2018

Semiparametric Statistical Estimation and Inference with Latent Information

Qianqian Wang
University of South Carolina

Semiparametric Statistical Estimation and Inference with Latent Information

by

Qianqian Wang

Bachelor of Science
Lanzhou University 2011

Master of Science
Michigan State University 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2018

Accepted by:

Yanyuan Ma, Major Professor

John Grego, Committe Member

Yen-Yi Ho, Committe Member

Lianming Wang, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I own a sincere thank you to professor Yanyuan Ma, for being my research advisor and for spending countless of hours discussing research work with me and encouraging me throughout my PhD study. I would like to thank professor John Grego, Yen-Yi Ho and Lianming Wang for being my committee members and for their guidance and support. I would like to thank professor Guangren Yang and Xia Cui for their collaboration of my research project. I would especially like to thank my family and friends for their endless amounts of love and support.

Abstract

In Chapter 1, we predicted disease risk by transformation models in the presence of missing subgroup identifiers. When a discrete covariate defining subgroup membership is missing for some of the subjects in a study, the distribution of the outcome follows a mixture distribution of the subgroup-specific distributions. Taking into account the uncertain distribution of the group membership and the covariates, we model the relation between the disease onset time and the covariates through transformation models in each sub-population, and develop a nonparametric maximum likelihood based estimation implemented through EM algorithm along with its inference procedure. We further propose methods to identify the covariates that have different effects or common effects in distinct populations, which enables parsimonious modeling and better understanding of the difference across populations. The methods are illustrated through extensive simulation studies and a real data example.

In Chapter 2, we discussed a generalized partially linear single index model with measurement error, instruments and binary response. Instrumental variables are important elements in studying many errors-in-variables problems. We use the relation between the unobservable variables and the instruments to devise consistent estimators for partially linear generalized single index models with binary response. We establish the consistency, asymptotic normality of the estimator and illustrate the numerical performance of the method through simulation studies and a data example. Despite the connection to Xu et al. (2015) in its general layout, the mathematical derivations are much more challenging in the context studied here.

In Chapter 3, we investigated the errors in covariates issues in a generalized par-

tially linear model. Different from the usual literature (Ma & Carroll 2006), we consider the case where the measurement error occurs to the covariate that enters the model nonparametrically, while the covariates precisely observed enter the model parametrically. To avoid the deconvolution type operations, which can suffer from very low convergence rate, we use the B-splines representation to approximate the nonparametric function and convert the problem into a parametric form for operational purpose. We then use a parametric working model to replace the distribution of the unobservable variable, and devise an estimating equation to estimate both the model parameters and the functional dependence of the response on the latent variable. The estimation procedure is devised under the functional model framework without assuming any distribution structure of the latent variable. We further derive theories on the large sample properties of our estimator. Numerical simulation studies are carried out to evaluate the finite sample performance, and the practical performance of the method is illustrated through a data example.

Table of Contents

List of Tables

List of Figures

Chapter 1

Predicting disease Risk by Transformation Models in the Presence of Missing Subgroup Identifiers[1]

1.1 Introduction

Biomedical studies can lead to mixture data. When a discrete covariate defining subgroup membership is missing for some of the subjects in a study, the distribution of the outcome is a mixture of the subgroup-specific distributions. One example is the kin-cohort study Wacholder et al. (1998) with the goal of estimating the cumulative risk of disease for mutation carriers Khoury et al. (1993). However, mutation status is only collected in the initial sample of participants, referred as probands, not in their relatives. For example, genetic mutation status is not available for deceased relatives or those who have not undergone genetic testing due to resource constraints. The disease phenotype information for such relatives is available from other sources, such as interviewing the proband in a family Marder et al. (2003). For a late-onset disease, such as Parkinson's disease (PD), parents of study participants are often deceased. Therefore even though age-at-onset of PD is provided by a family member, no genotyping can be performed on deceased parents. When estimating the disease risk distribution for mutation carriers and non-carriers using these relatives' disease onset information, the unknown mutation status needs to be accounted for by using

[1] Wang Q., Ma, Y., and Wang, Y. 2017. Statistica Sinica. 27, 4, p. 1857-1878 22 p. Reprinted here with permission of publisher.

the distribution of mutation status in such relatives as estimated from living relatives who provide blood sample Wang et al. (2012), Ma & Wang (2014).

We consider estimating the subgroup-specific distribution for outcomes that are subject to censoring and with missing subgroup identifiers. The nonparametric models in Wacholder et al. (1998), Wang et al. (2012), and Ma & Wang (2014) do not include any covariates other than the mutation status. We consider how to include covariates that can have identical or different effects across subgroups. Popular semi-parametric models for censored outcomes, such as the Cox proportional hazards model, accelerated failure time model, and transformation model have been studied extensively in the literature, but less so in a mixture data setting. Recently, Altstein & Li (2013) proposed a latent subgroup analysis for a semiparametric accelerated failure time model in a clinical trials setting. Our work differs from Altstein & Li (2013) in that the distribution of the subgroup identifiers is available in our problem, and we assume a semiparametric transformation model in each subgroup. A transformation model is applied to analyze neurological disorder data (e.g, Huntington's disease [HD] as in our motivating study) due to its useful biological and clinical interpretations; see for example Zhang et al. (2012).

We propose a semiparametric transformation model for mixture data. Compared to parametric transformation model in the literature Zhang et al. (2012), we allow for greater flexibility to account for subgroup heterogeneity. This is achieved in our model through characterizing the outcome in each subpopulation using a different distribution, indexed by both parameters and error distributions. They can also have both as shared covariate effect and/or a subgroup-specific covariate effect. In addition, we assume an unknown transformation to avoid the difficulty of specifying a parametric transformation. When assuming a homogeneous covariate effect, we account for a missing population identifier by taking advantage of the distribution of the mixing proportion and using a weighted least-square type estimator, which greatly simplifies the procedure. When we assume a subgroup-specific covariate effect, the weighted least-square estimator no longer applies, and we use the EM algorithm. We have performed extensive simulation studies to examine performance of the proposed approach and applied it to estimating the survival function for HD mutation carriers in a large genetic epidemiology study Dorsey & The Huntington Study Group COHORT Investigators (2012).

1.2 Modeling, Estimation, and Asymptotic Properties

Assume there are n observations from p populations. Here p is usually determined by the research purpose. For genetic studies, populations are defined by mutation carrier status. Throughout, we assume p is pre-determined. Denote the data from the ith observation as $\mathbf{O}_i = (\mathbf{q}_i, \mathbf{x}_i, \mathbf{z}_i, y_i, \delta_i)$, where $\mathbf{q}_i$ is a length p vector, with the jth entry q_{ij} being the probability that the ith observation is randomly sampled from the jth population. We also allow a subject's population membership to be known by allowing $\mathbf{q}_i$ to be a vector with 1 in one component and zero in all others. Let t_i be the time to event and c_i be the censoring time, $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i \le c_i)$. Let $\mathbf{x}_i$ denote the covariate vector that has a common effect on the event time across different populations, while $\mathbf{z}_i$ denotes the covariate vector that has a different effect in different populations. For simplicity, we sort the data so that $y_i \le y_k$ for all $i < k$.

1.2.1 Model

For the jth population, the linear transformation model we propose has the form

$$H(T) = -\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\alpha}_j + \epsilon_j. \tag{1.1}$$

Here H is an unknown, monotonically increasing function and, without loss of generality, we assume $H(0) = -\infty$. We assume ϵ_j is independent of $\mathbf{X}$, $\mathbf{Z}$, and has a known population-specific distribution $f_j(\epsilon_j)$. Here, in each population, this is a classical linear transformation model, in which the baseline population distribution can be heterogeneous due to the different choices of f_j. Selection of f_j for each population can be based on scientific or biological knowledge of a particular population. The covariate effect is also allowed to vary, reflected in the population-specific $\boldsymbol{\alpha}_j$. By including the term $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, we also allow the possibility that some covariates have a homogeneous effect across populations. We develop a test to assess whether a covariate exhibits evidence of deviation from a homogeneous effect model.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_p^{\mathrm{T}})^{\mathrm{T}}$, $\Phi(t) = \exp\{H(t)\}$, and $\phi(t) = \exp\{H(t)\}h(t)$. The conditional distribution function of the ith relative from (1.1) is then

$$
\begin{aligned}
& f(y_i, \delta_i \mid \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}, \Phi, \phi) \\
= & \left[h(y_i) \sum_{j=1}^{n} q_{ij} f_j\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} \right]^{\delta_i} \\
& \times \left[1 - \sum_{j=1}^{n} q_{ij} F_j\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} \right]^{1-\delta_i} \\
= & \phi(y_i)^{\delta_i} \Psi(O_i; \boldsymbol{\theta}, \Phi),
\end{aligned}
$$

where Φ is a function that depends only on $\boldsymbol{\theta}$ and Φ, but not on ϕ. The model can not be viewed as a transformation model, hence existing estimation procedures do not apply. To ensure identifiability, we require that the $\mathbf{q}_i$ variable takes m different vector values, denoted $\mathbf{u}_1, \ldots, \mathbf{u}_m$, so that the matrix $(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ has rank p. We point out that the identifiability here excludes any permutation. This identifiability is stronger than that up to a permutation in most classical mixture models Holzmann et al. (2006). We can achieve the stronger form of identifiability because the mixture probabilities, while different for different observations, are known.

1.2.2 Estimation

We propose a nonparametric maximum likelihood estimator (NPMLE) to estimate $\boldsymbol{\theta}$ and $\Phi(\cdot)$. Specifically, we obtain $\widehat{\boldsymbol{\theta}}$ and $\widehat{H} = \log(\widehat{\Phi})$ through maximizing

$$l(\boldsymbol{\theta}, \Phi) = \sum_{i=1}^{n} \delta_i \log\{\phi(y_i)\} + \sum_{i=1}^{n} \log\{\Psi(O_i; \boldsymbol{\theta}, \Phi)\}$$

with respect to $\boldsymbol{\theta}$ and Φ, where we restrict Φ, hence H, to be a piecewise constant non-decreasing function with non-negative jumps only at the observed event times. Following existing literature Wacholder et al. (1998), Wang et al. (2012), We exclude the probands from the analysis sample and the likelihood to protect against potential ascertainment bias from unknown sources that may be difficult to adjust (e.g., convenience sample of patients visiting a clinic). Given the mutation carrier status, we also assume the relatives' phenotypes are conditionally independent of probands' phenotypes, which is an assumption satisfied by a monogenic disorder with a known genetic cause controlled in the model (e.g., HD in our application).

Although conceptually simple, the computation of NPMLE is not straightforward because the maximization is with respect to not only $\boldsymbol{\gamma}$, but also $\Phi(\cdot)$ at all the y_i's that are not censored. As sample size increases, the potential number of parameters increases as well, hence the computational problem does not simplify in the asymptotic sense. To overcome the computational difficulty, we use an EM algorithm. To this end, we first use Laplace transformation in each population to obtain

$$1 - F_j(x) = \int_0^{\infty} \exp(-r_j e^x)\psi_j(r_j) dr_j,$$

where $\psi_j(\cdot)$ is the inverse Laplace transformation of $1 - F_j(x)$ as a function of e^x, consequently

$$
\begin{aligned}
& 1-\sum_{j=1}^{n} q_{ij} F_j\{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} \\
= & \sum_{j=1}^{n} q_{ij}\int_0^{\infty} \exp\{-r_{ij}e^{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\psi_j(r_{ij})dr_{ij} \\
= & \sum_{j=1}^{n} q_{ij}\int_0^{\infty} \exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\psi_j(r_{ij})dr_{ij}
\end{aligned}
$$

and

$$
\begin{aligned}
& h(y_i)\sum_{j=1}^{n} q_{ij} f_j\{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} \\
= & \sum_{j=1}^{n} q_{ij}\int_0^{\infty} \exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\phi(y_i)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)r_{ij}\psi_j(r_{ij})dr_{ij}.
\end{aligned}
$$

The ith observation here is $\mathbf{O}_i$, let $\mathbf{D} = (\mathbf{O}_1, \cdots, \mathbf{O}_n)$. Let $0 < t_1 < \cdots < t_K < \tau$ be the distinct event times, and write the quantities to be estimated as $\boldsymbol{\gamma} = \{\boldsymbol{\theta}^{\mathrm{T}}, H(t_1), \dots H(t_K)\}^{\mathrm{T}}$. The log-likelihood is then $l(\boldsymbol{\gamma}; \mathbf{D}) = \sum_{i=1}^{n} l_i(\boldsymbol{\gamma}; \mathbf{O}_i)$, where

$$
\begin{aligned}
l_i(\boldsymbol{\gamma}; \mathbf{O}_i) \;=\; & \log\sum_{j=1}^{n}\int_0^{\infty}\{\phi(y_i)r_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)\}^{\delta_i} \\
& \times\exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}q_{ij}\psi_j(r_{ij})dr_{ij}.
\end{aligned}
$$

We take advantage of this special data structure and view the population identifiers $\mathbf{G} = (G_1, \dots, G_n)$ and $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_n)$ as the missing variable, where $G_i = I_j$ represents that the ith observation is a random sample from the jth population, and $\mathbf{r}_i = (r_{i1}, \dots, r_{ip})^{\mathrm{T}}$ is the introduced random effects to facilitate computation. Then the complete data loglikelihood is $l(\boldsymbol{\gamma} \mid \mathbf{D}, \mathbf{G}, \mathbf{r}) = \sum_{i=1}^{n} l_i(\boldsymbol{\gamma} \mid O_i, G_i, \mathbf{r}_i)$, where

$$
\begin{aligned}
& l_i(\boldsymbol{\gamma} \mid O_i, G_i = I_j, r_{ij}) \\
= & \log\left[\{\phi(y_i)r_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)\}^{\delta_i}\exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\right] \\
= & \delta_i\log\{\phi(y_i)r_{ij}\}+\delta_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}.
\end{aligned}
$$

This is a Cox model log-likelihood. Thus, in the E-step, we calculate

$$Q(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(u)}, \mathbf{D}) \equiv E_{\boldsymbol{\gamma}^{(u)}}\{l(\boldsymbol{\gamma} \mid \mathbf{D}, \mathbf{G}, \mathbf{r}) \mid \mathbf{D}\} = \sum_{i=1}^{n} \frac{\int \sum_{j=1}^{n} l_i(\boldsymbol{\gamma} \mid O_i, \mathbf{G}_i = I_j, r_{ij}) a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}},$$

where

$$a_{ij}^{(u)} = \{\phi^{(u)}(y_i) r_{ij} \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j^{(u)})\}^{\delta_i} \exp\{-r_{ij} \Phi^{(u)}(y_i) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j^{(u)}}\} q_{ij} \psi_j(r_{ij}).$$

In the M-step, we maximize $Q(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(u)}, \mathbf{D})$ with respect to $\boldsymbol{\gamma}$ subject to the constraints $0 < H(t_1) < \cdots < H(t_K) \leq 1$ to obtain $\boldsymbol{\gamma}^{(u+1)}$. Specifically, taking derivative with respect to $\boldsymbol{\gamma}$, we obtain estimating equations

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} \frac{\int \sum_{j=1}^{n} \{\delta_i \mathbf{x}_i - \mathbf{x}_i r_{ij} \Phi(y_i) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j}\} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \\ &= \sum_{i=1}^{n} \delta_i \mathbf{x}_i - \mathbf{x}_i \Phi(y_i) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \frac{\sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}}. \end{aligned}$$

For $j = 1, \ldots, p$,

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} \frac{\int (\delta_i \mathbf{z}_i - \mathbf{z}_i r_{ij} e^{H(y_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j}) a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \\ &= \sum_{i=1}^{n} \frac{\delta_i \mathbf{z}_i \int a_{ij}^{(u)} dr_{ij} - \mathbf{z}_i \Phi(y_i) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}}. \end{aligned}$$

For $k = 1, \ldots, K$,

$$\begin{aligned} 0 &= \sum_{y_i \geq t_k} \frac{\int \sum_{j=1}^{n} \left\{ \frac{I(y_i = t_k)}{\phi_k} - r_{ij} e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \right\} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \\ &= \frac{1}{\phi_k} - \sum_{y_i \geq t_k} \frac{e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}}. \end{aligned}$$

This yields

$$\phi_k = \left(\sum_{y_i \geq t_k} \frac{e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \right)^{-1},$$

or in general

$$\begin{aligned} \phi(y_k; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \delta_k \left(\sum_{i=1}^{n} \frac{I(y_i \geq y_k) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \right)^{-1} \qquad (1.2) \\ \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{k=1}^{n} I(y_k \leq y_i) \delta_k \left(\sum_{i=1}^{n} \frac{I(y_i \geq y_k) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} \right)^{-1}. \end{aligned}$$

Plugging into the estimating equation for $\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$, we obtain

$$\begin{aligned}
\sum_{i=1}^{n} \delta_i \mathbf{x}_i - \mathbf{x}_i \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) e^{\mathbf{x}_i^{\mathrm{T}} \beta} \frac{\sum_{j=1}^{n} e^{\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} &= \mathbf{0} \qquad (1.3)\\
\sum_{i=1}^{n} \frac{\delta_i \mathbf{z}_i \int a_{ij}^{(u)} dr_{ij} - \mathbf{z}_i \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{n} a_{ij}^{(u)} dr_{ij}} &= \mathbf{0}
\end{aligned}$$

at $j = 1, \ldots, p$.

We solve the estimating equations (1.3) to obtain $\widehat{\boldsymbol{\beta}}^{(u+1)}, \widehat{\boldsymbol{\alpha}}^{(u+1)}, j = 1, \ldots, p$, and then substitute into (1.2) to obtain $\Phi^{(u+1)}(t)$, and hence also $H^{(u+1)}(t) = \log\{\Phi^{(u+1)}(t)\}$. The procedure iterates between the E-step and the M-step until convergence.

We point out that, although the functions $\psi_j(r)$'s are left as unknown, we can still calculate $\int a_{ij}^{(u)} dr_{ij}$ and $\int r_{ij} a_{ij}^{(u)} dr_{ij}$ in the M-step. Specifically,

$$\begin{aligned}
& \int a_{ij}^{(u)} dr_{ij} \\
= \;& q_{ij} \left\{1 - F_j(t)\right\}^{1-\delta_i} \left\{h^{(u)}(y_i) f_j(t)\right\}^{\delta_i} \Big|_{t = H^{(u)}(y_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j^{(u)}}, \\
& \int r_{ij} a_{ij}^{(u)} dr_{ij} \\
= \;& \left\{e^{-t} q_{ij} f_j(t)\right\}^{1-\delta} \left[e^{-t} q_{ij} h^{(u)}(y_i) \{f_j(t) - f_j'(t)\}\right]^{\delta} \Big|_{t = H^{(u)}(y_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j^{(u)}},
\end{aligned}$$

as shown in Appendix B.1, by taking advantage of the Laplace/inverse Laplace transform relation. In fact, even if an explicit form of $\psi_j(r)$ can be obtained, it is not necessary to go through the calculation because $\psi_j(r)$ itself is not needed. Finally, because ψ_j is defined as the inverse Laplace transform of a bounded function, it always exists for any ϵ distribution.

1.2.3 Theoretical properties

Although (1.1) is not a transformation model, under the list of conditions imposed in Appendix B.2, it can be cast into the general framework, Zeng & Lin (2007). To this end, we can verify that our Conditions (a), (b), (c) lead to their conditions (C1), (C2), (C3), respectively. Our Conditions (d) and (e) jointly ensure their conditions

(C4) and (C8). Our Condition (f) leads to their condition (C6), and our Condition (g) leads to their conditions (C5), (C7). These are mild conditions mainly imposing identifiability, sufficient smoothness, and boundedness of various functions; They are usually satisfied in practice. Having verified the regularity conditions C1-C7 of Zeng & Lin (2007), we can use their results to obtain the asymptotic properties of the NPMLE in the linear transformation model in the mixture data setting. We state the results in Theorem 1 and provide the proof in Appendix B.3.

Theorem 1. *Let* $\boldsymbol{\theta}_0, \Phi_0$ *denote the true value of* $\boldsymbol{\theta}, \Phi$*, and write* $\Phi = \{\Phi(t_1), \Phi(t_2), \ldots, \Phi(t_K)\}^{\mathrm{T}}$*. Under conditions (a)-(g) of Appendix B.2,* $\widehat{\boldsymbol{\theta}}, \widehat{\Phi}$ *are consistent, and have the asymptotic property that* $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \widehat{\Phi} - \Phi)$ *converges weakly to a zero mean Gaussian process. Then, for any function* $a_1(s)$ *with bounded total variation and any vector* $\boldsymbol{a}_2$*,* $\sqrt{n} \int a_1(s) d\{\widehat{\Phi}(s) - \Phi(s)\} + \sqrt{n} \boldsymbol{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ *converges to a zero mean normal distribution whose variance can be approximated by*

$$\boldsymbol{v}\{a_1(\cdot), \boldsymbol{a}_2\} \equiv -(\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}}) \left\{ \frac{\partial^2 l(\widehat{\Phi}, \widehat{\boldsymbol{\theta}})}{\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}) \partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}} \right\}^{-1} (\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}})^{\mathrm{T}},$$

where $\boldsymbol{a}_1 = \{a_1(t_1), \ldots, a_1(t_K)\}^{\mathrm{T}}$.

1.2.4 Inference

The main interest is often in the covariate effects described by $\boldsymbol{\theta}$. In such cases, we can perform inference using the results of a profiling procedure: at any $\boldsymbol{\theta}$, we use the same EM algorithm to calculate $\widehat{H}(T, \boldsymbol{\theta})$ except that we hold $\boldsymbol{\theta}$ fixed, and then calculate the information matrix using numerical derivatives. This is a simplification because it bypasses the need to invert a potentially high-dimensional matrix. For example, the $\alpha 100\%$ confidence interval for the jth component of $\boldsymbol{\theta}$, θ_j is

$$
\begin{aligned}
& \widehat{\theta}_j \pm Z_{(1+\alpha)/2} \left[-\sum_{i=1}^{n} \frac{\partial^2 l_i\{\boldsymbol{\theta}, \widehat{H}(t_1, \boldsymbol{\theta}), \ldots, \widehat{H}(t_K, \boldsymbol{\theta})\}}{\partial \theta_j^2} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right]^{-1/2} \\
\approx \quad & \widehat{\theta}_j \pm Z_{(1+\alpha)/2} \left[\sum_{i=1}^{n} \frac{-l_i\{\widehat{\boldsymbol{\theta}} + b\mathbf{e}_j, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j)\}}{b^2} \right. \\
& + \frac{2 l_i\{\widehat{\boldsymbol{\theta}}, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}}), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}})\}}{b^2} \\
& \left. - \frac{l_i\{\widehat{\boldsymbol{\theta}} - b\mathbf{e}_j, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j)\}}{b^2} \right]^{-1/2},
\end{aligned}
$$

where $Z_{(1+\alpha)/2}$ is the $(1+\alpha)/2$ quantile of the standard normal distribution, l_i is the likelihood evaluated at the ith observation, $\mathbf{e}_j$ is the vector with zero components everywhere except the jth component being 1, and b is a small number that facilitates the numerical derivative.

Likewise, for hypothesis testing of the form $H_0 : \boldsymbol{\theta} = \mathbf{c}$, we can construct the test statistic

$$
\begin{aligned}
\mathbf{Z} \quad = \quad & \left[-\sum_{i=1}^{n} \frac{\partial^2 l_i\{\boldsymbol{\theta}, \widehat{H}(t_1, \boldsymbol{\theta}), \ldots, \widehat{H}(t_K, \boldsymbol{\theta})\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right]^{1/2} (\boldsymbol{\theta} - \mathbf{c}) \\
\approx \quad & \left[\left(\sum_{i=1}^{n} \frac{-l_i\{\widehat{\boldsymbol{\theta}} + b\mathbf{e}_j + b\mathbf{e}_k, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j + b\mathbf{e}_k), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j + b\mathbf{e}_k)\}}{4b^2} \right. \right. \\
& + \frac{l_i\{\widehat{\boldsymbol{\theta}} + b\mathbf{e}_j - b\mathbf{e}_k, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j - b\mathbf{e}_k), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j - b\mathbf{e}_k)\}}{4b^2} \\
& + \frac{l_i\{\widehat{\boldsymbol{\theta}} - b\mathbf{e}_j + b\mathbf{e}_k, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j + b\mathbf{e}_k), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j + b\mathbf{e}_k)\}}{4b^2} \\
& \left. \left. - \frac{l_i\{\widehat{\boldsymbol{\theta}} - b\mathbf{e}_j - b\mathbf{e}_k, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j - b\mathbf{e}_k), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j - b\mathbf{e}_k)\}}{4b^2} \right)_{jk} \right]^{1/2} \\
& \times (\boldsymbol{\theta} - \mathbf{c}),
\end{aligned}
$$

and note that $\mathbf{Z}$ is approximately a standard multivariate normal random variable under H_0. Here, we use the notation (A_{jk}) to denote the square matrix $\mathbf{A}$ with size the length of $\boldsymbol{\theta}$ and (j,k) entry A_{jk}.

1.3 Homogeneous and no covariate effect model

When either $\boldsymbol{\beta}$ or $\boldsymbol{\alpha}_j$ does not appear in (1.1), the model is more restrictive and the computation simplifies. If $\boldsymbol{\beta}$ does not appear, then there is no homogeneous covariate effect in the transformation model. In terms of estimation, the procedures follows the same line with some minor simplifications. However, if $\boldsymbol{\alpha}_j$ does not appear, (1.1) greatly simplifies and can be treated quite differently, as we now explain.

The common-effect covariate effect model for the jth population is

$$H(T) = -\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_j,$$

where all the components in the model retain the same interpretation as in (1.1). The implication of the model is that the heterogeneity between subpopulations is due to the different variability of measurement errors, but not the heterogeneous effect of covariates. The conditional distribution is then simplified to

$$\begin{aligned} f(Y, \Delta \mid \mathbf{X}) &= \left[h(y)\sum_{j=1}^{n} q_j f_j\{H(y) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{\delta}\left[1 - \sum_{j=1}^{n} q_j F_j\{H(y) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{1-\delta} \\ &= \left[h(y)\mathbf{q}^{\mathrm{T}}\mathbf{f}\{H(y) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{\delta}\left[1 - \mathbf{q}^{\mathrm{T}}\mathbf{F}\left\{H(y) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right\}\right]^{1-\delta}, \end{aligned}$$

where $\mathbf{f} = (f_1, \ldots, f_p)^{\mathrm{T}}$, $\mathbf{F} = (F_1, \ldots, F_p)^{\mathrm{T}}$, and $h(y) \equiv H'(y)$, because the same transformation H and the same parameter $\boldsymbol{\beta}$ are assumed across all p populations. The population difference is only reflected in the distribution of ϵ_j, which is assumed to be f_j. We can however still use the different f_j's of the model to account for unexplained residual population heterogeneity, for example, different variances.

As before, estimating the distribution in each population is equivalent to estimating H and $\boldsymbol{\beta}$. As the $\mathbf{q}_i$'s have $m \geq p$ different vector values $\mathbf{u}_1, \ldots, \mathbf{u}_m$, assign the n observations to these m groups according to their $\mathbf{q}$ values. Assume there are, respectively, $r_1, \ldots, r_m$ observations in each group. In group k, we can view the model as a transformation model with the same transformation H, the same parameter $\boldsymbol{\beta}$, but a new distribution for ϵ, which has the mixture form $\mathbf{u}_k^{\mathrm{T}}\mathbf{f}(\epsilon)$. Thus, we can use the existing estimation method for transformation models to obtain the estimators of H and $\boldsymbol{\beta}$, using exclusively the kth group data. Denote the resulting estimators as $\widehat{H}_k$ and $\widehat{\boldsymbol{\beta}}_k$. We can then take the weighted average to obtain the final estimator $\widehat{H}(t) = \sum_{k=1}^m w_k(t)\widehat{H}_k(t)$ and $\widehat{\boldsymbol{\beta}} = \sum_{k=1}^m \mathbf{w}_k\widehat{\boldsymbol{\beta}}_k$. To be consistent with the estimation in the general model (1.1), we use the NPMLE proposed by Zeng & Lin (2006). Thus, we obtain $\widehat{\boldsymbol{\beta}}_k$, $\widehat{H}_k$ via maximizing

$$\begin{aligned} l_k(H, \boldsymbol{\beta}) \;&=\; n^{-1}\sum_{i=1}^n I(\mathbf{q}_i = \mathbf{u}_k)\Big(\delta_i \log\Big[h(y_i)\mathbf{u}_k^{\mathrm{T}}\mathbf{f}\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}\Big] \\ &\quad + (1-\delta_i)\log\Big[1 - \mathbf{u}_k^{\mathrm{T}}\mathbf{F}\Big\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\Big\}\Big]\Big) \end{aligned}$$

with respect to $\boldsymbol{\beta}$ and H. Here, we restrict $H(y)$ to be a piecewise constant non-decreasing function with nonnegative jumps only at the y_i's where $\mathbf{q}_i = \mathbf{u}_k$ and $\delta_i = 1$. We write these jump points $t_1, \ldots, t_K$, and write $\mathbf{H}_k = \{H(t_1), \ldots, H(t_K)\}^{\mathrm{T}}$. Zeng & Lin (2006) showed that the resulting $\widehat{\boldsymbol{\beta}}_k, \widehat{H}_k$ are consistent, and that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}, \widehat{H}_k - H)$ converges weakly to a zero mean Gaussian process. Thus, for any function $a_1(s)$ with bounded total variation and any vector $\mathbf{a}_2$, $\sqrt{n}\int a_1(s)d\{\widehat{H}_k(s) - H(s)\} + \sqrt{n}\mathbf{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})$ converges to a zero mean normal distribution whose variance can be approximated by

$$\mathbf{v}_k\{a_1(\cdot), \mathbf{a}_2\} \equiv -(\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}})\left\{\frac{\partial^2 l_k(\widehat{H}_k, \widehat{\boldsymbol{\beta}}_k)}{\partial(\mathbf{H}_k^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})\partial(\mathbf{H}_k^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}}\right\}^{-1}(\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}})^{\mathrm{T}},$$

where $\mathbf{a}_1 = \{a_1(t_1), \ldots, a_1(t_K)\}^{\mathrm{T}}$.

It remains to determine the choice of weights $\mathbf{w}_k$. Because the estimation in different group is based on different subjects, they are independent. Hence the optimal weights are proportional to the inverse of the variance of the estimators. The optimal weights for $\widehat{H}(t)$ are then $w_k(t) = \mathbf{v}_k\{I(s \le t), \mathbf{0}\}^{-1}/[\sum_{k=1}^m \mathbf{v}_k\{I(s \le t), \mathbf{0}\}^{-1}]$. $\mathbf{w}_k$ is a diagonal matrix with the jth diagonal element $w_{kj} = \mathbf{v}_k(0, \mathbf{e}_j)^{-1}/\{\sum_{k=1}^m \mathbf{v}_k(0, \mathbf{e}_j)^{-1}\}$. In practice, this may not work well since it relies on asymptotic results. Based on prior work in Ma & Wang (2014), a simple choice of $w_k(t) = \mathbf{w}_k = r_k^{-1}$ has satisfactory performance.

Because the within group NPMLE already guarantees the monotonicity of each $\widehat{H}_k$, the final weighted average estimator for $\widehat{H}$ is monotone. The asymptotic property of $\widehat{H}$ and $\boldsymbol{\beta}$ is standard: $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}, \widehat{H}-H)$ converges weakly to a zero mean Gaussian process. Then, for any function $a_1(t)$ with bounded total variation and any vector $\mathbf{a}_2$, $\sqrt{n}\int a_1(s)d\{\widehat{H}(s) - H(s)\} + \sqrt{n}\mathbf{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a zero mean normal distribution whose variance can be approximated with

$$\mathbf{v}\{a_1(\cdot), \mathbf{a}_2\} \equiv \sum_{k=1}^{m} \mathbf{v}_k\{a_1(\cdot)w_k(\cdot), \mathbf{w}_k\mathbf{a}_2\}$$

where $t_1, \ldots, t_K$ are the observed event times.

Testing whether population heterogeneity in the covariate effects is present in (1.1) is equivalent to testing $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_p$. This can be written as testing $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, $\mathbf{A}$ a $(p-1)d_z \times (d_x + pd_z)$ block matrix in which the (j, j) block is $\mathbf{I}$ and the $(2, j)$ block is $-\mathbf{I}$ for $j = 3, \ldots, p+1$. All other blocks are zero. Based on the asymptotic results in Section 3.2, we can conveniently use a Wald test: under Φ_0, $n(\mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{A}\boldsymbol{\theta}$ has χ^2 distribution with $(p-1)d_z$ degrees of freedom, where

$$\mathbf{V} = -(\mathbf{0}_{(p-1)d_z \times K}, \mathbf{A}) \left\{ \frac{\partial^2 l(\widehat{\Phi}, \widehat{\boldsymbol{\theta}})}{\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}} \right\}^{-1} (\mathbf{0}_{(p-1)d_z \times K}, \mathbf{A})^{\mathrm{T}}.$$

When no covariate is included in the model, $\boldsymbol{\beta}$ does not appear. The procedure can then be directly applied with the simplification of deleting all the steps concerning estimating $\boldsymbol{\beta}$: we estimate $H(\cdot)$ from each of the m groups, then combine the results via a weighted average. This is similar to the approaches in Wacholder et al. (1998) and in Ma & Wang (2014), except that the estimation of $H(\cdot)$ in each group is carried out via MLE instead of least squares, and the weight selection is different from that in Wacholder et al. (1998).

1.4 Simulation Studies

We performed six sets of simulation studies to demonstrate the performance of the proposed method for the transformation model in the mixture data context. We present three of the simulation studies here and relegate the remaining three to Appendix B.4. Our first set of simulations contain homogeneous covariate effects. We generated data using $p = 2$, without $\boldsymbol{\alpha}_j$, and $\mathbf{X}$ a bivariate random vector. The first component of $\mathbf{X}$ was a binary variable, taking values 1 or 0 each with probability 0.5, the second component was uniform on -1 to 1. The transformation H was a logarithm function. We set f_1 to be the extreme value distribution, f_2 to be the logistic distribution. The censoring distribution was exponential, resulting in an overall censoring rate about 25%. The results are in the first block of Table 1.1 and upper-left plot of Figure 1.1. For comparison, we also did the estimation treating the homogeneous effect as heterigeneous, and estimated β_1, β_2 as $\alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22}$ instead. The results are in the second block of Table 1.1 and upper-right plot of Figure 1.1. These estimations are still consistent, yet the variability roughly doubled.

The second set of simulations studied heterogeneous covariate effects. It included $\boldsymbol{\alpha}_j$, but not $\boldsymbol{\beta}$. We generated data using $p = 2$. $\mathbf{Z}$ was of the same structure as $\mathbf{X}$ in the first simulation for the first two terms and an intercept term for the third term. We kept H the same as in the first simulation. Usually, in transformation models, the intercept term is not identifiable. In our case, the difference of the intercepts in different populations is identifiable, and hence was estimated. Here we set f_1 to be standard normal and f_2 to be a t distribution with 5 degrees of freedom. The censoring distribution was still exponential to achieve a 20% overall censoring rate. Results are in the second block of Table 1.1 and lower-left plot of Figure 1.1.

Our third simulation included both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_j$. We generated data using $p = 2$. $\mathbf{X}$ is bivariate with the first component either 1 or 0 with equal probability, and the second component a standard normal. Z was a uniform covariate on [-1 1] and a constant 1 to capture the intercept. The true H was still the log transformation. We took both f_1, f_2 to be normal with mean zero, but the second population had four times the variance as the first. The censoring distribution was exponential yielding a 20% overall censoring rate. The results are in the third block of Table 1.1 and the lower-right plot of Figure 1.1.

The simulation studies suggest that the proposed method has satisfactory finite sample performance: the parameter estimation yields small biases in all three simulations, measured by the mean and median of the 1000 estimates; Inference results are precise, in that the sample standard deviation from the 1000 simulations are closely matched by the average and the median of the 1000 estimated standard deviations calculated from the asymptotic results. The overall distribution of the estimated parameters are close to normal, as indicated by the empirical coverage of the 95% confidence intervals, which are close to their nominal levels. The estimation of the transformation function H, as shown in Figure 1.1, is within expectations. Overall, the average of the curve estimation approximately overlays the true H curve, while the 95% confidence bands have better performance than the typical nonparametric curve estimation. This is because H is estimated as the root-n rate, instead of the usual nonparametric rate. We also tired different transformations than H, with the overall performance similar. The details of these simulations are in Appendix B.4.

1.5 Application to Huntington's Disease Study

HD is the most prevalent monogenic neurodegenerative disorder caused by expansion of C-A-G repeats at the HD gene on chromosome 4 MacDonald et al. (1993). Typically neurological, cognitive, and physical symptoms begin to exhibit around 30-50 years of age for affected individuals, and eventually death is from pneumonia, heart failure, or other complications 15-20 years after the diagnosis Foroud et al. (1999). The subjects analyzed here were recruited in the Cooperative Huntington's Observational Research Trial (COHORT, Dorsey & The Huntington Study Group COHORT Investigators 2012), an epidemiological study of the natural history of HD. The probands were recruited primarily at academic research centers from 50 sites in the United States, Canada, and Australia. Probands were either clinically diagnosed with HD or the individuals who pursued HD genetic testing and carried a mutation but who were not clinically diagnosed. The initial probands underwent clinical examination and genotyping for HD mutation, and reported family history information on their first-degree relatives. The relatives were not genotyped because there was no resource for in-person collection of blood samples. Thus the relatives' HD mutation status was unknown, while the distribution of their mutation status could be estimated from the pedigree structure and the probands' carrier status. The full details of the COHORT study design are described in Dorsey & The Huntington Study Group COHORT Investigators (2012) and in Wang et al. (2012).

There were 4105 subjects included in the COHORT analysis, and they were either mutation carriers or not, hence $p = 2$. The heterogeneous covariate effect model (1.1) was used to study the effect of several covariates on mortality in HD mutation carriers where, for carriers, f_1 was normal with mean zero standard deviation 0.2, and for non-carriers, f_2 was $0.2T_5$, with T_5 a student t with 5 degrees-of-freedom. The main research interest is to predict age at death based on CAG repeats length, adjusting for gender, proband's HD clinical diagnosis status and a relative's relationship to the probands. We assumed all covariates to have differential effect in each mutation group to allow for maximal flexibility. The covariates included in the model were: CAG repeats length at the HD gene, gender, and proband's HD diagnosis status.

The results are reported in Table 1.2. As expected, the effects of CAG repeats length has a significant effect on age-at-death with an estimated effect of -0.76 (SE: 0.09, p-value< 0.001). The results suggest that if all covariates are the same, the subjects with one unit CAG longer repeat are expected to have a 2.38 years shorter lifespan. Here 2.38 is calculated as the average of $\widehat{H}^{-1}(U) - \widehat{H}^{-1}(U - 0.76)$ for a random U, where $\widehat{H}$ is the estimated transformation function and is close to a linear function (See Figure 1.2). This finding is consistent with the clinical literature which indicates an inverse association between CAG repeats length and HD age at diagnosis and death, Foroud et al. (1999), Langbehn et al. (2004). Proband's HD diagnosis also has a significant effect after adjusting for CAG repeats and other covariates: having a positive HD diagnosis in a family member is associated with an earlier mean age-at-death in carrier, potentially due to other shared familial risk factors.

The estimated transformation $H(\cdot)$ and its bootstrap confidence interval are presented in Figure 1.2. The nonparametric function suggests that a linear transformation may fit the data adequately and, under a parametric approximation, predictions formula for the age-at-death in a mutation carrier subject can be obtained. The approximated linear function is $\widehat{H}(t) = -24.35 + 0.32t$, see Figure 1.3.

A limitation of our analysis is that probands data were not included to protect against potential bias resulting from unknown sources in the COHORT study that did not use a population-based ascertainment scheme for probands. When the proband ascertainment is population-based, for example, probands are randomly selected from diseased population (case-family design), their data may be included through a retrospective likelihood. It would be interesting to replicate our analysis in an independent study using such a design, including probands data in the analyses.

1.6 Discussion

A potentially interesting extension of our method is to further parametrize the mixing distributions and estimate the parameters from data. If the q_{ij}'s are modeled parametrically, semiparametrically, or nonparametrically and estimated as $\widehat{q}_{ij}$, it would be interesting to develop methods to account for the discrepancy between $\widehat{q}_{ij}$ and q_{ij} and to deliver appropriate estimation of the survival function and covariate effect using the $\widehat{q}_{ij}$.

Our method has the flexibility to account for cross-population heterogeneity by characterizing the outcome in each population using different distributions specified by covariate parameters and error distributions (e.g., distinct scale or shape parameter; population-specific covariate effect), while simultaneously allow for common components across populations (e.g., shared covariate effect). Whether or not to adopt population-specific effects or shared effects is often determined by the purpose of the analysis and prior knowledge. In many cases, covariates whose effects are of particular research interest might be assumed to be population-specific as a precaution, while covariates that are not of interest be modeled across population.

We have assumed that the relative observations are independent, and excluded probands from the analyses. In proband-relative studies, multiple relatives from the same family may be collected and thus could have residual familial correlation. Our current approach is still consistent if the probands are representative samples of the probands population, but the inferences developed would no longer be valid. When probands are not representative and there is residual familial aggregation, ascertainment schemes may need to be modeled and probands and relative data analyzed jointly. How to best accommodate familial correlation and adjust for probands ascertainment schemes is highly challenging, and interesting.

Acknowledgements

This work is supported by NIH grants NS073671, NS082062 and NSF grant DMS1608-540. The authors wish to thank CHDI Foundation and COHORT study investigators. Samples and/or data from the COHORT study, which received support from HP Therapeutics, Inc., were used in this study. The authors also thank an associate editor and the reviewers for their constructive comments that have helped improving the quality of the paper.

Table 1.1: Simulation results based on 1000 repetitions.

	true	mean	median	sd	mean($\widehat{sd}$)	median($\widehat{sd}$)	95% CI
				simulation 1.1			
β_1	1.0000	0.9834	0.9703	0.4384	0.4474	0.4472	0.9570
β_2	2.0000	1.9734	1.9626	0.3845	0.3958	0.3954	0.9570
				simulation 1.2			
α_{11}	1.0000	0.9958	0.9992	1.0400	0.9623	0.9414	0.9410
α_{12}	2.0000	2.0420	2.0456	0.8916	0.8539	0.8199	0.9310
α_{21}	1.0000	0.9915	1.0140	0.8581	0.8395	0.8378	0.9420
α_{22}	2.0000	1.9684	1.9879	0.7328	0.7436	0.7350	0.9530
				simulation 2			
α_{11}	1.0000	1.0644	1.0584	1.1017	1.1758	1.1264	0.9530
α_{12}	2.0000	2.0767	2.0493	1.2519	1.3178	1.2870	0.9620
α_{21}	1.5000	1.4353	1.4306	0.7582	0.8072	0.7918	0.9640
α_{22}	3.0000	2.9344	2.9167	0.8787	0.9039	0.8852	0.9490
				simulation 3			
β_1	1.0000	0.9895	0.9915	0.3944	0.3976	0.3974	0.9520
β_2	1.5000	1.4974	1.4894	0.1983	0.2083	0.2079	0.9560
α_1	2.0000	1.9007	1.9443	1.1372	1.1737	1.1683	0.9600
α_2	3.0000	3.0040	2.9988	0.5071	0.5071	0.5028	0.9420

Table 1.2: COHORT analysis results: estimated covariate effects (age, gender, proband's diagnosis of HD), their standard errors, and p-values.

	Carriers				Non-carriers			
	$\alpha_{1intercept}$	α_{1Age}	$\alpha_{1Gender}$	$\alpha_{1ProDiag}$	$\alpha_{2intercept}$	α_{2Age}	$\alpha_{2Gender}$	$\alpha_{2ProDiag}$
est	-33.65	0.76	-0.67	1.79	-7.07	0.18	2.82	-2.30
se	4.28	0.09	0.70	1.00	1.25	0.03	0.67	0.84
p-value	< 0.001	< 0.001	0.34	0.07	< 0.001	< 0.001	< 0.001	0.006

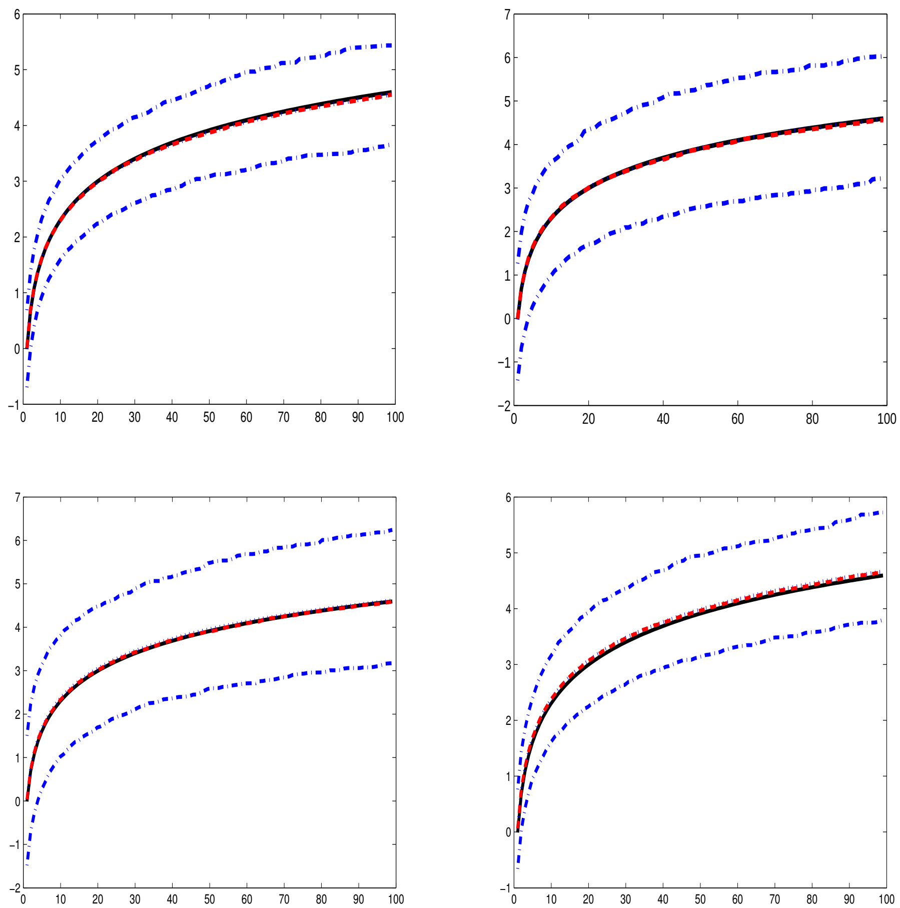

Figure 1.1 True function (solid line), median estimation (dashed line), mean estimation (dotted line) and 95% confidence band (dash-dotted line) of $H(T)$ in simulations 1.1 (upper-left), 1.2 (upper-right), 2 (lower-left), and 3 (lower-right) .

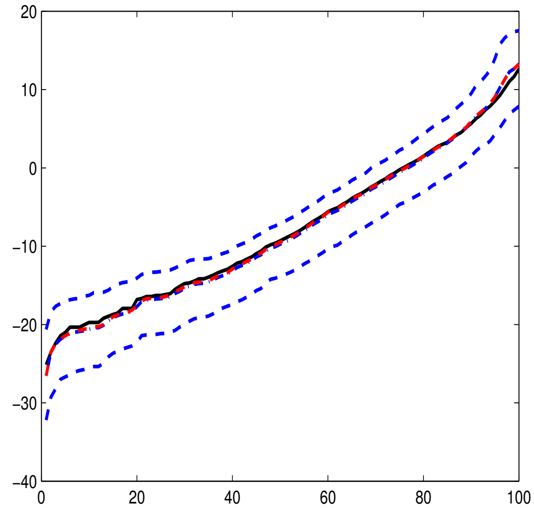

Figure 1.2 Estimated H function (solid line), median estimation (dashed line), mean estimation (dash-dotted line) and 95% confidence band (dashed line) of $H(T)$ in data analysis. Median, mean and 95% confidence band are based on 1000 bootstrapped samples.

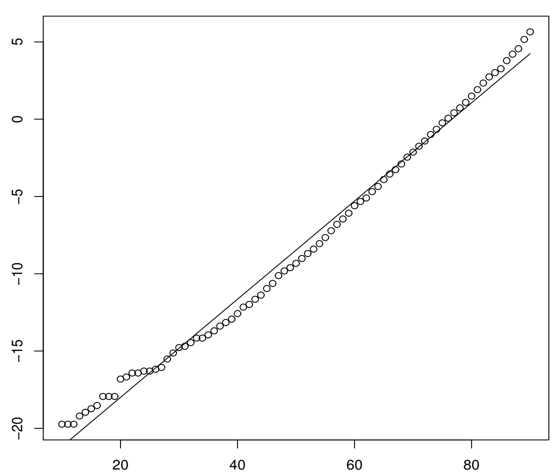

Figure 1.3 Fitted linear function $\widehat{H}(t)$ versus age t for HD data analysis.

Chapter 2

Generalized Partially Linear Single Index Model with Measurement Error, Instruments and Binary Response[1]

2.1 Introduction

Generalized linear models are familiar tools that are widely used in statistical applications. The model becomes complicated when the dependence of the response to some covariates, even after the transformation with a suitable link function, is not linear. A feasible and flexible approach to this is through introducing a partially linear single index structure, so that some covariates are modeled linearly, while some other covariates are summarized into an index, and the relation of the index to the response is modeled nonparametrically. This leads to the generalized partially linear single index model. A further complexity is when some of the covariates are measured with errors. Ignoring the measurement errors can generally lead to biased results, while taking the measurement error into account is also hard without specifying the measurement error variability exactly. Specifically, we denote the binary response variable Y, and let the $q \times 1$ covariate vector observed without error be $\mathbf{Z}$. We further let $\mathbf{X}$ be a $p \times 1$ latent variable. The model we study then is explicitly written as

$$\mathrm{pr}(Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = H\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma})\} \tag{2.1}$$

[1] Yang, G., Wang, Q., Cui, X and Ma, Y. Submitted to Computational Statistics and Data Analysis, 07/02/2018.

where $\boldsymbol{\beta} \in R^p$ and $\boldsymbol{\gamma} \in R^q$ are unknown parameters of interest, $H(\cdot)$ is a known inverse link function, for example, the inverse logit link function $H(\cdot) = 1 - 1/\{\exp(\cdot) + 1\}$ or the inverse probit link function $H(\cdot) = \Phi(\cdot)$, and $g(\cdot)$ is an unknown function. Because $\boldsymbol{\gamma}$ is not identifiable when incorporated with an unspecified g, the constraint $\|\boldsymbol{\gamma}\| = 1$ or the first component of $\boldsymbol{\gamma}$ is positive is often imposed. Here, we use the latter choice, which fixes the first component of $\boldsymbol{\gamma}$ to be 1 and leave the remaining components arbitrary. We denote the vector formed by the second to last components of $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}_{-1}$.

When $\mathbf{X}$ is latent or observed with error, the parameters in model (2.1) is generally hard to identify in practice. However, the existence of instruments is often very helpful and can save the situation. Instead of observing $\mathbf{X}$, we observe an erroneous version of $\mathbf{X}$, written as $\mathbf{W}$ and an instrumental variable $\mathbf{S}$. The variables $\mathbf{W}$ and $\mathbf{S}$ are linked to $\mathbf{X}$ through

$$\mathbf{W} = \mathbf{X} + \mathbf{U} \quad \text{and} \quad \mathbf{X} = \mathbf{m}(\mathbf{S}, \mathbf{Z}; \boldsymbol{\alpha}) + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\mathbf{m}(\cdot)$ is a known function up to an unknown parameter $\boldsymbol{\alpha}$. Here, we assume the conditional mean of $\boldsymbol{\varepsilon}$ and the marginal mean of $\mathbf{U}$ to be zero, that is, $E(\boldsymbol{\varepsilon}|\mathbf{S}, \mathbf{Z}) = \mathbf{0}$, $E(\mathbf{U}) = \mathbf{0}$. Further assume that $(\mathbf{X}, \mathbf{S}, \mathbf{Z})$ is independent of $\mathbf{U}$, $\mathbf{U}$ is independent of $\boldsymbol{\varepsilon}$, $\mathbf{W}$ is independent of $(\mathbf{S}, \mathbf{Z})$ given $\mathbf{X}$, and Y is independent of $(\mathbf{W}, \mathbf{S})$ given $(\mathbf{X}, \mathbf{Z})$. The model in (2.1), in combination with the instrumental variable condition studied here, has much resemblance with the problem setting in Xu et al. (2015). However, the critical difference lies in the presence of the unknown function g as well as the unknown index vector $\boldsymbol{\gamma}$. This seemingly small change actually brings much more complexity in all aspects of the analysis, including the method development, the theoretical proofs and the numerical implementation. To appreciate this fact, one can link to the additional hurdles encountered and overcome in the literature when moving from linear regression to single index models.

As a field of much practical importance, measurement error models in general have been extensively studied. However, as far as we are aware, no work exists in studying measurement error models when the experiment model is of the generalized partially linear single index type with binary response, while an instrumental variable exists to provide additional information. In fact, the only works in handling binary response models with measurement errors that we are aware are Stefanski & Carroll (1985), Stefanski & Carroll (1987), Buzas & Stefanski (1996), Huang & Wang (2001), Ma & Tsiatis (2006), in addition to Xu et al. (2015) mentioned above. However, none of these works contains a partially linear single index component, and most of these works do not consider instruments.

In this chapter, we demonstrate that by employing a prediction relation for the unobserved covariates using available instruments, we can construct consistent estimators for all the parameters in the generalized linear single index model. In addition, we also provide a nonparametric estimator for the unspecified function of the estimated index. The method we devise incorporates instrumental variables in a creative and different way from most traditional method in handling instruments. In fact, our work is the first in using instruments in handling the generalized linear single index regression models with measurement error and binary response.

The rest of chapter is organized as follows. We describe our main methodology and the asymptotic properties of our estimator in Section 2.2. Simulation studies are given in Section 2.3 to provide finite sample performance of our method. We analyze an AIDs study data in Section 2.4 and conclude the chapter in Section 2.5.

2.2 Estimation procedure via profiling and the asymptotic properties

Denote the ith observed data $\mathbf{O}_i = (Y_i, \mathbf{W}_i, \mathbf{S}_i, \mathbf{Z}_i)$, for $i = 1, \dots, n$. These observations are independent and identically distributed (i.i.d.) according to the model described in (2.1) and (2.2). Our main interest is in estimating $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}_{-1}^{\mathrm{T}})^{\mathrm{T}}$. However, $g(\cdot)$ is a nuisance unknown function.

First of all, we have

$$\mathbf{W} = \mathbf{m}(\mathbf{S}, \mathbf{Z}; \boldsymbol{\alpha}) + \mathbf{U} + \boldsymbol{\varepsilon},$$

where $E(\mathbf{U} + \boldsymbol{\varepsilon}|\mathbf{S}, \mathbf{Z}) = \mathbf{0}$. We can use least squares method to estimate $\widehat{\boldsymbol{\alpha}}$. (2.3) is the estimating equation to obtain $\widehat{\boldsymbol{\alpha}}$.

$$\sum_{i=1}^{n} \mathcal{S}_{\boldsymbol{\alpha}}(\mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial \mathbf{m}^{\mathrm{T}}(\mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \boldsymbol{\Omega}(\mathbf{S}_i, \mathbf{Z}_i)\{\mathbf{W}_i - \mathbf{m}(\mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\alpha})\} = \mathbf{0}, \quad (2.3)$$

where $\boldsymbol{\Omega}(\mathbf{S}, \mathbf{Z})$ is any weight matrix. We can choose to use ordinary least squares (OLS) or weighted least squares (WLS) method by using different weight matrix. Specifically, we can use identity matrix as weight matrix to obtain OLS estimator and use the inverse of the error variance-covariance matrix conditional on $(\mathbf{S}, \mathbf{Z})$ as weight matrix to obtain WLS estimator.

After we have an estimate $\widehat{\boldsymbol{\alpha}}$, we can write $\mathbf{X}$ in the form of $\widehat{\boldsymbol{\alpha}}$ and $(\mathbf{S}, \mathbf{Z})$ and plug into model (2.1) to obtain the joint distribution of $(Y, \mathbf{S}, \mathbf{Z})$ as

$$\begin{aligned} &\mathrm{pr}(Y = y, \mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z}) = f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z}) \\ &\times \int \left[1 - y + (2y-1) H\left\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\right\}\right] \\ &\times f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon}), \end{aligned} \quad (2.4)$$

where $f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z})$ is a conditional probability density function that satisfies $\int \boldsymbol{\varepsilon} f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $f_{\mathbf{S},\mathbf{Z}}(\mathbf{s}, \mathbf{z})$ is the joint pdf of $(\mathbf{S}, \mathbf{Z})$.

Now we move to construct the estimation procedure for $\boldsymbol{\theta}$ and $g(\cdot)$. Borrowing the ideas in Ma & Carroll (2006) and Xu et al. (2015), we will construct two sets of estimating equations in order to estimate $\boldsymbol{\theta}$ and $g(\cdot)$.

Treating (2.4) as a semiparametric model, the nuisance tangent space is

$$
\begin{aligned}
\Lambda &= \Lambda_1 \oplus \Lambda_2 \\
&= \{\mathbf{f}(\mathbf{S}, \mathbf{Z}) : E(\mathbf{f}) = 0, E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty, \forall \mathbf{f} \in \mathcal{R}^{p+q-1}\} \\
&\quad \oplus \{E\{\mathbf{f}(\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}) | Y, \mathbf{S}, \mathbf{Z}\} : E(\mathbf{f}|\mathbf{S}, \mathbf{Z}) = \mathbf{0}, E(\boldsymbol{\varepsilon}\mathbf{f}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z}) = \mathbf{0}, \\
&\quad E(\mathbf{f}^{\mathrm{T}}\mathbf{f}) < \infty, \forall \mathbf{f} \in \mathcal{R}^{p+q-1}\}.
\end{aligned}
$$

Notation $\oplus$ is used to emphasize that an arbitrary function $\mathbf{f}_1(\mathbf{S}, \mathbf{Z})$ in Λ_1 and an arbitrary function $\mathbf{f}_2(\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z})$ in Λ_2 satisfy $E\{\mathbf{f}_1(\mathbf{S}, \mathbf{Z})\mathbf{f}_2{}^{\mathrm{T}}(\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z})\} = \mathbf{0}$. The orthogonal complement of Λ is

$$
\begin{aligned}
\Lambda^{\perp} &= \{\mathbf{f}(Y, \mathbf{S}, \mathbf{Z}) : E(\mathbf{f}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}) = \boldsymbol{\alpha}(\mathbf{S}, \mathbf{Z})\boldsymbol{\varepsilon}, \|E\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}\|_{\infty} < \infty, \forall \mathbf{f} \in \mathcal{R}^{p+q-1}, \\
&\quad \forall \boldsymbol{\alpha} \in \mathcal{R}^{(p+q-1)\times p}\}.
\end{aligned}
$$

Let $\mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}$, and $\mathcal{S}_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}$ be the scores for $\boldsymbol{\theta}$ and $g(\cdot)$ respectively. Specifically,

$$
\begin{aligned}
&\mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} \\
&= (2Y-1) \\
&\quad \times \frac{\int \begin{pmatrix} \mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}}) + \boldsymbol{\varepsilon} \\ g'(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\mathbf{Z}_{-1} \end{pmatrix} H'\Big\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\Big\} f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon})}{\int \Big[1 - Y + (2Y-1)H\Big\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\Big\}\Big] f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon})}, \\
&S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} \\
&= (2Y-1) \\
&\quad \times \frac{\int H'\Big\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\Big\} f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon})}{\int \Big[1 - Y + (2Y-1)H\Big\{\mathbf{m}(\mathbf{S}, \mathbf{Z}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\beta} + g(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma})\Big\}\Big] f_{\varepsilon}(\boldsymbol{\varepsilon}|\mathbf{s}, \mathbf{z}) d\mu(\boldsymbol{\varepsilon})}.
\end{aligned}
$$

We get the efficient score by projecting $\mathcal{S}_{\boldsymbol{\theta}}$ and S_g to $\Lambda^{\perp}$

$$
\begin{aligned}
\mathcal{L}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} &= \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} - E[\boldsymbol{\beta}_{\boldsymbol{\theta}}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}], \\
\Phi\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} &= S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}],
\end{aligned}
$$

where $\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} \in \mathcal{R}^{p+q-1}$ and $b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} \in \mathcal{R}$ satisfy

$$
\begin{aligned}
E\{\ \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} - E[\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} &= \boldsymbol{\alpha_\theta}(\mathbf{S}, \mathbf{Z})\boldsymbol{\varepsilon}, \\
E\{S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} &= \boldsymbol{\alpha}_g(\mathbf{S}, \mathbf{Z})\boldsymbol{\varepsilon},
\end{aligned}
$$

where $\boldsymbol{\alpha_\theta}(\mathbf{S}, \mathbf{Z}) \in \mathcal{R}^{(p+q-1)\times p}$ and $\boldsymbol{\alpha}_g(\mathbf{S}, \mathbf{Z}) \in \mathcal{R}^{1\times p}$. Here we have to specify the following terms $\boldsymbol{\beta_\theta}, \boldsymbol{\alpha_\theta}, b_g$ and $\boldsymbol{\alpha}_g$. By multiplying $\boldsymbol{\varepsilon}$ on both sides of the above formulas and taking expectation conditional on $(\mathbf{S}, \mathbf{Z})$, we obtain

$$
\begin{aligned}
& E\{\ \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
= \ & \boldsymbol{\alpha_\theta}(\mathbf{S}, \mathbf{Z})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z}), \\
& E\{S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]\boldsymbol{\varepsilon}^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
= \ & \boldsymbol{\alpha}_g(\mathbf{S}, \mathbf{Z})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z}).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\boldsymbol{\alpha_\theta}(\mathbf{S}, \mathbf{Z}) &= E\{\ \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
&\quad \times\{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z})\}^{-1}, \\
\boldsymbol{\alpha}_g(\mathbf{S}, \mathbf{Z}) &= E\{S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]\boldsymbol{\varepsilon}^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
&\quad \times\{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z})\}^{-1}.
\end{aligned}
$$

Inserting the form of $\boldsymbol{\alpha_\theta}(\mathbf{S}, \mathbf{Z})$ and $\boldsymbol{\alpha}_g(\mathbf{S}, \mathbf{Z})$ respectively, we obtain the following equations

$$
\begin{aligned}
& E\{\ \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
= \ & E\{\ \mathcal{S}_{\boldsymbol{\theta}}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[\boldsymbol{\beta_\theta}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z} \\
& \times\{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z})\}^{-1}\boldsymbol{\varepsilon}, \\
& E\{S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]\boldsymbol{\varepsilon}^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \\
= \ & E\{S_g\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}\boldsymbol{\varepsilon}^{\mathrm{T}} - E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]\boldsymbol{\varepsilon}^{\mathrm{T}}|\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}\} \quad (2.5) \\
& \times\{E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}|\mathbf{S}, \mathbf{Z})\}^{-1})\boldsymbol{\varepsilon}.
\end{aligned}
$$

Then we can obtain the terms $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ and b_g by solving the equations in (2.5). Unfortunately, the integral equations in (2.5) do not have a closed form solution hence numerical methods are required to obtain approximate solutions. In fact, these are first type Fredholm integral equations and require regularization to obtain stable solutions. Nevertheless, such integral equations are well studied in numerical analysis and many methods exist. Here, our final goal is to obtain $E[\boldsymbol{\beta}_{\boldsymbol{\theta}}\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]$ and $E[b_g\{\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|Y, \mathbf{S}, \mathbf{Z}]$ instead of $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ and b_g, hence the numerical problem is an easier one to handle than the typical Type I Fredholm integral equations. For details on how to solve the integral equations in (2.5), we refer to Kress (1991).

Obviously, it follows that

$$
\begin{aligned}
\mathbf{0} &= E[\mathcal{L}\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|\mathbf{S}, \mathbf{Z}], \qquad (2.6)\\
0 &= E[\Phi\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\theta}, g(\cdot)\}|\mathbf{S}, \mathbf{Z}]. \qquad (2.7)
\end{aligned}
$$

Equations (2.6) and (2.7) form the backbone of our method that allows for a general unknown function $g(\cdot)$. Because $g(\cdot)$ is modeled nonparametrically, we use the local linear method to estimate $\widehat{g}(\cdot)$. Let $K(z)$ be a smooth symmetric density function, let h be a bandwidth, and define $K_h(z) = h^{-1}K(z/h)$. Let $U = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}$ and $u_0 = \mathbf{z}_0^{\mathrm{T}}\boldsymbol{\gamma}$. We approximate $g(\cdot)$ locally by a linear function

$$
g(U, \boldsymbol{\beta}) \approx g(u_0, \boldsymbol{\beta}) + g'(u_0, \boldsymbol{\beta})(U - u_0).
$$

The nonparametric function estimator $\widehat{g}(\cdot)$ is then defined as the solution to $g(u_0, \boldsymbol{\beta})$ of the local linear estimating equation

$$
\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{n} \begin{Bmatrix} 1 \\ (U_i - u_0)/h \end{Bmatrix} K_h(U_i - u_0)\Phi\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\theta}, g(u_0, \boldsymbol{\beta}) \\
&\quad + g'(u_0, \boldsymbol{\beta})(U_i - u_0)\}. \qquad (2.8)
\end{aligned}
$$

Note that although $g(\cdot)$ depends on $\boldsymbol{\beta}$, its estimator depends on $\boldsymbol{\theta}$, hence we write it as $\widehat{g}(\cdot, \boldsymbol{\theta})$. The estimate $\widehat{\boldsymbol{\theta}}$ is subsequently obtained as the solution to

$$
\mathbf{0} = \sum_{i=1}^{n} \mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\theta}, \widehat{g}(U_i; \boldsymbol{\beta})\}, \qquad (2.9)
$$

In the above description, we have developed the whole methodology as if the computation can be carried through. However, a closer look at the expressions involving conditional expectation given $\mathbf{S}, \mathbf{Z}$ reveals that these quantities are not computable without knowing the distribution of $\boldsymbol{\varepsilon}$ given $\mathbf{S}, \mathbf{Z}$. Instead of estimating the error distribution $f_{\varepsilon|\mathbf{S},\mathbf{Z}}(\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z})$, which is difficult, we propose to use a working model and carry out all the calculations under this working model. Thus, in summary, we first estimate the function g through the local linear method, by treating $\boldsymbol{\theta}$ as parameters that are held fixed. The set of estimating equations are exactly (2.8) at $\mathbf{z}_0 = \mathbf{z}_1, \ldots, \mathbf{z}_n$, and the solutions are $\widehat{g}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_1, \boldsymbol{\theta}), \ldots, \widehat{g}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_n, \boldsymbol{\theta})$. We then estimate $\boldsymbol{\theta}$ through solving (2.9). Obviously, this is a type of profiling estimation procedure.

We now study the asymptotic properties of the proposed estimator, which is computed under the working model of $f_{\varepsilon|\mathbf{S},\mathbf{Z}}(\boldsymbol{\varepsilon}, \mathbf{S}, \mathbf{Z})$. We first list the regularity conditions required.

(C1) The kernel function $K(\cdot)$ is non-negative, has compact support, and satisfies $\int K(s)ds = 1$, $\int sK(s)ds = 0$, $0 < \mu_2 = \int s^2K(s)ds < \infty$ and $\int sK^2(s)ds < \infty$.

(C2) The bandwidth h in the kernel smoothing satisfies $nh^2 \to \infty$ and $nh^4 \to 0$ when $n \to \infty$.

(C3) The link function $H(\cdot)$ is differentiable.

(C4) The nonparametric function $g(\cdot)$ has continuous first order derivative.

(C5) The random variable $U = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}$ has compact support and its marginal density function $f_U(\cdot)$ is bounded away from zero on the support.

Let $\boldsymbol{\alpha}^{\otimes 2} = \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}$ for all matrix or vector $\boldsymbol{\alpha}$ throughout the text. Then we have the following result.

Theorem 2. *Under the regularity conditions (C1)/(C5), $\widehat{\boldsymbol{\theta}}$ satisfy*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1^{\mathrm{T}}})$$

in distribution when $n \to \infty$, where

$$
\begin{aligned}
\boldsymbol{A} &= \left\{ \begin{array}{c} E\left(\frac{\partial \mathcal{L}\{Y_i, \boldsymbol{S}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, g(U_i;\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} - \frac{\partial \mathcal{L}\{Y_i, \boldsymbol{S}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, g(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})} \frac{E\left[\partial \Phi\{Y_j, \boldsymbol{S}_j, \boldsymbol{Z}_j; \boldsymbol{\beta}, g(U_i;\boldsymbol{\beta})\}/\partial \boldsymbol{\beta}^{\mathrm{T}} | U_i\right]}{E[\partial \Phi\{Y_j, \boldsymbol{S}_j, \boldsymbol{Z}_j; \boldsymbol{\beta}, g(U_i;\boldsymbol{\beta})\}/\partial g(U_i;\boldsymbol{\beta}) | U_i]} \right) \\ -E\left(\frac{\partial \mathcal{L}\{Y_j, \boldsymbol{S}_j, \boldsymbol{Z}_j; \boldsymbol{\beta}, g(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})} \frac{E[(\boldsymbol{Z}_j - \boldsymbol{Z}_i)_{-1}^{\mathrm{T}} \partial \Phi\{Y_j, \boldsymbol{S}_j, \boldsymbol{Z}_j; \boldsymbol{\beta}, g(U_i,\boldsymbol{\beta})\}/\partial g(U_i,\boldsymbol{\beta}) g'(U_i,\boldsymbol{\beta}) | U_i]}{E[\partial \Phi\{Y_j, \boldsymbol{S}_j, \boldsymbol{Z}_j; \boldsymbol{\beta}, g(U_i,\boldsymbol{\beta})\}/\partial g(U_i,\boldsymbol{\beta}) | U_i]} \right) \end{array} \right\} \\
\boldsymbol{B} &= E\Bigg\{ \Bigg(\mathcal{L}\{Y_i, \boldsymbol{S}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}, g(U_i;\boldsymbol{\theta})\} - E\Big[\partial \mathcal{L}\{Y, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\theta}, g(U;\boldsymbol{\theta})\}/\partial g(U;\boldsymbol{\theta}) \mid U = U_i\Big] \\
&\quad \times \Big[E\left\{\partial \Phi\{Y, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\theta}, g(U;\boldsymbol{\theta})\}/\partial g(U;\boldsymbol{\theta}) \mid U = U_i\right\} \Big]^{-1} \\
&\quad \times \Phi\{Y_i, \boldsymbol{S}_i, \boldsymbol{Z}_i; \boldsymbol{\theta}, g(U_i;\boldsymbol{\theta})\} \Bigg)^{\otimes 2} \Bigg\}.
\end{aligned}
$$

2.3 Numerical study

We performed three sets of simulation studies to evaluate the finite sample performance of the proposed estimator. In all the simulations, we set the sample size $n = 1000$ and we repeated the experiments 1000 times. In the first simulation, we generated the observations (Y_i, W_i, S_i, Z_i) from the model

$$
\begin{aligned}
\mathrm{pr}(Y_i = 1 | X_i = x_i, Z_i = z_i) &= H\{\beta x_i + g(\gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i})\}, \\
W_i &= X_i + U_i, \\
X_i &= \alpha_1 + \alpha_2 S_i + \epsilon_i, \qquad (2.10)
\end{aligned}
$$

where $H(t)$ is the inverse logit link function and the invers probit link function, and $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta = 0.3$, $\gamma_1 = 1$, $\gamma_2 = 0.5$, $\gamma_3 = 1$, $\gamma_4 = -0.3$. The true function $g(t) = t$, i.e. we experiment with a simple linear function with slope 1 and intercept 0 for g. The observable covariates Z_{1i}, Z_{2i}, Z_{4i} and the instrument variable S_i are generated from the standard normal distribution. The observable covariate Z_{3i} is generated from uniform distribution on domain $[-1, 1]$. U_i is generated from a normal distribution with mean zero and variance 0.6. ϵ_i is generated from a standard normal distribution with mean 0 and variance ${S_i}^2/2$ and a t_5 distribution multiplied by $|S_i|/\sqrt{2}$. Since our working model for ϵ_i is set to be normal, it corresponds to a correct working model in the simulation where ϵ_i's are normally distributed, and corresponds to a misspecified working model in the simulation where ϵ_i's are non-normally distributed.

In our second simulation, we experimented with different parameters values. Here, we set $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta = 0.3$, $\gamma_1 = 1$, $\gamma_2 = 0.2$, $\gamma_3 = 0.3$, $\gamma_4 = -0.4$. The observable covariates Z_{1i}, Z_{2i}, Z_{4i} and the instrument variable S_i are generated from 0.5 times a standard normal distribution. The observable covariate Z_{3i} is generated from uniform distribution on domain $[-1, 1]$. U_i and ϵ_i are generated similarly as in the first simulation. Our true g function in the second simulation is nonlinear and it is the quadratic function $g(t) = 1.5t^2$.

Further, in the third simulation, we generated the data similarly as in the first simulation, except that the true function form of g is now $g(t) = 1.5\sin(t)$ for normal distributed ϵ_i and $g(t) = \sin(t)$ for t distributed ϵ_i. Thus we also experiment with a nonlinear function form for g.

We used respectively OLS and WLS to estimate α_1 and α_2, and compared the subsequent performance with the estimation result under the known α for all three simulation studies described above. The results of the three simulations are summarized in Tables C.1 to C.6 and Figure C.1. From these results, it is quite clear that the proposed estimators indeed yield consistent estimation, regardless a correct or a mis-specified working model is used, in that the biases are quite small and the mean and median estimated curves track the true curves very well. Even though WLS produces more efficient estimators for α_1 and α_2, the efficiency in the parameter estimation of $\boldsymbol{\alpha}$ does not really translate to the efficiency difference in estimating the main parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In fact, even when $\boldsymbol{\alpha}$ is completely known, we do not see a significant advantage in estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. This is quite encouraging since this confirms our theoretical discovery. This result also provides practical significance since $\boldsymbol{\alpha}$ is typically not known in reality.

2.4 Real data analysis

The data set we analyze here is from the popular AIDS Clinical Trials Group (ACTG) study. In this study four different treatments 'ZDV', 'ZDV+ddI', 'ZDV+ddC', and 'ddC' were used on HIV infected adults whose CD4 cell counts were between 200 and 500 per cubic millimetre. 'ZDV' is a standard treatment, and is considered as the reference treatment. For convenience, we name 'ZDV' treatment1, 'ZDV+ddI' treatment2, 'ZDV+ddC' treatment3 and 'ddC' treatment4. Age was included as an explanatory variable. There were 1036 patients in our sample who had no antiretroviral therapy prior to the study. The purpose of this study is to see whether there is any difference among the four treatments in terms of preventing a patient's CD4 count from dropping below 50%. CD4 count is an important indicator for HIV positive patients to develop AIDS or to die from HIV caused disease. This is considered an endpoint event and when it occurs, our response variable Y is set to 1. We use Z_1, Z_2, Z_3 as three treatment indicators besides the reference treatment.

Let X be the baseline log(CD4 count) before the start of the treatment. Here, we treat X as a latent variable, since it can not be measured precisely. Instead of observing X, we observe W, which is the average of two available measurements of X. We thus assume W is X plus a random noise. We also have an instrumental variable S, which is the screening log(CD4 count). Figure C.2 suggests that there is a linear relationship between W and S, thus we further assume a linear regression model to link X and S. Finally, to model the relation between the occurrence of AIDS or death with the covariates and treatments, we used the familiar logistic model. The complete form of the model that is used to describe the ACTG data is

$$\begin{aligned} \mathrm{pr}(Y_i = 1 | X_i = x_i, Z_i = z_i) &= H\{\beta x_i + g(\gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_{age} z_{age})\}, \\ W_i &= X_i + U_i, \\ X_i &= \alpha_1 + \alpha_2 S_i + \epsilon_i. \end{aligned} \tag{2.11}$$

We used the methodology developed earlier in the chapter to analyze the data. Using OLS method, we got the estimates for α_1 and α_2 to be (0.001, 0.674). The estimates for the main parameters β and γ's are shown in Table C.7. These results indicate that there is significant difference between the four different treatments.

We further fixed the Age variable at 41 years old, which is the mid point on the range of Age variable to compare the four treatments. The estimated g function of treatments 1, 2, 3 and 4 are -1.23, -1.78, -1.94 and -2.20 respectively. Their corresponding 90% confidence intervals are (-1.65, -0.96), (-2.01, -1.63), (-2.17, -1.78) and (-2.26, -1.80). It indicates that treatments 2, 3, 4 are more efficient than treatment 1 for 41 year old patients in general.

We also plot the estimated g as a function of the estimated index in Figure C.3, together with its 95% confidence band. We can see that g is decreasing and nonlinear and has a general decreasing trend, indicating a protective effect of the index in terms of risk of CD4 counts decreasing or death.

2.5 Discussion

Measurement error issue is a widely encountered problem in statistical applications. When the magnitude of the error is known or estimable, either from multiple measurements or from validation data, many methods are available to proceed with the subsequent analysis that adjust for the known measurement error issue. However, when the measurement error magnitude is unknown and un-estimable, which is often the case in practice, instruments are often indispensable. In this chapter, we demonstrate that instrumental variable can be used in estimation in the generalized linear single index model context with binary response, which is unsolved in the literature before. In addition, the estimation of the model parameters is conducted without making any parametric assumption for the distribution of the unobserved variables in the model, i.e. we have worked in the functional model framework.

The simulation studies show satisfactory performance of the proposed estimator in finite sample situation. Further, despite the fact that we present our main estimator in the context of logistic and probit models, the method is not restricted to these models only. In fact, any generalized partially linear regression model of Y conditional on X and $\mathbf{Z}$ can be handled by our method via a suitable link function H, thus Y is not restricted to binary variables and the method can be further extended to arbitrary generalized semiparametric single index models in terms of methodology. However, we foresee computational challenges in the more general cases.

Acknowledgement

Guangren Yang's research was supported by the National Social Science Foundation of China grant 16BTJ032, the National Statistical Scientific Research Center Projects 2015LD02, Science and Technology Program of Guangzhou 2016201604030074 and Science and Technology Planning Project of Guangdong 2016A050503033. Xia Cui is a College Talent Cultivated by 'Thousand-Hundred-Ten' Program of Guangdong Province and her research was supported by grants from National Natural Science Foundation of China grant 11471086, the China Scholarship Council 201408440400, Humans and Social Science Research Team of Guangzhou University 201503XSTD and the Training Program for Excellent Young College Teachers of Guangdong Province Yq201404. Ma's research is partially supported by the national science foundation and the national institute of neurological disorder and stroke.

Chapter 3

Locally efficient estimation in generalized partially linear model with measurement error in nonlinear function[1]

3.1 Introduction

Generalized partially linear models have been widely used in statistics. Such models enrich the more classic generalized linear models by allowing a covariate to enter the link function through a nonparametric form. This is useful when the dependence of the response to some covariates, even after transformation through a suitable link function, is still not linear and difficult to specify. At the same time, the model also allows the more classic generalized linear dependence on some other covariates. Many works exist in the literature for estimation and inference for generalized partially linear models, see, for example Carroll et al. (1995), Liang et al. (2009), Yu & Ruppert (2012).

When one of the covariates involved in the generalized partially linear model cannot be measured precisely, the problem becomes much more difficult. In fact, most of the works in handling measurement error issues in the generalized partially linear model considered only the case that measurement error occurs to a covariate involved in the linear component (Ma & Carroll 2006, Liu et al. 2017, Liang & Ren 2005, Liu 2007, Liang & Thurston 2008). When the model degenerates to simply the general-

[1]Wang, Q., Ma, Y. and Yang, G. Submitted to TEST, 07/12/2018.

ized linear model, even more literatures exist to handle the measurement error issues (Stefanski & Carroll 1985, 1987, Huang & Wang 2001, Ma & Tsiatis 2006, Carroll & Crainiceanu 2006, Buonaccorsi 2010, Xu et al. 2015). When handled properly, it can be shown that the parameters can be estimated at the root-n convergence rate despite of the presence of the measurement error and the possible presence of the nonparametric function in the model. However, it is a different story when the covariate inside the nonparametric function itself is measured with error. We conjecture that this is because as soon as the covariate inside an unknown function is subject to error, the problem falls into the general framework of nonparametric measurement error models and the standard practice for estimation and inference is through deconvolution. Deconvolution method is widely used in handling latent components and has been used to show that nonparametric regression with errors in covariates can have very slow convergence rate. Possibly due to these inherent difficulties generalized partially linear models with errors in the covariate inside the nonparametric function has not been studied systematically.

We tackle this difficult problem where the error occurs to the covariate inside the nonparametric component of the generalized partially linear model through a novel approach that avoids the deconvolution treatment completely. Two key ideas lead to our success in this endeavor. The first is the idea of using B-splines expansion to approximate the nonparametric function of the latent covariate. The B-spline nature allows us to write out the approximation form without having to perform the estimation simultaneously. This is different from nonparametric estimation via kernel method, where the approximation and estimation is integrated and inseparable. The second idea is the recognition that after the B-spline approximation, the error-free model is effectively a parametric model, or at least a parametric model in terms of operation, hence the only nonparametric component in the measurement error model is the distribution of the latent covariate. This implies that the semiparametric approach in Tsiatis & Ma (2004) can be adopted here to help establishing the estimation procedure. The encouraging discovery is that we not only can bypass the difficulties caused by nonparametric function of a covariate measured with error in terms of estimation, we also prove that the procedure can retain the root-n convergence rate of the parameter estimation in the original model.

The structure of this chapter is as follows. We describe the model and the estimation methodology in Section 3.2, following with establishing the large sample properties of the parameter estimation in Section 3.3. Two simulation studies are conducted in Section 3.4 and we analyzed the AIDS Clinical Trials Group (ACTG) study in Section 3.5. We finish the chapter with some discussions in Section 3.6. All the technical details and proofs are provided in an Appendix.

3.2 Main results

3.2.1 The model

The generalized partially linear model we study is

$$f_{Y|X,\mathbf{Z}}(y, x, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, g) = f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + g(x), \boldsymbol{\alpha}\}, \tag{3.1}$$

where f is a known link function up to the unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and unknown function $g(\cdot)$. For example, $f(\cdot)$ can be the inverse logit link function $f(\cdot) = 1 - 1/\{\exp(\cdot) + 1\}$. The response variable Y is an observable variable and $X, \mathbf{Z}$ are covariates. Here $\mathbf{Z}$ is observable, while X is a random variable measured with error, thus it is not directly observable. Instead of observing X, we observe W, where

$$W = X + U, \tag{3.2}$$

and U is a normal random error independent of $X, \mathbf{Z}$ with mean zero and variance σ_U^2. For ease the presentation of the main methodology, we assume σ_U^2 is known. When σ_U^2 is not known, a common approach is to use repeated measurements to estimate

σ_U^2 first and then plug in. The observed data are $(W_i, \mathbf{Z}_i, Y_i), i = 1, \ldots, n$, which are independent and identity distributed (iid). Our goal is to estimate $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $g(\cdot)$ hence to understand the dependence of Y on the covariates $(X, \mathbf{Z})$.

3.2.2 Efficient Score Derivation

For preparation, we first approximate $g(x)$ with a B-spline representation, i.e. $g(x) \approx \mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}$. Under this approximation, model (3.1) becomes

$$f_{Y|X,\mathbf{Z}}(y, x, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, g) \approx f_{Y|X,\mathbf{Z}}(y, x, \mathbf{z}, \boldsymbol{\theta}) \equiv f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}, \boldsymbol{\alpha}\},$$

which is a complete parametric model with unknown parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. This model falls in the general framework of Tsiatis & Ma (2004) hence the estimation procedure there can be adopted here. Specifically, the joint distribution of the observed variables conditional on $\mathbf{Z}$ is

$$f_{W,Y|\mathbf{Z}}(y, w, \mathbf{z}, \boldsymbol{\theta}) = \int f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}, \boldsymbol{\alpha}\} f_{W|X}(w, x) f_{X|\mathbf{Z}}(x, \mathbf{z}) d\mu(x).$$

with the condition distribution function $f_{X|\mathbf{Z}}(x, \mathbf{z})$ being a nuisance parameter. The nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$ can be written as

$$\begin{aligned} \Lambda &= [E\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\} : E\{\mathbf{a}(X, \mathbf{Z}) \mid \mathbf{Z}\} = 0], \\ \Lambda^{\perp} &= [\mathbf{h}(Y, W, \mathbf{Z}) : E\{\mathbf{h}(Y, W, \mathbf{Z}) \mid X, \mathbf{Z}\} = \mathbf{0} \text{ almost everywhere}]. \end{aligned}$$

The efficient score for $\boldsymbol{\theta}$ is the residual of its score vector $\mathbf{S}_{\boldsymbol{\theta}}(y, w, \mathbf{z})$ after projecting it on to the nuisance tangent space Λ, denoted by

$$\mathbf{S}_{\mathrm{res}}(y, w, \mathbf{z}, \boldsymbol{\theta}) \equiv \mathbf{S}_{\boldsymbol{\theta}}(y, w, \mathbf{z}, \boldsymbol{\theta}) - \Pi\{\mathbf{S}_{\boldsymbol{\theta}}(Y, W, \mathbf{Z}, \boldsymbol{\theta})|\Lambda\},$$

where

$$\mathbf{S}_{\boldsymbol{\theta}}(y, w, \mathbf{z}, \boldsymbol{\theta}) \equiv \frac{\partial \log f_{W,Y|\mathbf{Z}}(y, w, \mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Here "$_{res}$" stands for residual. The detailed form of $\mathbf{S}_{res}(y, w, \mathbf{z}, \boldsymbol{\theta})$ is given as

$$\mathbf{S}_{\mathrm{res}}(Y, W, \mathbf{Z}, \boldsymbol{\theta}) = \mathbf{S}_{\boldsymbol{\theta}}(Y, W, \mathbf{Z}, \boldsymbol{\theta}) - E\{\mathbf{a}(X, \mathbf{Z}, \boldsymbol{\theta})|Y, W, \mathbf{Z}\}, \tag{3.3}$$

where $\mathbf{a}(X, \mathbf{Z}, \boldsymbol{\theta})$ satisfies

$$E\{\mathbf{S}_{\boldsymbol{\theta}}(Y, W, \mathbf{Z}, \boldsymbol{\theta}) \mid X, \mathbf{Z}\} = E[E\{\mathbf{a}(X, \mathbf{Z}, \boldsymbol{\theta})|Y, W, \mathbf{Z}\} \mid X, \mathbf{Z}]. \tag{3.4}$$

Now, noting that the above derivation is obtained from the approximate model (3.3), we hence perform some further analysis. Separating the components corresponding to $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in $\boldsymbol{\theta}$, we can write the $\mathbf{S}_{\boldsymbol{\theta}}(y, w, z, \boldsymbol{\theta}) \equiv \{\mathbf{S}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(y, w, z, \boldsymbol{\theta})^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}}(y, w, z, \boldsymbol{\theta})^{\mathrm{T}}\}^{\mathrm{T}}$, which leads to the corresponding relation $\mathbf{S}_{res}(y, w, \mathbf{z}, \boldsymbol{\theta}) \equiv \{\mathbf{S}_{res1}(y, w, \mathbf{z}, \boldsymbol{\theta})^{\mathrm{T}}, \mathbf{S}_{\mathrm{res}2}(y, w, \mathbf{z}, \boldsymbol{\theta})^{\mathrm{T}}\}^{\mathrm{T}}$. The estimating equation of the approximate model can be written as

$$\sum_{i=1}^{n} \mathbf{S}_{\mathrm{res}}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \{\mathbf{S}_{res1}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta})^{\mathrm{T}}, \mathbf{S}_{\mathrm{res}2}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta})^{\mathrm{T}}\}^{\mathrm{T}} = \mathbf{0}. \tag{3.5}$$

Remember that our original model contains an unknown function $g(z)$. Thus, for the estimation of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, it is beneficial to treat g as a nuisance parameter as well first, and estimate $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using profiling. We then plug in the estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and estimate g via the B-spline approximation. Of course in addition to g, the distribution of the unobservable covariate conditional on the observable covariate $\mathbf{Z}$ is also a nuisance component and still has to be taken into account.

Let $\boldsymbol{\delta} \equiv (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ be a p-dimensional parameter. We propose to solve for $\boldsymbol{\gamma}$ from $\sum_{i=1}^{n} \mathbf{S}_{\mathrm{res}2}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \mathbf{0}$ to obtain $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\delta})$ first. Now from

$$f_{W,Y|\mathbf{Z}}(w, \mathbf{z}, y, \boldsymbol{\delta}, g, f_X) = \int f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + g(x), \boldsymbol{\alpha}\} f_{W|X}(w, x) f_{X|\mathbf{Z}}(x, \mathbf{z}) d\mu(x).$$

we can construct the nuisance tangent space as $\Lambda = \Lambda_{f_X} + \Lambda_g$, where

$$\begin{aligned} \Lambda_{f_X} &= [E\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\} : E\{\mathbf{a}(X, \mathbf{Z}) \mid \mathbf{Z}\} = \mathbf{0}] \\ \Lambda_g &= \left(E\left[s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(X), \boldsymbol{\alpha}\}\mathbf{b}(X)|Y, W, \mathbf{Z}\right] : \forall \mathbf{b}(X)\right), \end{aligned}$$

where $s(y, t, \boldsymbol{\alpha}) \equiv \partial \log f(y, t, \boldsymbol{\alpha})/\partial t$. Note that Λ_{f_X} and Λ_g are not orthogonal to each other. We can further verify that

$$\begin{aligned}
\Lambda_{f_X}^{\perp} &= [\mathbf{h}(Y,W,\mathbf{Z}) : E\{\mathbf{h}(Y,W,\mathbf{Z}) \mid X,\mathbf{Z}\} = \mathbf{0} \text{ almost everywhere}], \\
\Lambda_{g}^{\perp} &= \Big(\mathbf{h}(Y,W,\mathbf{Z}) : E\left[\mathbf{h}(Y,W,\mathbf{Z}) s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\} \mid X,\mathbf{Z}\right] = \mathbf{0} \\
&\quad \text{almost everywhere}\Big).
\end{aligned}$$

The efficient score for $\boldsymbol{\delta}$ is now the residual of the score vector $\mathbf{S}_{\boldsymbol{\delta}}$ after projecting it on to the nuisance tangent space Λ, denoted by

$$\mathbf{S}_{eff}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) = \mathbf{S}_{\boldsymbol{\delta}}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) - \Pi\{\mathbf{S}_{\boldsymbol{\delta}}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) \mid \Lambda\}. \tag{3.6}$$

Its explicit form is given as

$$\begin{aligned}
\mathbf{S}_{eff}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) &= \mathbf{S}_{\boldsymbol{\delta}}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) - E\{\mathbf{a}(X,\mathbf{Z})|Y,W,\mathbf{Z}\} \\
&\quad - E\left[s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\}\mathbf{b}(X)|Y,W,\mathbf{Z}\right],
\end{aligned}$$

where $\mathbf{a}(X,\mathbf{Z})$ and $\mathbf{b}(X)$ satisfy

$$\begin{aligned}
& E\{\mathbf{S}_{\boldsymbol{\delta}}(Y,W,\mathbf{Z},\boldsymbol{\delta},g) \mid X,\mathbf{Z}\} \\
=\; & E[E\{\mathbf{a}(X,\mathbf{Z})|Y,W,\mathbf{Z}\} \mid X,\mathbf{Z}] \\
& +E(E[s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\}\mathbf{b}(X)|Y,W,\mathbf{Z}] \mid X,\mathbf{Z}) \\
& \text{and} \\
& E[\mathbf{S}_{\boldsymbol{\delta}}(Y,W,\mathbf{Z},\boldsymbol{\delta},g)s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\} \mid X,\mathbf{Z}] \\
=\; & E[E\{\mathbf{a}(X,\mathbf{Z})|Y,W,\mathbf{Z}\}s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\} \mid X,\mathbf{Z}] \\
& +E(E[s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\}\mathbf{b}(X)|Y,W,\mathbf{Z}] \\
& \times s\{Y,\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}+g(X),\boldsymbol{\alpha}\} \mid X,\mathbf{Z}).
\end{aligned} \tag{3.7}$$

We can then form the estimating quation $\sum_{i=1}^{n}\mathbf{S}_{eff}\{Y_i,W_i,\mathbf{Z}_i,\boldsymbol{\delta},\widehat{\gamma}(\boldsymbol{\delta})\} = \mathbf{0}$ to solve for $\widehat{\boldsymbol{\delta}}$ as the estimator, where $\mathbf{a}(X,\mathbf{Z}),\mathbf{b}(X)$ are the solutions to the integral equations (3.7).

3.2.3 Estimation under working model

The above derivations are based on efficient score calculation and hence will yield the efficient estimator. However, a close look at the procedure reveals that the procedure is not practical because the implementation relies on the unknown function $f_{X|\mathbf{Z}}(x, \mathbf{z})$. Thus, our estimator needs to be calculated under a posited working model of $\mathbf{f}^*_{X|\mathbf{Z}}(x, \mathbf{z})$. The procedure is described below, where we use * to denote a quantity whose calculation is carried out using $\mathbf{f}^*_{X|\mathbf{Z}}(x, \mathbf{z})$ instead of $\mathbf{f}_{X|\mathbf{Z}}(x, \mathbf{z})$.

1. Posit a working model $\mathbf{f}^*_{X|\mathbf{Z}}(x, \mathbf{z})$.
2. Solving for $\boldsymbol{\gamma}$ from $\sum_{i=1}^n \mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) = \mathbf{0}$ to obtain $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\delta})$.
3. Calculate the score function $\mathbf{S}^*_{\boldsymbol{\delta}}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, g)$ under the working model $\mathbf{f}^*_{X|\mathbf{Z}}(x, \mathbf{z})$.
4. Solve the integral equation (3.7) to get $\mathbf{a}(X, \mathbf{Z})$ and $\mathbf{b}(X)$.
5. Calculate the approximate efficient score function $\mathbf{S}^*_{\mathrm{e}ff}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, \widehat{g})$ following (3.6), where $\widehat{g}(\cdot) = \mathbf{B}(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}(\boldsymbol{\delta})$.
6. Solve the estimating equation $\sum_{i=1}^n \mathbf{S}^*_{\mathrm{e}ff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{g}) = \mathbf{0}$ to obtain $\widehat{\boldsymbol{\delta}}$.

When we calculate $\mathbf{a}(X, \mathbf{Z})$ at each observed $\mathbf{z}$ value and calculate $\mathbf{b}(\mathbf{X})$, we discretize the distribution of X on m equally spaced points on the support of $f_{X|\mathbf{Z}}(x, \mathbf{z})$ and calculate the probability mass function $\pi_j(\mathbf{Z})$ at each of the m points. We of course normalize the $\pi_j(\mathbf{Z})$ in order to ensure $\sum_{j=1}^m \pi_j(\mathbf{Z}) = 1$. Note that using the descritization, $f^*_{X,Y,W|\mathbf{Z}}(x_j, y, w, \mathbf{z}) \approx f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_j), \boldsymbol{\alpha}\} f_{W|X=x_j}(w, x_j)\pi_j(\mathbf{Z})$. Further, $\mathbf{S}^*_{\boldsymbol{\delta}}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, g)$, $E^*\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\}$ and $E^*[s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(X, \boldsymbol{\delta}), \boldsymbol{\alpha}\}\mathbf{b}(X)|Y, W, \mathbf{Z}]$

can be approximated by

$$
\begin{aligned}
&\mathbf{S}^*_{\boldsymbol{\delta}}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, g) \\
\approx\ & \frac{\partial \log[\sum_{i=1}^m f\{y, \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_i), \boldsymbol{\alpha}\} f_{W|X}(w, x_i)\pi_i(\mathbf{Z})]}{\partial \boldsymbol{\delta}}, \\
&E^*\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\} \\
\approx\ & \frac{\sum_{i=1}^m \mathbf{a}(x_i, \mathbf{Z}) f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})}, \\
&E^*[s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(X), \boldsymbol{\alpha}\}\mathbf{b}(X)|Y, W, \mathbf{Z}] \\
\approx\ & \frac{\sum_{i=1}^m s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_i), \boldsymbol{\alpha}\}\mathbf{b}(x_i) f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})}.
\end{aligned}
$$

Let $\mathbf{A}(X, \mathbf{Z}) \equiv \{\mathbf{a}(x_1, \mathbf{Z}), \mathbf{a}(x_2, \mathbf{Z}), \ldots, \mathbf{a}(x_m, \mathbf{Z})\}^{\mathrm{T}}$ and $\mathbf{B}(X) \equiv \{\mathbf{b}(x_1), \mathbf{b}(x_2), \ldots, \mathbf{b}(x_m)\}^{\mathrm{T}}$. Let $\mathcal{M}_1(X, \mathbf{Z}) \equiv \{\mathbf{m}_1(x_1, \mathbf{Z}), \mathbf{m}_1(x_2, \mathbf{Z}), \ldots, \mathbf{m}_1(x_m, \mathbf{Z})\}^{\mathrm{T}}$ be a $m \times p_{\boldsymbol{\delta}}$ matrix, where $p_{\boldsymbol{\delta}}$ is the length of $\boldsymbol{\delta}$ and $\mathbf{m}_1(x_i, \mathbf{Z}) \equiv E\{\mathbf{S}^*_{\boldsymbol{\delta}}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, g) \mid x_i, \mathbf{Z}\}$. Further, let $\mathcal{M}_2(X, \mathbf{Z}) \equiv \{\mathbf{m}_2(x_1, \mathbf{Z}), \mathbf{m}_2(x_2, \mathbf{Z}), \ldots, \mathbf{m}_2(x_m, \mathbf{Z})\}^{\mathrm{T}}$ be a $m \times p_{\boldsymbol{\delta}}$ matrix, where $\mathbf{m}_2(x_i, \mathbf{Z}) \equiv E\left[\mathbf{S}^*_{\boldsymbol{\delta}}(Y, W, \mathbf{Z}, \boldsymbol{\delta}, g) s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_i)\} \mid x_i, \mathbf{Z}\right]$. Finally, let $\mathbf{C}(X, \mathbf{Z})$ be a $m \times m$ matrix with the (i, j) block equal to

$$
E\left\{\frac{f^*_{X,Y,W|\mathbf{Z}}(x_j, Y, W, \mathbf{Z})}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})} \mid x_i, \mathbf{Z}\right\},
$$

let $\mathbf{D}(X, \mathbf{Z})$ be an $m \times m$ matrix with the (i, j) block equal to

$$
E\left[\frac{s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_j), \boldsymbol{\alpha}\} f^*_{X,Y,W|\mathbf{Z}}(x_j, Y, W, \mathbf{Z})}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})} \mid x_i, \mathbf{Z}\right],
$$

let $\mathbf{F}(X, \mathbf{Z})$ be an $m \times m$ matrix with the (i, j) block equal to

$$
E\left[\frac{f^*_{X,Y,W|\mathbf{Z}}(x_j, Y, W, \mathbf{Z}) s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_i)\}}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})} \mid x_i, \mathbf{Z}\right],
$$

and let $\mathbf{G}(X, \mathbf{Z})$ be an $m \times m$ matrix with the (i, j) block

$$
E\left[\frac{s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_j), \boldsymbol{\alpha}\} f^*_{X,Y,W|\mathbf{Z}}(x_j, Y, W, \mathbf{Z}) s\{Y, \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta} + g(x_i)\}}{\sum_{i=1}^m f^*_{X,Y,W|\mathbf{Z}}(x_i, Y, W, \mathbf{Z})} \mid x_i, \mathbf{Z}\right].
$$

We can get $\mathbf{a}(x_i, \mathbf{Z})$ and $\mathbf{b}(x_i)$ by solving

$$
\begin{bmatrix} \mathbf{C}(X, \mathbf{Z}) & \mathbf{D}(X, \mathbf{Z}) \\ \mathbf{F}(X, \mathbf{Z}) & \mathbf{G}(X, \mathbf{Z}) \end{bmatrix} \begin{bmatrix} \mathbf{A}(X, \mathbf{Z}) \\ \mathbf{B}(X) \end{bmatrix} = \begin{bmatrix} \mathcal{M}_1(X, \mathbf{Z}) \\ \mathcal{M}_2(X, \mathbf{Z}) \end{bmatrix}.
$$

3.3 Asymptotic properties

Let $\mathbf{S}_{res2}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, g)$ be $\mathbf{S}_{res2}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ with all the appearance of $\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}$ in it replaced by $g(X)$.

We first list the set of regularity conditions required for establishing the large sample properties of our estimator.

(C1) The true density $f_X(x)$ is bounded with compact support. Without loss of generality, we assume the support of $f_X(x)$ is $[0, 1]$.

(C2) The function $g(x) \in C^q([0, 1])$, $q > 1$, is bounded with compact support.

(C3) The spline order $r \geq q$.

(C4) Define the knots $t_{-r+1} = \cdots = t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1} = \cdots = t_{N+r}$, where N is the number of interior knots that satisfies $N \to \infty$, $N^{-1}n(\log n)^{-1} \to \infty$ and $Nn^{-1/(2q)} \to \infty$ as $n \to \infty$. Denote the number of spline bases d_γ, i.e. $d_\gamma = N + r$.

(C5) Let h_j be the distance between the jth and $(j-1)$th interior knots. Let $h_b = \max_{1\leq j\leq N} h_j$ and $h_s = \min_{1\leq j\leq N} h_j$. There exists a constant $c_h \in (0, \infty)$ such that $h_b/h_s < c_h$. Hence, $h_b = O_p(N^{-1})$ and $h_s = O_p(N^{-1})$.

(C6) $\boldsymbol{\gamma}_0$ is a d_γ-dimensional spline coefficient vector such that $\sup_{x\in[0,1]} |\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}_0 - g(x)| = O_p(h_b^q)$.

(C7) The equation set

$$\begin{aligned} E\{\mathbf{S}_{eff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \boldsymbol{\gamma})\} &= \mathbf{0}, \\ E\{\mathbf{S}_{res2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \boldsymbol{\gamma})\} &= \mathbf{0} \end{aligned}$$

has unique root for $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_0$. Recall that $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. The derivatives with respect to $\boldsymbol{\theta}$ of the left hand side are smooth functions of

$\boldsymbol{\theta}$, with its singular values bounded and bounded away from $\mathbf{0}$. Let the unique root be $\boldsymbol{\theta}^*$. Note that $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}^*$ are functions of N, that is, for any sufficiently large N, there is a unique root $\boldsymbol{\theta}^*$ in the neighborhood of $\boldsymbol{\theta}_0$.

(C8) The maximum absolute row sum of the matrix $\partial \mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma}_0^{\mathrm{T}}$, i.e. $\|\partial \mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma}_0^{\mathrm{T}}\|_\infty$, is integrable.

The conditions listed above are all standard bounded, smoothness conditions on functions and some classical conditions imposed on the spline order and number of knots. These are commonly used conditions in spline approximation and semiparametric regression literature. We now establish the consistency of $\widehat{\boldsymbol{\delta}}_n$ and $\widehat{\boldsymbol{\gamma}}_n$ as well as the asymptotic distribution property of $\widehat{\boldsymbol{\delta}}_n$.

Theorem 3. *Assume Conditions* (C1) – (C7) *to hold. Let* $\widehat{\boldsymbol{\theta}}_n$ *satisfy*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \boldsymbol{S}^*_{\mathrm{eff}}(Y_i, W_i, \boldsymbol{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\boldsymbol{\gamma}}_n) &= \boldsymbol{0} \\ \frac{1}{n}\sum_{i=1}^n \boldsymbol{S}_{\mathrm{res}_2}^*(Y_i, W_i, \boldsymbol{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\boldsymbol{\gamma}}_n) &= \boldsymbol{0}.\end{aligned}$$

Then $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_p(1)$ *element-wise.*

The result in Theorem 3 is used to further establish the asymptotic properties of the estimator of the parameters of interest $\widehat{\boldsymbol{\delta}}_n$ and estimator of the function of interest $\mathbf{B}(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_n$.

Theorem 4. *Assume Conditions* (C1) – (C8) *to hold and let*

$$\boldsymbol{Q} \equiv E\left\{ \frac{\partial \boldsymbol{S}^*_{\mathrm{eff}}(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\delta}_0^{\mathrm{T}}} \Bigg|_{\boldsymbol{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma} = g(\cdot)} \right\}.$$

Then

$$\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = -\boldsymbol{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{S}^*_{\mathrm{eff}}(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, g) + o_p(1).$$

Consequently, $\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \to N(\boldsymbol{0}, \boldsymbol{V})$ *in distribution when* $n \to \infty$*, where*

$$\boldsymbol{V} = \boldsymbol{Q}^{-1} var\{\boldsymbol{S}^*_{\mathrm{eff}}(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, g)\}(\boldsymbol{Q}^{-1})^{\mathrm{T}}.$$

Theorem 4 indicates that $\boldsymbol{\delta}$ is estimated at the root-n rate. The proofs of Theorems 3 and 4 are given in the Appendix. Because the B-spline estimation of $g(\cdot)$ is at a slower rate than root-n, the estimation of $\boldsymbol{\delta}$ does not have any impact on the first order asymptotic properties of $\widehat{g}$. Thus, for the analysis of the asymptotic properties of $\widehat{g}$, we can treat $\boldsymbol{\delta}$ as known. Then, the proof of Theorem 2 in Jiang & Ma (2017) can be directly used. We skip the details of the proof and provide the specific convergence property of the estimation of g in Theorem 5.

Theorem 5. *Assume Conditions* (C1) – (C8) *to hold and let*

$$\boldsymbol{P} \equiv E\left\{ \frac{\partial \boldsymbol{S}_{\mathrm{res}2}^*(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \bigg|_{\boldsymbol{B}(\cdot)^{\mathrm{T}}\gamma = g(\cdot)} \right\}.$$

Then $\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|_2 = O_p\{(nh_b)^{-1/2}\}$. *Further,*

$$\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0 = -\boldsymbol{P}^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_{\mathrm{res}2}^*(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})\{1 + o_p(1)\}.$$

This leads to that $\widehat{g}(x)$*, which equals* $\boldsymbol{B}(x)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_n$*, satisfies*

$$\sup_{x \in [0,1]} |\widehat{g}(x) - g(x)| = O_p\{(nh_b)^{-1/2}\}.$$

Specifically, $bias\{\widehat{g}(x)\} = E\{\widehat{g}(x) - g(x)\} = O(h_b^{q-1/2})$ *and*

$$\begin{aligned} & \sqrt{nh_b}[\widehat{g}(x) - g(x) - bias\{\widehat{g}(x)\}] \\ = \; & \sqrt{nh_b}\boldsymbol{B}(x)^{\mathrm{T}} \left\{ -\boldsymbol{P}^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_{\mathrm{res}2}^*(Y_i, W_i, \boldsymbol{Z}_i, \boldsymbol{\delta}_0, g) \right\} + o_p(1). \end{aligned}$$

3.4 Numerical Study

In our first simulation, we generated the observations $(W_i, \mathbf{Z}_i, Y_i)$ from the model

$$\mathrm{pr}(Y_i = 1 | X_i = x_i, Z_i = z_i) = H\{g(x_i) + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i}\}, \qquad (3.8)$$

where $W = X + U$ and $U = \mathrm{n}ormal(0, 0.03)$. The true function is: $g(x) = -5\exp\{-0.8(x - 2.5)^2\}$ and $H(t)$ is the inverse logit link function. We set $\beta_1 = 1$,

$\beta_2 = 0.5$, $\beta_3 = 1$ and $\beta_4 = -0.3$. The sample size is 1000 and we ran 1000 simulations. X_i is generated from a truncated normal distribution with mean 0.5 and variance 1/36 on [0,1] independently of $\mathbf{Z}_i$. We implemented our method using a normal working model, corresponding to a correct working model case. In order to investigate the performance of our method under a misspecified working model, we also performed another study, in which we have X_i generated from a truncated student-t distribution with degrees of freedom 5. Covariates Z_{1i}, Z_{2i} and Z_{4i} are generated from the standard normal distribution. The covariate Z_{3i} is generated from a uniform distribution on $[-1, 1]$. In both studies, we estimated both the parameters β_1, β_2, β_3, β_4 and the function $g(x)$.

In the second simulation, we set the true g function to be $g(x) = -5exp(-0.2x^2)+5$, while all other settings remain the same. Similarly to the first simulation, we compared the performance of a correct working model and a misspecified working model in terms of estimating both β_1, β_2, β_3, β_4 and $g(x)$.

In both simulations 1 and 2, we discretized the distribution of X on $[0, 1]$ to $m = 15$ equal segments and we use the truncated normal distribution discussed earlier as our working model. We used quadratic splines with 7 knots to estimate $g(x)$. The simulation results are shown in Tables D.1, D.2 and Figures D.1, D.2.

The results in Tables D.1 and D.2 show little bias for the $\boldsymbol{\beta}$ estimation, regardless a correct working model or a misspecified working model is used. Figures D.1 and D.2 show that the estimators of $g(x)$ have somewhat large bias on the boundary in both methods, which is within our expectation when factoring in the boundary effect. The performance of $g(x)$ estimation is satisfactory in the interior of the function domain. The simulation results show no big difference between the performance of the correct working model of $f_X(x)$ and a misspecified one, confirming our theory on consistency in both cases.

3.5 Data Analysis

The data set we analyzed is from an AIDS Clinical Trials Group (ACTG) study. The goal of this study is to compare four different treatments, 'ZDV', 'ZDV+ddI', 'ZDV+ddC' and 'ddC', on HIV infected adults whose CD4 cell counts were from 200 to 500 per cubic millimeter. We labelled those treatments as treatment 1, treatment 2, treatment 3 and treatment 4. We used treatment 1 as the base treatment because it is a standard treatment. There were 1036 patients enrolled in the study and they had no antiretroviral therapy at enrollment. The criteria that we used to compare the four treatments is whether a patient has his or her CD4 count drop below 50%, which is an important indicator for HIV infected patients to develop AIDS or die. We have $Y = 1$ if a patient has his or her CD4 count drop below 50%, and $Y = 0$ otherwise.

Our model has the form:

$$\mathrm{pr}(Y_i = 1|X_i = x_i, Z_i = z_i) = H\{g(x_i) + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i}\}, \tag{3.9}$$

where $W = X + U$ and $U = \mathrm{n}ormal(0, \sigma_U^2)$. The covariates Z_1, Z_2, and Z_3 are dichotomous variables. $Z_{1i} = Z_{2i} = Z_{3i} = 0$ indicates the ith individual receives treatment 1, the base treatment; $Z_{1i} = 1$ and $Z_{2i} = Z_{3i} = 0$ indicates the ith individual receives treatment 2; $Z_{1i} = 0$, $Z_{2i} = 1$ and $Z_{3i} = 0$ indicates the ith individual receives treatment 3; $Z_{1i} = Z_{2i} = 0$ and $Z_{3i} = 1$ indicates the ith individual receives treatment 4. The covariate X is the baseline log(CD4 count) prior to the start of treatment. Because CD4 count cannot be measured precisely, X is considered as our unobservable covariate. We use the average of two available measurements of log(CD4 count) as W.

First, we estimated the variance of U using the two repeated measurements and we got $\widehat{\sigma}_U^2 = 0.3$. Then, we constructed our working model of unobservable variance X. We assume that X follows a truncated normal distribution and estimated its variance by $\widehat{\sigma}_X^2 = \widehat{\sigma}_W^2 - \widehat{\sigma}_U^2$.

Table D.3 shows that Treatment 2, Treatment 3 and Treatment 4 are more efficient than the baseline treatment, i.e. Treatment 1, at 90% confidence level according to the P-values of β_1, β_2 and β_3. The estimated index function $g(x)$ is in Figure D.3. We generated 1000 bootstrapped samples and calculated the bootstrapped mean, median and 90% confidence band for $g(x)$. It shows that $g(x)$ is an decreasing function, indicating that a large baseline CD4 count leads to a smaller risk of developing AIDS or having his/her CD4 counts drop below 50%. Thus, our analysis indicates that in general, the alternative treatments and a higher baseline CD4 count are beneficial to a patient.

3.6 Discussion

We devised a consistent and locally efficient estimation procedure to estimate both parameters and functions in a generalized partially linear model where the covariate inside the nonparametric function is subject to measurement error. The method does not make any assumption on the distribution of the covariate measured with error other than its finite support, which is easily satisfied in practice. The method is efficient in terms of estimating the model parameters if a correct working model is used, and retains its consistency even if this working model is misspecified. The estimation procedure breaks free from the deconvolution approach, which is the norm of practice in handling nonparametric problems with measurement errors. Instead, a novel usage of B-spline approach in combination with semiparametric method is exploited to push through the analysis.

Many possible extensions can be explored further. Possibilities include handling multi-variate covariates measured with error, via multivariate B-splines, or incorporating index modeling approach or additive structures. Although our method is developed conceptually for generalized linear models, we did not really make use of the linear structure, hence any model of the form $f(Y, g(X), \mathbf{Z}, \boldsymbol{\beta})$ can be treated in a similar way. To this end continuous Y typically involves normal error and has been widely studied, while binary response is studied in the main text of this work. When Y is count data, many computational issues arise, and is worth careful investigation further.

We have assumed the measurement error U to either have a known distribution, or to have its model parameters estimable from multiple observations. Of course, any other available information to identify the measurement error distributional model parameter also works and the plug-in procedure is largely "blind" to how the parameter is estimated. Of course, the estimated distributional model parameter will alter the estimation variability of $\boldsymbol{\delta}$, which can be take into account in a standard way (Yi et al. 2015).

Bibliography

Altstein, L. & Li, G. (2013), 'Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model', *Biometrics* **6**9, 52–61.

Buonaccorsi, J. P. (2010), 'Measurement error: Models, methods, and applications', *New York: Chapman & Hall/CRC.* .

Buzas, J. S. & Stefanski, L. A. (1996), 'Instrumental variable estimation in generalized linear measurement error models.', *Journal of the American Statistical Association* **9**1, 999–1006.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1995), 'Generalized partially linear single-index models', *Journal of the American Statistical Association* **9**2, 477–489.

Carroll, R. J., R. D. S. L. A. & Crainiceanu, C. (2006), 'Measurement error in non-linear models: A modern perspective (2nd ed.)', *London: CRC Press* .

Dorsey, E. R. & The Huntington Study Group COHORT Investigators (2012), 'Characterization of a large group of individuals with huntington disease and their relatives enrolled in the cohort study', *PLos One* **7**, e29522.

Foroud, T., Gary, J., Ivashina, J. & Conneally, P. (1999), 'Differences in duration of huntington's disease based on age at onset', *Journal of Neurology, Neurosurgery and Psychiatry* **6**6, 52–56.

Holzmann, H., Munk, A. & Stratmann, B. (2006), 'Identifiability of finite mixtures of elliptical distributions', *Scandinavian Journal of Statistics* **3**3, 753–763.

Huang, Y. & Wang, C. Y. (2001), 'Consistent functional methods for logistic regression with errors in covariates', *Journal of the American Statistical Association* **9**6, 1469–1482.

Jiang, F. & Ma, Y. (2017), 'A spline-assisted semiparametric approach to nonparametric measurement error models', *COBRA Preprint Series* .

Khoury, M. J., Beaty, T. H. & Cohen, B. H. (1993), *Fundamentals of Genetic Epidemiology*, Oxford University Press, New York, NY.

Kress, R. (1991), *Linear Integral Equations*, Springer, New York.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S. & Hayden, M. R. (2004), 'A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length', *Clinical genetics* **6**5(4), 267–277.

Liang, H., Qin, Y., Zhang, X. & Ruppert, D. (2009), 'Empirical likelihood-based inferences for generalized partially linear models', *Scand Stat Theory Appl* **3**6, 433–443.

Liang, H. & Ren, H. (2005), 'Generalized partially linear measurement error models', *Journal of Computational and Graphical Statistics* **1**4, 237–250.

Liang, H. & Thurston, S. W. (2008), 'Additive partial linear models with measurement errors', *Biometrika* **9**5, 667–678.

Liu, J., Ma, Y., Zhu, L. & Carroll, R. J. (2017), 'Estimation and inference of error-prone covariate effect in the presence of confounding variables', *Electronic Journal of Statistics* **1**1, 480–501.

Liu, L. (2007), 'Estimation of generalized partially linear models with measurement error using sufficiency scores', *Statistics & Probability Letters* **7**7, 1580–1588.

Ma, Y. & Carroll, R. J. (2006), 'Locally efficient estimators for semiparametric models with measurement error', *Journal of the American Statistical Association* **1**01, 1465–1474.

Ma, Y. & Tsiatis, A. A. (2006), 'Closed form semiparametric estimators for measurement error models', *Statistica Sinica* **1**6, 183–193.

Ma, Y. & Wang, Y. (2014), 'Estimating disease onset distribution functions in mutation carriers with censored mixture data', *Journal of the Royal Statistical Society, Series C* **6**3, 1–23.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L. et al. (1993), 'A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes', *Cell* **7**2(6), 971–983.

Marder, L., Sulzbach, G. O., Bernardes, A. M. & Ferreira, J. Z. (2003), 'Removal of cadmium and cyanide from aqueous solutions through electrodialysis', *Journal of the Brazilian Chemical Society* **1**4, 610–615.

Stefanski, L. A. & Carroll, R. J. (1985), 'Covariate measurement error in logistic regression', *Annals of Statistics* **1**3, 1335–1351.

Stefanski, L. A. & Carroll, R. J. (1987), 'Conditional scores and optimal scores in gener alized linear measurement error models', *Biometrika* **7**4, 703–716.

Tsiatis, A. & Ma, Y. (2004), 'Locally efficient semiparametric estimators for functional measurement error models', *Biometrika* **9**1, 835–848.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. & Tucker, M. (1998), 'The kin-cohort study for estimationg penetrance', *American Journal of Epidemiology* **1**48, 623–630.

Wang, Y., Garcia, T. P. & Ma, Y. (2012), 'Nonparametric estimation for censored mixture data with application to the cooperative huntington's observational research trial', *Journal of the American Statistical Association* **1**07, 1324–1338.

Xu, K., Ma, Y. & Wang, L. (2015), 'Instrument assisted regression for errors in variables models with binary response', *Scandinavian Journal of Statistics* **4**2, 104–117.

Yi, G., Ma, Y., Spiegelman, D. & Carroll, R. J. (2015), 'Functional and structural methods with mixed measurement error and misclassification in covariates', *Journal of the American Statistical Association* **1**10, 681–696.

Yu, Y. & Ruppert, D. (2012), 'Penalized spline estimation for partially linear single-index models', *Journal of the American Statistical Association* **9**7, 1042–1054.

Zeng, D. & Lin, D. Y. (2006), 'Efficient estimation of semiparametric transformation models for counting process', *Biometrika* **9**3, 627–640.

Zeng, D. & Lin, D. Y. (2007), 'Maximum likelihood estimation in semiparametric regression models with censored data', *Journal of the Royal Statistical Society* **6**9, 507–564.

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W. paulsen, J. S. & the PREDICT-HD Investigators and Coordinators of the Huntington Study Group (2012), 'Indexing disease progression at study entry with individuals at-risk for huntington disease', *American journal of medical genetics: neuropsychiatric genetics* .

Appendix A

Permission to Reprint

Dear Dr. Wang,

In terms of the permission, we do not have the exclusive copyright of your paper. You may use your article at will when it is non-commercial. Please let me know if I have misunderstood your request. Thank you.

Sincerely yours,

Anna Chiang

Statistica Sinica

Editorial Assistant

Institute of Statistical Science,

Academia Sinica 128, Academia Road Section 2

Taipei 115, Taiwan

Tel: 886-2-2783-5611

E-mail: ss.office@stat.sinica.edu.tw

Website: http://www3.stat.sinica.edu.tw/statistica/

Appendix B

Chapter 1 Appendix

B.1 Derivation of $\int a_{ij}^{(u)} dr_{ij}$ and $\int r_{ij} a_{ij}^{(u)} dr_{ij}$

Here we show the derivation of the relationships

$$\begin{aligned}
\int a_{ij}^{(u)} dr_{ij} &= q_{ij} \left\{1 - F_j(t)\right\}^{1-\delta_i} \left\{h^{(u)}(y_i) f_j(t)\right\}^{\delta_i} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \\
\int r_{ij} a_{ij}^{(u)} dr_{ij} &= e^{-t} q_{ij} f_j(t) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \qquad \text{if } \delta_i = 0 \\
\int r_{ij} a_{ij}^{(u)} dr_{ij} &= -e^{-t} q_{ij} h^{(u)}(y_i) \{f_j'(t) - f_j(t)\} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \qquad \text{if } \delta_i = 1.
\end{aligned}$$

Let $t = H^{(u)}(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}$. Then

$$a_{ij}^{(u)} = \{h^{(u)}(y_i) r_{ij} \exp(t)\}^{\delta_i} \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}.$$

When $\delta_i = 0$,

$$\begin{aligned}
\frac{da_{ij}^{(u)}}{dt} &= -r_{ij} e^t \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \\
\frac{d^2 a_{ij}^{(u)}}{dt^2} &= -r_{ij} e^t \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \\
&\quad + r_{ij}^2 e^{2t} \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}.
\end{aligned}$$

When $\delta_i = 1$,

$$\begin{aligned}
\frac{da_{ij}^{(u)}}{dt} &= -h^{(u)}(y_i) r_{ij}^2 e^{2t} \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \\
&\quad + h^{(u)}(y_i) r_{ij} e^t \exp(-r_{ij} e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}
\end{aligned}$$

Thus, when $\delta_i = 0$,

$$r_{ij} a_{ij}^{(u)} = -e^{-t} \frac{da_{ij}^{(u)}}{dt} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}, \delta_i = 0},$$

and when $\delta_i = 1$,

$$r_{ij} a_{ij}^{(u)} = h^{(u)}(y_i) e^{-t} \left(\frac{d^2 a_{ij}^{(u)}}{dt^2} - \frac{d a_{ij}^{(u)}}{dt} \right) \Big|_{t = H^{(u)}(y_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\alpha}_j^{(u)}, \delta_i = 0}.$$

B.2 List of regularity conditions

(a) The parameter value $\boldsymbol{\theta}_0$ belongs to the interior of a compact set $\boldsymbol{\Theta} \in \mathbb{R}^{d_\theta}$, and $\phi_0(t) > 0$ for all $t \in [0, \tau]$. (C1).

(b) With probability 1, $\mathrm{pr}(Y_i \geq \tau \mid \mathbf{X}_i, \mathbf{Z}_i) > \delta_0 > 0$ for some constant $\delta_0 > 0$. (C2).

(c) $f_j(s)$ is bounded away from zero and infinity on its support for $j = 1, \ldots, p$. (C3).

(d) $f_j(s)$ is three times continuously differentiable and, the $f_j^{(v)}(s)/\exp(ks), v = 0, \ldots, 3, k = 2, \ldots, 4$, are square integrable on $(-\infty, \log(\tau)]$ for $j = 1, \ldots, p$. (C4), (C8).

(e) The covariates $\mathbf{X}, \mathbf{Z}$ have finite kth moments, $k = 1, \ldots, 6$. (C4), (C8).

(f) The first moment of $\log f_j(s)$ exists for $j = 1, \ldots, p$. (C6).

(g) $m \geq p$ and the matrix $(\mathbf{u}_1, \ldots \mathbf{u}_m)$ has rank p. (C5), (C7).

B.3 Proof of Theorem 1

Because NPMLE for the linear transformation model in the mixture model setting we consider can be cast into the general framework established in Zeng & Lin (2007), we prove Theorem 1 through verifying the conditions (C1)-(C8) required by them.

Condition (a) ensures that the true parameter value is in the interior of a compact set of the parameter space, with Conditions (c) and (d), we further guarantee the differentiability and positivity of the hazard function. This leads to condition (C1) of Zeng & Lin (2007).

Condition (b) is equivalent to their (C2).

Condition (c) guarantees that (C3) of Zeng & Lin (2007) is satisfied.

Condition (C4) of Zeng & Lin (2007) is a type of Lipschitz condition, with respect to both parameter and function; It is guaranteed by the stronger differentiability conditions in our Condition (d) and and the moment conditions in (e).

Our Condition (g) is stated in their (C5).

Condition (C6) of Zeng & Lin (2007) requires sufficient smoothness and boundedness of the hazard functions and some functions derived from them, as do our Conditions (c), (d) and (f).

Condition (C7) there is an identifiability condition that arises due to the generality of the framework they consider; It is guaranteed to hold under our Condition (g) and the parameterization requiring $H(0) = -\infty$.

Condition (C8) of Zeng & Lin (2007) strengthen their (C4) to hold along each path in a neighborhood of the true parameter value, while our Conditions (d) and (e) are imposed for all the parameter values in a compact set jointly ensuring that this holds.

$\square$

B.4 Additional simulations

Our fourth simulation is the same as the first, except that the true transformation H is $\log\{t/(1-t)\}$. In this case, the overall censoring rate is about 25%. The results are in Table B.1 and Figure B.1.

Similarly, the fifth simulationis the same as the second but with $H = \log\{t/(1-t)\}$, and an overall censoring rate of about 20%. The results are in Table B.1 and Figure B.1.

The sixth simulation is the same as the third except that $H = \log\{t/(1-t)\}$, with an overall censoring rate of about 25%. The results are in Table B.1 and Figure B.1.

Table B.1: Simulation results. Results based on 1000 simulations.

	true	mean	median	sd	mean($\widehat{\mathrm{s}d}$)	median($\widehat{\mathrm{s}d}$)	95% CI
				simulation 4			
β_1	1.0000	0.9809	0.9776	0.4393	0.4605	0.4601	0.9650
β_2	2.0000	1.9693	1.9565	0.3974	0.4088	0.4084	0.9540
				simulation 5			
α_{11}	1.0000	0.9893	0.9986	0.6229	0.6363	0.6351	0.9590
α_{12}	2.0000	1.9895	1.9988	0.5339	0.5552	0.5535	0.9550
α_{21}	1.5000	1.4764	1.4410	1.1660	1.1346	1.1292	0.9530
α_{22}	3.0000	2.9565	2.9681	0.9947	0.9971	0.9933	0.9460
				simulation 6			
β_1	1.0000	0.9973	0.9914	0.2951	0.2982	0.2978	0.9590
β_2	1.5000	1.5038	1.4982	0.1551	0.1569	0.1567	0.9590
α_1	2.0000	1.8943	1.9186	0.7693	0.7955	0.7945	0.9510
α_2	3.0000	3.0311	3.0257	0.3728	0.3609	0.3595	0.9560

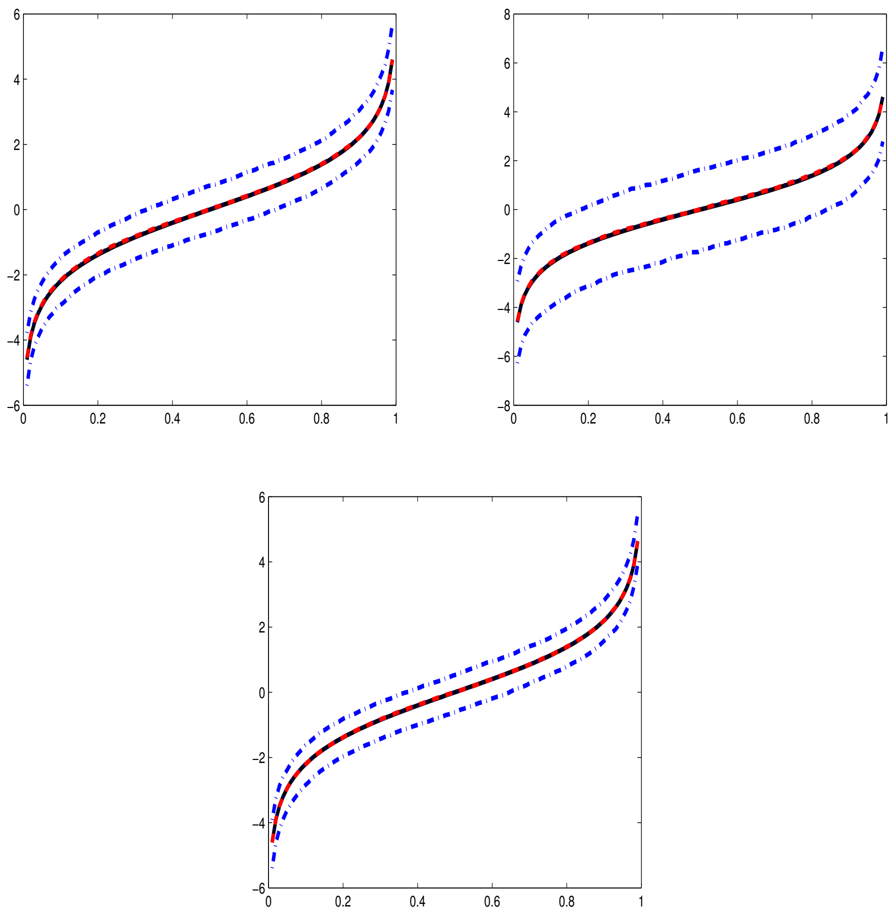

Figure B.1 True function (solid line), median estimation (dashed line), mean estimation (dotted line) and 95% confidence band (dash-dotted line) of $H(T)$ in simulations 4 (upper-left), 5 (upper-right), and 6 (lower) .

Appendix C

Chapter 2 Appendix

C.1 Proof of Thoerem 1

Let $\widehat{\boldsymbol\beta}$, $\widehat{\boldsymbol\gamma}$ and $\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol\gamma}, \widehat{\boldsymbol\beta})$, $i = 1, \ldots, n$ solve

$$\mathbf{0} = n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \widehat{\boldsymbol\beta}, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol\gamma}, \widehat{\boldsymbol\beta})\}.$$

Then we have

$$\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \widehat{\boldsymbol\beta}, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol\gamma}, \widehat{\boldsymbol\beta})\} \\
&= n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma, \boldsymbol\beta)\} + \Big[n^{-1}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma, \boldsymbol\beta)\}}{\partial\boldsymbol\beta^{\mathrm{T}}} \\
&\quad + n^{-1}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\}}{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}\frac{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}{\partial\boldsymbol\beta^{\mathrm{T}}}\Big]\Big|_{\boldsymbol\theta=\boldsymbol\theta^*}\sqrt{n}(\widehat{\boldsymbol\beta} - \boldsymbol\beta) \\
&\quad + n^{-1}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\}}{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}\frac{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}{\partial\boldsymbol\gamma^{\mathrm{T}}}\Big|_{\boldsymbol\theta=\boldsymbol\theta^*}\sqrt{n}(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma) \\
&= n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\} + \Big\{E\Big[\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\}}{\partial\boldsymbol\beta^{\mathrm{T}}} \\
&\quad + \frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\}}{\partial g(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}\frac{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}{\partial\boldsymbol\beta^{\mathrm{T}}}\Big] + o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol\beta} - \boldsymbol\beta) \\
&\quad + \Big\{E\Big[\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)\}}{\partial g(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}\frac{\partial\widehat{g}(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol\gamma; \boldsymbol\beta)}{\partial\boldsymbol\gamma^{\mathrm{T}}}\Big] + o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma) \\
&= n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, \widehat{g}(U_i; \boldsymbol\beta)\} + \Big\{E\Big[\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(U_i; \boldsymbol\beta)\}}{\partial\boldsymbol\beta^{\mathrm{T}}} \\
&\quad + \frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(U_i; \boldsymbol\beta)\}}{\partial g(U_i; \boldsymbol\beta)}\frac{\partial\widehat{g}(U_i; \boldsymbol\beta)}{\partial\boldsymbol\beta^{\mathrm{T}}}\Big] + o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol\beta} - \boldsymbol\beta) \\
&\quad + \Big\{E\Big[\frac{\partial\mathcal{L}\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol\beta, g(U_i; \boldsymbol\beta)\}}{\partial g(U_i; \boldsymbol\beta)}\frac{\partial\widehat{g}(U_i; \boldsymbol\beta)}{\partial\boldsymbol\gamma^{\mathrm{T}}}\Big] + o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma) \qquad \text{(C.1)}
\end{aligned}$$

where $U_i = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}$. Because of the estimation process, we have

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n}\left\{\begin{array}{c}1\\(U_i-u_0)/h\end{array}\right\}K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})\\&\quad+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\},\end{aligned}\tag{C.2}$$

at all $u_0 = \mathbf{z}_0^{\mathrm{T}}\boldsymbol{\gamma}$ and all parameter values of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^*$. Thus, we have

$$\begin{aligned}\mathbf{0} &= n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c}1\\(U_i-u_0)/h\end{array}\right\}K_h(U_i-u_0)\\&\quad\times\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta}^*,\widehat{g}(u_0,\boldsymbol{\beta}^*)+\widehat{g}'(u_0,\boldsymbol{\beta}^*)(U_i-u_0)\}}{\partial\boldsymbol{\beta}^{*\mathrm{T}}}\\&\quad+\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta}^*,\widehat{g}(u_0;\boldsymbol{\beta}^*)+\widehat{g}'(u_0,\boldsymbol{\beta}^*)(U_i-u_0)\}}{\partial\{\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}\frac{\partial\widehat{g}(u_0;\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^{*\mathrm{T}}}\\&\quad+\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta}^*,\widehat{g}(u_0,\boldsymbol{\beta}^*)+\widehat{g}'(u_0,\boldsymbol{\beta}^*)(U_i-u_0)\}}{\partial\}\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}\frac{\partial\widehat{g}'(u_0;\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^{*\mathrm{T}}}(U_i-u_0)\Big]\end{aligned}$$

at all $\boldsymbol{\beta}^*$. Let $f_U(\cdot)$ be the probability density function of U. Then

$$\frac{\partial\widehat{g}(u_0;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\frac{E\left[\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}\mid U_i=u_0\right]}{E\left[\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}/\partial g(u_0;\boldsymbol{\beta})\mid U_i=u_0\right]}+o_p(1).\tag{C.3}$$

On the other hand, we have

$$\begin{aligned}\mathbf{0} &= n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c}1\\(U_i-u_0)/h\end{array}\right\}\Big[\frac{\partial K_h(U_i-u_0)}{\partial(U_i-u_0)}(\mathbf{Z}_i-\mathbf{z}_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})\\&\quad+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}\\&\quad+K_h(U_i-u_0)\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}\frac{\partial\widehat{g}(u_0,\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}}\\&\quad+K_h(U_i-u_0)\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}\\&\quad\times\frac{\partial\widehat{g}'(u_0,\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}}(U_i-u_0)\\&\quad+K_h(U_i-u_0)\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}}\\&\quad\times\widehat{g}'(u_0,\boldsymbol{\beta})(\mathbf{Z}_i-\mathbf{z}_0)\Big]_{-1}^{\mathrm{T}}\\&\quad+n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c}\mathbf{0}_{1\times(q-1)}\\(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\end{array}\right\}K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})\\&\quad+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}\end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n}\begin{Bmatrix} 1 \\ (U_i-u_0)/h \end{Bmatrix}\frac{\partial K_h(U_i-u_0)}{\partial(U_i-u_0)}(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta}) \\
&\quad +\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\} \\
&\quad +\Big\{E\begin{pmatrix} 1 \\ 0 \end{pmatrix}\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial g(u_0,\boldsymbol{\beta})}\mid U_i=u_0\Big]\Big\}\frac{\partial\widehat{g}(u_0,\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}_{-1}^{\mathrm{T}}}f_U(u_0) \\
&\quad +\Big\{E\begin{pmatrix} 1 \\ 0 \end{pmatrix}\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial g(u_0,\boldsymbol{\beta})}g'(u_0,\boldsymbol{\beta})(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\mid U_i=u_0\Big]\Big\}f_U(u_0) \\
&\quad +E\Big[\begin{Bmatrix} \mathbf{0}_{1\times(q-1)} \\ (\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}/h \end{Bmatrix}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}\mid U_i=u_0\Big]f_U(u_0)+o_p(1)
\end{aligned}
$$

Now, to analyze the first term above, we introduce $f_{U,\mathbf{Z}_{-1}}(u,\mathbf{z}_{-1})$ as the joint pdf of $U,\mathbf{Z}_{-1}$. Note that $f_{U,\mathbf{Z}_{-1}}(u,\mathbf{z}_{-1})=f_{\mathbf{Z}}(\mathbf{z})$ for $u=\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}$ where $\boldsymbol{\gamma}$ is any parameter with the first component 1. Then

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\frac{\partial K_h(U_i-u_0)}{\partial(U_i-u_0)}(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})+\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\} \\
&= n^{-1}\sum_{i=1}^{n}h^{-2}K'\{(U_i-u_0)/h\}(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta}) \\
&\quad +\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\} \\
&= \int h^{-2}K'\{(u-u_0)/h\}(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})(u-u_0)\} \\
&\quad \times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u,\mathbf{z}_{-1},\mathbf{s},y)dud\mathbf{z}_{-1}d\mathbf{s}dy\{1+o_p(1)\} \\
&= \int h^{-1}K'(t)(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u_0+ht,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})ht\} \\
&\quad \times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+ht,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy\{1+o_p(1)\} \\
&= \Big(\int h^{-1}K'(t)(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\} \\
&\quad \times\ f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy \\
&\quad +\int K'(t)(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Big[\frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial u_0}f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y) \\
&\quad +\frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial g(u_0,\boldsymbol{\beta})}g'(u_0,\boldsymbol{\beta})f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)
\end{aligned}
$$

$$
\begin{aligned}
& +\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\} \frac{\partial f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y)}{\partial u_0}\Bigg] t dt d\mathbf{z}_{-1} d\mathbf{s} dy + O_p(h)\Bigg) \\
& \times\{1 + o_p(1)\} \\
= & -\left(\int (\mathbf{z} - \mathbf{z}_0)_{-1}^{\mathrm{T}} \left[\frac{\partial\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\}}{\partial u_0} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y)\right.\right. \\
& + \frac{\partial\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\}}{\partial g(u_0, \boldsymbol{\beta})} g'(u_0, \boldsymbol{\beta}) f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y) \\
& \left.\left. +\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\} \frac{\partial f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y)}{\partial u_0}\right] d\mathbf{z}_{-1} d\mathbf{s} dy + O_p(h)\right) \\
& \times\{1 + o_p(1)\} \\
= & -\left(\int (\mathbf{z} - \mathbf{z}_0)_{-1}^{\mathrm{T}} \frac{d\left[\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y)\right]}{du_0} d\mathbf{z}_{-1} d\mathbf{s} dy\right. \\
& \left. +O_p(h)\right) \times \{1 + o_p(1)\} \\
= & -\frac{d}{du_0}\left(\int (\mathbf{z} - \mathbf{z}_0)_{-1}^{\mathrm{T}} \left[\Phi\{y, \mathbf{s}, u_0, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta})\} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0, \mathbf{z}_{-1}, \mathbf{s}, y)\right]\right. \\
& \left. d\mathbf{z}_{-1} d\mathbf{s} dy\right) \{1 + o_p(1)\} + O_p(h)
\end{aligned}
$$

Continue from the last page, we have

$$
\begin{aligned}
= & -\frac{d}{du_0} E((\mathbf{Z} - \mathbf{z}_0)_{-1}^{\mathrm{T}} E[\Phi\{Y, \mathbf{S}, \mathbf{Z}; \boldsymbol{\beta}, g(U, \boldsymbol{\beta})\} \mid \mathbf{Z}, \mathbf{S}] \mid U = u_0) \\
& \times\{1 + o_p(1)\} + O_p(h) \\
= & \ o_p(1),
\end{aligned}
$$

where the last equality is because Φ has expectation zero conditional on $\mathbf{S}, \mathbf{Z}$. Further

$$
\begin{aligned}
& n^{-1} \sum_{i=1}^{n} (U_i - u_0)/h \frac{\partial K_h(U_i - u_0)}{\partial(U_i - u_0)} (\mathbf{Z}_i - \mathbf{z}_0)_{-1}^{\mathrm{T}} \Phi\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\beta}, \widehat{g}(u_0, \boldsymbol{\beta}) \\
& +\widehat{g}'(u_0, \boldsymbol{\beta})(U_i - u_0)\} \\
= & \ n^{-1} \sum_{i=1}^{n} h^{-1}(U_i - u_0) K'\{(U_i - u_0)/h\} (\mathbf{Z}_i - \mathbf{z}_0)_{-1}^{\mathrm{T}} \Phi\{Y_i, \mathbf{S}_i, \mathbf{Z}_i; \boldsymbol{\beta}, \widehat{g}(u_0, \boldsymbol{\beta}) \\
& +\widehat{g}'(u_0, \boldsymbol{\beta})(U_i - u_0)\} \\
= & \int h^{-1}(U_i - u_0) K'\{(u - u_0)/h\} (\mathbf{z} - \mathbf{z}_0)_{-1}^{\mathrm{T}} \Phi\{y, \mathbf{s}, u, \mathbf{z}_{-1}; \boldsymbol{\beta}, g(u_0, \boldsymbol{\beta}) \\
& +g'(u_0, \boldsymbol{\beta})(u - u_0)\} \\
& f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u, \mathbf{z}_{-1}, \mathbf{s}, y) du d\mathbf{z}_{-1} d\mathbf{s} dy \{1 + o_p(1)\}
\end{aligned}
$$

$$
\begin{aligned}
&= \int tK'(t)(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u_0+ht,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})ht\}\\
&\quad \times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+ht,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy\{1+o_p(1)\}\\
&= \Big(\int tK'(t)(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}\\
&\quad \times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy\Big)\{1+o_p(1)\}+O_p(h)\\
&= -\Big(\int(\mathbf{z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)d\mathbf{z}_{-1}d\mathbf{s}dy\Big)\\
&\quad \times\{1+o_p(1)\}+O_p(h)\\
&= -E((\mathbf{Z}-\mathbf{z}_0)_{-1}^{\mathrm{T}}E[\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g(U,\boldsymbol{\beta})\}\mid\mathbf{Z},\mathbf{S}]\mid U=u_0)\{1+o_p(1)\}+O_p(h)\\
&= o_p(1),
\end{aligned}
$$

where the last equality is because Φ has expectation zero conditional on $\mathbf{S},\mathbf{Z}$. Similarly

$$
\begin{aligned}
&E\Big[\begin{Bmatrix}\mathbf{0}_{1\times(q-1)}\\(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}/h\end{Bmatrix}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}\mid U_i=u_0\Big]\\
&= E\Big(\begin{Bmatrix}\mathbf{0}_{1\times(q-1)}\\(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}/h\end{Bmatrix}E[\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(U_i,\boldsymbol{\beta})\}\mid\mathbf{S}_i,\mathbf{Z}_i]\mid U_i=u_0\Big)\\
&= \mathbf{0}.
\end{aligned}
$$

We thus obtain

$$
\begin{aligned}
\mathbf{0} &= \Big\{E\begin{pmatrix}1\\0\end{pmatrix}\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial g(u_0,\boldsymbol{\beta})}\mid U_i=u_0\Big]\Big\}\frac{\partial\widehat{g}(u_0,\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}_{-1}^{\mathrm{T}}}f_U(u_0)\\
&\quad +\Big\{E\begin{pmatrix}1\\0\end{pmatrix}\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}}{\partial g(u_0,\boldsymbol{\beta})}g'(u_0,\boldsymbol{\beta})(\mathbf{Z}_i-\mathbf{z}_0)_{-1}^{\mathrm{T}}\mid U_i=u_0\Big]\Big\}f_U(u_0)\\
&\quad +o_p(1).
\end{aligned}
$$

Hence

$$
\begin{aligned}
\frac{\partial\widehat{g}(u_0,\boldsymbol{\beta})}{\partial\boldsymbol{\gamma}_{-1}} &= -\frac{E[(\mathbf{Z}_i-\mathbf{z}_0)_{-1}\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}/\partial g(u_0,\boldsymbol{\beta})g'(u_0,\boldsymbol{\beta})\mid U_i=u_0]}{E[\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0,\boldsymbol{\beta})\}/\partial g(u_0,\boldsymbol{\beta})\mid U_i=u_0]}\\
&\quad +o_p(1) \qquad \text{(C.4)}
\end{aligned}
$$

Continue from (C.1) by inserting (C.3) and (C.4), this leads to

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\} + n^{-1/2}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})}\\
&\quad \times\{\widehat{g}(U_i;\boldsymbol{\beta})-g(U_i;\boldsymbol{\beta})\} + \Big\{E\Big(\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\}}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\\
&\quad -\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})}\frac{E\left[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}^{\mathrm{T}}\mid U_i\right]}{E\left[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\}/\partial g(U_i;\boldsymbol{\beta})\mid U_j=U_i\right]}\Big)\\
&\quad +o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\
&\quad -\Big\{E\Big(\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})}\\
&\quad \frac{E[(\mathbf{Z}_j-\mathbf{Z}_i)_{-1}^{\mathrm{T}}\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i,\boldsymbol{\beta})\}/\partial g(U_i,\boldsymbol{\beta})g'(U_i,\boldsymbol{\beta})\mid U_i]}{E[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i,\boldsymbol{\beta})\}/\partial g(U_i,\boldsymbol{\beta})\mid U_i]}\Big)\\
&\quad +o_p(1)\Big\}\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) \qquad\qquad (\mathrm{C}.5)
\end{aligned}
$$

where $g^*(u;\boldsymbol{\beta})$ lies on the line connecting $g(u;\boldsymbol{\beta})$ and $\widehat{g}(u;\boldsymbol{\beta})$. Let $g^{'*}(u_0,\boldsymbol{\beta})$ be on the line connecting $\widehat{g}'(u_0,\boldsymbol{\beta})$ and $g'(u_0,\boldsymbol{\beta})$. We now rewrite (C.2) as

$$
\begin{aligned}
\mathbf{0} &= n^{-1}\sum_{i=1}^{n}\left\{\begin{matrix}1\\(U_i-u_0)/h\end{matrix}\right\}K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},\widehat{g}(u_0,\boldsymbol{\beta})\\
&\quad +\widehat{g}'(u_0,\boldsymbol{\beta})(U_i-u_0)\}\\
&= n^{-1}\sum_{i=1}^{n}\left\{\begin{matrix}1\\(U_i-u_0)/h\end{matrix}\right\}K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\\
&\quad +g'(u_0,\boldsymbol{\beta})(U_i-u_0)\}\\
&\quad +\Big[n^{-1}\sum_{i=1}^{n}\left\{\begin{matrix}1\\(U_i-u_0)/h\end{matrix}\right\}K_h(U_i-u_0)\\
&\quad \frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}\Big]\{\widehat{g}(u_0;\boldsymbol{\beta})-g(u_0;\boldsymbol{\beta})\}\\
&\quad +\Big[n^{-1}\sum_{i=1}^{n}\left\{\begin{matrix}1\\(U_i-u_0)/h\end{matrix}\right\}K_h(U_i-u_0)(U_i-u_0)\\
&\quad \times\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}\Big]\{\widehat{g}'(u_0;\boldsymbol{\beta})-g'(u_0;\boldsymbol{\beta})\},
\end{aligned}
$$

$$
\begin{aligned}
= \;& n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c} 1 \\ (U_i-u_0)/h \end{array}\right\} K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta}) \\
& +g'(u_0,\boldsymbol{\beta})(U_i-u_0)\} + \left[n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{cc} 1 & (U_i-u_0)/h \\ (U_i-u_0)/h & (U_i-u_0)^2/h^2 \end{array}\right\} K_h(U_i-u_0)\right. \\
& \left.\times\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}{\partial\{g^*(u_0;\boldsymbol{\beta})+g^{'*}(u_0;\boldsymbol{\beta})(U_i-u_0)\}}\right] \\
& \times\left[\begin{array}{c} \widehat{g}(u_0;\boldsymbol{\beta})-g(u_0;\boldsymbol{\beta}) \\ h\{\widehat{g}'(u_0;\boldsymbol{\beta})-g'(u_0;\boldsymbol{\beta})\} \end{array}\right] \\
= \;& n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c} 1 \\ (U_i-u_0)/h \end{array}\right\} K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta}) \\
& +g'(u_0,\boldsymbol{\beta})(U_i-u_0)\} + \left[\int h^{-1}\left\{\begin{array}{cc} 1 & (u-u_0)/h \\ (u-u_0)/h & (u-u_0)^2/h^2 \end{array}\right\} K\Big(\frac{u-u_0}{h}\Big)\right. \\
& \times\frac{\partial\Phi\{y,\mathbf{s},u,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})(u-u_0)\}}{\partial\{g(u_0;\boldsymbol{\beta})+g'(u_0;\boldsymbol{\beta})(u-u_0)\}} \\
& \left.\times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u,\mathbf{z}_{-1},\mathbf{s},y)dud\mathbf{z}_{-1}d\mathbf{s}dy+o_p(1)\right]\left[\begin{array}{c} \widehat{g}(u_0;\boldsymbol{\beta})-g(u_0;\boldsymbol{\beta}) \\ h\{\widehat{g}'(u_0;\boldsymbol{\beta})-g'(u_0;\boldsymbol{\beta})\} \end{array}\right] \\
= \;& n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c} 1 \\ (U_i-u_0)/h \end{array}\right\} K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta}) \\
& +g'(u_0,\boldsymbol{\beta})(U_i-u_0)\} + \left[\int \begin{pmatrix} 1 & t \\ t & t^2 \end{pmatrix} K(t)\right. \\
& \times\frac{\partial\Phi\{y,\mathbf{s},u_0+th,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}{\partial\{g(u_0;\boldsymbol{\beta})+g'(u_0;\boldsymbol{\beta})th\}} \\
& \left.\times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+th,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy+o_p(1)\right]\left[\begin{array}{c} \widehat{g}(u_0;\boldsymbol{\beta})-g(u_0;\boldsymbol{\beta}) \\ h\{\widehat{g}'(u_0;\boldsymbol{\beta})-g'(u_0;\boldsymbol{\beta})\} \end{array}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
& \int K(t)\frac{\partial\Phi\{y,\mathbf{s},u_0+th,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}{\partial\{g(u_0;\boldsymbol{\beta})+g'(u_0;\boldsymbol{\beta})th\}} \\
& \times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+th,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy
\end{aligned}
$$

$$
\begin{aligned}
&= \int K(t)\frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy + O(h^2)\\
&= \int \frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)d\mathbf{z}_{-1}d\mathbf{s}dy + O(h^2)\\
&= E\Big[\frac{\partial\Phi\{Y,\mathbf{S},U,\mathbf{Z}_{-1};\boldsymbol{\beta},g(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U=u_0\Big] f_U(u_0) + O(h^2).
\end{aligned}
$$

Similarly

$$
\begin{aligned}
&\int tK(t)\frac{\partial\Phi\{y,\mathbf{s},u_0+th,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}{\partial\{g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}\\
&\times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+th,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy = O(h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&\int t^2K(t)\frac{\partial\Phi\{y,\mathbf{s},u_0+th,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}{\partial\{g(u_0;\boldsymbol{\beta})+g'(u_0,\boldsymbol{\beta})th\}}\\
&\times f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0+th,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy\\
= &\int t^2K(t)\frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)dtd\mathbf{z}_{-1}d\mathbf{s}dy\\
&+O(h^4)\\
= &\int \mu_2\frac{\partial\Phi\{y,\mathbf{s},u_0,\mathbf{z}_{-1};\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} f_{U,\mathbf{Z}_{-1},\mathbf{S},Y}(u_0,\mathbf{z}_{-1},\mathbf{s},y)d\mathbf{z}_{-1}d\mathbf{s}dy\{1+O(h^2)\}\\
= &E\Big[\frac{\partial\Phi\{Y,\mathbf{S},U,\mathbf{Z}_{-1};\boldsymbol{\beta},g(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})}\mu_2 \mid U=u_0\Big] f_U(u_0)\{1+O_p(h^2)\}.
\end{aligned}
$$

So

$$
\begin{aligned}
\mathbf{0} = &\; n^{-1}\sum_{i=1}^n \left\{\begin{matrix} 1 \\ (U_i-u_0)/h\end{matrix}\right\} K_h(U_i-u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\\
&+g'(u_0,\boldsymbol{\beta})(U_i-u_0)\}\\
&+\left(E\Big[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} \mid U_i=u_0\Big]\begin{pmatrix} 1 & 0\\ 0 & \mu_2\end{pmatrix} f_U(u_0)+o_p(1)\right)\\
&\times\begin{bmatrix} \widehat{g}(u_0;\boldsymbol{\beta})-g(u_0;\boldsymbol{\beta})\\ h\{\widehat{g}'(u_0;\boldsymbol{\beta})-g'(u_0;\boldsymbol{\beta})\}\end{bmatrix}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& \widehat{g}(u_0;\boldsymbol{\beta}) - g(u_0;\boldsymbol{\beta}) \\
= \quad & -\left(E\left[\frac{\partial\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta})\}}{\partial g(u_0;\boldsymbol{\beta})} \mid U_i = u_0\right] f_U(u_0) + o_p(1)\right)^{-1} \\
& \times\left[n^{-1}\sum_{i=1}^{n} K_h(U_i - u_0)\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g(u_0;\boldsymbol{\beta}) + g'(u_0,\boldsymbol{\beta})(U_i - u_0)\}\right].
\end{aligned}
$$

Therefore

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})}\{\widehat{g}(U_i;\boldsymbol{\beta}) - g(U_i;\boldsymbol{\beta})\} \\
= \quad & -n^{-3/2}\sum_{i,j=1}^{n} K_h(U_j - U_i)\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})} \\
& \left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g(U,\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U = U_i\right\} f_U(U_i) + o_p(1)\right]^{-1} \\
& \times\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta}) + g'(U_i,\boldsymbol{\beta})(U_j - U_i)\}. \qquad \text{(C.6)}
\end{aligned}
$$

Note that, under the conditions (C1) and (C2),

$$
\begin{aligned}
& n^{-1/2}\sum_{j=1}^{n} E\Big\{K_h(U_j - U_i)\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})} \\
& \left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U = U_i\right\} f_U(U_i) + o_p(1)\right]^{-1} \\
& \times\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta}) + g'(U_i,\boldsymbol{\beta})(U_j - U_i)\} \mid \mathbf{S}_j,\mathbf{Z}_j,Y_j\Big\} \\
& = n^{-1/2}\sum_{j=1}^{n}\Big\{E\Big[\frac{\partial\mathcal{L}\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g^*(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U = U_j\Big] \\
& \left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U = U_j\right\}\right]^{-1} + o_p(1)\Big\}\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_j;\boldsymbol{\beta})\},
\end{aligned}
$$

and

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n} E\Big\{K_h(U_j - U_i)\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol{\beta},g^*(U_i;\boldsymbol{\beta})\}}{\partial g(U_i;\boldsymbol{\beta})} \\
& \left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},g(U;\boldsymbol{\beta})\}}{\partial g(U;\boldsymbol{\beta})} \mid U = U_i\right\} f_U(U_i) + o_p(1)\right]^{-1} \\
& \times\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol{\beta},g(U_i;\boldsymbol{\beta}) + g'(U_i,\boldsymbol{\beta})(U_j - U_i)\} \mid \mathbf{S}_i,\mathbf{Z}_i,Y_i\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g^*(U_i;\boldsymbol\beta)\}}{\partial g(U_i;\boldsymbol\beta)}\Bigg[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right\}\\
&\quad f_U(U_i)+o_p(1)\Bigg]^{-1}E\left[K_h(U_j-U_i)\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\right.\\
&\quad \left.+g'(U_i,\boldsymbol\beta)(U_j-U_i)\}\mid\mathbf{S}_i,\mathbf{Z}_i,Y_i\right]\\
&= n^{-1/2}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g^*(U_i;\boldsymbol\beta)\}}{\partial g(U_i;\boldsymbol\beta)}\Bigg[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right\}\\
&\quad f_U(U_i)+o_p(1)\Bigg]^{-1}\Big(E\left[\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}\mid U=U_i\right]f_U(U_i)+O(h^2)\Big)\\
&= O_p(n^{1/2}h^2)=o_p(1).
\end{aligned}
$$

Inserting these results to (C.6), in combination with U-statistic properties, we have

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g^*(U_i;\boldsymbol\beta)\}}{\partial g(U_i;\boldsymbol\beta)}\{\widehat{g}(U_i;\boldsymbol\beta)-g(U_i;\boldsymbol\beta)\}\\
&= -n^{-1/2}\sum_{i=1}^{n}\Bigg(E\Bigg[\frac{\partial\mathcal{L}\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g^*(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\Bigg]\\
&\quad\Bigg[E\Bigg\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\Bigg\}\Bigg]^{-1}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}\Bigg)+o_p(1)\\
&= -n^{-1/2}\sum_{i=1}^{n}E\left[\frac{\partial\mathcal{L}\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right]\\
&\quad\left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right\}\right]^{-1}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}+o_p(1).
\end{aligned}
$$

Continuing from (C.5), using the property that $nh^4\to 0$, we obtain

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\Bigg(\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}-E\left[\frac{\partial\mathcal{L}\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right]\\
&\quad\left[E\left\{\frac{\partial\Phi\{Y,\mathbf{S},\mathbf{Z};\boldsymbol\beta,g(U;\boldsymbol\beta)\}}{\partial g(U;\boldsymbol\beta)}\mid U=U_i\right\}\right]^{-1}\Phi\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}\Bigg)\\
&\quad+\Bigg\{E\Bigg[\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}}{\partial\boldsymbol\beta^{\mathrm{T}}}\\
&\quad-\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}}{\partial g(U_i;\boldsymbol\beta)}\frac{E\left[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}/\partial\boldsymbol\beta^{\mathrm{T}}\mid U_i\right]}{E\left[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}/\partial g(U_i;\boldsymbol\beta)\mid U_i\right]}\Bigg]\\
&\quad+o_p(1)\Bigg\}\sqrt{n}(\widehat{\boldsymbol\beta}-\boldsymbol\beta)-\Bigg\{E\Bigg[\frac{\partial\mathcal{L}\{Y_i,\mathbf{S}_i,\mathbf{Z}_i;\boldsymbol\beta,g(U_i;\boldsymbol\beta)\}}{\partial g(U_i;\boldsymbol\beta)}\\
&\quad\times\frac{E[(\mathbf{Z}_j-\mathbf{Z}_i)_{-1}^{\mathrm{T}}\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol\beta,g(U_i,\boldsymbol\beta)\}/\partial g(U_i,\boldsymbol\beta)g'(U_i,\boldsymbol\beta)\mid U_i]}{E[\partial\Phi\{Y_j,\mathbf{S}_j,\mathbf{Z}_j;\boldsymbol\beta,g(U_i,\boldsymbol\beta)\}/\partial\{g(U_i,\boldsymbol\beta)\}\mid U_i]}\Bigg]\\
&\quad+o_p(1)\Bigg\}\sqrt{n}(\widehat{\boldsymbol\gamma}-\boldsymbol\gamma)+o_p(1).
\end{aligned}
$$

Table C.1: Simulation 1 results (link function: logit)

	truth	α_1 1.0	α_2 1.0	β 0.3	γ_2 0.5	γ_3 1.0	γ_4 -0.3
ϵ: Normal distribution							
α_0	mean			0.3126	0.5085	1.0291	-0.2706
	median			0.3003	0.5005	1.0008	-0.2957
	se			0.1277	0.0750	0.1684	0.0798
OLS	mean	1.0023	0.9993	0.3237	0.5063	1.0381	-0.2710
	median	1.0026	1.0014	0.3005	0.5005	1.0010	-0.2946
	se	0.0337	0.0461	0.1437	0.0729	0.1729	0.0750
WLS	mean	0.9982	0.9996	0.3160	0.5067	1.0335	-0.2682
	median	0.9983	1.0003	0.2997	0.5008	1.0013	-0.2935
	se	0.0300	0.0421	0.1514	0.0709	0.1652	0.0768
ϵ: Student t distribution t_5							
α_0	mean			0.3125	0.5037	1.0305	-0.2717
	median			0.3004	0.5005	1.0008	-0.2961
	se			0.1355	0.0704	0.1667	0.0714
OLS	mean	1.0014	1.0003	0.3250	0.5041	1.0315	-0.2709
	median	1.0007	1.0000	0.3006	0.5005	1.0008	-0.2954
	se	0.0397	0.0544	0.1577	0.0694	0.1607	0.0711
WLS	mean	0.9997	1.0002	0.3143	0.5090	1.0393	-0.2733
	median	0.9985	1.0010	0.2995	0.5007	1.0017	-0.2961
	se	0.0318	0.0473	0.1497	0.07132	0.1761	0.0717

This leads to the result stated in the theorem. $\square$

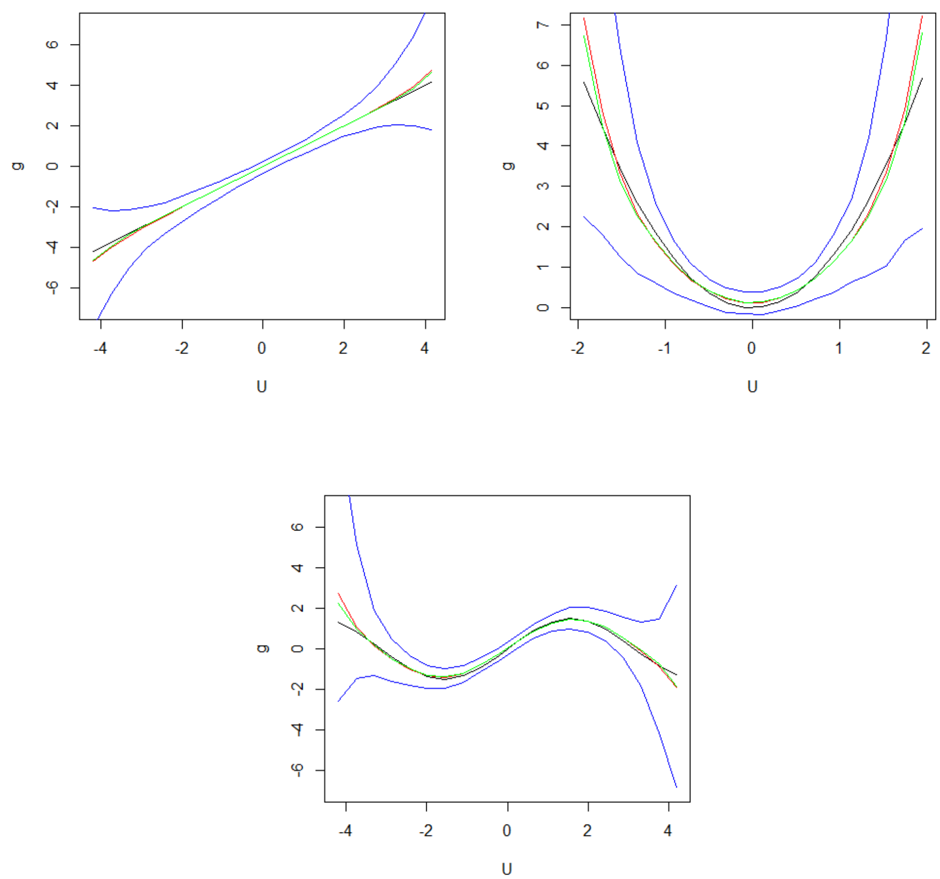

Figure C.1 True function (black line), median estimation (gree line), mean estimation (red line) and 95% confidence band (blue line) of $g(u)$ in simulations 1 (upper-left), 2 (upper-right) and 3 (lower) when link function is inverse logit, ϵ is normal distributed and with OLS method applied.

Table C.2: Simulation 2 results (link function: logit)

	truth	α_1 1.0	α_2 1.0	β 0.3	γ_2 0.2	γ_3 0.3	γ_4 -0.4
ϵ: Normal distribution							
α_0	mean			0.3276	0.1954	0.2978	-0.4011
	median			0.3013	0.2000	0.3006	-0.3995
	se			0.2583	0.0768	0.1216	0.1452
OLS	mean	0.9991	0.9979	0.3318	0.1962	0.3018	-0.4010
	median	0.9987	0.9985	0.3022	0.2002	0.3007	-0.3997
	se	0.0319	0.0434	0.2493	0.0914	0.1133	0.1483
WLS	mean	0.9990	0.9990	0.3416	0.1926	0.3004	-0.4062
	median	0.9997	0.9974	0.3029	0.2004	0.3007	-0.3999
	se	0.0295	0.0411	0.2146	0.0863	0.1211	0.1317
ϵ: Student t distribution t_5							
α_0	mean			0.3243	0.1907	0.3062	-0.4017
	median			0.3036	0.1996	0.3009	-0.3997
	se			0.1601	0.0843	0.1209	0.1290
OLS	mean	0.9997	1.0006	0.3330	0.1913	0.3080	-0.4085
	median	0.9999	1.0008	0.3009	0.1998	0.3005	-0.3993
	se	0.0376	0.0550	0.2461	0.0958	0.1190	0.1519
WLS	mean	0.9979	1.0002	0.3258	0.1983	0.3029	-0.4050
	median	0.9979	1.0000	0.3030	0.2003	0.3004	-0.4003
	se	0.0324	0.0471	0.2153	0.0804	0.105	0.1175

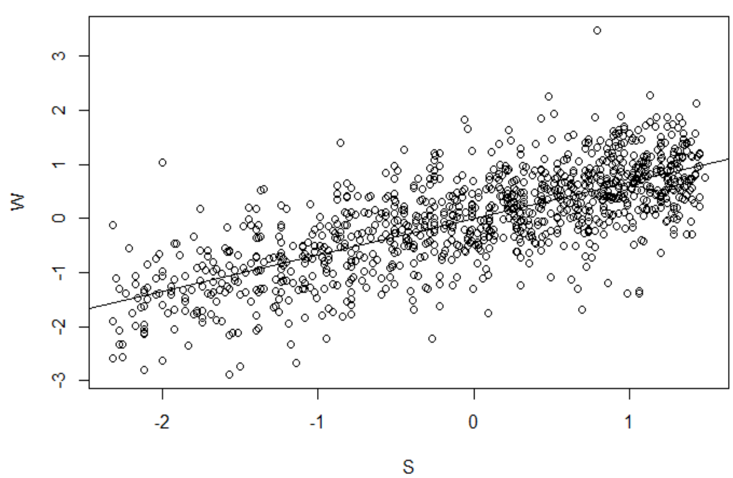

Figure C.2 Plot of averaged baseline CD4 count versus screening CD4 count.

Table C.3: Simulation 3 results (link function: logit)

		α_1	α_2	β	γ_2	γ_3	γ_4
	truth	1.0	1.0	0.3	0.5	1.0	-0.3
ϵ: Normal distribution							
α_0	mean			0.2616	0.5060	1.0506	-0.2467
	median			0.2571	0.4997	1.0189	-0.2572
	se			0.1041	0.0883	0.1749	0.0775
OLS	mean	0.9988	0.9988	0.2666	0.5161	1.0624	-0.2500
	median	0.9984	0.9994	0.2581	0.5017	1.0115	-0.2606
	se	0.0344	0.0449	0.1292	0.1042	0.2111	0.0924
WLS	mean	0.9998	0.9998	0.2625	0.5161	1.0717	-0.2499
	median	1.0005	0.9983	0.2586	0.5011	1.0135	-0.2594
	se	0.0296	0.0415	0.1143	0.1013	0.20575	0.0803
ϵ: Student t distribution t_5							
α_0	mean			0.2642	0.5114	1.0616	-0.2481
	median			0.2579	0.5003	1.0216	-0.2612
	se			0.1133	0.0982	0.1893	0.0817
OLS	mean	1.0017	0.9995	0.2613	0.5178	1.0709	-0.2530
	median	1.0010	0.9989	0.2552	0.5013	1.0220	-0.2624
	se	0.0395	0.0556	0.1199	0.1134	0.2112	0.0990
WLS	mean	0.9981	1.0003	0.2560	0.5127	1.0658	-0.2487
	median	0.9979	0.9990	0.2496	0.4999	1.0129	-0.2571
	se	0.0329	0.0480	0.1089	0.0955	0.2046	0.0818

Table C.4: Simulation 1 results (link function: probit)

		α_1	α_2	β	γ_2	γ_3	γ_4
	truth	1.0	1.0	1.0	0.5	1.0	0.3
ϵ: Normal distribution							
α_0	mean			1.0611	0.5080	1.0845	0.3365
	median			0.9916	0.4999	1.0251	0.3114
	se			0.2825	0.0894	0.1881	0.0788
OLS	mean	0.9989	0.9993	1.0788	0.5089	1.0874	0.3363
	median	0.9992	0.9998	0.9980	0.4996	1.0298	0.3111
	se	0.0340	0.0444	0.2905	0.09074	0.1922	0.0808
WLS	mean	1.0001	1.0004	1.0759	0.5060	1.0736	0.3347
	median	0.9995	1.0009	1.0084	0.4993	1.0292	0.3120
	se	0.0313	0.0409	0.2462	0.0765	0.1616	0.0724
ϵ: Student t distribution t_5							
α_0	mean			1.0591	0.5073	1.0917	0.3374
	median			0.9846	0.4981	1.0350	0.3131
	se			0.2986	0.0867	0.1907	0.0792
OLS	mean	1.0010	0.9986	1.0641	0.5083	1.0901	0.3351
	median	1.0004	0.9985	0.9931	0.4984	1.0305	0.3103
	se	0.0385	0.0568	0.2948	0.0908	0.1891	0.0772
WLS	mean	1.0000	1.0022	1.0669	0.5058	1.0751	0.3365
	median	1.0004	1.0030	1.0110	0.4995	1.0230	0.3147
	se	0.0342	0.0473	0.2510	0.0762	0.1609	0.0728

Table C.5: Simulation 2 results (link function: probit)

		α_1	α_2	β	γ_2	γ_3	γ_4
	truth	1.0	1.0	0.3	0.2	0.3	-0.4
ϵ: Normal distribution							
α_0	mean			0.3746	0.1891	0.3032	-0.4081
	median			0.3113	0.1997	0.3009	-0.4009
	se			0.1986	0.0657	0.0914	0.0945
OLS	mean	1.0006	0.9980	0.3828	0.1951	0.3011	-0.4049
	median	0.9998	0.9986	0.3115	0.2002	0.3008	-0.4006
	se	0.0325	0.0449	0.2322	0.0656	0.0979	0.1081
WLS	mean	0.9997	1.0004	0.3798	0.1935	0.3022	-0.4103
	median	1.0002	1.0007	0.3158	0.1997	0.3014	-0.4006
	se	0.0294	0.0412	0.1881	0.0602	0.0757	0.0892
ϵ: Student t distribution t_5							
α_0	mean			0.3760	0.1934	0.3060	-0.4085
	median			0.3087	0.2001	0.3010	-0.4013
	se			0.2084	0.0635	0.0793	0.1001
OLS	mean	0.9983	1.0003	0.4009	0.1963	0.3002	-0.4065
	median	0.9992	1.0016	0.3127	0.2005	0.3003	-0.4016
	se	0.0371	0.0548	0.2514	0.0672	0.0867	0.1037
WLS	mean	0.9978	1.0009	0.3740	0.1967	0.3027	-0.4144
	median	0.9985	1.0005	0.3105	0.2002	0.3012	-0.4028
	se	0.0330	0.0477	0.1973	0.0577	0.0812	0.0946

Table C.6: Simulation 3 results (link function: probit)

		α_1	α_2	β	γ_2	γ_3	γ_4
	truth	1.0	1.0	0.3	0.5	1.0	-0.3
ϵ: Normal distribution							
α_0	mean			0.2744	0.5108	1.0560	-0.2502
	median			0.2594	0.5011	1.0200	-0.2613
	se			0.1228	0.0828	0.1773	0.0783
OLS	mean	1.0003	0.9987	0.2776	0.5077	1.0545	-0.2507
	median	0.9998	1.0002	0.2624	0.5011	1.0216	-0.2628
	se	0.0331	0.0453	0.0692	0.0833	0.1737	0.0784
WLS	mean	0.9988	0.9990	0.2763	0.5091	1.0473	-0.2503
	median	0.9991	0.9979	0.2651	0.50015	1.0149	-0.2609
	se	0.033	0.0470	0.1291	0.0808	0.1787	0.0792
ϵ: Student t distribution t_5							
α_0	mean			0.2627	0.5065	1.0549	-0.2531
	median			0.2566	0.5002	1.0222	-0.2659
	se			0.0904	0.0776	0.1600	0.0784
OLS	mean	1.0008	0.9987	0.2721	0.5066	1.0540	-0.2502
	median	1.0007	0.9995	0.2604	0.5005	1.0120	-0.2606
	se	0.0378	0.0554	0.1181	0.0805	0.1785	0.0817
WLS	mean	1.0008	0.9990	0.2749	0.5093	1.0549	-0.2569
	median	1.0006	0.9994	0.2602	0.5005	1.0170	-0.2687
	se	0.0331	0.0462	0.1330	0.0800	0.1738	0.0795

Table C.7: Realdata analysis results

	β	γ_1	γ_2	γ_3	γ_{age}
realdata estimates	-0.72808	1.0	1.511	2.4915	-2.3035
bootstrapped mean	-0.7219		1.7897	2.6411	-2.3819
bootstrapped median	-0.7082		1.6383	2.5455	-2.3029
bootstrapped se	0.1274		0.3737	0.2919	0.41061
95% CI	(-0.9779,-0.4783)		(0.8186,2.2836)	(1.9194,3.0636)	(-3.0996,-1.5075)

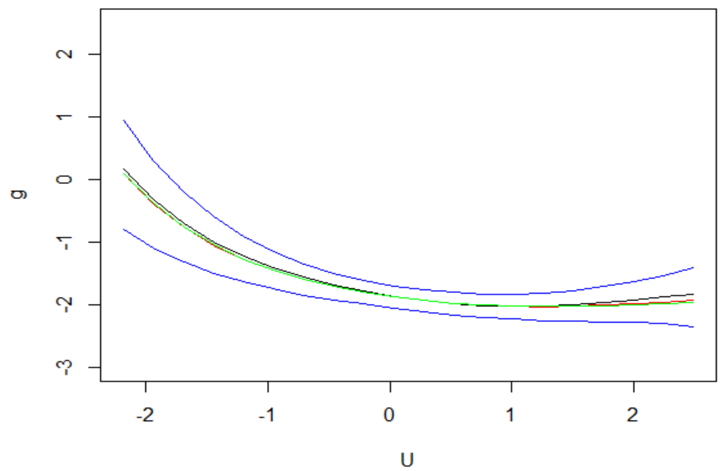

Figure C.3 Estimated $g(u)$ for real data (black line), median estimation (green line), mean estimation (red line) and 90% confidence band (blue line) of $g(u)$ using 1000 bootstrapped samples .

Appendix D

Chapter 3 Appendix

D.1 Proof of Thoerem 3

From the definitions of $\mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, g)$ and $\mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \boldsymbol{\gamma})$, we have

$$
\begin{aligned}
E\{\mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g)|X_i, \mathbf{Z}_i\} &= \mathbf{0}, \\
E_a\{\mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)|X_i, \mathbf{Z}_i\} &= \mathbf{0},
\end{aligned}
$$

where $_a$ here and throughout the text stands for “approximate”, and E_a indicates the expectation calculated with $g(\cdot)$ replaced by the approximate model $\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}_0$. Taking another expectation, we get

$$
\begin{aligned}
E\{\mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g)\} &= \mathbf{0}, \\
E_a\{\mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)\} &= \mathbf{0}.
\end{aligned}
$$

Using Condition (C6), we further get

$$
\begin{aligned}
E\{\mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)\} &= o(1), \\
E\{\mathbf{S}_{\mathrm{r}es2}^*\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)\} &= o(1),
\end{aligned}
$$

component-wise. Condition (C7) ensures that $[E\{\mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \boldsymbol{\gamma})\}^{\mathrm{T}}, E\{\mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \boldsymbol{\gamma})\}^{\mathrm{T}}]^{\mathrm{T}}$ is invertible near its zero $\boldsymbol{\theta}^*$ as a vector function of $\boldsymbol{\theta}$, and the first derivative of the inverse function is bounded in the neighborhood of $\boldsymbol{\theta}^*$. Therefore, $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 = o_p(1)$. On the other hand, since

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{S}^*_{eff}(Y_i, W_i, \mathbf{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\boldsymbol{\gamma}}_n) &= \mathbf{0}, \\
\frac{1}{n}\sum_{i=1}^n \mathbf{S}_{\mathrm{r}es2}^*(Y_i, W_i, \mathbf{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\boldsymbol{\gamma}}_n) &= \mathbf{0},
\end{aligned}
$$

we have

$$
\begin{aligned}
E\{\mathbf{S}^*_{\mathrm{e}ff}(Y_i, W_i, \mathbf{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\gamma}_n)\} &= o(1), \\
E\{\mathbf{S}_{\mathrm{r}es2}^{*}(Y_i, W_i, \mathbf{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\gamma}_n)\} &= o(1)
\end{aligned}
$$

element-wise. Using exactly the same argument as above, we can also obtain $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 = o_p(1)$. Hence combining the two results, we get $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = o_p(1)$. □

D.2 Proof of Theorem 4

We first write

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{e}ff}\{Y_i, W_i, \mathbf{Z}_i, \widehat{\boldsymbol{\delta}}_n, \widehat{\boldsymbol{\gamma}}_n(\widehat{\boldsymbol{\delta}}_n)\} \\
&= \mathbf{T}_1 + \mathbf{T}_2(\widetilde{\boldsymbol{\delta}}_n)\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{T}_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{e}ff}\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)\}, \\
\mathbf{T}_2(\boldsymbol{\delta}) &= \mathbf{T}_{21}(\boldsymbol{\delta}) + \mathbf{T}_{22}(\boldsymbol{\delta})\frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^{\mathrm{T}}},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{T}_{21}(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n)}{\partial \boldsymbol{\delta}^{\mathrm{T}}}, \\
\mathbf{T}_{22}(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{\mathrm{e}ff}\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})^{\mathrm{T}}},
\end{aligned}
$$

and $\widetilde{\boldsymbol{\delta}}_n$ is on the line connecting $\boldsymbol{\delta}_0$ and $\widehat{\boldsymbol{\delta}}_n$.

We further expand $\mathbf{T}_1$ as a function of $\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)$ about $\boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)$ to obtain

$$
\mathbf{T}_1 = \mathbf{T}_{11} + \mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)\}\sqrt{n}\{\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)\},
$$

where

$$
\begin{aligned}
\mathbf{T}_{11} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{e}ff}\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)\}, \\
\mathbf{T}_{12}\{\boldsymbol{\gamma}(\boldsymbol{\delta}_0)\} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{\mathrm{e}ff}\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}(\boldsymbol{\delta}_0)\}}{\partial \boldsymbol{\gamma}(\boldsymbol{\delta}_0)^{\mathrm{T}}},
\end{aligned}
$$

and $\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)$ is on the line connects $\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)$ and $\boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)$.

Because of the consistency of $\mathbf{B}(x)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_n$ to $g(x)$ derived from Condition (C6) and Theorem 3, and the weak law of large numbers, for arbitrary $d_\gamma \times p$ matrix $\mathbf{G}$ with $\|\mathbf{G}\|_2 = 1$, we have

$$\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)\}\mathbf{G} = E\left\{\frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{G}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\gamma = g(\cdot)}\right\}\{1 + o_p(1)\},$$

where

$$\begin{aligned}
& E\left\{\frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{G}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\gamma = g(\cdot)}\right\} \\
= & \int \left\{\frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{G}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\gamma = g(\cdot)}\right\} f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g, f_X) dy_i dw_i d\mathbf{z}_i \\
= & \int \left\{\frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\mathbf{G} + O_p(h_b^q)\right\}\{f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0, f_X) \\
& + O_p(h_b^q)\}\, dy_i dw_i d\mathbf{z}_i \\
= & \int \frac{\partial \mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\mathbf{G} f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0, f_X) dy_i dw_i d\mathbf{z}_i + O_p(h_b^q) \\
= & \frac{\partial}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\int \{\mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g) + O_P(h_b^q)\}\mathbf{G}\{f(y_i, w_i, \mathbf{z}_i, \beta_0, g, f_X) \\
& + O_p(h_b^q)\} dy_i dw_i d\mathbf{z}_i \\
& - \int \{\mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g) + O_P(h_b^q)\}\mathbf{G}\frac{\partial f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0, f_X)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} dy_i dw_i d\mathbf{z}_i \\
& + O_P(h_b^q) \\
= & -\int \mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g)\left\{\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\gamma}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)\right\}^{\mathrm{T}} \\
& f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g, f_X) dy_i dw_i d\mathbf{z}_i + O_p(h_b^p) \\
= & O_p(h_b^q). \qquad \text{(D.1)}
\end{aligned}$$

Here, like before, $f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}, f_X)$ stands for $f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g, f_X)$ with $g(\cdot)$ replaced by $\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}$, and $\mathbf{S}_{a,\gamma}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0) \equiv \partial \log f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}, f_X)/\partial \boldsymbol{\gamma}$. The second equality holds by condition (C6). The third equality holds because $\|\partial \mathbf{S}^*_{\mathrm{e}ff}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0)/\partial \boldsymbol{\gamma}_0^{\mathrm{T}}\|_\infty$ is integrable by condition (C8) and $f(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0, f_X)$ is absolutely integrable. The fourth equality holds also by condition (C6). The fifth equality holds because $E\{\mathbf{S}_{eff}^*(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}, g)\} = \mathbf{0}$. For the last equality, we note that $\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\gamma}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0) = E[s\{y_i, \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}_0, \boldsymbol{\alpha}_0\}\mathbf{G}^{\mathrm{T}}\mathbf{B}(X) \mid y_i, w_i, \mathbf{z}_i]$. By Condition (C6) and definitions of Λ_g and $\Lambda_{a,\gamma}$, for any $d_\gamma \times p$ matrix $\mathbf{G}$, there exists a function $\mathbf{h}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g) \equiv E[s\{y_i, \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g(X), \boldsymbol{\alpha}_0\}\mathbf{G}^{\mathrm{T}}\mathbf{B}(X) \mid y_i, w_i, \mathbf{z}_i] \in \Lambda_g$ such that $\sup|\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\gamma}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma}_0) - \mathbf{h}(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g)| = O_P(h_b^q)$. Further, $\mathbf{S}_{eff}^*(y_i, w_i, \mathbf{z}_i, \boldsymbol{\delta}_0, g)$ is orthogonal to any function in Λ_g, thus the last equality holds. Hence, we obtain $\|\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\delta}_0)\}\|_2 = O_p(h_b^q)$.

Based on the asymptotic results of Proposition 4 in Jiang & Ma (2017), we have $\|\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)\|_2 = O_p\{(nh_b)^{-1/2}\}$. Then we have

$$\|\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0)\}\sqrt{n}\{\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\delta}_0)\}\|_2 = O_p(h_b^{q-1/2}).$$

Further, by (C6) we have $\mathbf{T}_{11} = n^{-1/2}\sum_{i=1}^n \mathbf{S}_{eff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g) + O_p(n^{1/2}h_b^q)$. Since $h_b^{q-1/2} = o_p(n^{1/2}h_b^q)$, and $n^{1/2}h_b^q = o_p(1)$ by conditions (C4) and (C5), then

$$\mathbf{T}_1 = n^{-1/2}\sum_{i=1}^n \mathbf{S}_{eff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g) + o_p(1). \tag{D.2}$$

We next consider each term in $\mathbf{T}_2(\widetilde{\boldsymbol{\delta}}_n)$. Since $\widehat{\boldsymbol{\gamma}}_n(\cdot)$ satisfies $n^{-1}\sum_{i=1}^n \mathbf{S}_{\mathrm{res}2}^*\{Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})\} = \mathbf{0}$ for any $\boldsymbol{\delta}$,

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial \mathbf{S}_{\mathrm{res}2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n)}{\partial \boldsymbol{\delta}^{\mathrm{T}}} + \frac{1}{n}\sum_{i=1}^n \frac{\partial \mathbf{S}_{\mathrm{res}2}^*\{Y_i, W_i, \mathbf{Z}_i, , \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})^{\mathrm{T}}}\frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^{\mathrm{T}}} = \mathbf{0}.$$

Then

$$\frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^{\mathrm{T}}} = -\{\mathbf{T}_{23}(\boldsymbol{\delta})\}^{-1}\mathbf{T}_{24}(\boldsymbol{\delta}),$$

where

$$\begin{aligned}
\mathbf{T}_{23}(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^n \frac{\partial \mathbf{S}_{\mathrm{res}2}^*\{Y_i, W_i, \mathbf{Z}_i, , \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\delta})^{\mathrm{T}}}, \\
\mathbf{T}_{24}(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^n \frac{\partial \mathbf{S}_{\mathrm{res}2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}, \widehat{\boldsymbol{\gamma}}_n)}{\partial \boldsymbol{\delta}^{\mathrm{T}}}.
\end{aligned}$$

Hence

$$\mathbf{T}_2(\widetilde{\boldsymbol{\delta}}_n) = \mathbf{T}_{21}(\widetilde{\boldsymbol{\delta}}_n) - \mathbf{T}_{22}(\widetilde{\boldsymbol{\delta}}_n)\{\mathbf{T}_{23}(\widetilde{\boldsymbol{\delta}}_n)\}^{-1}\mathbf{T}_{24}(\widetilde{\boldsymbol{\delta}}_n).$$

By the consistency of $\widetilde{\boldsymbol{\delta}}_n$ to $\boldsymbol{\delta}_0$ and $\mathbf{B}(x)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_n$ to $g(x)$, we have

$$\mathbf{T}_{21}(\widetilde{\boldsymbol{\delta}}_n) = E\left\{ \frac{\partial \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\delta}_0^{\mathrm{T}}} \bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=g(\cdot)} \right\} \{1 + o_p(1)\},$$

and

$$\mathbf{T}_{24}(\widetilde{\boldsymbol{\delta}}_n) = E\left\{ \frac{\partial \mathbf{S}_{\mathrm{res}2}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\delta}_0^{\mathrm{T}}} \bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=g(\cdot)} \right\} \{1 + o_p(1)\}.$$

From (D.1), we also have

$$\mathbf{T}_{22}(\widetilde{\boldsymbol{\delta}}_n) = E\left\{ \frac{\partial \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=g(\cdot)} \right\} \{1 + o_p(1)\} = O_p(h_b^q).$$

Based on the proof of Proposition 4 in Jiang & Ma (2017), we have $\|\mathbf{T}_{23}(\widetilde{\boldsymbol{\delta}}_n)^{-1}\|_2 = O_p(h_b^{-1})$. Therefore we have $\mathbf{T}_{22}(\widetilde{\boldsymbol{\delta}}_n)\{\mathbf{T}_{23}(\widetilde{\boldsymbol{\delta}}_n)\}^{-1}\mathbf{T}_{24}(\widetilde{\boldsymbol{\delta}}_n) = O_p(h_b^{q-1})$, where $q > 1$ by condition (C2). Thus

$$\mathbf{T}_2(\widetilde{\boldsymbol{\delta}}_n) = E\left\{ \frac{\partial \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\delta}_0^{\mathrm{T}}} \bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=g(\cdot)} \right\} \{1 + o_p(1)\} + O(h_b^{q-1}).$$

Therefore,

$$\begin{aligned} & \sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \\ = & -\left[E\left\{ \frac{\partial \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, \boldsymbol{\gamma})}{\partial \boldsymbol{\delta}_0^{\mathrm{T}}} \bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=g(\cdot)} \right\} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g) \\ & + o_p(1). \end{aligned}$$

Since $n^{-1/2}\sum_{i=1}^n \mathbf{S}_{\mathrm{e}ff}^*(Y_i, W_i, \mathbf{Z}_i, \boldsymbol{\delta}_0, g)$ is the sum of zero-mean random vectors, this will converge in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$ given in Theorem 4. $\square$

Table D.1: Simulation results under a correct working model

		β_1	β_2	β_3	β_4
	truth	1.0	0.5	1.0	-0.3
Simulation 1	mean	1.0183	0.5096	1.0106	-0.3103
	median	1.0125	0.5075	1.010	-0.3086
	se	0.0955	0.0817	0.1231	0.0792
	mse	0.0095	0.0068	0.0153	0.0064
Simulation 2	mean	1.0177	0.5052	1.0074	-0.3051
	median	1.0136	0.5043	1.0068	-0.3029
	se	0.0858	0.0467	0.0917	0.0425
	mse	0.0077	0.0022	0.0085	0.0018

Table D.2: Simulation results under a misspecified working model

		β_1	β_2	β_3	β_4
	truth	1.0	0.5	1.0	-0.3
Simulation 1	mean	1.0145	0.5122	1.0147	-0.3100
	median	1.0113	0.5117	1.0172	-0.3087
	se	0.0941	0.0823	0.1240	0.0822
	mse	0.0091	0.0069	0.0156	0.0069
Simulation 2	mean	1.0149	0.5051	1.0063	-0.3035
	median	1.0083	0.5038	1.0059	-0.3017
	se	0.0762	0.0463	0.0882	0.0320
	mse	0.0060	0.0022	0.0078	0.0010

Table D.3: Realdata analysis results

	β_1	β_2	β_3
estimates	-0.8076	-1.0970	-0.5150
bootstrap mean	-0.7876	-1.1012	-0.4896
bootstrap median	-0.7770	-1.0953	-0.4827
bootstrap se	0.3575	0.3133	0.2947
P-value	0.0239	<0.0001	0.0805

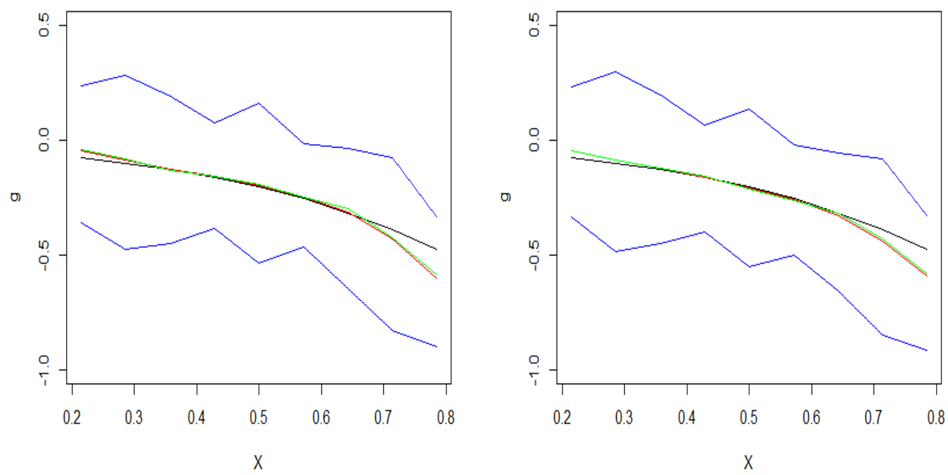

Figure D.1 True function (black line), median estimation (green line), mean estimation (red line) and 90% confidence band (blue line) of $g(x)$ in simulation 1. Correct working model on the left and misspecified working model on the right.

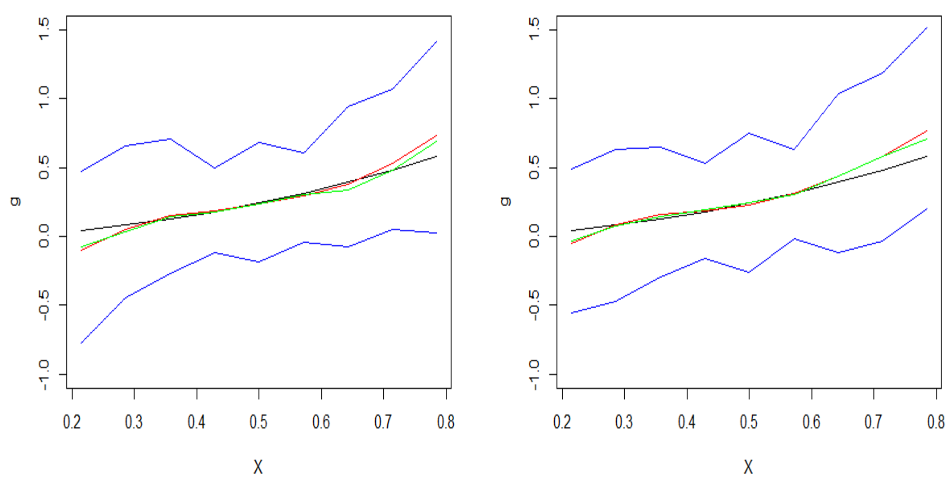

Figure D.2 True function (black line), median estimation (green line), mean estimation (red line) and 90% confidence band (blue line) of $g(x)$ in simulation 2. Correct working model on the left and misspecified working model on the right.

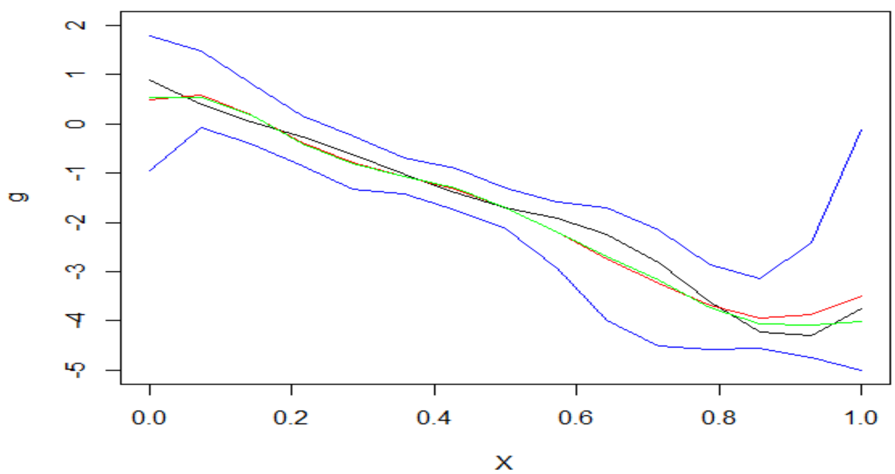

Figure D.3 Estimated $g(x)$ for real data (black line), median estimation (green line), mean estimation (red line) and 90% confidence band (blue line) of $g(x)$ from 1000 bootstrapped samples.