new}(\cdot)$ using EM algorithm and smoothing methods.

Note that, in the generalized partially linear model (3.1.1), Y_i is independently distributed with the density function

$$f(y_i|\boldsymbol{\beta}) = \exp\{y_i\theta(\eta_i) - b(\theta(\eta_i)) + c(y_i)\} \quad \text{for } i = 1, 2, \cdots, N, \tag{3.3.1}$$

but we only observe $\mathbf{z} = \mathbf{z}(\mathbf{y})$ and $\mathbf{w}$ and $\mathbf{v}$. Denote by $n \times 1$ vectors $\mathbf{g} = (g(v_1), g(v_2), \cdots, g(v_N))^T$ and $\mathbf{g}' = (g'(v_1), g'(v_2), \cdots, g'(v_N))^T$ its derivative with respect to $\boldsymbol{\beta}$. Also, denote by $\hat{\mathbf{g}}$ and $\hat{\mathbf{g}}'$ their estimators, respectively. Then, in the first module in which we update $\hat{\boldsymbol{\beta}}$, given $\hat{\boldsymbol{\beta}}^{old}$, $\hat{\mathbf{g}}^{old}$, $\hat{\mathbf{g}}'^{old}$, the Q function in the EM algorithm can be written as:

$$\begin{aligned} Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{old}) &= E\left\{\log f(\mathbf{y})\Big|\mathbf{z}, \hat{\boldsymbol{\beta}}^{old}, \hat{\mathbf{g}}^{old}, \hat{\mathbf{g}}'^{old}\right\} \\ &= E\left[\sum_{i=1}^{N}\{y_i\theta(\eta_i) - b(\theta(\eta_i)) + c(y_i)\}\Big|\mathbf{z}, \hat{\boldsymbol{\beta}}^{old}, \hat{\mathbf{g}}^{old}, \hat{\mathbf{g}}'^{old}\right] \\ &= \sum_{i=1}^{N}\theta(\eta_i)\tilde{y}_i - \sum_{i=1}^{N} b(\theta(\eta_i)) + \sum_{i=1}^{N} E\left\{c(y_i)\Big|\mathbf{z}, \hat{\boldsymbol{\beta}}^{old}, \hat{\mathbf{g}}^{old}, \hat{\mathbf{g}}'^{old}\right\}, \end{aligned}$$

where $\tilde{y}_i = E(y_i|\mathbf{z}, \hat{\boldsymbol{\beta}}^{old}, \hat{\mathbf{g}}^{old}, \hat{\mathbf{g}}'^{old})$. Define the working response u_i and weight T_i for the ith observation:

$$u_i = \eta_i^{old} + (\tilde{y}_i - \mu_i^{old})\frac{d\eta_i}{d\mu_i}\Big|_{old} \quad \text{and} \quad T_i = \frac{1}{\mathrm{Var}(u_i)}\Big|_{old} = \frac{[k'(\eta_i)]^2}{\mathrm{Var}(\tilde{y}_i)}\Big|_{old},$$

where $\eta_i^{old} = \mathbf{w}_i{}^T\hat{\boldsymbol{\beta}}^{old} + \hat{g}^{old}(v_i)$, $\mu_i^{old} = k(\eta_i^{old})$ and $k(\cdot) = h^{-1}(\cdot)$. We solve the following estimating equation to update $\hat{\boldsymbol{\beta}}^{new}$,

$$\sum_{i=1}^{N}\left\{u_i - \mathbf{w}_i^T\boldsymbol{\beta} - \hat{g}^{old}(v_i)\right\}\left\{\mathbf{w}_i + \hat{g}'^{old}(v_i)\right\}T_i = 0.$$

We then replace $\hat{\boldsymbol{\beta}}^{old}$ with $\hat{\boldsymbol{\beta}}^{new}$ in the expression of $\tilde{y}_i$, u_i and T_i, and solve the above estimation equation again to get a new $\hat{\boldsymbol{\beta}}^{new}$. Repeating this procedure until $\hat{\boldsymbol{\beta}}^{new}$ numerically converges.

In the second module, we estimate the non-parametric function $g(\cdot)$ and its derivative function $g'(\cdot)$ when given $\hat{\boldsymbol{\beta}}^{new}$, $\hat{\mathbf{g}}^{old}$. We use a B-spline method to smooth the functions $g(\cdot)$ and $g'(\cdot)$. The Q function which we are trying to maximize is the same as the Q function in the first module except that the $\tilde{y}_i$ there is replaced by $\tilde{y}_i = E(y_i|\mathbf{z}, \hat{\boldsymbol{\beta}}^{new}, \hat{\mathbf{g}}^{old}, \hat{\mathbf{g}}'^{old})$. Now, define the working response u_i for the ith observation as:

$$u_i = \eta_i^{old} + (\tilde{y}_i - \mu_i^{old})\frac{d\eta_i}{d\mu_i}\Big|_{old},$$

where $\eta_i^{old} = \mathbf{w}_i^T \hat{\boldsymbol{\beta}}^{new} + \hat{g}^{old}(v_i)$ and $\mu_i^{old} = g(\eta_i^{old})$. We update $\hat{\mathbf{g}}^{new}$ and $\hat{\mathbf{g}}'^{new}$ by smoothing

$$\hat{\mathbf{g}}^{new} = M\tilde{\mathbf{g}} \quad \text{and} \quad \hat{\mathbf{g}}'^{new} = M\tilde{\mathbf{g}}',$$

where $\tilde{\mathbf{g}}=(u_1-\mathbf{w}_1^T\hat{\boldsymbol{\beta}}^{new}, \cdots, u_N-\mathbf{w}_N^T\hat{\boldsymbol{\beta}}^{new})^T$ and $\tilde{\mathbf{g}}'^{new}=(\frac{\partial}{\partial\hat{\boldsymbol{\beta}}^{new}}u_1-\mathbf{w}_1, \cdots, \frac{\partial}{\partial\hat{\boldsymbol{\beta}}^{new}}u_N-\mathbf{w}_N)^T$. Here, the matrix M is a projection matrix defined by $M = P(P^TP)^{-1}P^T$, where P is an $N \times s$ matrix, i^{th} row of which is s B-spline basis functions of v_i. We then replace $\hat{\mathbf{g}}^{old}$ and $\hat{\mathbf{g}}'^{old}$ by $\hat{\mathbf{g}}^{new}$ and $\hat{\mathbf{g}}'^{new}$ in the expression of $\tilde{y}_i$, u_i and get a new set of $\hat{\mathbf{g}}^{new}$ and $\hat{\mathbf{g}}'^{new}$. Repeating this procedure until $\hat{\mathbf{g}}^{new}$ numerically converge.

Finally, we iteratively cycle between the above Modules I and II until both $\|\hat{\boldsymbol{\beta}}^{new} - \hat{\boldsymbol{\beta}}^{old}\|$ and $\|\hat{\mathbf{g}}^{new} - \hat{\mathbf{g}}^{old}\|$ are very small.

3.3.2 Gibbs Sampling Method for Computing Expectations

In the EM algorithms described above, we need to calculate the expectation $E(y_i|\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$ and its derivative $\frac{\partial}{\partial\boldsymbol{\beta}}E(y_i|\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$. When these two quantities have explicit expressions, the algorithms can be implemented directly. Sometimes, the missing pattern may be very complicated and the expectations do not have any explicit expressions. In this case, if the conditional density distributions $f(y_i|\mathbf{y}_{-i}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$, for each $i = 1, \cdots, N$ can be simulated from, we can utilize a Gibbs sampling method to numerically calculate the expectation and its derivative by Monte-Carlo approximations. Here, $\mathbf{y}_{-i}$ represents the individual response vector but without the ith response value.

Suppose $\mathbf{y}^* = (y_1^*, \cdots, y_N^*)^T$ is a set of Gibbs samples from $f(y_1, \cdots, y_N|\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$ and we repeat the simulation M times to get M sets of such Gibbs samples. Then the expectation $E(y_i|\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$ can be approximated by $\sum_* y_i^*/M$, where the sum is over the M sets of Gibbs samples. To calculate $\frac{\partial}{\partial\boldsymbol{\beta}}E(y_i|\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}, \mathbf{g}')$, we

note that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} E(y_i|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}') &= \int y_i \frac{\partial}{\partial \boldsymbol{\beta}} f(\mathbf{y}|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')d\mathbf{y} \\
&= \int y_i \left\{ \frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}')}{\partial \boldsymbol{\beta}} - \frac{\partial \log f(\mathbf{z}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}')}{\partial \boldsymbol{\beta}} \right\} f(\mathbf{y}|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')d\mathbf{y} \\
&= \int y_i \frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}')}{\partial \boldsymbol{\beta}} f(\mathbf{y}|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')d\mathbf{y} \\
&\qquad - \int y_i E\left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}') \Big| \mathbf{z} \right\} f(\mathbf{y}|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')d\mathbf{y} \\
&= E\left\{ y_i \frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}')}{\partial \boldsymbol{\beta}} \Big| \mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}' \right\} \\
&\qquad -E\left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{g},\mathbf{g}') \Big| \mathbf{z} \right\} E(y_i|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')
\end{aligned}$$

Thus, $\partial/\partial\boldsymbol{\beta} E(y_i|\mathbf{z},\boldsymbol{\beta},\mathbf{g},\mathbf{g}')$ can be approximated by

$$\frac{1}{M}\sum_{*} y_i^* \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}^*|\boldsymbol{\beta},\mathbf{g},\mathbf{g}') - \left\{ \frac{1}{M}\sum_{*} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}^*|\boldsymbol{\beta},\mathbf{g},\mathbf{g}') \right\} \left\{ \frac{1}{M}\sum_{*} y_i^* \right\}.$$

Here, again, the sum's are over the M sets of Gibbs samples.

3.3.3 Computing the Asymptotic Variance of $\hat{\boldsymbol{\beta}}$

The asymptotic variance of the parameter estimators often needs to be computed for statistical inference. Note that

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z},\mathbf{w},\mathbf{v}\big) &= E\left\{ \frac{\partial^2 \log f_{\boldsymbol{\beta},g_{\boldsymbol{\beta}}}\big(\mathbf{y}|\mathbf{w},\mathbf{v}\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \Big| \mathbf{z},\mathbf{w},\mathbf{v},\boldsymbol{\beta},g_{\boldsymbol{\beta}}(\cdot) \right\} \\
&\qquad -E\left\{ \frac{\partial^2 \log f_{\boldsymbol{\beta},g_{\boldsymbol{\beta}}}\big(\mathbf{y}|\mathbf{z},\mathbf{w},\mathbf{v}\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \Big| \mathbf{z},\mathbf{w},\mathbf{v},\boldsymbol{\beta},g_{\boldsymbol{\beta}}(\cdot) \right\} \\
&= E\left\{ \frac{\partial^2 \log f_{\boldsymbol{\beta},g_{\boldsymbol{\beta}}}\big(\mathbf{y}|\mathbf{w},\mathbf{v}\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \Big| \mathbf{z},\mathbf{w},\mathbf{v},\boldsymbol{\beta},g_{\boldsymbol{\beta}}(\cdot) \right\} \\
&\qquad +\mathrm{Var}\left\{ \frac{\partial \log f_{\boldsymbol{\beta},g_{\boldsymbol{\beta}}}\big(\mathbf{y}|\mathbf{w},\mathbf{v}\big)}{\partial \boldsymbol{\beta}} \Big| \mathbf{z},\mathbf{w},\mathbf{v},\boldsymbol{\beta},g_{\boldsymbol{\beta}}(\cdot) \right\} \\
&\doteq I_1 + I_2. \qquad (3.3.2)
\end{aligned}$$

Base on the density function (3.3.1) and after some straightforward algebraic calculations, we have

$$I_1 = \sum_{i=1}^{N} \left[\{\tilde{y}_i - b'(\theta(\eta_i))\}\, \theta''(\eta_i) - b''(\theta(\eta_i))\{\theta'(\eta_i)\}^2 \right] \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_i\right) \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_i\right)^T,$$

$$I_2 = \sum_{i=1}^{N} \theta'^2(\eta_i)\tilde{\sigma}_i^2 \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_i\right) \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_i\right)^T + 2\sum_{i<j} \theta'(\eta_i)\theta'(\eta_j)\tilde{\sigma}_{ij} \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_i\right) \left(\frac{\partial}{\partial \boldsymbol{\beta}}\eta_j\right)^T,$$

where $\tilde{y}_i = E(y_i|\mathbf{z}, \boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot))$, $\tilde{\sigma}_i^2 = \mathrm{Var}(y_i|\mathbf{z}, \mathbf{w}, \mathbf{v}, \boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot))$, $\tilde{\sigma}_{ij} = \mathrm{Cov}\big(y_i, y_j|\mathbf{z}, \mathbf{w}, \mathbf{v}, \boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot)\big)$. The conditional expectation $\tilde{y}_i$ and covariance $\tilde{\sigma}_i^2, \tilde{\sigma}_{ij}$ in the above formula can be calculated either by explicit expressions (if the missing pattern is simple) or by utilizing the Gibbs Sampling method mentioned in Section 3.3.2. To calculate $\hat{i}_{\boldsymbol{\beta}}$ in Theorem 3.2.2, we use the plug-in method, substituting the true parameters in I_1 and I_2 with their corresponding estimates:

$$\hat{i}_{\boldsymbol{\beta}} = -\frac{1}{n}\left(I_1\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, g_{\boldsymbol{\beta}}=\hat{g}_{\hat{\boldsymbol{\beta}}}} + I_2\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, g_{\boldsymbol{\beta}}=\hat{g}_{\hat{\boldsymbol{\beta}}}} \right)$$

The estimator of asymptotic variance is just the inverse matrix of $\hat{i}_{\boldsymbol{\beta}}$. Note that, in the calculation, we only need to know the likelihood function of the complete data (which is readily available), in stead of the likelihood function of the observed data (which is often hard to get).

3.3.4 Generalized Cross-Validation for Missing data

In numerical examples studied in Sections 3.5 and 3.5.2 later, we will use a univariate cubic B-spline basis function to carry out the smoothing task in the second module of proposed algorithm. The univariate cubic B-spline basis function is defined by

$$B(v|t_0, \cdots, t_4) = \frac{4}{3}\sum_{i=0}^{4} c_{i,3}\left\{\max(0, v - t_i)\right\}^3,$$

where $c_{j,3} = \prod_{j=0, j\neq i}^{4} \frac{1}{t_j - t_i}$ and $t_0, \cdots, t_4$ are the evenly-spaced design knots. The number of interior knots r of the B-spline is a tuning parameter which controls

the smoothness of our estimator. The usual generalized cross-validation criterion (GCV) such as those discussed in [Gu (2002)] can not be directly applied to the missing data. We propose a modified GCV criterion to select r, which minimizes the generalized cross-validation criterion for missing data (GCVM) defined as following:

$$\begin{aligned} GCVM(r) &= N \times \frac{\text{residual sum of squares of working response}}{(\text{equivalent degrees of freedom})^2} \\ &= N \times \frac{\| \, T^{\frac{1}{2}}(\mathbf{u} - \mathbf{w}^T\boldsymbol{\beta} - \hat{\mathbf{g}}_{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \, \|^2}{\{N - (r+4+q)\}^2}. \end{aligned}$$

Here, N is the number of complete observations, and $\mathbf{u}$ is the working response and T is the weight defined in the algorithm of Section 3.3.1. The degrees of freedom here is $N-(r+4+q)$, noting that the univariate cubic B-spline basis function with r interior knots has $r+4$ free parameters and $\boldsymbol{\beta}$ is q dimension. We consider this GCVM as a modification of the GCV defined in [Mao and Zhao (2003)].

3.4 Connection to the efficient estimators from complete data

The estimators obtained in Section 3.3 can often be written as the conditional expectation of the semiparametric efficient estimator obtained from the complete dataset, conditional on the observed data. For instance, consider the standard Gaussian partially linear model,

$$Y_i = \mathbf{W}_i^T\boldsymbol{\beta} + g(V_i) + e_i, \quad e_i \sim N(0,\sigma^2),\, i = 1, 2, \cdots, N, \tag{3.4.1}$$

where σ^2 is the unknown variance of the normal error e_i. Model (3.4.1) is a special case of (3.1.1) and also a special case of the Gaussian partially linear varying coefficient model considered by [Ahmad, Leelahanon and Li (2005)]. If we observe all N of y_i's with no missing data, we can use the method proposed by [Ahmad, Leelahanon and Li (2005)] to obtain a semiparametrically efficient estimators, say $\hat{\boldsymbol{\beta}}^*(\cdot)$, for the regression parameter $\boldsymbol{\beta}$ and the corresponding

estimator of nonparametric component, say $\hat{\mathbf{g}}^*$. When there are missing data as outlined in Sections 3.1 and 3.2, we can use the method proposed in Section 3.3 to obtain an estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, and the corresponding estimator of the nonparametric component, say $\hat{\mathbf{g}}(\cdot)$. The theorem below shows that the set of estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{g}}(\cdot)$ obtained from the method using the algorithm in Section 3.3 and the set of estimators $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{g}}^*(\cdot)$ using the method proposed by [Ahmad, Leelahanon and Li (2005)] are closely related. The proof is given in Appendix.

Theorem 3.4.1 (**Connection to the efficient estimator without missing data**). *Under the Gaussian partially linear model (3.4.1), the estimators* $\hat{\boldsymbol{\beta}}$ *and* $\hat{\mathbf{g}}_{\hat{\beta}}(\cdot)$ *from the iterative algorithm in Section 3.3 are the solutions of*

$$\begin{cases} \boldsymbol{\beta} = E[\hat{\boldsymbol{\beta}}^*|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \mathbf{g}] \\ \mathbf{g} = E[\hat{\mathbf{g}}^*|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \mathbf{g}], \end{cases}$$

where $\hat{\boldsymbol{\beta}}^*$ *is the complete data semiparametric efficient estimator proposed by [Ahmad, Leelahanon and Li (2005)] , and* $\hat{\mathbf{g}}^*$ *is the corresponding estimator of nonparametric component.*

3.5 Numerical Studies

3.5.1 Simulation Studies Using Gaussian and Logistic Models

In this section, we carry out several simulation studies to examine the performance of the proposed estimation methodology. Two model settings are considered: one is a Gaussian partially linear model and the other is a logistic partially linear model. Under each model, two different missing data patterns are studies. In one of them the resulting observed z_i's are independent and in the other the z_i's are correlated.

More specifically, let us consider the following Gaussian model:

$$y_i = w_i\boldsymbol{\beta} + g(v_i) + e_i, \quad i = 1, \cdots, N, \tag{3.5.1}$$

where e_i's are independent normal random error with mean 0 and standard deviation σ. We assume that the true unknown parameters $\boldsymbol{\beta}_0 = 4$, $\sigma_0 = 0.25$ and the true function $g_0(t) = \sin(2\pi t)$, and we simulate a set of $N = 500$ pairs of (w_i, v_i) from $w_i \sim \text{Uniform}[-0.375, 0.375]$ and $v_i \sim \text{Uniform}[0, 1]$. Based on Model (3.5.1), we simulate a set of $N = 500$ complete responses y_i. To obtain the set of $n = 125$ observed responses $z_j, j = 1, \ldots, 125$, that are independent, we group y_i's by the size of $k = 4$ with no overlap, and the observed responses z_j's are the averages of the y_i's in each of these $n = N/k = 500/4 = 125$ independent groups. To obtain the set of $n = 125$ observed responses $z_j, j = 1, \ldots, 125$, that are correlated, y_i's are grouped by the size of $k = 5$ with one overlapping component in each pair of neighboring groups, and the observed responses z_j's are the averages of the y_i's in each of these 125 overlapping groups. That is, $z_1 = y_1 + y_2 + \cdots + y_5$, $z_2 = y_5 + y_6 + y_7 + y_8 + y_9$, etc. Unless specified otherwise, in our data analysis, we only assume that we know only one set of $n = 125$ z_j's, the set of 500 individual (w_i, v_i)'s and the form of the semiparametric model (3.5.1).

The logistic partially linear model used in our study is

$$E(y_i) = \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \quad \eta_i = w_i\boldsymbol{\beta} + g(v_i), \quad i = 1, \cdots, N, \tag{3.5.2}$$

where y_i is independently bernoulli(μ_i) distributed. Again, we assume that the true parameter $\boldsymbol{\beta}_0 = 4$ and the true function $g_0(v_i) = \sin(2\pi v_i)$, and simulate $w_i \sim \text{Uniform}[-0.375, 0.375]$ and $v_i \sim \text{Uniform}[0, 1]$ but for $i = 1, \ldots, N = 5000$. We also simulate a set of complete responses y_i, $i = 1, \ldots, N$, according to (3.5.2). To obtain the set of $n = 1250$ independent observed responses $z_j, j = 1, \cdots, 1250$, that are independent, all the y_i's are grouped by size $k = 4$ and the observed responses z_j, $j = 1 \ldots, 1250$, are just the the maximums of the y_i's in each of the $n = N/k = 5000/4 = 1250$ independent groups. To obtain a set of $n = 1250$

correlated observed responses z_j, $j = 1 \ldots, 1250$, the 5000 y_i's are grouped of size $k = 5$ but with one overlapping component and z_j's are just the maximum values of the y_i's in each of these 1250 overlapping groups, i.e. $z_1 = \max(y_1, \cdots, y_5), z_2 = \max(y_5, \cdots, y_9)$, etc. Again, unless specified otherwise, in our model fitting and data analysis, we only assume that we know only one set of $n = 1250$ z_j's, the set of 5000 (w_i, v_i)'s and the form of the semiparametric model (3.5.2).

Each of the above two simulations are repeated 4000 times. The performance of the estimators using the proposed algorithm in Section 3.3 is summarized in Table 3.1 and Figure 3.1 in the rows labeled as "Proposed" or "Prop.". To get a better sense of their performance, we included in Table 3.1 and Figure 3.1 the estimation results using a conventional but inefficient back-fitting approach (labeled as "Backfitting") and the model fitting results pretending that in addition we know all about the complete data y_i's (labeled as "Com.lik."). The true model curve (labeled as "True") is also plotted in Figure 3.1. In the table, the estimated mean squared error (MSE) is defined by $MSE(\hat{\boldsymbol{\beta}}) = \sum_r (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^2/4000$, where sum $\sum_r$ is over the 4000 replications. For the Gaussian model, the estimates of σ are also provided.

Table 3.1: Summary of the parametric estimation results in the simulation studies

		Gaussian Regression			Logistic Regression	
		bias($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\boldsymbol{\beta}}$)	$\hat{\sigma}$	bias($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\boldsymbol{\beta}}$)
nonoverlap	proposed	.0032	.0239	.2409	.0152	.2175
	backfitting	-.0139	.0281	.2431	.0216	.2496
	com.lik.	-.0012	.0054	.2474	-.0029	.0408
overlap	proposed	.0073	.0248	.2409	.0053	.2149
	backfitting	-.0128	.0264	.2416	.0639	.2715
	com.lik.	-.0043	.0049	.2479	.0050	.0350

From Table 3.1, we can see that our proposed method outperforms the conventional backfitting algorithm in terms of bias and MSE. Our proposed estimator has larger MSE than the estimator using complete dataset. This is expected since

the number of observed z_i's is only one-fourth (1/4) of the number of the complete data y_i's and the large fraction of missing information ($3/4 = 75\%$ missing) results in the loss of efficiency. We also can see from Figure 3.1 that the non-parametric component $g(\cdot)$ can be estimated fairly accurately in all four cases. In terms of variance, our proposed method performs similar to the backfitting algorithm. This is also not surprising since our semiparametric efficiency result only applies to the parametric component, but not the non-parametric component. In conclusion, the above simulation studies have demonstrated that our proposed estimation algorithm works well.

3.5.2 Nebraska IPP Data Analysis

We have also applied our method to two group testing settings for data collected in the Nebraska Infertility Prevention Project (IPP). The state of Nebraska takes part in the nationwide IPPs through its Sexually Transmitted Diseases and Infertility Control Program. Urine or swab (cervical or male urethra) specimens are collected on each individual and are transported to the Nebraska Public Health Laboratory in Omaha for chlamydia and gonorrhea testing; see, e.g., [Chen, Tebbs and Bilder (2009)]. More than 30,000 individual tests are performed annually by the Nebraska Public Health Laboratory. The sample prevalence for chlamydia and gonorrhea is about 8.1% and 2.2%, respectively. This result makes group testing strategies potentially attractive as a means for surveillance while saving money. For our study, the goal is to explore and study the possibility of implementing potential group testing strategies in the Nebraska IPPs study. We obtained a set of randomly selected 9000 individual testing results from the Nebraska Department of Health and Human Services in 2009, along with a set of covariates of risk factors. We first explore a simple group testing strategy, where individual samples are grouped in groups of size k. In each group, the testing result of the group is set to be positive ("1"), if there is at least one

subject infected in the group. Otherwise, the group response is negative ("0"). In our study, we have considered three group sizes $k = 2, 5, 10$, respectively, and we assume that we only know the group testing responses but not the individual responses. For each infection and with each group size, we fit the group testing data to the latent logistic regression model

$$\text{logit}\,\{P(Y_i) = 1\} = \boldsymbol{\beta}_1 \text{Gender}_i + \boldsymbol{\beta}_2 \text{Symptoms}_i + \boldsymbol{\beta}_3 \text{Urethritis}_i + g(\text{Age}_i) \quad (3.5.3)$$

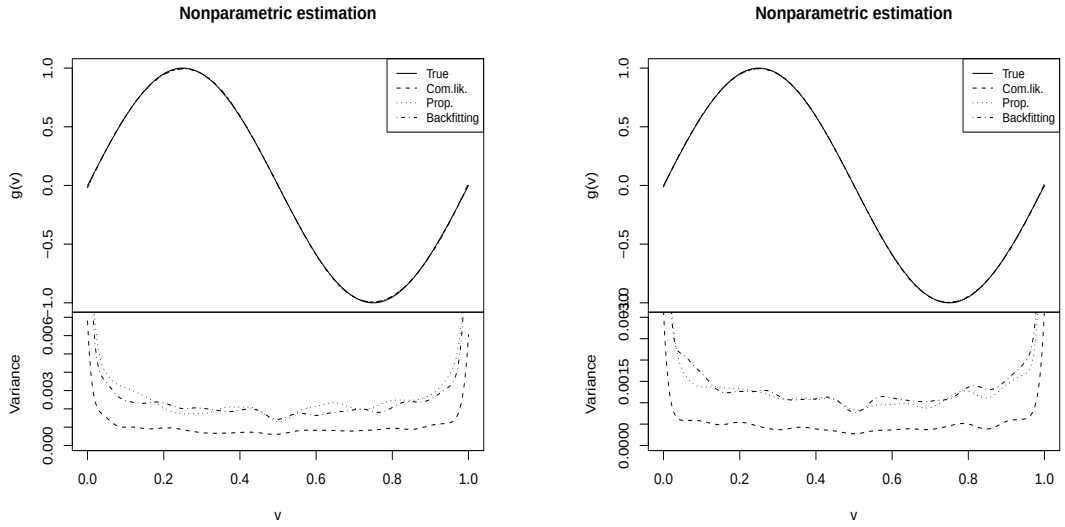
where y_i is the ith subject's infection status (infected=1, not infected=0) (which is assumed to unobserved). Here, following [Chen, Tebbs and Bilder (2009)], we only use the Age, Gender, Urethritis Status and Infection Symptoms Status as the covariates in the model. But different from [Chen, Tebbs and Bilder (2009)], the variable 'Age' is modeled non-parametrically, where the 'Age' variable is scaled to lie between 0 and 1.

In practice, the testing methods are not 100% accurate.
[Gastwirth and Hammick (1989)] proposed a two layer group testing method, that consists of one (first) round of screening test and another (second) round of confirmatory test. Generally speaking, the screening test is much cheaper but not quite accurate, while the confirmatory test is almost perfect but with a much higher cost. In our study, we also explore the possibility of using this GH group testing strategy. To implement the strategy in our study, first, all 9000 subjects are grouped without overlap in group size $k = 2, 5, 10$, respectively, and a screening test is applied to each group. Then, for those groups tested positive in the screening tests, confirmatory tests are applied. In our study, both of the screening tests and confirmatory tests are not 100 percent accurate where the sensitivity and specificity of the screening and confirmatory testing method are $(\gamma_1^{(s)} = 95\%, \gamma_2^{(s)} = 98\%)$ and $(\gamma_1^{(c)} = 99.99\%, \gamma_2^{(c)} = 99.99\%)$, respectively. In this study, the observed responses are the outcomes from both the screening tests and the confirmatory tests. We fit them to the same Model (3.5.3), again treating the individual outcomes y_i's as missing. Note that, the result of a confirmatory test

is correlated to the result of the screening test for the same group, thus part of the z_i's are correlated.

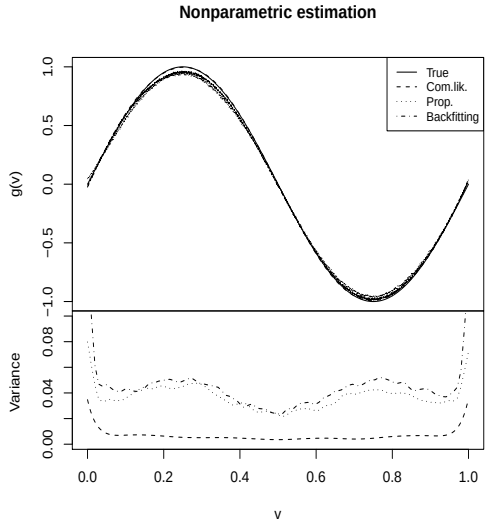
The first half of Table 3.2 displays the parameter estimation results of this first study. And the second half of Table 3.2 shows the results from the second study. To calculate the total costs needed for each testing strategy with difference group sizes, we assume that both the cost of a single test used in the first study and a single confirmatory test used in the second study are D dollars, and the cost of a single screening test used in the second strategy is just $D/10$ dollars. We can see that the estimates of the regression coefficients are pretty close across all the values of k in both testing strategies. As the group size k increases, the standard error increases due to the loss of information from missing data, but both the number of tests needed and the total costs decrease drastically. Clearly, the group testing approaches can potentially be very useful in this practice to cut the cost with only a slight loss of efficiency.

Finally, Figure 3.2 shows the estimated $\hat{g}(\text{Age})$ for Chlamydia and Gonorrhea respectively, for the second study with group size $k = 2, 5$. We can see that both of the plots indicate non-linearity of $\hat{g}(\text{Age})$ which validates our idea that modeling via linear model is insufficient.

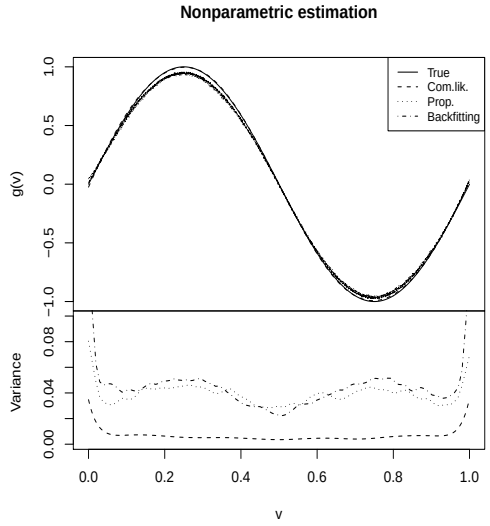


(a) Gaussian without overlap

(b) Gaussian with overlap



(c) Logistic without overlap



(d) Logistic with overlap

Figure 3.1: Simulation results for $\hat{g}(\cdot)$.

'True', the true function used to simulate the dataset; 'Com.lik', estimation using complete dataset, i.e. no grouping; 'Prop' and 'Backfitting', our proposed algorithm and backfitting algorithm using the incomplete dataset, i.e. grouping with size $= 4$, respectively. The panels under each nonparametric estimation are the pointwise variance estimation: $\widehat{\mathrm{Var}}(\hat{g}(v))$.

Table 3.2: Summary of the parametric estimation and testing results for the Nebraska IPP data.

Part I. Simple random group testing method										
k	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	Num	Cost	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	Num	Cost
	Chlamydia					Gonorrhea				
1	.34	.59	1.19	9000	9000	.05	1.38	2.67	9000	9000
	(.07)	(.06)	(.08)			(.12)	(.09)	(.10)		
2	.33	.58	1.17	4500	4500	.05	1.41	2.68	4500	4500
	(.08)	(.08)	(.13)			.05(.15)	(.12)	(.15)		
5	.32	.64	1.17	1800	1800	.06	1.40	2.62	1800	1500
	(.11)	(.11)	(.21)			(.25)	(.23)	(.26)		
10	.30	.64	1.16	900	900	.04	1.42	2.63	900	900
	(.15)	(.14)	(.32)			(.34)	(.28)	(.31)		
Part II. Gastwirth-Hammick two layer group testing method										
k	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	Num-S/C	Cost	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	Num-S/C	Cost
	Chlamydia					Gonorrhea				
1	.31	.60	1.15	9000/841	1741	.03	1.37	2.64	9000/342	1142
	(.07)	(.06)	(.09)			(.13)	(.09)	(.12)		
2	.35	.58	1.20	4500/733	1183	.06	1.39	2.65	4500/259	709
	(.08)	(.08)	(.14)			(.17)	(.14)	(.17)		
5	.37	.61	1.12	1800/603	783	.03	1.39	2.69	1800/193	373
	(.11)	(.11)	(.25)			(.26)	(.25)	(.26)		
10	.36	.61	1.15	900/490	580	.04	1.46	2.72	900/171	261
	(.16)	(.14)	(.33)			(.35)	(.28)	(.33)		

Note: the numbers in the brackets are the corresponding standard errors. The case $k = 1$ indicates we observe complete data. For $k >1$, estimates and standard errors are averaged over 100 sets of pools. Since for $k > 1$, different grouping yields different testing results, we randomly group patients 100 times and take the average of the estimates. The column 'Num' represents the number of tests required for each scenario. For two layers grouping testing case, 'Num-S' and 'Num-C' represents the number of screening tests and confirmatory tests required, respectively.

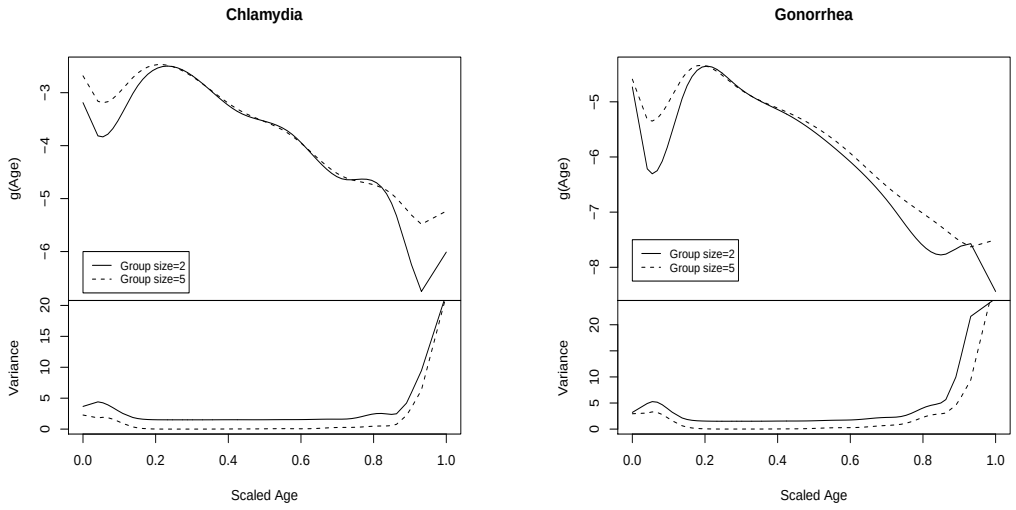


Figure 3.2: Estimated $\hat{g}(Age)$ for Chlamydia and Gonorrhea with group size$= 2, 5$ respectively.

The upper two panels are non-parametric estimation of the unknown function $\hat{g}(\text{Age})$. The lower two panels are the pointwise variance estimates: $\hat{\text{Var}}(\hat{g}(\text{Age}))$.

Chapter 4

Quadruple paired robust Estimation Of Treatment Effect In Observational Studies With Missing Responses

4.1 Background and Motivation

In observational or randomized treatment comparative studies, missing data in the responses of interest poses a challenge for the estimation of the treatment effect. When the missing mechanism of the response depends on treatment, baseline covariates, and post-baseline variables, missing at random (MAR) ([Roderick et al. (1987)]) is often assumed. Such missing mechanism is ignorable when likelihood-based methods are used, but the resulting estimators are often sensitive to the fully parametric specification of the model. An estimator derived from the efficient influence function of the semiparametric observed data model was proposed by [Davidian et al. (2005)] in the context of pretest-posttest studies, which enjoys the desired property of double-robustness, i.e. the correct model specification of either the missing data mechanism, or the conditional distribution of response given covariates ensures the consistency of the resulting estimator.

One assumption made by [Davidian et al. (2005)] in the construction of the semiparametric model is the independence between treatment indicator and baseline covariates (including pretest response). This is true only for randomized experiments. In this chapter, we specifically focus on observational studies where the treatment exposures are associated with demographic and physiologic characteristics of the study subjects. In addition, the missing responses considered in

[Davidian et al. (2005)] leads to further complications in the observational setting: the post-baseline covariates, which will likely have different "outcomes" given different treatment exposures, are in turn confounded by the imbalance of baseline covariates.

Our motivation comes from the burden of illness study conducted by Duke University Medical Center for patients with coronary artery disease. One primary object of this observational study is to estimate the difference in the 5-year medical expenses between two treatment groups in the presence of missing data(see [Mark et al (1994)]). Specifically, a total of 2284 out of 3214 patients have complete cost records. Some literatures including but not limited to ([Anstrom and Tsiatis(2001), Eisenstein et al. (2001)]) have investigated this problem without considering post-baseline variables. Nevertheless, it is reasonable to assume that the missing in the response is related to both baseline and post-baseline measurements. For instance, some patients might stop the treatment in the middle of the study because the intermediate diagnosis shows benign results. Some other patients might opt-out because they are no longer able to afford the current treatment. These factors are not observable at the beginning of the study and thus defined as post-baseline covariates. The challenge is how to utilize these information in the statistical inference.

The rest of the chapter is organized as follows. In Section 4.2, we briefly review the potential outcome framework and three existing estimators. In Section 4.3, we discuss the assumptions for the missing data and treatment exposure mechanisms. Then, we derive the semiparametric efficiency bound of RAL estimators under these assumptions. In Section 4.4, we propose a semiparametric efficient estimator inspired by the formulation of the efficiency bound, and show that this estimator is robust against four types of model mis-specifications. Numerical studies including simulations and a real data analysis on burden of illness dataset are carried out in Section 4.5 to demonstrate the performance of the proposed estimator, in terms

of efficiency and robustness.

4.2 Potential Outcome Framework and Existing Estimators

Suppose an observational study of n subjects is carried out to compare the treatment effect of response Y under treatment ($Z = 1$) relative to control ($Z = 0$). The treatment effect can be expressed as the difference between two potential outcomes:

$$\beta = E[Y^1 - Y^0], \tag{4.2.1}$$

where Y^1 and Y^0 are two hypothetical and unobservable responses for a typical subject exposed to treatment ($Z = 1$) or control ($Z = 0$). The response observed Y can be written as $Y = ZY^1 + (1 - Z)Y^0$.

Let M be the missing indicator. We observe Y only if $M = 0$. In addition, two categories of covariates are observed for all subjects in the study:

- *Baseline covariates* $\boldsymbol{X}$: Variables recorded before treatment exposure, or those variables not affected by treatments, such as the demographic information, medical history, prior medications and measurements of clinical endpoints at baseline.

- *Post-baseline variables* $\boldsymbol{W}$: Post-baseline measurements, such as the adverse experience to a drug, the vital signs measured during treatment, the progressive co-morbidity conditions, concomitant medications, or secondary outcomes of interest ([Frangakis and Rubin (2002)]). These variables are often associated with baseline covariates, and may also be causally influenced by treatment exposure. Similar to the response Y, $\boldsymbol{W}$ can be expressed as function of two potential outcomes: $\boldsymbol{W} = Z\boldsymbol{W}^1 + (1 - Z)\boldsymbol{W}^0$.

Given the baseline and post-baseline variables, two propensity functions can be defined for the treatment exposure and missing data mechanism:

$$p(\boldsymbol{X}) = \mathrm{P}(Z = 1 \,|\boldsymbol{X}\,) \tag{4.2.2}$$

$$q(\boldsymbol{W}, \boldsymbol{X}, Z) = \mathrm{P}(M = 1 \mid \boldsymbol{W}, \boldsymbol{X}, Z). \tag{4.2.3}$$

Based on these propensity functions, two inverse-weighting estimators of the Horvitz-Thompson type are:

$$\begin{aligned} \hat{\beta}_{HT1} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(1-M_i)Z_iY_i}{\hat{p}(\boldsymbol{X}_i)\left\{1-\hat{q}(\boldsymbol{W}_i,\boldsymbol{X}_i,Z_i)\right\}} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n} \frac{(1-M_i)(1-Z_i)Y_i}{\left\{1-\hat{p}(\boldsymbol{X}_i)\right\}\left\{1-\hat{q}(\boldsymbol{W}_i,\boldsymbol{X}_i,Z_i)\right\}}, \end{aligned} \tag{4.2.4}$$

and

$$\begin{aligned} \hat{\beta}_{HT2} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i\hat{\gamma}_1(\boldsymbol{W}_i,\boldsymbol{X}_i)}{\hat{p}(\boldsymbol{X}_i)} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)\hat{\gamma}_0(\boldsymbol{W}_i,\boldsymbol{X}_i)}{1-\hat{p}(\boldsymbol{X}_i)}, \end{aligned} \tag{4.2.5}$$

where $\hat{p}(\boldsymbol{X})$ and $\hat{q}(\boldsymbol{W}, \boldsymbol{X}, Z)$ are estimators of the propensity functions, and $\hat{\gamma}_j$'s are estimators for

$$\gamma_j(\boldsymbol{x}, \boldsymbol{w}) = E(Y^j \,\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W}^j = \boldsymbol{w}\,) \quad j = 0, 1. \tag{4.2.6}$$

In addition, an outcome regression estimator based on baseline covariates only is:

$$\hat{\beta}_{ORB} = \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{\beta}_1(\boldsymbol{X}_i) - \hat{\beta}_0(\boldsymbol{X}_i)\right\}, \tag{4.2.7}$$

where $\hat{\beta}_j$'s are estimators for

$$\beta_j(\boldsymbol{x}) = E(Y^j \,|\boldsymbol{X} = \boldsymbol{x}\,) \quad j = 0, 1. \tag{4.2.8}$$

We will discuss in the next section about the conditions under which the aforementioned estimators (4.2.4), (4.2.5) and (4.2.7) are consistent.

4.3 Semiparametric Efficiency

4.3.1 Assumptions

Suppose $\boldsymbol{X}$ contains all the confounding factors that may impact the exposures of treatment to individual patients, the following assumption can be reasonably made:

Assumption 1 (Unconfounded Treatment Assignment).

$$(\boldsymbol{W}^1, \boldsymbol{W}^0, Y^1, Y^0) \perp Z \,|\boldsymbol{X}$$

Under Assumption 1, treatment exposures of subjects sharing similar baseline characteristics would be independent of the potential outcomes. Therefore, we have for $j = 0, 1$:

$$\begin{aligned} E(Y|Z = j, \boldsymbol{X}) = E(Y^j|Z = j, \boldsymbol{X}) = E(Y^j|\boldsymbol{X}), \\ E(\boldsymbol{W}|Z = j, \boldsymbol{X}) = E(\boldsymbol{W}^j|Z = j, \boldsymbol{X}) = E(\boldsymbol{W}^j|\boldsymbol{X}). \end{aligned} \tag{4.3.1}$$

If the missing mechanism of Y is associated with observed data for a subject, and not dependent on the potentially unobservable Y, the MAR assumption can be postulated as follows:

Assumption 2 (Missing at Random).

$$M \perp Y \mid \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z},$$

Under Assumption 1 and 2, we have:

$$\begin{aligned} E(Y|\boldsymbol{X}, \boldsymbol{W}, Z = j, M = 0) &= E(Y^j|\boldsymbol{X}, \boldsymbol{W}^j, Z = j) \\ &= E(Y^j|\boldsymbol{X}, \boldsymbol{W}^j) \quad j = 0, 1. \end{aligned} \tag{4.3.2}$$

Therefore it can be shown that estimator (4.2.4) is consistent if $p(\boldsymbol{x})$ and $q(\boldsymbol{w}, \boldsymbol{x}, z)$ are consistently estimated, and estimator (4.2.5) is consistent if $p(\boldsymbol{x})$

and $\gamma_j(\boldsymbol{x}, \boldsymbol{w}), j = 0, 1$ are consistently estimated. For estimator (4.2.7), it is consistent if $\beta_j(\boldsymbol{x}), j = 0, 1$ are consistently estimated.

It's natural to ask whether a double-robust estimator of β exists that ensures the consistency when either the propensity functions or the conditional distributions of potential outcomes are correctly modeled. In causal inference, there is an evident link between the construction of double-robust estimators ([Lunceford and Davidian (2004)]), and the semiparametric efficiency bound of regular asymptotic linear estimators ([Hahn (1998)]). Inspired by this observation, we first derive the semiparametric efficiency bound for RAL estimators of β by constructing the efficient influence function, then propose an estimator that achieve this efficiency bound.

4.3.2 Semiparametric Efficiency Bound

Let $f(\boldsymbol{w}^1, \boldsymbol{w}^0, y^1, y^0 \mid \boldsymbol{x})$ be the joint conditional distribution of $\boldsymbol{W}^1$, $\boldsymbol{W}^0$, Y^1 and Y^0 given $\boldsymbol{X} = \boldsymbol{x}$, and $f(\boldsymbol{x})$ be the density function of baseline covariates $\boldsymbol{X}$. The density or probability mass function of the observed data $\boldsymbol{L} = (YI(M = 0), M, Z, \boldsymbol{X}, \boldsymbol{W})$ can be expressed as follows:

$$\begin{aligned} f(\boldsymbol{L}) \;=\; & [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z)\} p(\boldsymbol{X}) f_1(\boldsymbol{W}, Y \,|\boldsymbol{X})]^{Z(1-M)} \\ & \times [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z)\}\{1 - p(\boldsymbol{X})\} f_0(\boldsymbol{W}, Y \,|\boldsymbol{X})]^{(1-Z)(1-M)} \\ & \times \{q(\boldsymbol{W}, \boldsymbol{X}, Z) p(\boldsymbol{X}) f_1^*(\boldsymbol{W} \,|\boldsymbol{X})\}^{ZM} \\ & \times [q(\boldsymbol{W}, \boldsymbol{X}, Z)\{1 - p(\boldsymbol{X})\} f_0^*(\boldsymbol{W} \,|\boldsymbol{X})]^{(1-Z)M} \\ & \times f(\boldsymbol{X}) \end{aligned} \tag{4.3.3}$$

where

$$\begin{aligned} f_1(\boldsymbol{w}, y \mid \boldsymbol{x}) &= \iint f(\boldsymbol{w}, \boldsymbol{w}^0, y, y^0 \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^0\, \mathrm{d}y^0 \\ f_0(\boldsymbol{w}, y \mid \boldsymbol{x}) &= \iint f(\boldsymbol{w}^1, \boldsymbol{w}, y^1, y \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^1\, \mathrm{d}y^1 \\ f_1^*(\boldsymbol{w} \mid \boldsymbol{x}) &= \iiint f(\boldsymbol{w}, \boldsymbol{w}^0, y^1, y^0 \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^0\, \mathrm{d}y^1\, \mathrm{d}y^0 \\ f_0^*(\boldsymbol{w} \mid \boldsymbol{x}) &= \iiint f(\boldsymbol{w}^1, \boldsymbol{w}, y^1, y^0 \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{w}^1\, \mathrm{d}y^1\, \mathrm{d}y^0. \end{aligned}$$

A semiparametric model $f(\boldsymbol{L}; \eta)$ will assume some components in the distribution $f(\boldsymbol{L})$ to be unrestricted except for the obvious constraint of being a proper probability density or probability mass function:

$$\begin{aligned} f(\boldsymbol{L}; \eta) &= [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z; \eta_1)\} p(\boldsymbol{X}; \eta_2) f_1(\boldsymbol{W}, Y \mid \boldsymbol{X}; \eta_3)]^{Z(1-M)} \\ &\quad \times [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z; \eta_1)\}\{1 - p(\boldsymbol{X}; \eta_2)\} f_0(\boldsymbol{W}, Y \mid \boldsymbol{X}; \eta_3)]^{(1-Z)(1-M)} \\ &\quad \times \{q(\boldsymbol{W}, \boldsymbol{X}, Z; \eta_1) p(\boldsymbol{X}; \eta_2) f_1^*(\boldsymbol{W} \mid \boldsymbol{X}; \eta_3)\}^{ZM} \\ &\quad \times [q(\boldsymbol{W}, \boldsymbol{X}, Z; \eta_1)\{1 - p(\boldsymbol{X}; \eta_2)\} f_0^*(\boldsymbol{W} \mid \boldsymbol{X}; \eta_3)]^{(1-Z)M} \\ &\quad \times f(\boldsymbol{X}; \eta_4) \end{aligned} \tag{4.3.4}$$

where $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$ is the set of parameters for the semiparametric model, with some components being infinite dimensional. The flexibility of semiparametric models allows inference for the treatment effect without making additional distributional assumptions about the covariates. At the true parameter values $\eta^0 = (\eta_1^0, \eta_2^0, \eta_3^0, \eta_4^0)$, the model is equal to the density of $\boldsymbol{L}$:

$$f(\boldsymbol{L}; \eta^0) = f(\boldsymbol{L}).$$

It is well known that the semiparametric efficiency bound is the supremum of Cramer-Rao bounds for all parametric submodels ([Begun et al. (1983)], [Bickel et al. (1993)] and [Newey (1990)]). In the following theorem, we calculate the semiparametric efficiency bound for RAL estimators of the treatment effect β in model (4.3.4) under suitable assumptions. Here, the regularity restriction

excludes super-efficient estimators. Since the distribution of baseline covariates is generally not the focus of the inference, we'll assume that $\dim(\eta_4) = \infty$.

Theorem 4.3.1. *Under Assumption 1 and 2, the semiparametric efficiency bound for RAL estimators of β is:*

$$V = E\left[\delta^2\right],$$

where, expectation is taken under the true model and

$$\begin{aligned}\delta = &\frac{Z(1-M)\{Y-\gamma_1(\boldsymbol{W},\boldsymbol{X})\}}{p(\boldsymbol{X})\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}} - \frac{(1-Z)(1-M)\{Y-\gamma_0(\boldsymbol{W},\boldsymbol{X})\}}{\{1-p(\boldsymbol{X})\}\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}} \\ &+\frac{Z\{\gamma_1(\boldsymbol{W},\boldsymbol{X})-\beta_1(\boldsymbol{X})\}}{p(\boldsymbol{X})} - \frac{(1-Z)\{\gamma_0(\boldsymbol{W},\boldsymbol{X})-\beta_0(\boldsymbol{X})\}}{1-p(\boldsymbol{X})} \\ &+\beta_1(\boldsymbol{X})-\beta_0(\boldsymbol{X})-\beta. \end{aligned} \tag{4.3.5}$$

The efficiency bound remains the same regardless of the model specifications for $p(\boldsymbol{x})$, $q(\boldsymbol{w},\boldsymbol{x},z)$, $\beta_j(\boldsymbol{x})$ and $\gamma_j(\boldsymbol{w},\boldsymbol{x})$.

The proof of this theorem involves construction of the tangent space for model (4.3.4) and the projection of influence function onto the tangent space. The detail is given in the Appendix.

Theorem 4.3.1 implies the ancillarity of the propensity functions in efficient treatment effect estimation. In other words, the knowledge of propensity scores does not decrease the variance bound. A parallel result in the absence of missing data was given by [Hahn (1998)] for the propensity of treatment assignment's role in efficiency consideration.

4.4 Efficient and Quadruple paired robust Estimator

4.4.1 Semiparametric Efficient Estimation

Inspired by the formulation of efficiency bound, we propose an estimator $\hat{\beta}_{QR}$ as follows:

$$\begin{aligned}\hat{\beta}_{QR} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ & \frac{Z_i(1-M_i)\{Y_i - \hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\hat{p}_n(\boldsymbol{X}_i)\{1-\hat{q}_n(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}} \\ & - \frac{(1-Z_i)(1-M_i)\{Y_i - \hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\{1-\hat{p}_n(\boldsymbol{X}_i)\}\{1-\hat{q}_n(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}} \\ & + \frac{Z_i}{\hat{p}_n(\boldsymbol{X}_i)}\{\hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_1^{(n)}(\boldsymbol{X}_i)\} + \hat{\beta}_1^{(n)}(\boldsymbol{X}_i) \\ & - \frac{1-Z_i}{1-\hat{p}_n(\boldsymbol{X}_i)}\{\hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\} - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\Bigg\}. \end{aligned} \tag{4.4.1}$$

The following Theorem states that the estimator $\hat{\beta}_{QR}$ is consistent and semiparametrically efficient if the unknown components can be estimated with uniform consistency:

Theorem 4.4.1. *Suppose the models for $p(\boldsymbol{x})$, $q(\boldsymbol{w}, \boldsymbol{x}, z)$, $\beta_j(\boldsymbol{x})$ and $\gamma_j(\boldsymbol{w}, \boldsymbol{x})$ are correctly specified. Under suitable regularity conditions, estimator $\hat{\beta}_{QR}$ for the treatment effect is consistent and semiparametrically efficient:*

$$\sqrt{n}(\hat{\beta}_{QR} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

where V is the semiparametric efficiency bound from Theorem 4.3.1.

The proof of Theorem 4.4.1 involves replacing the individual estimates $\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}$ in (4.4.1) with the true functions and showing the difference is $o_p(n^{-1/2})$. The calculation is straightforward, thus omitted here.

The regularity conditions in Theorem 4.4.1 are the ones such that $\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}$ converge to the truth when the models are correct. The efficiency of $\hat{\beta}_{QR}$ can be attributed to the efficient use of baseline and post-baseline information. Although the observed $\boldsymbol{W}$ were confounded with treatment Z, the conditional distribution

of potential outcome $\boldsymbol{W}^j$ given $\boldsymbol{X}$ can be modeled based on observed $\boldsymbol{W}$ in treatment group j according to (4.3.1). Similarly, although Y are not observed for all subjects, the conditional distribution of Y^j given $\boldsymbol{W^j}$ and $\boldsymbol{X}$ can be modeled based on observed Y in treatment group j according to (4.3.2). These two conditional distributions together decide the joint conditional distribution of $(Y^j, \boldsymbol{W}^j)$ given $\boldsymbol{X}$.

4.4.2 Quadruple paired robustness

In practice, it is often difficult to ensure that all of the unknown components are correctly modeled and estimated. In particular, since $E(Y|\boldsymbol{X}, Z = j, M = 0)$ does not always equal to $E(Y^j|X)$, the regression of observed Y on $\boldsymbol{X}$ in treatment group j does not typically give a consistent estimator for β_j. External evidence, such as a piloting study with fully observed responses, can sometimes be used to construct estimators for β_j. In other cases, $E(Y^j|X)$ can be derived from $E(Y^j|\boldsymbol{W}^j, \boldsymbol{X})$ and $E(\boldsymbol{W}^j|\boldsymbol{X})$ using the chain rule for conditional expectations, which causes the model of β_j to be dependent on the correct specification of γ_j.

Fortunately, due to the special structure of the estimator, the sufficient conditions for the consistency of $\hat{\beta}_{QR}$ is less stringent:

Theorem 4.4.2. *Estimator $\hat{\beta}_{QR}$ is consistent under suitable regularity conditions when one of the following conditions is true:*

- *models for $p(\boldsymbol{x})$ and $q(\boldsymbol{w}, \boldsymbol{x}, z)$ are correctly specified;*
- *models for $p(\boldsymbol{x})$ and $\gamma_j(\boldsymbol{w}, \boldsymbol{x}), j = 0, 1$ are correctly specified;*
- *models for $q(\boldsymbol{w}, \boldsymbol{x}, z)$ and $\beta_j(\boldsymbol{x}), j = 0, 1$ are correctly specified;*
- *models for $\gamma_j(\boldsymbol{w}, \boldsymbol{x}), j = 0, 1$ and $\beta_j(\boldsymbol{x}), j = 0, 1$ are correctly specified;*

The proof of Theorem 4.4.2 is given in the Appendix.

The regularity conditions in Theorem 4.4.2 are the ones such that $\hat{p}, \hat{q}, \hat{\beta}, \hat{\gamma}$ converge to some function $p^*, q^*, \beta^*, \gamma^*$ regardless of model specification and $p^*, q^*, \beta^*, \gamma^*$ equal to the truth if the models are correctly specified. The fact that the consistency for any of these four particular combinations of model components ensures the consistency of $\hat{\beta}_{QR}$ is not surprising. Consider two simpler double-robust estimators, one for the inference with missing data but randomized treatment exposure, and the other for the causal inference with non-randomized treatment exposure, but no missing data - for the missing data problem, a double-robust estimator would have either correctly specified model for the missing data mechanism, or correct model for the conditional model of response given observed data; for the observational data problem, a double-robust estimator would have either correctly specified model for the treatment assignment propensity, or a correct conditional model of response given baseline covariates. When treatment exposures are not randomly assigned, and the responses are missing at random, we are presented with these two related inference problems simultaneously. Unsurprisingly, a robust estimator in this case would require one of the two components from the missing data model, and one of the two components from the causal inference model to be correctly specified. The four combinations of these model components therefore yield the four configurations given in Theorem 4.4.2.

As we shall demonstrate through simulations in the next section, estimator $\hat{\beta}_{QR}$ is more robust against model mis-specifications compared to $\hat{\beta}_{HT1}$, $\hat{\beta}_{HT2}$ and $\hat{\beta}_{ORB}$. In addition, the efficiency of $\hat{\beta}_{QR}$ is often superior to others, due to its more efficient use of baseline and post-baseline covariates.

4.5 Numerical Studies

4.5.1 Simulations

Simulations are carried out to demonstrate the performance of $\hat{\beta}_{QR}$ in terms of efficiency and robustness. For each of the simulations, we generate 2000 Monte-Carlo samples with 1000 observations in each sample. Different estimators are compared in terms of the sample distributions.

In a typical observational study, baseline covariates are usually multivariate. In our simulations we consider two independent covariates $\boldsymbol{X} = (X_1, X_2)$, each following standard normal distribution.

The treatment indicator Z is a binary variable generated from the underlying true propensity of treatment assignment:

$$\text{logit}\left\{P(Z=1|\boldsymbol{X})\right\} = \text{logit}\left\{p(\boldsymbol{X})\right\} = -0.5 + 0.5I_1 - 0.5I_2 + 0.5I_1I_2$$

where

$$I_i = I(X_i > 0) \quad \text{for} \quad i = 1, 2$$

The underlying potential responses $Y^j, W^j, j = 0, 1$ are generated as follows:

$$\begin{aligned}
W^1 &= 1 + 3X_1 - 2X_2 + \zeta_1, \\
W^0 &= -1 + 3X_1 - X_2 + \zeta_0, \\
Y^1 &= 2 - X_1 + 2.5X_2 - 2X_1X_2 - 1.5W^1 + \eta_1, \\
Y^0 &= 1 - 0.5X_1 + 1.5X_2 - X_1X_2 - 1.5W^0 + \eta_0.
\end{aligned}$$

where ζ_j and $\eta_j, j = 0, 1$ are measurement errors generated by

$$\zeta_1 = 2\varepsilon_0 + \varepsilon_1, \quad \zeta_0 = 2\varepsilon_0 + \varepsilon_2, \quad \eta_1 = \varepsilon_0 + \varepsilon_3, \quad \eta_0 = \varepsilon_0 + \varepsilon_4$$

where $\varepsilon_i, i = 0, \cdots, 4$ are mutually independent standard normal random variables. In reality, the auxiliary outcomes $\boldsymbol{W}$ could be multivariate, but for computational and notational convenience we are going to use univariate W in the simulations.

Table 4.1: Models for estimation

	Model
$p(\boldsymbol{x})$	
C	$\text{logit}(p(\boldsymbol{X})) = a_0 + a_1 I_1 + a_2 I_2 + a_3 I_1 I_2$
I	$\text{logit}(p(\boldsymbol{X})) = a_0 + a_1 I_1 + a_2 I_2$
$q(\boldsymbol{x}, \boldsymbol{w}, z)$	
C	$\text{logit}(q(\boldsymbol{X}, \boldsymbol{W}, Z)) = b_0 + b_1 I_1 + b_2 I_2 + b_3 I_w + b_4 I_1 I_2 + b_5 Z + b_6 Z I_2$
I	$\text{logit}(q(\boldsymbol{X}, \boldsymbol{W}, Z)) = b_0 + b_1 I_1 + b_2 I_2 + b_3 I_w + b_4 Z$
$\gamma_j(\boldsymbol{x}, \boldsymbol{w}), j = 0, 1$	
C	$\gamma_j(\boldsymbol{X}, \boldsymbol{W}) = c_0 + c_1 X_1 + c_2 X_2 + c_3 X_1 X_2 + c_4 W_j$
I	$\gamma_j(\boldsymbol{X}, \boldsymbol{W}) = c_0 + c_1 X_1 + c_2 X_2 + c_3 W_j$
$\beta_j(\boldsymbol{x}), j = 0, 1$	
C	$\beta_j(\boldsymbol{X}) = d_0 + d_1 X_1 + d_2 X_2 + d_3 X_1 X_2$
I	$\beta_j(\boldsymbol{X}) = d_0 + d_1 X_1 + d_2 X_2$

The response W and Y are derived from Y^1, Y^0, W^1, W^0, and Z:

$$\begin{aligned} W &= W^1 Z + W^0 (1 - Z) \\ Y &= Y^1 Z + Y^0 (1 - Z). \end{aligned}$$

Missing indicator $M = 0, 1$ is generated according to the underlying true propensity of missing $q(w, \boldsymbol{x}, z)$:

$$\text{logit}\left\{q(W, \boldsymbol{X}, Z)\right\} = 1 + 0.5 I_1 - 0.7 I_2 - 0.6 I_w + 0.5 I_1 I_2 + 0.5 Z + 0.6 I_2 Z,$$

where $I_w = I(W > 0)$.

In order to demonstrate the performance of the proposed estimator under different settings and assess their robustness against model mis-specifications, we will use a pair of working models for each of the four categories of functional components in the estimation process - one correctly specified model (C), and one incorrectly specified model (I). The models to be used in the simulations are listed in Table 4.1.

In Simulation 1, we used the correct model specifications for all parametric

Table 4.2: Simulation scenarios

Simulation	$p(\boldsymbol{x})$	$q(\boldsymbol{x}, \boldsymbol{w}, z)$	$\gamma_j(\boldsymbol{x}, \boldsymbol{w})$	$\beta_j(\boldsymbol{x})$
1	C	C	C	C
2	I	I	C	C
3	I	C	I	C
4	C	I	C	I
5	C	C	I	I
6	I	I	I	I

components of the semiparametric model to demonstrate the consistency of $\hat{\beta}_{QR}$. In addition, we carried out 5 additional groups of simulations (simulation 2-6), each accounts for a different scenario of model mis-specifications (see Table 4.2 for details).

In each simulation, we compared the proposed estimator $\hat{\beta}_{QR}$ with the Horwitz-Thompson estimators $\hat{\beta}_{HT1}$, $\hat{\beta}_{HT2}$ and outcome regression estimator $\hat{\beta}_{ORB}$.

The boxplots for the Monte-Carlo distributions of the four estimators are given in Figure 4.1. The summary statistics of the distributions (mean squared error and bias) are listed in Table 4.3.

Simulation 1 shows that all four estimators are consistent when the parametric components of the observed data model are correctly specified. Our proposed estimator $\hat{\beta}_{QR}$ has the smallest MSE, indicating its efficiency. In Simulation 2-5, when two out of the four parametric components were misspecified, the advantage of $\hat{\beta}_{QR}$ is clearly demonstrated by its significantly smaller MSE and bias compared to other estimators. In contrast, $\hat{\beta}_{HT1}$, $\hat{\beta}_{HT2}$ and $\hat{\beta}_{ORB}$ are all biased when one or more of the contributing components of the estimators are misspecified. Simulation 6 demonstrates the case when all parametric components are incorrectly modeled. Moderate increase in both bias and MSE is seen for $\hat{\beta}_{QR}$, as expected.

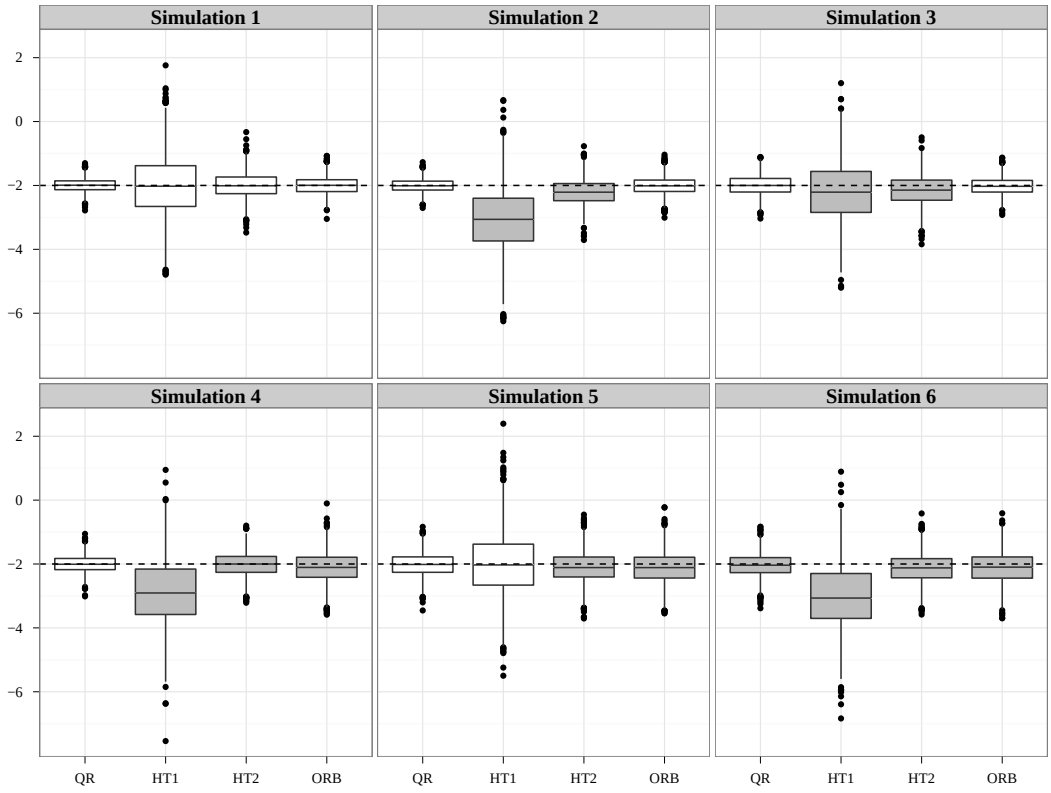


Figure 4.1: Simulation 1: $p, q, \gamma_j's, \beta_j's$ all correctly specified; Simulation 2: p, q mis-specified; Simulation 3: $p, \gamma_j's$ mis-specified; Simulation 4: $q, \beta_j's$ mis-specified; Simulation 5: $\beta_j's, \gamma_j's$ mis-specified; Simulation 6: $p, q, \gamma_j's, \beta_j's$ all mis-specified. Grey boxes indicate the estimators that are susceptible to model mis-specifications in a simulation.

Table 4.3: Summary statistics of simulation study

Simulation	Estimator	Mean Squared Error	Bias
1	QR	0.057	0.001
	HT1	0.979	-0.043
	HT2	0.138	0.005
	ORB	0.152	-0.033
2	QR	0.058	-0.005
	HT1	2.094	-1.030
	HT2	0.187	-0.199
	ORB	0.153	-0.025
3	QR	0.122	0.015
	HT1	0.890	-0.183
	HT2	0.226	-0.147
	ORB	0.147	-0.014
4	QR	0.067	-0.004
	HT1	1.866	-0.871
	HT2	0.146	0.006
	ORB	0.234	-0.120
5	QR	0.123	-0.013
	HT1	0.963	-0.007
	HT2	0.229	-0.079
	ORB	0.228	-0.117
6	QR	0.133	-0.053
	HT1	2.070	-1.001
	HT2	0.235	-0.138
	ORB	0.238	-0.105

4.5.2 Burden of Illness Data Analysis

We use the burden of illness dataset to further demonstrate the efficiency of the proposed estimator. This dataset is recorded from an observational study carried out by Duke University Medical Center. The study population included individuals with one or two vessel coronary artery diseases, ejection fraction $\geq 30\%$, and no history of congestive heart failure([Anstrom and Tsiatis(2001)]). The structure of this dataset fits in our framework because of the following two features. First, this study is observational, thus the treatment assignment is not randomized but depends on the baseline demographics, medical history of the patients. Second, there are missing data in response. The missing could related to the types of treatment, the baseline covariates and more importantly, some post-baseline covariates like safety issues or intermediate diagnosis. In this section, we will show numerically that the post-baseline variables actually play an important role in the estimation procedure.

Due to the limited availability of the original dataset. We simulate the dataset according to the descriptive statistics given in ([Anstrom and Tsiatis(2001)]) which can be summarized as

- 2284 out of 3214 patients have complete data.
- The mean medical cost was \$41,793 for PCI and \$26,801 for MED within complete data.
- For PCI patients with complete data, the median cost was \$32,226, whereas the 95th and 99th percentiles were \$104,105 and \$170,971.
- For MED patients with complete data, the median cost was \$16,219, whereas the 95th and 99th percentiles were \$80,442 and \$173,204.
- Three strongest predictors in the treatment propensity model is *hypertension, history of smoking* and *ejection fraction(with range 0.3–0.92).*

- Among the 51% of individuals that received PCI, the estimated propensity scores ranged from 0.18 to 0.84 with median 0.54. For MED patients, the scores ranged from 0.17 to 0.80 with median 0.49.
- The probability of missing was 0.39 for PCI patients, and 0.23 for MED ones.

The simulation is carried out to mimic those statistics: let X_1, X_2, X_3 represent ejection fraction, smoking history and hypertension for each patient, where $X_1 \sim U[0.3, 0.92], X_2 \sim \text{Bernulli}(0.5), X_3 \sim U[0, 1]$. Let Z be the treatment indicator, where $Z = 1$ means receiving PCI. Assume the propensity score follows a logistic regression model

$$P(Z = 1) = \frac{e^{\eta}}{1 + e^{\eta}}, \quad \text{where} \quad \eta = 1.2(X_1 - X_2 + X_3 - 0.6)$$

The post-baseline variable W is simulated as follows

$$W = \begin{cases} (2X_1 + X_2 + 2X_3)/5 + N(0, 0.1) & \text{for MED patient} \\ (X_1 + 2X_2 + X_3)/4 + N(0, 0.1) & \text{for PCI patient} \end{cases}$$

The missing probability also follows a logistic regression model

$$P(\text{missing}) = \frac{e^{\lambda}}{1 + e^{\lambda}}, \quad \text{where} \quad \lambda = 1_{\{W \geq 0.8\}} + X_1 - X_2 + X_3 + 0.4Z - 1.8$$

The dependence of the post-baseline variable W can be interpreted as when safety index or intermediate disease measurement is over a threshold, the patient will more likely have missing data. Finally, the 5 year medical cost Y is simulated as follows

$$Y = \begin{cases} 18000 + 25000X_1X_3 + 30000W + N(0, 10000) & \text{PCI} \\ 9000 + 20000X_1X_3 + 22000W + N(0, 5000) & \text{MED} \end{cases}$$

The above models are simple, however, sufficient to illustrate our idea.

We simulate the dataset 2000 times and use HT1, HT2 and ORB estimator as benchmarks. The results are summarized in Table 4.4.

Table 4.4: Summary statistics of real data analysis

Estimator	Mean Squared Error	Bias
QR	158376	7.9
HT1	381102	13.2
HT2	191325	8.4
ORB	165671	-73.1

Straightforward calculation shows that the average 5-year cost of PCI is $\$14,382$ more than that of MED based on the simulation model we are using. Table 4.4 shows that our proposed estimator outperforms the other three in terms of bias and MSE.

Chapter 5
Concluding Remarks

In this dissertation, we develop new approaches to handle two specific problems in recent statistical research: 1) model selection with hierarchical structure; 2) semi-parametric inference with missing data. For the first problem, we explicitly incorporate the hierarchical structure among the predictors in model selection problem. "Hierarchical scores" are constructed from the known hierarchical tree and used as weights on the penalty function in Elastic net approach. The resulting estimator is proved to have the hierarchical grouping property and provide consistent model selection. We present our idea through Elastic net penalty. However, the construction of hierarchical score is independent of the choice of penalty functions. Thus we can combine the hierarchical scores with any types of penalty function which encourages grouping. For example, the aforementioned OSCAR approach is a good candidate. Because of the special structure of OSCAR penalty, its model selection consistency remain unclear. Nevertheless, the hierarchical grouping property is still valid. We can show that if we replace the Enet with OSCAR in (2.2.6), then $\hat{\beta}_i = \hat{\beta}_j$ if the tuning parameter is over a threshold of order $O(\sqrt{1-\varphi_{ij}\phi_{ij}})$.

For the second problem, we consider two frameworks separately: generalized partially linear model and causal inference with observation study. We discuss the estimation method and algorithm for generalized partially linear regression model with missing response variables. The missing pattern we consider is a generalization of many different missing pattern discussed in the existing literatures. For the estimation method, we show that the estimator of parametric part, which

maximizes the generalized profile likelihood, is consistent and semiparametric efficient under some regularity conditions. In addition, we proposed an iterative algorithm to obtain the estimators. The algorithm runs iteratively between two modules, one of which uses EM algorithm and WILS to get the estimator of parametric component; the other uses EM algorithm and smoothing methods to obtain the estimator of nonparametric component. Simulation studies were performed to illustrate the proposed methodology and the simulation results showed that our algorithm works well in finite sample cases. Under group testing setting, condition C_6 and C_7 can be easily verified if the number of patient in each testing group is bounded as the total number of patients goes to infinity.

The semiparametric efficiency bound inspired estimator proposed under the causal inference framework is an important addition to the family of double-robust estimators. It specifically addresses the causal effect estimation problem in observational studies when responses were missing at random conditional on auxiliary variables. Under the assumption of joint ignorability for the responses and these auxiliary variables, the estimator is shown to be consistent for the causal effect. The proposed estimator is semiparametric in nature because only part of the full-data model needs to be specified parametrically in the form of commonly used regression models. As a result we circumvent the difficulties associated with the modeling of auxiliary variable distributions. Furthermore, we have shown both in theory and through simulations that the quadruple paired robustness property of this estimator provides additional protections against model misspecifications of the parametric parts of the full-data model. For the robustness property, it's important to note its uniqueness compared to the usual double-robustness due to the particular setup of our problem. In a typical double robust estimation problem, the estimator would remain consistent when one of the two components, either the propensity part of the model or the outcome regression part of the model is correctly specified. This will no longer be sufficient when two

types of "missing data" processes - the imbalance of baseline covariates between treatments, and the missing of responses are present simultaneously. In this case, we need to correctly specify at least one component from each of the two missing data processes in order to consistently estimate the causal effect. This is exactly what the quadruple paired robustness property has shown.

Appendix A
Proofs

A.1 Proof of Theorem 2.2.1

First, it is easy to see that if s_i's satisfies Condition 1 and 2 when $\alpha = 1$, then for any $\alpha > 0$, the two conditions still hold. So for the proof below, we set $\alpha = 1$. Thus, we have

$$s_i = \sum_{l=1}^{p-1} \mathbf{v}_i(l)\tau^{-l}, \quad \text{for} \quad i = 1, \cdots, p$$

By the construction of s_i, Condition 1 holds because of the two facts: first, if x_j and x_k have same parents, $\mathbf{v}_j = \mathbf{v}_k$, thus $s_j = s_k$. Second, if $s_j = s_k$, from the unique representation of the number system in base $\tau \geq 3$, $\mathbf{v}_j = \mathbf{v}_k$, thus since the binary representation is injective, x_j and x_k have same parents.

For condition 2, assume that the ancestors of x_i and x_k is a subset of the ancestors of x_i and x_j. Then there exists two integer $1 \leq L_2 < L_1 \leq p-1$ such that

$$\mathbf{v}_i(l) = \mathbf{v}_j(l), l = 1, \cdots, L_1 - 1 \quad \text{and} \quad \mathbf{v}_i(l) = \mathbf{v}_k(l), l = 1, \cdots, L_2 - 1$$

$$\mathbf{v}_i(L_1) = \mathbf{v}_j(L_1) = 1 \quad \text{and} \quad \mathbf{v}_i(L_2) = \mathbf{v}_k(L_2) = 1$$

$$\{L_2 < l \leq L_1 | \mathbf{v}_i(l) \neq \mathbf{v}_k(l)\} \text{is not empty}$$

First, we have following inequality:

$$\begin{aligned} \min(\frac{s_i}{s_j}, \frac{s_j}{s_i}) &\geq \frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_1+1}^{p-1} \tau^{-l}} \\ &> \frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}} \end{aligned}$$

Let $L_* = \inf\{L_2 < l \leq L_1 | \mathbf{v}_i(l) \neq \mathbf{v}_k(l)\}$

$$\begin{aligned}
\min(\frac{s_i}{s_k}, \frac{s_k}{s_i}) &\leq \frac{\sum_{l=1}^{L_*-1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_*+1}^{p-1} \tau^{-l}}{\sum_{l=1}^{L_*-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_*}} \\
&\leq \frac{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \sum_{l=L_1+1}^{p-1} \tau^{-l}}{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_1}} \\
&< \frac{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}}{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_1}}
\end{aligned}$$

The second inequality is basic algebra. It is easy to show that if $\tau \geq 3$, we have

$$\frac{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l}}{\sum_{l=1}^{L_1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}} \geq \frac{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \frac{1}{\tau-1}\tau^{-L_1}}{\sum_{l=1}^{L_1-1} \mathbf{v}_i(l)\tau^{-l} + \tau^{-L_1}}$$

Thus, we have

$$\min(\frac{s_i}{s_j}, \frac{s_j}{s_i}) > \min(\frac{s_i}{s_k}, \frac{s_k}{s_i})$$

A.2 Proof of Theorem 2.3.1

By taking derivatives on (2.3.5) with respect to γ_i and γ_j respectively, we have

$$0 = -2s_i\tilde{\mathbf{x}}_i'(\mathbf{y} - \sum_k \hat{\gamma}_k\tilde{\mathbf{x}}_k) + \lambda_2\|\hat{\theta}_i\|_1 + \lambda_1(1 + 2\delta\hat{\gamma}_i)$$

$$0 = -2s_j\tilde{\mathbf{x}}_j'(\mathbf{y} - \sum_k \hat{\gamma}_k\tilde{\mathbf{x}}_k) + \lambda_2\|\hat{\theta}_j\|_1 + \lambda_1(1 + 2\delta\hat{\gamma}_j)$$

Subtracting above two equation and taking the absolute value on two sides, we have

$$\begin{aligned}
|\hat{\gamma}_i - \hat{\gamma}_j| &\leq \frac{1}{\lambda_1\delta}\|\mathbf{y} - \sum_k \hat{\gamma}_k\tilde{\mathbf{x}}_k\|_2 \cdot \|s_i\tilde{\mathbf{x}}_i - s_j\tilde{\mathbf{x}}_j\|_2 + \frac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_i\|_1 - \|\hat{\theta}_j\|_1\Big| \\
&\leq \frac{\sqrt{n}\|\mathbf{y}\|_2 s^{(K)}}{\lambda_1\delta}\sqrt{2(1 - \varphi_{ij}\phi_{ij})} + \frac{\lambda_2}{2\lambda_1\delta}\Big|\|\hat{\theta}_i\|_1 - \|\hat{\theta}_j\|_1\Big|
\end{aligned}$$

Using inequality $\|\hat{\theta}_i\|_1 \leq \sqrt{p_i}\|\theta_i\|_2$ and $n = \theta_i'\mathbf{x}_i'\mathbf{x}_i\theta_i \geq nc_*\|\theta_i\|_2^2$, we have $\|\hat{\theta}_i\|_1 \leq O(\sqrt{p_i})$, where c_* is the lower bound of the eigenvalue of $\Sigma_k = \frac{1}{n}\mathbf{X}_k'\mathbf{X}_k$ for $k = 1, \cdots, K$. As a result, if $\lambda_2 \max\sqrt{p_k} = o(n)$, combining with the fact that $\|\mathbf{y}\|_2 = O(\sqrt{n})$, the last equation holds.

A.3 Proof of Theorem 2.3.2

First, we prove 1). Define event $E = \{\|\hat{\beta}_A - \beta_A^0\|_\infty \leq (1-\alpha)\beta_*, \hat{\beta}_B = 0\}$ for $0 < \alpha < 1$. By observation, $\{\text{sgn}(\hat{\beta}_A) = \text{sgn}(\beta_A^0), \hat{\beta}_B = 0\} \supseteq E$. For any local minimizer $\hat{\beta}$ of (2.3.6) on the subspace $\{\beta : \beta_{(A\cup B)^c} = 0\}$, by taking differentiation with respect to β, we have:

$$-2\mathbf{x}_{kj}'(\mathbf{y} - \mathbf{X}\hat{\beta}) \quad + \quad \lambda_2 \frac{z_{kj}}{s_k} + \frac{\lambda_1}{\sqrt{n}}\left(\frac{\mathbf{x}_{kj}'\mathbf{X}_k\hat{\beta}_k}{s_k\|\mathbf{X}_k\hat{\beta}_k\|_2} + \frac{2\delta\mathbf{x}_{kj}'\mathbf{X}_k\hat{\beta}_k}{\sqrt{n}s_k}\right) = 0$$

where $(k,j) \in A \cup B$, $z_{kj} = \text{sgn}(\hat{\beta}_{kj})$ if $\hat{\beta}_{kj} \neq 0$ and $|z_{kj}| \leq 1$ if $\hat{\beta}_{kj} = 0$. The existence of $\hat{\beta}$ in E is implied by the following two KKT conditions:

$$-2\mathbf{X}_A'\varepsilon - 2\mathbf{X}_A'\mathbf{X}_A\beta_A^0 + 2n\tilde{\Sigma}\hat{\beta}_A + \lambda_1\bar{P}_1(\hat{\beta}_A) + \lambda_2\bar{P}_2 = 0 \tag{A.3.1}$$

$$-2S_B\mathbf{X}_B'(\varepsilon + \mathbf{X}_A(\beta_A^0 - \hat{\beta}_A)) + \lambda_2\bar{P}_3 + \lambda_1\bar{P}_4(\hat{\beta}_A) + \lambda_1\delta\bar{P}_5(\hat{\beta}_A) = 0 \tag{A.3.2}$$

where $\bar{P}_1(\hat{\beta}_A) = \left(\frac{\mathbf{X}_{A_k}'\mathbf{X}_{A_k}\hat{\beta}_{A_k}}{s_k\sqrt{n}\|\mathbf{X}_{A_k}\hat{\beta}_{A_k}\|_2}\right)_{1\leq k\leq r}$, $\bar{P}_2 = \left(\frac{\text{sgn}(\beta_{A_k}^0)}{s_k}\right)_{1\leq k\leq r}$, $\bar{P}_3 = (\mathbf{u}_k)_{1\leq k\leq r}$, $\mathbf{u}_i = (u_{ij}, (i,j) \in B_i)$ with $|u_{ij}| \leq 1$, $\bar{P}_4(\hat{\beta}_A) = \left(\frac{\mathbf{X}_{B_k}'\mathbf{X}_{A_k}\hat{\beta}_{A_k}}{\sqrt{n}\|\mathbf{X}_{A_k}\hat{\beta}_{A_k}\|_2}\right)_{1\leq k\leq r}$, $\bar{P}_5(\hat{\beta}_A) = \left(\frac{2\mathbf{X}_{B_k}'\mathbf{X}_{A_k}\hat{\beta}_{A_k}}{n}\right)_{1\leq k\leq r}$

We can rewrite condition (A.3.1) as

$$\beta_A^0 - \hat{\beta}_A = \tilde{\Sigma}^{-1}\left(-\mathbf{X}_A'\varepsilon/n + \lambda_1\delta S_A^{-1}\bar{\Sigma}\beta_A^0/n + \frac{1}{2}\lambda_1\bar{P}_1/n + \frac{1}{2}\lambda_2\bar{P}_2/n\right) \tag{A.3.3}$$

Notice that $\bar{P}_2$ can be bounded using the following inequality:

$$\|\mathbf{X}_k'\mathbf{X}_k\hat{\beta}_k\|_\infty \leq \sqrt{n}\|\mathbf{X}_k\hat{\beta}_k\|_2$$

The right hand side of (A.3.3) is bounded by $(1-\alpha)\beta_*$ given the following inequality:

$$\|\mathbf{X}_A'\varepsilon/\sqrt{n}\|_\infty \leq C_1^{-1}(1-\alpha)\beta_*\sqrt{n} - C_1^{-1}\zeta\lambda_1/\sqrt{n} - \frac{\lambda_1+\lambda_2}{2\sqrt{n}\min s_k} \tag{A.3.4}$$

Combining with condition (e) and Miranda's existence theorem, (A.3.4) further implies the existence of $\hat{\beta}$ in $\{\|\hat{\beta}_A - \beta_A^0\|_\infty \leq (1-\alpha)\beta_*\}$.

Condition (A.3.2) is equivalent to

$$\left\| -2S_B\mathbf{X}'_B\varepsilon - 2S_B\mathbf{X}'_B\mathbf{X}_A(\beta^0_A - \hat{\beta}_A) + \lambda_1\bar{P}_4(\hat{\beta}_A) + \lambda_1\delta\bar{P}_5(\hat{\beta}_A)\right\|_\infty \leq \lambda_1$$

By plugging (A.3.3) we have

$$\begin{aligned}\left\| -2S_B\mathbf{X}'_B\left(I - \frac{\mathbf{X}_A\tilde{\Sigma}^{-1}\mathbf{X}'_A}{n}\right)\varepsilon \right. &- 2S_B\mathbf{X}'_B\mathbf{X}_A\tilde{\Sigma}^{-1}\left(\lambda_1\delta S_A^{-1}\bar{\Sigma}\beta^0_A/n + \frac{\lambda_2\bar{P}_2}{2n}\right) \\ &\left. - 2S_B\mathbf{X}'_B\mathbf{X}_A\tilde{\Sigma}^{-1}\frac{\lambda_1\bar{P}_1}{2n} + \lambda_1\bar{P}_4 + \lambda_1\delta\bar{P}_5\right\|_\infty \leq \lambda_2\end{aligned}$$

We will bound the second to fifth term of above inequality. By G-HEIC condition, we have

$$\left\|2S_B\mathbf{X}'_B\mathbf{X}_A\tilde{\Sigma}^{-1}\left(\lambda_1\delta S_A^{-1}\bar{\Sigma}\beta^0_A/n + \frac{\lambda_2\bar{P}_2}{2n}\right)\right\|_\infty \leq (1-\eta)\lambda_2$$

The rest terms can be bounded as

$$\left\|2S_B\mathbf{X}'_B\mathbf{X}_A\tilde{\Sigma}^{-1}\frac{\lambda_1\bar{P}_1}{2n}\right\|_\infty \leq \lambda_1 C_3 C_1\frac{\max s_k}{\min s_k} \qquad \left\|\lambda_1\frac{\mathbf{X}'_{B_k}(\mathbf{X}_{A_k}\hat{\beta}_{A_k})}{\sqrt{n}\|\mathbf{X}_{A_k}\hat{\beta}_{A_k}\|_2}\right\|_\infty \leq \lambda_1$$

$$\begin{aligned}\left\|\frac{2}{n}\mathbf{X}'_{B_k}\mathbf{X}_{A_k}\hat{\beta}_{A_k}\right\|_\infty &\leq \left\|\frac{2}{n}\mathbf{X}'_{B_k}\mathbf{X}_{A_k}(\hat{\beta}_{A_k} - \beta^0_{A_k})\right\|_\infty + \left\|\frac{2}{n}\mathbf{X}'_{B_k}\mathbf{X}_{A_k}\beta^0_{A_k}\right\|_\infty \\ &\leq \frac{2}{n}\left\{nC_3\beta_* + nC_3\|\beta^0_A\|_\infty\right\}\end{aligned}$$

Thus (A.3.2) is implied by

$$\begin{aligned}\left\|\mathbf{X}'_B\left(I - \frac{\mathbf{X}_A\tilde{\Sigma}^{-1}\mathbf{X}'_A}{n}\right)\varepsilon\right\|_\infty/\sqrt{n} &\leq \frac{\eta\lambda_2}{2\sqrt{n}\max s_k} - \frac{\lambda_1 C_1 C_3}{2\sqrt{n}\min s_k} - \frac{\lambda_1}{2\sqrt{n}\max s_k} \\ &- \frac{\lambda_1\delta}{\sqrt{n}\max s_k}C_3(\beta_* + \|\beta^0_A\|_\infty) \qquad \text{(A.3.5)}\end{aligned}$$

Direct calculation shows that $\mathbf{x}'_{ij}\left(I - \frac{\mathbf{X}_A\tilde{\Sigma}^{-1}\mathbf{X}'_A}{n}\right)\varepsilon/\sqrt{n}$ is normally distributed with mean 0 and variance no larger than σ^2.

So by condition (e), the above two inequality (A.3.4) and (A.3.5) hold in probability using the following classical standard Gaussian tail bound and Bonferroni's inequality. Let $A_n = \{\|\mathbf{X}'_A\varepsilon/\sqrt{n}\|_\infty \leq O(\sqrt{n}\beta_*), \|\mathbf{X}'_B\left(I - \frac{\mathbf{X}_A\tilde{\Sigma}^{-1}\mathbf{X}'_A}{n}\right)\varepsilon\|_\infty/\sqrt{n} \leq$

$O(\sqrt{n}\beta_*)$, we have

$$\begin{aligned} P(A_n^c) &\leq \sum_{(k,j)\in A} P(|\mathbf{X}'_{kj}\varepsilon/\sqrt{n}| > O(\sqrt{n}\beta_*)) \\ &\quad + \sum_{(k,j)\in B} P(\left|\mathbf{X}'_{kj}\left(I - \frac{\mathbf{X}_A\tilde{\Sigma}^{-1}\mathbf{X}'_A}{n}\right)\varepsilon/\sqrt{n}\right| > O(\sqrt{n}\beta_*)) \\ &\leq pO(e^{-\frac{n\beta_*^2}{2\sigma^2}}/(\sqrt{n}\beta_*/\sigma)) \to 0 \end{aligned}$$

For 2), without loss of generality, assuming $\tilde{\beta} = \hat{\beta}_A + \sum_{k=r+1}^K t_k\theta_k$, where $t_k \geq 0$, $\|\mathbf{X}_k\theta_k\|_2^2 = n$ for $k = r+1, \cdots, K$. $P_n(\tilde{\beta}) \geq P_n(\hat{\beta}_A)$ is equivalent to

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}_A\hat{\beta}_A - \sum_{k=r+1}^K t_k\mathbf{X}_k\theta_k\|_2^2 + \lambda_1\sum_{k=r+1}^K \frac{t_k + \delta t_k^2}{s_k} + \lambda_2\sum_{k=r+1}^K \frac{t_k\|\theta_k\|_1}{s_k} \\ \geq \|\mathbf{y} - \mathbf{X}_A\hat{\beta}_A\|_2^2 \end{aligned}$$

which is further implied by

$$|(\mathbf{X}_k\theta_k)'(\mathbf{y} - \mathbf{X}_A\hat{\beta}_A)| \leq \frac{\lambda_2\|\theta_k\|_1}{2s_k} + \frac{\lambda_1}{2s_k} \quad \text{for } k = r+1, \cdots, K.$$

Replacing $\mathbf{y}$ by $\varepsilon + \mathbf{X}_A\beta_A^0$, the above inequality is implied by

$$\left|(\frac{\mathbf{X}_k\theta_k}{\sqrt{n}})'\varepsilon\right| \leq \left(\frac{\lambda_2}{2\sqrt{n}\max s_k} - C_2\sqrt{n}\beta_*\right)\|\theta_k\|_1 + \frac{\lambda_1}{2\sqrt{n}\max s_k} \tag{A.3.6}$$

The above two inequality (A.3.6) holds in probability using the following classical standard Gaussian tail bound and Bonferroni's inequality. Let $B_n = \{|(\frac{\mathbf{X}_k\theta_k}{\sqrt{n}})'\varepsilon| \leq O(\sqrt{n}\beta_*)\}, k = r+1, \cdots, K$, then

$$\begin{aligned} P(B_n^c) &\leq \sum_{k=r+1}^K P\left(\left|(\frac{\mathbf{X}_k\theta_k}{\sqrt{n}})'\varepsilon\right| > O(\sqrt{n}\beta_*)\right) \\ &\leq pO(e^{-\frac{n\beta_*^2}{2\sigma^2}}/(\sqrt{n}\beta_*/\sigma)) \to 0 \end{aligned}$$

A.4 Proof of Theorem 3.2.1(a)

Under the regularity conditions, $l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)$ is continuous in $\boldsymbol{\beta}$ and a measurable function of $\mathbf{z}$ and $\mathbf{v}$ for each $\boldsymbol{\beta}$. Therefore, it follows that $\hat{\boldsymbol{\beta}}$ is measurable.

By Conditions I and (3.2.4),

$$n^{-1}l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \to_p l_0(\boldsymbol{\beta}) \quad \text{for each } \boldsymbol{\beta} \in \mathcal{B}.$$

Furthermore, by Condition C, for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}$,

$$\begin{aligned} & n^{-1}\big|l\big(\boldsymbol{\beta}_1, g_{\boldsymbol{\beta}_1}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\boldsymbol{\beta}_2, g_{\boldsymbol{\beta}_2}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big| \\ \leq \;& A_n \left|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right| + B_n \left\|g_{\boldsymbol{\beta}_1} - g_{\boldsymbol{\beta}_2}\right\| \\ \leq \;& A_n \left|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right| + B_n \sup_{\boldsymbol{\beta}} \left\|g'_{\boldsymbol{\beta}}\right\| \left|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right| \\ = \;& C_n|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2|. \end{aligned}$$

where $C_n = A_n + B_n \sup_{\boldsymbol{\beta}} \|g'_{\boldsymbol{\beta}}\|$ is bounded in probability. It follows from the Tightness Characterization Theorem in [Newey (1991)] that

$$\big\{n^{-1}l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big\}$$

is tight. Hence, by Theorem 2.1 in [Newey (1991)]

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \big|n^{-1}l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l_0(\boldsymbol{\beta})\big| \to_p 0 \quad \text{as} \quad n \to \infty. \tag{A.4.1}$$

Also, for each $\boldsymbol{\beta}$, by Conditions C,

$$\begin{aligned} & n^{-1}\big|l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big| \\ & \leq B_n \|\hat{g}_{\boldsymbol{\beta}} - g_{\boldsymbol{\beta}}\|. \end{aligned}$$

Thus, by Condition NP,

$$\sup_{\boldsymbol{\beta}} \frac{1}{n}\big|l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big| \to_p 0 \quad \text{as } n \to \infty. \tag{A.4.2}$$

Now combine (A.4.1) and (A.4.2), we have

$$\sup_{\boldsymbol{\beta}} \big|\frac{1}{n}l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l_0(\boldsymbol{\beta})\big| \to_p 0 \quad \text{as } n \to \infty. \tag{A.4.3}$$

Thus, by (A.4.3) and triangle inequalities,

$$\sup_{\boldsymbol{\beta}} \frac{1}{n}l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \to_p \sup_{\boldsymbol{\beta}} l_0(\boldsymbol{\beta}) = l_0(\boldsymbol{\beta}_0). \tag{A.4.4}$$

Note that $\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)$, by (A.4.3), we have

$$\Big|\frac{1}{n} l\big(\hat{\boldsymbol{\beta}}, \hat{g}_{\hat{\boldsymbol{\beta}}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l_0(\hat{\boldsymbol{\beta}})\Big| \to_p 0 \quad \text{as } n \to \infty.$$

It follows from (A.4.4) that

$$l_0(\hat{\boldsymbol{\beta}}) \to_p l_0(\boldsymbol{\beta}_0) \quad \text{as } n \to \infty.$$

Let $\mathcal{N}_0$ denote an open neighborhood of $\boldsymbol{\beta}_0$ and consider the compact set $\mathcal{B}_0 = \mathcal{B} \backslash \mathcal{N}_0$. For any fixed $\boldsymbol{\beta}_1 \in \mathcal{B}_0$, we can construct an open neighborhood $\mathcal{N}_{\beta_1}$ of $\boldsymbol{\beta}_1$ such that there exists an $\epsilon > 0$ satisfies

$$\inf_{\boldsymbol{\beta}_1 \in \mathcal{N}_{\boldsymbol{\beta}_1}} |l_0(\boldsymbol{\beta}_1) - l_0(\boldsymbol{\beta}_0)| > \epsilon,$$

which implies

$$P_0(\hat{\boldsymbol{\beta}} \in \mathcal{N}_{\boldsymbol{\beta}_1}) \leq P_0\Big(\Big|l_0(\hat{\boldsymbol{\beta}}) - l_0(\boldsymbol{\beta}_0)\Big| > \epsilon \Big) \to 0 \quad \text{as } n \to \infty.$$

Note that $\cup_{\beta \in \mathcal{B}_0} \mathcal{N}_\beta$ provides an open cover of $\mathcal{B}_0$. By compactness of $\mathcal{B}_0$, there exists a finite subcover $\{\mathcal{N}_{\boldsymbol{\beta}_1}, \cdots, \mathcal{N}_{\boldsymbol{\beta}_k}\}$. Then

$$P_0(\hat{\boldsymbol{\beta}} \notin \mathcal{N}_0) = P_0(\hat{\boldsymbol{\beta}} \in \mathcal{B}_0) \leq \sum_{j=1}^{k} P_0(\hat{\boldsymbol{\beta}} \in \mathcal{N}_{\boldsymbol{\beta}_j}) \to 0 \quad \text{as } n \to \infty.$$

Therefore,

$$\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}_0 \quad \text{as } n \to \infty.$$

The convergence of non-parametric part will be proved after the proof of Theorem 2.

A.5 Proof of Theorem 3.2.2

Denote by $\tilde{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta} \in \mathcal{B}} l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot)\big)$. Then by the conditions $C_1 - C_5$ and Theorem 3 in [Bhat (1974)], we have

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_{\mathscr{D}} N(0, i_{\boldsymbol{\beta}}^{-1})$$

Now for $\hat{\boldsymbol{\beta}}$, using a Taylor's expansion,

$$\begin{aligned} 0 &= \frac{\partial l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\ &= \frac{\partial l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2 l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \end{aligned}$$

where $\hat{\boldsymbol{\beta}}^*$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Hence,

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= -\left\{\frac{1}{n}\frac{\partial^2 l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*}\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\frac{\partial l\big(\boldsymbol{\beta}, \hat{g}_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\} \\ &= -\left\{\frac{1}{n}\frac{\partial^2 l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\frac{\partial l\big(\boldsymbol{\beta}, g_{\boldsymbol{\beta}}(\cdot)\big)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\} + o_p(1) \\ &= \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1) \end{aligned}$$

The second equality follows from conditions C_6 and C_7 and the fact that $\hat{\boldsymbol{\beta}}^* \to_p \boldsymbol{\beta}_0$. The last equality follows from the definition of $\tilde{\boldsymbol{\beta}}$. Thus $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_{\mathscr{D}} N(0, i_{\boldsymbol{\beta}}^{-1})$.

The result

$$\hat{i}_{\boldsymbol{\beta}} \to_p i_{\boldsymbol{\beta}} \quad \text{as } n \to \infty,$$

follows from Condition C_7 and Theorem 1.

A.6 Proof of Theorem 3.2.1(b)

The large sample properties for the non-parametric part is relatively easy to verify given the results of Theorem 2.

For the first equation:

$$\begin{aligned} \|\hat{g}_{\hat{\boldsymbol{\beta}}} - g_{\boldsymbol{\beta}_0}\| &\leq \|\hat{g}_{\hat{\boldsymbol{\beta}}} - \hat{g}_{\boldsymbol{\beta}_0}\| + \|\hat{g}_{\boldsymbol{\beta}_0} - g_{\boldsymbol{\beta}_0}\| \\ &\leq \sup_{\boldsymbol{\beta}^* \in \mathcal{B}} \sup_{v \in \mathcal{V}} \Big|\frac{\partial \hat{g}_{\boldsymbol{\beta}}(v)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \cdot |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0| + o_p(1) \\ &\leq C|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0| + o_p(1) = o_p(1) \end{aligned}$$

The second inequality is derived from the definition of the norm and Taylor expansion together with the conditions NP. The last equality is a direct implication

from Theorem 1.

For the second equation:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{g}_{\hat{\boldsymbol{\beta}}}(v_i)-g_{\boldsymbol{\beta}_0}(v_i)\right\}^2 \leq \|\hat{g}_{\hat{\boldsymbol{\beta}}}-g_{\boldsymbol{\beta}_0}\|^2 = o_p(1)$$

A.7 Proof of Theorem 3.4.1

By Ahmad et al. (2005), $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{g}}^*$ are given by

$$\begin{cases} \hat{\boldsymbol{\beta}}^* = \left\{(\mathbf{w}-M\mathbf{w})^T(\mathbf{w}-M\mathbf{w})\right\}^{-}(\mathbf{w}-M\mathbf{w})^T(\mathbf{y}-M\mathbf{y}) \\ \hat{\mathbf{g}}^* = M(\mathbf{y}-\mathbf{w}\hat{\boldsymbol{\beta}}^*). \end{cases}$$

Under Gaussian regression model, $\mu_i = \eta_i$; the link function h is just the identity function.

For any fixed $\boldsymbol{\beta}$, the corresponding estimator of nonparametric component from the iterative algorithm satisfies the following equation:

$$\begin{aligned} \hat{\mathbf{g}}_{\boldsymbol{\beta}} &= ME(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\beta},\hat{\mathbf{g}}_{\boldsymbol{\beta}}) - M\mathbf{w}\boldsymbol{\beta} \\ &\equiv M\mathbf{E1} - M\mathbf{w}\boldsymbol{\beta}, \end{aligned}$$

and its derivative to $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\mathbf{g}}'_{\boldsymbol{\beta}} &= M\frac{\partial E(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\beta},\hat{\mathbf{g}}_{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}} - M\mathbf{w} \\ &\equiv M\mathbf{E2} - M\mathbf{w}, \end{aligned}$$

where $\mathbf{E1} = E(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\beta},\hat{\mathbf{g}}_{\boldsymbol{\beta}})$ and $\mathbf{E2} = \partial E(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\beta},\hat{\mathbf{g}}_{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}$. The estimator $\hat{\boldsymbol{\beta}}$ is the solution of

$$(\mathbf{E1} - \mathbf{w}\boldsymbol{\beta} - \hat{\mathbf{g}}_{\boldsymbol{\beta}})^T(\mathbf{w} + \hat{\mathbf{g}}'_{\boldsymbol{\beta}}) = 0,$$

that is

$$\begin{aligned} 0 &= \left\{(\mathbf{E1}-M\mathbf{E1}) - (\mathbf{w}-M\mathbf{w})\boldsymbol{\beta}\right\}^T(\mathbf{w}-M\mathbf{w}+M\mathbf{E2}) \\ &= (\mathbf{E1}-M\mathbf{E1})^TM\mathbf{E2} - \boldsymbol{\beta}^T(\mathbf{w}-M\mathbf{w})^TM\mathbf{E2} \\ &\quad +(\mathbf{E1}-M\mathbf{E1})^T(\mathbf{w}-M\mathbf{w}) - \boldsymbol{\beta}^T(\mathbf{w}-M\mathbf{w})^T(\mathbf{w}-M\mathbf{w}) \\ &= (\mathbf{E1}-M\mathbf{E1})^T(\mathbf{w}-M\mathbf{w}) - \boldsymbol{\beta}^T(\mathbf{w}-M\mathbf{w})^T(\mathbf{w}-M\mathbf{w}). \end{aligned}$$

The third equation is from $M^T = M$ and $M^T M = M$. Therefore, $\hat{\boldsymbol{\beta}}$ is the solution of

$$\begin{aligned}\boldsymbol{\beta} &= \left\{(\mathbf{w} - M\mathbf{w})^T(\mathbf{w} - M\mathbf{w})\right\}^{-1}\left\{(\mathbf{w} - M\mathbf{w})^T(\mathbf{E1} - M\mathbf{E1})\right\} \\ &= E[\hat{\boldsymbol{\beta}}^*|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \hat{\mathbf{g}}_{\boldsymbol{\beta}}],\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{g}}_{\boldsymbol{\beta}} &= M\mathbf{E1} - M\mathbf{w}\boldsymbol{\beta} \\ &= E(M\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \hat{\mathbf{g}}_{\boldsymbol{\beta}}) - E(M\mathbf{w}\hat{\boldsymbol{\beta}}^*|\mathbf{z}, \boldsymbol{\beta}, \hat{\mathbf{g}}_{\boldsymbol{\beta}}) \\ &= E[\hat{\mathbf{g}}^*|\mathbf{z}, \boldsymbol{\beta}, \hat{\mathbf{g}}_{\boldsymbol{\beta}}].\end{aligned}$$

Hence, the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{g}}_{\hat{\beta}}(\cdot)$ from iterative algorithm are the solution of

$$\begin{cases} \boldsymbol{\beta} = E[\hat{\boldsymbol{\beta}}^*|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \mathbf{g}] \\ \mathbf{g} = E[\hat{\mathbf{g}}^*|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \mathbf{g}]. \end{cases}$$

A.8 Proof of Theorem 4.3.1

The following derivation of semiparametric efficiency bounds was based on the theories of [Begun et al. (1983), Newey (1990)], also see [Hahn (1998)] for a similar application.

Suppose $f(\boldsymbol{L} \mid \theta)$ is a parametric submodel of $f(\boldsymbol{L} \mid \eta)$:

$$\begin{aligned} f(\boldsymbol{L} \mid \theta) &= [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\}p(\boldsymbol{X}; \theta)f_1(\boldsymbol{W}, Y \mid \boldsymbol{X}; \theta)]^{Z(1-M)} \\ &\quad \times [\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\}\{1 - p(\boldsymbol{X}; \theta)\}f_0(\boldsymbol{W}, Y \mid \boldsymbol{X}; \theta)]^{(1-Z)(1-M)} \\ &\quad \times \{q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)p(\boldsymbol{X}; \theta)f_1^*(\boldsymbol{W} \mid \boldsymbol{X}; \theta)\}^{ZM} \\ &\quad \times [q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\{1 - p(\boldsymbol{X}; \theta)\}f_0^*(\boldsymbol{W} \mid \boldsymbol{X}; \theta)]^{(1-Z)M} \\ &\quad \times g(\boldsymbol{X}; \theta), \end{aligned} \tag{A.8.1}$$

and $f(\boldsymbol{L} \mid \theta_0)$ is the true distribution of observed data $\boldsymbol{L}$.

The score function of $f(\boldsymbol{L} \mid \theta)$ is:

$$\begin{aligned} S_\theta(\boldsymbol{L}) = & \frac{\{M - q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\}\dot{q}(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)}{q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\{1 - q(\boldsymbol{W}, \boldsymbol{X}, Z; \theta)\}} + \frac{\{Z - p(\boldsymbol{X}; \theta)\}\dot{p}(\boldsymbol{X}; \theta)}{p(\boldsymbol{X}; \theta)\{1 - p(\boldsymbol{X}; \theta)\}} \\ & + Z(1 - M)r_1(\boldsymbol{W}, Y \mid \boldsymbol{X}; \theta) + (1 - Z)(1 - M)r_0(\boldsymbol{W}, Y \mid \boldsymbol{X}; \theta) \\ & + ZMs_1(\boldsymbol{W} \mid \boldsymbol{X}; \theta) + (1 - Z)Ms_0(\boldsymbol{W} \mid \boldsymbol{X}; \theta) + t(\boldsymbol{X}; \theta) \end{aligned} \tag{A.8.2}$$

where

$$\begin{aligned} \dot{q}(\boldsymbol{w}, \boldsymbol{x}, z; \theta) &= \frac{\partial}{\partial \theta} q(\boldsymbol{w}, \boldsymbol{x}, z; \theta) \\ \dot{p}(\boldsymbol{x}; \theta) &= \frac{\partial}{\partial \theta} p(\boldsymbol{x}; \theta) \\ r_j(\boldsymbol{w}, y \mid \boldsymbol{x}; \theta) &= \frac{\partial}{\partial \theta} \log f_j(\boldsymbol{w}, y \mid \boldsymbol{x}; \theta) \\ s_j(\boldsymbol{w} \mid \boldsymbol{x}; \theta) &= \frac{\partial}{\partial \theta} \log f_j^*(\boldsymbol{w} \mid \boldsymbol{x}; \theta) \\ t(\boldsymbol{x}; \theta) &= \frac{\partial}{\partial \theta} \log g(\boldsymbol{x}; \theta) \end{aligned} \tag{A.8.3}$$

Assume for the time being that the model is completely nonparametric, which means that $q(\boldsymbol{w}, \boldsymbol{x}, z)$, $p(\boldsymbol{x})$, $f_j(\boldsymbol{w}, y|\boldsymbol{x})$, $f_j^*(\boldsymbol{w}|\boldsymbol{x})$ and $g(\boldsymbol{x})$ are all unrestricted density or probability functions. Define the tangent set to be the mean square closure of all linear combinations of score $S_\theta(\boldsymbol{L})$ for smooth parametric submodels [Newey (1990)]:

$$\Phi = \{\omega \in \mathbb{R} : E[\|\omega\|^2] < \infty, \exists \boldsymbol{a}_k S_{\theta k} \text{ with } \lim_{k \to \infty} E[\|\omega - \boldsymbol{a}_k S_{\theta k}\|^2] = 0\}.$$

The tangent set for model $f(\boldsymbol{L} \mid \eta)$ is:

$$\begin{aligned} \Phi = & [Z(1 - M)r_1(\boldsymbol{W}, Y \mid \boldsymbol{X}) + (1 - Z)(1 - M)r_0(\boldsymbol{W}, Y \mid \boldsymbol{X}) \\ & + ZMs_1(\boldsymbol{W} \mid \boldsymbol{X}) + (1 - Z)Ms_0(\boldsymbol{W} \mid \boldsymbol{X}) \\ & + \{M - q(\boldsymbol{W}, \boldsymbol{X}, Z)\}a(\boldsymbol{W}, \boldsymbol{X}, Z) + \{Z - p(\boldsymbol{X})\}b(\boldsymbol{X}) + t(\boldsymbol{X})] \end{aligned} \tag{A.8.4}$$

where $r_j(\boldsymbol{w}, y \mid \boldsymbol{x})$, $s_j(\boldsymbol{w} \mid \boldsymbol{x})$ and $t(\boldsymbol{x})$ are any one dimensional functions that

satisfy the following conditions:

$$\iint r_j(\boldsymbol{w}, y \mid \boldsymbol{x}) f_j(\boldsymbol{w}, y \mid \boldsymbol{x}) \mathrm{d}\boldsymbol{w}\mathrm{d}y = 0, \tag{A.8.5}$$

$$\int s_j(\boldsymbol{w} \mid \boldsymbol{x}) f_j^*(\boldsymbol{w} \mid \boldsymbol{x}) \mathrm{d}\boldsymbol{w} = 0, \tag{A.8.6}$$

$$\int r_j(\boldsymbol{w}, y \mid \boldsymbol{x}) \frac{f_j(\boldsymbol{w}, y \mid \boldsymbol{x})}{f_j^*(\boldsymbol{w} \mid \boldsymbol{x})} \mathrm{d}y = s_j(\boldsymbol{w} \mid \boldsymbol{x}) \quad j = 0, 1, \forall \boldsymbol{w}, \boldsymbol{x}, \tag{A.8.7}$$

$$\int t(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0. \tag{A.8.8}$$

and $a(\boldsymbol{w}, \boldsymbol{x}, z)$ and $b(\boldsymbol{x})$ are any square-integrable measurable functions.

The first lemma gives the orthogonal structure of the tangent set Φ, which is useful in finding the efficiency bound.

Lemma 1. *The tangent set* Φ *can be expressed as:*

$$\Phi = \Phi_1 \oplus \Phi_2 \oplus \Phi_3 \oplus \Phi_4 \oplus \Phi_5, \tag{A.8.9}$$

where

$$\begin{aligned}
\Phi_1 &= \{t(\boldsymbol{X})\}, \\
\Phi_2 &= \{Z(1-M) r_1(\boldsymbol{W}, Y \mid \boldsymbol{X}) + ZM s_1(\boldsymbol{W} \mid \boldsymbol{X})\}, \\
\Phi_3 &= \{(1-Z)(1-M) r_0(\boldsymbol{W}, Y \mid \boldsymbol{X}) + (1-Z) M s_0(\boldsymbol{W} \mid \boldsymbol{X})\}, \\
\Phi_4 &= \{[M - q(\boldsymbol{W}, \boldsymbol{X}, Z)] a(\boldsymbol{W}, \boldsymbol{X}, Z)\}, \\
\Phi_5 &= \{[Z - p(\boldsymbol{X})] b(\boldsymbol{X})\}.
\end{aligned}$$

The proof of Lemma 1 is based on direct computation using the chain rule for conditional expectations, thus omitted here.

By the definition of β in (4.2.1) , for a parametric submodel $f(\boldsymbol{L} \mid \theta)$,

$$\begin{aligned}
\beta(\theta) = &\iiint y^1 f_1(\boldsymbol{w}^1, y^1 \mid \boldsymbol{x}; \theta) g(\boldsymbol{x}; \theta) \mathrm{d}\boldsymbol{w}^1 \mathrm{d}\boldsymbol{x} \mathrm{d}y^1 \\
&- \iiint y^0 f_0(\boldsymbol{w}^0, y^0 \mid \boldsymbol{x}; \theta) g(\boldsymbol{x}; \theta) \mathrm{d}\boldsymbol{w}^0 \mathrm{d}\boldsymbol{x} \mathrm{d}y^0
\end{aligned}$$

Take derivative with respect to θ, we have:

$$\begin{aligned}
\frac{\partial \beta(\theta)}{\partial \theta} \mid_{\theta=\theta_0} &= E\{Y^1 r_1(\boldsymbol{W}^1, Y^1 \mid \boldsymbol{X}; \theta_0)\} + E\{\beta_1(\boldsymbol{X}) t(\boldsymbol{X}; \theta_0)\} \\
&\quad - E\{Y^0 r_0(\boldsymbol{W}^0, Y^0 \mid \boldsymbol{X}; \theta_0)\} - E\{\beta_0(\boldsymbol{X}) t(\boldsymbol{X}; \theta_0)\}
\end{aligned} \tag{A.8.10}$$

In the next lemma we show that the estimand $\beta(\theta)$ is pathwise differentiable by finding a suitable function F_β. See [Newey (1990)] for a definition of pathwise differentiable parameter.

Lemma 2. *Denote*

$$\begin{aligned}
U_1 &= \frac{Z(1-M)\{Y-\beta_1(\boldsymbol{X})\}}{p(\boldsymbol{X})\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}} \\
U_0 &= \frac{(1-Z)(1-M)\{Y-\beta_0(\boldsymbol{X})\}}{\{1-p(\boldsymbol{X})\}\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}} \\
V_1 &= \frac{Z\{q(\boldsymbol{W},\boldsymbol{X},Z)-M\}}{p(\boldsymbol{X})\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}\{\beta_1(\boldsymbol{X})-\gamma_1(\boldsymbol{W},\boldsymbol{X})\} \\
V_0 &= \frac{(1-Z)\{q(\boldsymbol{W},\boldsymbol{X},Z)-M\}}{\{1-p(\boldsymbol{X})\}\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}\{\beta_0(\boldsymbol{X})-\gamma_0(\boldsymbol{W},\boldsymbol{X})\} \\
Q &= \beta(\boldsymbol{X})-\beta \\
F_\beta &= U_1-U_0+V_1-V_0+Q
\end{aligned}$$

For any parametric submodel $f(\boldsymbol{L}\mid\theta)$, we have:

$$\frac{\partial\beta(\theta)}{\partial\theta}\mid_{\theta=\theta_0} = E_{\theta_0}\{F_\beta S_{\theta_0}(\boldsymbol{L})\} \tag{A.8.11}$$

The proof of Lemma 2 is based on the chain rule for conditional expectations as well which is lengthy but straightforward. Thus we omit it here.

According to the semiparametric efficiency theory
[Begun et al. (1983), Newey (1990)], if β is a pathwise differentiable parameter, the asymptotic variance bound for regular consistent estimators of β is $E(\delta^2)$, where $\delta=\Pi(F_\beta\,|\Phi)$, the projection of F_β onto the tangent set Φ.

Consider the projection of U_1,U_0,V_1, V_0 and Q to the orthogonal subspaces of Φ:

$$\Pi(U_1\mid\Phi_1)=\Pi(U_0\mid\Phi_1)=\Pi(V_1\mid\Phi_1)=\Pi(V_0\mid\Phi_1)=0$$

$$\Pi(Q\mid\Phi_1)=Q$$

$$\Pi(U_1\mid\Phi_2)=\frac{Z(1-M)\{Y-\gamma_1(\boldsymbol{W},\boldsymbol{X})\}}{p(\boldsymbol{X})\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}+\frac{Z\{\gamma_1(\boldsymbol{W},\boldsymbol{X})-\beta_1(\boldsymbol{X})\}}{p(\boldsymbol{X})}$$

$$\Pi(U_0\mid\Phi_2)=\Pi(V_1\mid\Phi_2)=\Pi(V_0\mid\Phi_2)=\Pi(Q\mid\Phi_2)=\Pi(U_1\mid\Phi_3)=0$$

$$
\begin{aligned}
\Pi(U_0 \mid \Phi_3) &= (1-Z)(1-M)\frac{Y-\gamma_0(\boldsymbol{W},\boldsymbol{X})}{\{1-p(\boldsymbol{X})\}\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}} \\
&\quad +(1-Z)\frac{\gamma_0(\boldsymbol{W},\boldsymbol{X})-\beta_0(\boldsymbol{X})}{1-p(\boldsymbol{X})} \\
\Pi(V_1 \mid \Phi_3) &= \Pi(V_0 \mid \Phi_3) = \Pi(Q \mid \Phi_3) = 0 \\
\Pi(U_1 \mid \Phi_4) &= \frac{q(\boldsymbol{W},\boldsymbol{X},Z)-M}{\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}p(\boldsymbol{X})} E\left[Z\{Y-\beta_1(\boldsymbol{X})\}\middle|\, \boldsymbol{W},\boldsymbol{X}\right] \\
\Pi(U_0 \mid \Phi_4) &= \frac{q(\boldsymbol{W},\boldsymbol{X},Z)-M}{\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}\{1-p(\boldsymbol{X})\}} E\left[(1-Z)\{Y-\beta_0(\boldsymbol{X})\}\middle|\, \boldsymbol{W},\boldsymbol{X}\right] \\
\Pi(V_1 \mid \Phi_4) &= \frac{q(\boldsymbol{W},\boldsymbol{X},Z)-M}{\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}p(\boldsymbol{X})} E\left[Z\{\beta_1(\boldsymbol{X})-\gamma_1(\boldsymbol{W},\boldsymbol{X})\}\middle|\, \boldsymbol{W},\boldsymbol{X}\right] \\
\Pi(V_0 \mid \Phi_4) &= \frac{q(\boldsymbol{W},\boldsymbol{X},Z)-M}{\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}\{1-p(\boldsymbol{X})\}} E\left\{(1-Z)\beta_0(\boldsymbol{X})\middle|\, \boldsymbol{W},\boldsymbol{X}\right\} \\
&\quad -\frac{q(\boldsymbol{W},\boldsymbol{X},Z)-M}{\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}\{1-p(\boldsymbol{X})\}} E\left\{(1-Z)\gamma_0(\boldsymbol{W},\boldsymbol{X})\middle|\, \boldsymbol{W},\boldsymbol{X}\right\} \\
\Pi(Q \mid \Phi_4) &= 0 \\
\Pi(F_\beta \mid \Phi_5) &= 0
\end{aligned}
\tag{A.8.12}
$$

Also note that

$$
\begin{aligned}
E\left[\frac{Z}{p(\boldsymbol{X})}\{Y-\gamma_1(\boldsymbol{W},\boldsymbol{X})\}\middle|\, \boldsymbol{W},\boldsymbol{X}\right] &= 0 \\
E\left[\frac{1-Z}{1-p(\boldsymbol{X})}\{Y-\gamma_0(\boldsymbol{W},\boldsymbol{X})\}\middle|\, \boldsymbol{W},\boldsymbol{X}\right] &= 0
\end{aligned}
\tag{A.8.13}
$$

By Lemma 1 and (A.8.12)-(A.8.13), the asymptotic variance bound of regular treatment effect estimators is $E(\delta^2)$, where δ is the efficient influence function given in (4.3.5).

The above efficiency bound is derived under the nonparametric specification of $f(\boldsymbol{L};\eta)$. For alternative specifications, the derivation of efficiency bound is similar with the same F_β, although the tangent set $\tilde{\Phi}$ corresponding to $f(\boldsymbol{L};\eta)$ will be different. Similar to the result from lemma 1, $\tilde{\Phi}$ can also be expressed as the direct sum of orthogonal linear subspaces $\tilde{\Phi}_k, k = 1\ldots5$, which correspond to the specified models for $\boldsymbol{X}$, $(\boldsymbol{W}^1, Y^1)|\boldsymbol{X}$, $(\boldsymbol{W}^0, Y^0)|\boldsymbol{X}$, $M|\boldsymbol{W},\boldsymbol{X},Z$ and $Z|\boldsymbol{X}$ respectively. In general, as the specified component models are subsets of the corresponding nonparametric models previously discussed, we have $\tilde{\Phi}_k \subseteq \Phi_k$.

Recall that in any cases, model for the joint distribution of $\boldsymbol{W}$ and $\boldsymbol{X}$ is nonparametric in $f(\boldsymbol{L};\eta)$. As a consequence the corresponding tangent set always consists of three orthogonal linear subspaces Φ_1, Φ_2 and Φ_3, regardless of the model specifications for $Y^j \mid \boldsymbol{W}, \boldsymbol{X}, j = 0, 1$. For alternative specifications for the models of $p(\boldsymbol{x})$ and $q(\boldsymbol{W}, \boldsymbol{X}, Z)$, the corresponding linear subspaces in the tangent set will generally become subsets of Φ_4 and Φ_5. However note that $\Pi(F_\beta \mid \Phi_5) = 0$, so for any subset Φ_5' of Φ_5, $\Pi(F_\beta \mid \Phi_5') = 0$. Thus model specifications of $p(\boldsymbol{x})$ does not change the projection of F_β onto the tangent set. Therefore, only the change in model specifications for $q(\boldsymbol{W}, \boldsymbol{X}, Z)$ could potentially change the efficiency bound.

To derive efficiency bounds in alternative semiparametric models, we list all possible specifications of $q(\boldsymbol{W}, \boldsymbol{X}, Z)$, and discuss the structures of corresponding tangent sets:

1. Nonparametric $q(\boldsymbol{W}, \boldsymbol{X}, Z)$: corresponding linear subspace in the tangent set is Φ_4. Therefore the structure of tangent set is the same as in the case of completely nonparametric $f(\boldsymbol{L};\eta)$, except for the linear subspace corresponds to $p(\boldsymbol{x})$. Hence the projection of F_β onto the tangent set will be equal to (4.3.5).

2. Semiparametric $q(\boldsymbol{W}, \boldsymbol{X}, Z)$: Similar to the argument for Φ_1, Φ_2 and Φ_3, the corresponding linear subspace in the tangent set is still Φ_4. Thus the projection of F_β onto the tangent set will be equal to (4.3.5).

3. Parametric $q(\boldsymbol{W}, \boldsymbol{X}, Z)$: The case of parametrically modeled $q(\boldsymbol{W}, \boldsymbol{X}, Z)$ needs special treatment due to the special structure of the corresponding linear subspaces. For parametrically modeled $q(\boldsymbol{w}, \boldsymbol{x}, z; \eta_1)$ with $\dim(\eta_1) =$

d, the tangent set for the semiparametric model $f(\boldsymbol{L};\eta)$ is:

$$\begin{aligned}\Phi = \{&Z(1-M)r_1(\boldsymbol{W},Y \mid \boldsymbol{X}) + (1-Z)(1-M)r_0(\boldsymbol{W},Y \mid \boldsymbol{X}) \\ &+ ZMs_1(\boldsymbol{W} \mid \boldsymbol{X}) + (1-Z)Ms_0(\boldsymbol{W} \mid \boldsymbol{X}) + t(\boldsymbol{X}) \\ &+ [M - q(\boldsymbol{W},\boldsymbol{X},Z)]\tilde{a}(\boldsymbol{W},\boldsymbol{X},Z) + [Z - p(\boldsymbol{X})]b(\boldsymbol{X})\} \quad \text{(A.8.14)}\end{aligned}$$

$b(\boldsymbol{x})$, $r_i(\boldsymbol{w},y \mid \boldsymbol{X})$ and $s_i(y \mid \boldsymbol{X}), i = 0,1$ are defined in (A.8.4), and

$$\tilde{a}(\boldsymbol{w},\boldsymbol{x},z) \in \left\{ \frac{A_q^T \tilde{\boldsymbol{q}}(\boldsymbol{w},\boldsymbol{x},z)}{q(\boldsymbol{W},\boldsymbol{X},Z)\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}, A_q \in R^{d\times 1} \right\}$$

where $\tilde{\boldsymbol{q}}(\boldsymbol{w},\boldsymbol{x},z)$ is a function of $\boldsymbol{w}$, $\boldsymbol{x}$ and z that takes value in $R^{d\times 1}$:

$$\tilde{\boldsymbol{q}}(\boldsymbol{w},\boldsymbol{x},z) = \frac{\partial}{\partial \eta_1} q(\boldsymbol{w},\boldsymbol{x},z;\eta_1) \mid_{\eta_1=\eta_1^0} \quad \text{(A.8.15)}$$

The tangent set can be expressed as the direct sum of orthogonal subspaces:

$$\tilde{\Phi} = \Phi_1 \oplus \Phi_2 \oplus \Phi_3 \oplus \tilde{\Phi}_4 \oplus \Phi_5$$

where Φ_1, Φ_2, Φ_3, Φ_5 are defined as in Lemma 1, and
$\tilde{\Phi}_4 = \{[M - q(\boldsymbol{W},\boldsymbol{X},Z)]\tilde{a}(\boldsymbol{W},\boldsymbol{X},Z)\}$.

The efficient influence function is derived by projecting F_β from (A.8.11) onto the above orthogonal subspaces. Since only subspace $\tilde{\Phi}_4$ is different to the case of completely nonparametric $f(\boldsymbol{L};\eta)$, we have:

$$\begin{aligned}
\Pi(U_1 \mid \tilde{\Phi}_4) &= E\left[\frac{\tilde{\boldsymbol{q}}^T(\boldsymbol{W}^1,\boldsymbol{X},Z)}{q(\boldsymbol{W}^1,\boldsymbol{X},Z)}\{Y^1 - \beta_1(\boldsymbol{X})\}\right]\Delta_q^{-1}\boldsymbol{\Phi}_q \\
\Pi(U_0 \mid \tilde{\Phi}_4) &= E\left[\frac{\tilde{\boldsymbol{q}}^T(\boldsymbol{W}^0,\boldsymbol{X},Z)}{q(\boldsymbol{W}^0,\boldsymbol{X},Z)}\{Y^0 - \beta_0(\boldsymbol{X})\}\right]\Delta_q^{-1}\boldsymbol{\Phi}_q \\
\Pi(V_1 \mid \tilde{\Phi}_4) &= E\left[\frac{\tilde{\boldsymbol{q}}^T(\boldsymbol{W}^1,\boldsymbol{X},Z)}{q(\boldsymbol{W}^1,\boldsymbol{X},Z)}\{\beta_1(\boldsymbol{X}) - \gamma_1(\boldsymbol{W}^1,\boldsymbol{X})\}\right]\Delta_q^{-1}\boldsymbol{\Phi}_q \\
\Pi(V_0 \mid \tilde{\Phi}_4) &= E\left[\frac{\tilde{\boldsymbol{q}}^T(\boldsymbol{W}^0,\boldsymbol{X},Z)}{q(\boldsymbol{W}^0,\boldsymbol{X},Z)}\{\beta_0(\boldsymbol{X}) - \gamma_0(\boldsymbol{W}^0,\boldsymbol{X})\}\right]\Delta_q^{-1}\boldsymbol{\Phi}_q \\
\Pi(Q \mid \tilde{\Phi}_4) &= 0 \quad \text{(A.8.16)}
\end{aligned}$$

where

$$\begin{aligned}
\Delta_q &= E\left[\frac{\tilde{\boldsymbol{q}}(\boldsymbol{W}^0,\boldsymbol{X},Z)\tilde{\boldsymbol{q}}^T(\boldsymbol{W}^0,\boldsymbol{X},Z)}{q(\boldsymbol{W},\boldsymbol{X},Z)\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}\right] \\
\boldsymbol{\Phi}_q &= \frac{M - q(\boldsymbol{W},\boldsymbol{X},Z)}{q(\boldsymbol{W},\boldsymbol{X},Z)\{1-q(\boldsymbol{W},\boldsymbol{X},Z)\}}\tilde{\boldsymbol{q}}(\boldsymbol{W}^0,\boldsymbol{X},Z)
\end{aligned}$$

By (A.8.16) and the definitions of $\gamma_j(\boldsymbol{w}, \boldsymbol{x})$, we see that $\Pi(F_\beta \mid \tilde{\Phi}_4) = 0$. Together with (A.8.12), (A.8.12), (A.8.12) and (A.8.12), we see that the efficient influence function $\tilde{\delta}$ is as given in (4.3.5), and the asymptotic variance bound is $E(\delta^2)$.

4. $q(\boldsymbol{W}, \boldsymbol{X}, Z)$ given: In the special case where $q(\boldsymbol{W}, \boldsymbol{X}, Z)$ is a known function, the corresponding linear subspace is $\{0\}$. Therefore the projection of F_β onto the tangent set is the same as the case of parametric $q(\boldsymbol{W}, \boldsymbol{X}, Z)$. Thus δ is as given in (4.3.5).

Together, they prove theorem 4.3.1.

A.9 Proof of Theorem 4.4.2

We replace the estimated components of the estimators with their corresponding true functionals when the models are correctly specified. We use superscript * to denote the function to which the corresponding estimators will converge when the model is mis-specified. By the chain rule for conditional expectations, we have:

- When models for $p(\boldsymbol{x})$ and $q(\boldsymbol{w}, \boldsymbol{x}, z)$ are correctly specified,

$$
\begin{aligned}
\hat{\beta}_{QR} \quad = \quad & \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i(1-M_i)\{Y_i^1 - \hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{p(\boldsymbol{X}_i)\{1-q(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}}\right. \\
& -\frac{(1-Z_i)(1-M_i)\{Y_i^0 - \hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\{1-p(\boldsymbol{X}_i)\}\{1-q(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}} \\
& +\frac{Z_i}{p(\boldsymbol{X}_i)}\{\hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_1^{(n)}(\boldsymbol{X}_i)\} + \hat{\beta}_1^{(n)}(\boldsymbol{X}_i) \\
& \left. -\frac{1-Z_i}{1-p(\boldsymbol{X}_i)}\{\hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\} - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\right\} + o_p(1) \\
\xrightarrow{p} \quad & E\left\{\frac{Z(1-M)\{Y^1 - \gamma_1^*(\boldsymbol{W}, \boldsymbol{X})\}}{p(\boldsymbol{X})\{1-q(\boldsymbol{W}, \boldsymbol{X}, Z)\}} - \frac{(1-Z)(1-M)\{Y^0 - \gamma_0^*(\boldsymbol{W}, \boldsymbol{X})\}}{\{1-p(\boldsymbol{X})\}\{1-q(\boldsymbol{W}, \boldsymbol{X}, Z)\}}\right. \\
& +\frac{Z}{p(\boldsymbol{X})}\{\gamma_1^*(\boldsymbol{W}, \boldsymbol{X}) - \beta_1^*(\boldsymbol{X})\} + \beta_1^*(\boldsymbol{X}) \\
& \left. -\frac{1-Z}{1-p(\boldsymbol{X})}\{\gamma_0^*(\boldsymbol{W}, \boldsymbol{X}) - \beta_0^*(\boldsymbol{X})\} - \beta_0^*(\boldsymbol{X})\right\} \\
= \quad & E(Y^1 - Y^0) = \beta.
\end{aligned}
$$

- When models for $p(\boldsymbol{x})$ and $\gamma_j(\boldsymbol{w}, \boldsymbol{x}), j = 0, 1$ are correctly specified,

$$
\begin{aligned}
\hat{\beta}_{QR} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i(1-M_i)\{Y_i^1 - \gamma_1(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{p(\boldsymbol{X}_i)\{1-\hat{q}_n(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}}\right. \\
&\quad -\frac{(1-Z_i)(1-M_i)\{Y_i^0 - \gamma_0(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\{1-p(\boldsymbol{X}_i)\}\{1-\hat{q}_n(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}} \\
&\quad +\frac{Z_i}{p(\boldsymbol{X}_i)}\{\gamma_1(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_1^{(n)}(\boldsymbol{X}_i)\} + \hat{\beta}_1^{(n)}(\boldsymbol{X}_i) \\
&\quad \left. -\frac{1-Z_i}{1-p(\boldsymbol{X}_i)}\{\gamma_0(\boldsymbol{W}_i, \boldsymbol{X}_i) - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\} - \hat{\beta}_0^{(n)}(\boldsymbol{X}_i)\right\} + o_p(1) \\
&\xrightarrow{p} E\left\{\frac{Z(1-M)\{Y^1 - \gamma_1(\boldsymbol{W}, \boldsymbol{X})\}}{p(\boldsymbol{X})\{1-q^*(\boldsymbol{W}, \boldsymbol{X}, Z)\}} - \frac{(1-Z)(1-M)\{Y^0 - \gamma_0(\boldsymbol{W}, \boldsymbol{X})\}}{\{1-p(\boldsymbol{X})\}\{1-q^*(\boldsymbol{W}, \boldsymbol{X}, Z)\}}\right. \\
&\quad +\frac{Z}{p(\boldsymbol{X})}\{\gamma_1(\boldsymbol{W}, \boldsymbol{X}) - \beta_1^*(\boldsymbol{X})\} + \beta_1^*(\boldsymbol{X}) \\
&\quad \left. -\frac{1-Z}{1-p(\boldsymbol{X})}\{\gamma_0(\boldsymbol{W}, \boldsymbol{X}) - \beta_0^*(\boldsymbol{X})\} - \beta_0^*(\boldsymbol{X})\right\} \\
&= E(\gamma_1(\boldsymbol{W}, \boldsymbol{X}) - \gamma_0(\boldsymbol{W}, \boldsymbol{X})) = \beta.
\end{aligned}
$$

- When models for $q(\boldsymbol{w}, \boldsymbol{x}, z)$ and $\beta_j(\boldsymbol{x}), j = 0, 1$ are correctly specified,

$$
\begin{aligned}
\hat{\beta}_{QR} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i(1-M_i)\{Y_i^1 - \hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\hat{p}_n(\boldsymbol{X}_i)\{1-q(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}}\right. \\
&\quad -\frac{(1-Z_i)(1-M_i)\{Y_i^0 - \hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i)\}}{\{1-\hat{p}_n(\boldsymbol{X}_i)\}\{1-q(\boldsymbol{W}_i, \boldsymbol{X}_i, Z_i)\}} \\
&\quad +\frac{Z_i}{\hat{p}_n(\boldsymbol{X}_i)}\{\hat{\gamma}_1^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \beta_1(\boldsymbol{X}_i)\} + \beta_1(\boldsymbol{X}_i) \\
&\quad \left. -\frac{1-Z_i}{1-\hat{p}_n(\boldsymbol{X}_i)}\{\hat{\gamma}_0^{(n)}(\boldsymbol{W}_i, \boldsymbol{X}_i) - \beta_0(\boldsymbol{X}_i)\} - \beta_0(\boldsymbol{X}_i)\right\} + o_p(1) \\
&\xrightarrow{p} E\left\{\frac{Z(1-M)\{Y^1 - \gamma_1^*(\boldsymbol{W}, \boldsymbol{X})\}}{p^*(\boldsymbol{X})\{1-q(\boldsymbol{W}, \boldsymbol{X}, Z)\}} - \frac{(1-Z)(1-M)\{Y^0 - \gamma_0^*(\boldsymbol{W}, \boldsymbol{X})\}}{\{1-p^*(\boldsymbol{X})\}\{1-q(\boldsymbol{W}, \boldsymbol{X}, Z)\}}\right. \\
&\quad +\frac{Z}{p^*(\boldsymbol{X})}\{\gamma_1^*(\boldsymbol{W}, \boldsymbol{X}) - \beta_1(\boldsymbol{X})\} + \beta_1(\boldsymbol{X}) \\
&\quad \left. -\frac{1-Z}{1-p^*(\boldsymbol{X})}\{\gamma_0^*(\boldsymbol{W}, \boldsymbol{X}) - \beta_0(\boldsymbol{X})\} - \beta_0(\boldsymbol{X})\right\} \\
&= E(\beta_1(\boldsymbol{X}) - \beta_0(\boldsymbol{X})) = \beta.
\end{aligned}
$$

- When models for $\gamma_j(\boldsymbol{w}, \boldsymbol{x}), j = 0, 1$ and $\beta_j(\boldsymbol{x}), j = 0, 1$ are correctly specified,

$$
\begin{aligned}
\hat{\beta}_{QR} \quad = \quad & \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i(1-M_i)\{Y_i^1-\gamma_1(\boldsymbol{W}_i,\boldsymbol{X}_i)\}}{\hat{p}_n(\boldsymbol{X}_i)\{1-\hat{q}_n(\boldsymbol{W}_i,\boldsymbol{X}_i,Z_i)\}}\right. \\
& -\frac{(1-Z_i)(1-M_i)\{Y_i^0-\gamma_0(\boldsymbol{W}_i,\boldsymbol{X}_i)\}}{\{1-\hat{p}_n(\boldsymbol{X}_i)\}\{1-\hat{q}_n(\boldsymbol{W}_i,\boldsymbol{X}_i,Z_i)\}} \\
& +\frac{Z_i}{\hat{p}_n(\boldsymbol{X}_i)}\{\gamma_1(\boldsymbol{W}_i,\boldsymbol{X}_i)-\beta_1(\boldsymbol{X}_i)\}+\beta_1(\boldsymbol{X}_i) \\
& \left.-\frac{1-Z_i}{1-\hat{p}_n(\boldsymbol{X}_i)}\{\gamma_0(\boldsymbol{W}_i,\boldsymbol{X}_i)-\beta_0(\boldsymbol{X}_i)\}-\beta_0(\boldsymbol{X}_i)\right\}+o_p(1) \\
\xrightarrow{p} \quad & E\left\{\frac{Z(1-M)\{Y^1-\gamma_1(\boldsymbol{W},\boldsymbol{X})\}}{p^*(\boldsymbol{X})\{1-q^*(\boldsymbol{W},\boldsymbol{X},Z)\}}-\frac{(1-Z)(1-M)\{Y^0-\gamma_0(\boldsymbol{W},\boldsymbol{X})\}}{\{1-p^*(\boldsymbol{X})\}\{1-q^*(\boldsymbol{W},\boldsymbol{X},Z)\}}\right. \\
& +\frac{Z}{p^*(\boldsymbol{X})}\{\gamma_1(\boldsymbol{W},\boldsymbol{X})-\beta_1(\boldsymbol{X})\}+\beta_1(\boldsymbol{X}) \\
& \left.-\frac{1-Z}{1-p^*(\boldsymbol{X})}\{\gamma_0(\boldsymbol{W},\boldsymbol{X})-\beta_0(\boldsymbol{X})\}-\beta_0(\boldsymbol{X})\right\} \\
= \quad & E(\beta_1(\boldsymbol{X})-\beta_0(\boldsymbol{X}))=\beta.
\end{aligned}
$$

References

[Ahmad, Leelahanon and Li (2005)] Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model. *The Annals of Statistics*. **33**, 258–283.

[Anstrom and Tsiatis(2001)] Anstrom, K. J. and Tsiatis, A. A. (2001). Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data. *Biometrics* **57,** 1207–1218.

[Bang and Rubins (2005)] Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973.

[Begun et al. (1983)] Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11,** 432–452.

[Bhat (1974)] B.R. Bhat (1974), On the Method of Maximum-Likelihood for Dependent Observations. *Journal of the Royal Statistical Society. Series B*, **36**, 48–53.

[Bickel et al. (1993)] Bickel, P., Klaassen, A., Ritov, Y. and Wellner, J. (1993). *Efficient and adaptive inference in semi-parametric models.* Johns Hopkins University Press, Baltimore.

[Boente, He and Zhou (2006)] Boente, G.,He, X. and Zhou, J. (2006), Robust Estimates in Generalized Partially Linear Models. *The Annals of Statistics*, **34**, 2856–2878.

[Bondell and Reich (2008)] Bondell, H and Reich, B. (2008), Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*. **64**, 115–123.

[Breiman et al. (1984)] Breiman, L., Friedman, J., Olshen, R. and Stone, C.(1984), Classification and Regression Trees. *Wadsworth International Group*.

[Cai, Fan and Li (2000)] Cai, Z.,Fan, J. and Li, R. (2000). Effcient estimation and inferences for varying-coeffcient models. *Journal of the American Statistical Association*, **95** 888–902.

[Carroll et al. (1997)] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997) Generalized Partially Linear Single-Index Models, *Journal of the American Statistical Association*. **92** 477–489.

[Cardoso, Koerner and Kubanek (1998)] Cardoso, M., Koerner, K., and Kubanek, B. (1998). Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: Preliminary results. *Transfusion*, **38**, 905–907.

[Chaussabel et al. (2008)] Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I. et al. (2008), A Modular Analysis Framwork for Blood Genomic Studies: Application to Systemic Lupus Erythematosus. *Immunity*, **29**, 150–164.

[Chen, Tebbs and Bilder (2009)] Chen, P., Tebbs, J., Bilder, C. (2009), Group Testing Regression Models with Fixed and Random Effects. *Biometrics*, **65**, 1270–1278.

[Davidian et al. (2005)] Davidian, M., Tsiatis, A. and Leon, S. (2005). Semiparametric Estimation of Treatment Effect in a Pretest-Posttest Study with Missing Data. *Statistical Science* **20**, 261.

[Dempster, Laird and Rubin (1977)] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1–38.

[Dorfman (1943)] Dorfman, R. (1943) The detection of defective members of large populations. *Annals of Mathematical Statistics*, **14** 436–440.

[Efron et al. (2004)] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.(2004), Least angle regression. *Annals of Statistics*, **32**, 407-499.

[Eisenstein et al. (2001)] Eisenstein, E. E., Shaw, L. K., Anstrom, K. J., Nelson, C. L., Hakim, Z., Hasselblad, V., and Mark, D. B. (2001). Assessing the clinical and economic burden of coronary artery disease: 1986–1998. *Medical Care* **39,** 824–835.

[Engle et al. (1986)] Engle, R.F.,Granger, C.W.J, Rice, J. and Weiss, A. (1986), Semiparametric Estimates of the Relation between Weather and Electricity Sales. *Journal of the American Statistical Association*, **81**, 310–320.

[Fan and Huang (2005)] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying coeffcient partially linear models. *Bernoulli*, **11** 1031–1057.

[Fan and Li (2001)] 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 and Lv (2008)] Fan, J. and Lv, J. (2008), Sure independence screening for ultra-high dimensional feature space. *Journal of Royal Statistical Society B*, **70**, 849–911.

[Fan and Lv (2011)] Fan, J. and Lv, J. (2011), Non-concave penalized likelihood with NP-Dimensionality. *IEEE–Information Theory*, **57**, 5467–5484.

[Fan and Lv (2009)] Lv, J. and Fan, Y. (2009), A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, **37**, 3498–3528.

[Frangakis and Rubin (2002)] Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

[Frank and Friedman (1993)] Frank, I.E. and Friedman, J.H. (1993), A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

[Gastwirth and Hammick (1989)] Gastwirth, J.L., Hammick, P.A. (1989), Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *Journal of Statistical Planning and Inference*, **22**, 15-27.

[Gu (2002)] Gu, C. (2002), Smoothing Spline ANOVA Models. *Springer-Verlag*, New York.

[Green (1990)] Green, P.J. (1990), On Use of the EM for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society B*, **52**, 443–452.

[Hahn (1998)] Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66,** 315–332.

[Huang et al (2010)] Huang, J., Breheny, P., Ma, S. and Zhang, C.-H. (2010). The Mnet method for variable selection. *Technical report # 402*, Department of Statistics and Actuarial Science, Univeristy of Iowa.

[Huang et al (2010)] Huang, J., Ma, S., Li, H. and Zhang, C. (2010), The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression. *Technical Report*, **403**.

[Huang et al. (2009)] Huang, J., Ma, S., Xie, H. and Zhang, C. (2009), A group bridge approach for variable selection. *Biometrika*, **96**, 339-355.

[Jia and Yu (2010)] Jia, J. and Yu, B. (2010), On Model Selection Consistency of the Elastic Net When $p \gg n$. *Statistica Sinica*, **20**, 595–611.

[Lam and Fan (2008)] Lam, C., Fan, J. (2008), Profile-Kernel Likelihood Inference With Diverging Number of Parameters. *The Annals of Statistics*, **36**, 2232–2260.

[Li (2009)] Li, M. (2009), Nonparametric and Semiparametric Regression, Missing Data, and Related Algorithms. *Dissertation.* Rutgers University.

[Little and Rubin (2002)] Little, R.J.A. and Rubin, D.B (2002), *Statistical Analysis with missing data*. J.Wiley. New York, 2nd ed.

[Linda et al. (2005)] Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., and Moncada, J. (2005). Utility of pooled urine specimens for detection of Chlamydia trachomatis and Neisseria gonorrhoeae in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology*, **43**, 1674–1677.

[Lunceford and Davidian (2004)] Lunceford, J. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960.

[Ma, Chiou and Wang (2006)] Ma, Y., Chiou, J.-M., Wang, N. (2006), Efficient semiparametric estimator heteroscedastic partially linear models. *Biometrika*, **93**, 75–84.

[Mao and Zhao (2003)] Mao, W., Zhao, L.H. (2003), Free-knot polynomial splines with confidence intervals. *J. R. Statist. Soc. B* **65**, 901–919.

[Mark et al (1994)] Mark, D. B., Nelson, C. L., Califf, et al. (1994). Continuing evolution of therapy for coronary-artery disease–initial results for the ear of coronary angioplasty. *Circulation* **89,** 2015–2025.

[Newey (1990)] Newey, W.K. (1990), Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, **5**, 99–135.

[Newey (1991)] Newey, W.K. (1991), Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, **59**, 1161-1167.

[Pilcher et al. (2005)] Pilcher, C., Fiscus, S., Nguyen, T., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J., McPherson, J., Stalzer, B., Hightow, L., Miller, W., Eron, J., Cohen, M., and Leone, P. (2005), Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine* **352**, 1873–1883.

[Rours et al. (2005)] Rours, G., Verkooyen, R., Willemse, H., van der Zwaan, E., van Belkum, A., de Groot, R., Verbrugh, H., and Ossewaarde, J. (2005). Use of pooled urine samples and automated DNA isolation to achieve improved sensitivity and cost-effectiveness of large-scale testing for Chlamydia trachomatis in pregnant women. *Journal of Clinical Microbiology* **43**, 4684–4690.

[Roderick et al. (1987)] Roderick, J., Little, A. and Rubin, D. (1987). *Statistical analysis with missing data*. J. Wiley.

[Rubin (2004)] Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinanvian Journal of Statistics* **31,** 161–170.

[Ruppert, Wang and Carroll (2003)] Ruppert, D., Wang, M.P., Carroll, R.J. (2003), Semiparametric Regression. *Cambridge University Press*.

[Severini and Wong (1992)] Severini, T.A. and Wong, W.H. (1992), Profile Likelihood and Conditionally Parametric Models. *The Annals of Statistics*, **20**, 1768–1802.

[Shentu (2006)] Shentu, Y. (2006), *Dissertation*. Rutgers University.

[Silverman et al. (1990)] Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. (1990), A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J.R.Statist.Soc.* B, **52**, 271–324.

[Stein (1956)] Stein, C. (1956), Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab*, **1**, 187–195.

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

[U.S. Census Bureau (2010)] U.S. Census Bureau (2010), Income, Poverty, and Health Insurance Coverage in the United States: 2009.

[Wang et al. (2009)] Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009), Hierachically penalized Cox regression with grouped variables. *Biometrika*, **96**, 307-322.

[Wang, Linton and Härdle (2004)] Wang, Q., Linton, O. and Härdle, W. (2004) Semiparametric Regression Analysis With Missing Response at Random. *Journal of the American Statistical Association*, **99**, 334–345.

[Wang, Rotnitzky and Lin (2010)] Wang, L., Rotnitzky, A., Lin, X. (2010) Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, **105**, 1135–1146

[Wu (1983)] Wu, C.F.J (1983), On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95–103.

[Xie, Simpson and Carroll (2008)] Xie, M., Simpson, D.G. and Carroll, R.J (2008), Semiparametric analysis of heterogeneous data using varying scale glm. *Journal of the American Statistical Association*, **482**, 650–660.

[Yuan and Lin (2006)] Yuan, M and Lin, Y. (2006), Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B,* **68**, 49–67.

[Zou and Hastie (2005)] Zou, H and Hastie, T. (2005), Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B,* **67**, 301–320.

[Zhang (2010)] Zhang, C. (2010), Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics,* **38** 894–942.

Vita

Wenqian Qiao

- **EDUCATION**

2007-2012 Ph.D. in Statistics, Rutgers University, New Jersey, USA.

2010-2011 M.S. in Mathematical Finance, Rutgers University, New Jersey, USA.

2003-2007 B.S. in Mathematics, Fudan University, Shanghai, China.

- **PROFESSIONAL EXPERIENCE**

Summer 2011	Intern, Standard and Poor's, New York, NY
Summer 2010	Intern, Sanofi-Aventis, Bridgewater, NJ
Summer 2009	Intern, Johnson and Johnson, Raritan, NJ
09/2008-05/2012	Teaching assistant/Graduate assistant, Department of Statistics and Biostatistics, Rutgers University.

- **PUBLICATIONS**

2010 A Comparison of Several Methods for Analyzing Data from Thorough QT studies, *Journal of Biopharmaceutical Statistics*, 2010, Vol. 20, 604-612, Tian, H., Qiao, W. and Natarajan, J.