Spring 2017

Semiparametric Estimation and Inference in Causal Inference and Measurement Error Models

Jianxuan Liu
University of South Carolina

Semiparametric Estimation and Inference in Causal Inference and Measurement Error Models

by

Jianxuan Liu

Bachelor of Management
Capital University of Economics and Business China 2007

Master of Science
East Carolina University 2010

Master of Science
Indiana University Bloomington 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

2017

Accepted by:

Yanyuan Ma, Major Professor

John M. Grego, Committee Member

Edsel A. Pena, Committee Member

Lianming Wang, Committee Member

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

Acknowledgments

I would like to thank all the people who contributed in some way to the work described in this dissertation.

First and foremost, I thank my academic advisor, Professor Yanyuan Ma, for her excellent guidance, patience, encouragement and faithful support during my PhD journey. I appreciate all her contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm she has for her research was contagious and motivational for me, even during tough times in my research. I am also thankful for the excellent example she has provided as a successful woman in science and professorship.

Additionally, I would like to thank my committee members Professors Edsel Pena, John Grego and Lianming Wang for their interest in my work. I thank Professor Edsel Pena as a great mentor who infuses me with lots of statistical wisdom. I am grateful for Professor Lianming Wang for his encouragement during hardship and his faithfulness in my abilities. My work would never have been accomplished without the support from Professor John Grego after my major advisor moved forward for her new position at Penn State University. Although they are not in my committee, I like to thank my coauthors Professors Lan Wang, Ray Carroll and Liping Zhu for all their insightful inputs. I also thank Professors Joshua Tebbs, Timothy Hanson, Yen-Yi Ho, Xianzhen Huang, Paramita Chakraborty and Gabriel Terejanu for their wonderful teaching inside and outside classroom, for their willingness to offer me their best help. I also want to thank Dr. Jason Brinkley for the guidence and opportunities he gave me during my research internship at American Institutes for

Research. I sincerely appreciate his insights and generous suggestions on career.

I am indebted to all my friends who were always so helpful in numerous ways. I thank them for their accompany and willingness to help which made my PhD studies vividly. Friday Happy Hours with girls at Starbucks would never fade away as the time goes by. I can't thank my friends more who offered their help at the time I had a bike accident and two auto accidents which were caused by the other drivers.

This dissertation would not have been possible without the love and blessings bestowed on me from the Almighty God who gave me strength, talent, perseverance and endurance to finish my work. I am thankful for His miracle healing of my memory loss problem in June 2014 which was due to a sequence of life crisis started in August 2009. I thank God for giving me the hard time which allow me to grow, to defeat my most inner weakness which I might not be aware of and to become stronger and faithful in Him. Special thank goes to Professor John Bishop who has been a life mentor to me since August 2009. I am so blessed with the love and care from John and his lovely wife Mary. I want to thank them for their faithful support in times of good and bad. My PhD study was also enriched by the Harvest Chinese Christian Church, Columbia Chinese Christian Church and the USC Student Fellowship, First Baptist Church Columbia, Raleigh Chinese Christian Church and Chinese Family For Christ. I thank all my Pastors and Sunday School teachers for their wonderful teaching and sharing of God's words.

Finally, I would also like to say a heartfelt thank you to my mom, daughter, husband and parents-in-law for their understanding, love and support in whatever way they could during this challenging period.

Abstract

This dissertation research has focused on theoretical and practical developments of semiparametric modeling and statistical inference for high dimensional data and measurement error data. In causal inference framework, when evaluating the effectiveness of medical treatments or social intervention policies, the average treatment effect becomes fundamentally important. We focus on propensity score modelling in treatment effect problems and develop new robust tools to overcome the curse of dimensionality. Furthermore, estimating and testing the effect of covariates of interest while accommodating many other covariates is an important problem in many scientific practices, including but not limited to empirical economics, public health and medical research. However when the covariates of interest are measured with error, to evaluate the effect precisely, we must reduce the bias caused by measurement error and adjust for the confounding effects simultaneously. We design a general methodology for a general single index semiparametric measurement error model and for a class of Poisson models, and introduce a bias-correction approach to construct a class of locally efficient estimators. We derive the corresponding estimating procedures and examine the asymptotic properties. Extensive simulation studies have been conducted to verify the performance of our semiparametric approaches.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Semiparametric models became very popular in the last decade because they are flexible and in the meantime the estimators are efficient. It is well-known that parametric models rely heavily on the correctness of models which reside in a restrictive set of dimensions. Hence misspecification of the model will lead to severe bias and inconsistency in estimation. While nonparametric models, flexible in the sense that it allows infinitely many parameters in the model, often suffer from the curse of dimensionality which results in slow convergence rates and the lack of interpretability due to overly opaque structure of the model. A semiparametric model overcomes these drawbacks by permitting precise estimation of the (parametric) components of interest and maintains interpretability of the model while leaving out some features of the model completely unspecified (nonparametric component). Semiparametric models date back to Newey (1990) and have been further developed by Bickel et al. (1993) and Tsiatis (2006).

Consider a class of statistical models

$$\{p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

where $\mathbf{X}_1, X_2, \cdots, \mathbf{X}_n$ are independent identically distributed random vectors. The parameter $\boldsymbol{\theta}$ can be written as $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}$ is the parameter of interest while $\boldsymbol{\eta}$ is the nuisance parameter. Let us denote the true parameters by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\eta}_0^{\mathrm{T}})^{\mathrm{T}}$. Under the semiparametric theory framework, in order to construct a consistent, regular asymptotically linear (RAL) and locally efficient estimator , we utilize an influence

function $\varphi(\mathbf{X})$ such that $E\{\varphi(\mathbf{X})\} = \mathbf{0}$ and

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}} - \boldsymbol{\beta}_0) = n^{1/2}\sum_{i=1}^{n}\varphi(\mathbf{X}_i) + o_p(1),$$

where $\varphi(\mathbf{X}_i), i = 1, 2, \cdots, n$ are i.i.d. mean zero random vectors and $o_p(1)$ denotes a term that converges to zero in probability. The random vector $\varphi(\mathbf{X}_i)$ is referred to the i-th influence function which satisfies that $E(\varphi\varphi^{\mathrm{T}})$ is finite and nonsingular (Bickel et al. (1993)). From the geometry point of view, an influence function is orthogonal to the Hilbert space $\mathcal{H}$ which is a complete normed linear vector space equipped with an inner product. The asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ can be derived from the influence function $\varphi(\mathbf{X})$. See Tsiatis (2006) for further details.

The beauty of semiparametric method prevails over a wide range of applications, especially when other methods are overwhelmed by the dimensions of the parameter space. My dissertation work investigates the frequentist semiparametric methods thoroughly with applications on dimension reduction problem in causal inference and errors-in-variables (EIV) models.

The problem of estimating average treatment effect is of fundamental importance when evaluating the effectiveness of medical treatments or social intervention policies. Most of the existing methods for estimating average treatment effect rely on some parametric assumptions on the propensity score model or outcome regression model one way or the other. In reality, both models are prone to misspecification, which can have undue influence on the estimated average treatment effect. In Chapter 2, we propose a new robust approach to estimating the average treatment effect based on observational data in the challenging situation when neither a plausible parametric outcome model nor a reliable parametric propensity score model is available. Our estimator can be considered as a robust extension of the popular class of propensity score weighted estimators. The new approach has the advantage of being robust, flexible, data adaptive and it can handle many covariates simultaneously. Adapting a dimension reduction approach, we estimate the propensity score weights semipara-

metrically by using a nonparametric link function to relate the treatment assignment indicator to a low-dimensional structure of the covariates which are formed typically by several linear combinations of the covariates. We develop a class of consistent estimators for the average treatment effect and studied their theoretical properties. We demonstrate the robust performance of the new estimators on simulated data and a real data example of analyzing the effect of maternal smoking on babies' birth weight.

In Chapter 3, we introduce a general single index semiparametric measurement error model for the case that the main covariate of interest is measured with error and modeled parametrically, and where there are many other variables also important to the modeling. We propose a semiparametric bias-correction approach to estimate the effect of the covariate of interest. The resultant estimators are shown to be root-n consistent, asymptotically normal and locally efficient. Comprehensive simulations and an analysis of an empirical data set are performed to demonstrate the finite sample performance and the bias reduction of the locally efficient estimators.

We extend the work on measurement error problem to a wider range of counting response models in Chapter 4. We thoroughly examine Poisson models for a counting response Y where the covariates of interest are measured with normal additive error and possible confounding effect. We propose a class of constructive locally efficient semiparametric estimators for a class of Poisson models in the presence of functional measurement error. The estimators follow from the estimating equations that are based on the semiparametric efficient score derived under a possibly incorrect distributional assumption for the unobserved covariate. Our approach produces locally efficient estimator with significant bias reductions among current existing methods. We verify the asymptotic properties of the bias reductions estimators through extensive simulation studies.

Finally, I conclude with summary remarks in Chapter 5 about the work on semi-

parametric estimation and inference in causal inference and measurement error models.

Chapter 2

A Alternative Robust Estimator of Average Treatment Effect in Causal Inference

2.1 Introduction

Estimating average treatment effect is important in comparing different medical treatments, social programs and intervention policies. The problem is challenging when the data come from an observational study instead of a randomized experiment. Direct differencing of the sample average effects is susceptible to confounding bias, which is caused by imbalances in baseline covariate distributions between the treatment group and the control group.

Under the commonly imposed no unmeasured confounders assumption (Rosenbaum and Rubin (1983); De Luna et al. (2011)), a variety of methods have been proposed to consistently estimate the average treatment effect. The class of doubly robust (DR) estimators (Scharfstein et al. (1999); Robins and Rotnitzky (2001); Bang and Robins (2005); Rubin and van der Laan (2007); Cao et al. (2009); Tan (2010); Rotnitzky et al. (2012); Van der Laan and Rose (2011); Vansteelandt et al. (2012), among others) have been particularly popular due to their double protection against model misspecification. The standard practice in the DR estimation relies on parametric specification of the propensity score model and the outcome regression model. Let $\mathcal{M}_\beta$ denote the class of joint densities satisfying the parametric assumptions implied by the outcome regression model (indexed by β) and let $\mathcal{M}_\gamma$ denote the class of joint densities satisfying the parametric assumptions implied by the propensity score

model (indexed by β). The DR estimator is consistent under $\mathcal{M}_\beta \cup \mathcal{M}_\gamma$, that is, as long as one of these two classes of parametric models is correctly specified.

Despite the appealing theoretical properties of DR estimators, Carpenter et al. (2006), Kang and Schafer (2007) and Vansteelandt et al. (2012) observed in simulations that their finite-sample bias can be amplified when one of the working models is misspecified and the bias can be severe if both models are slightly misspecified. Vermeulen and Vansteelandt (2015) recently proposed a novel generic strategy for estimating the nuisance parameters of the working models of an arbitrary DR estimator by minimizing the squared first-order bias of the DR estimator under the misspecification of both working models. Their approach aimed at bias reduction under misspecification of both models. Vermeulen and Vansteelandt (2016) further explored the use of data-adaptive estimators in constructing bias-reduced doubly robust estimation. These estimators provide useful improvement over standard DR estimators, but still need at least one working model to be correctly specified using a parametric model.

Motivated by the practical concern of bias reduction, we propose an alternative approach by directly considering estimators of average treatment effects that are consistent in a larger class of semiparametric models. The semiparametric class we study imposes a semiparametric structure for the propensity score model while imposing no structure for the outcome regression model. It encompasses the commonly assumed parametric class $\mathcal{M}_\beta \cup \mathcal{M}_\gamma$. As a direct consequence, our proposed estimator is expected to be consistent for many distributions outside $\mathcal{M}_\beta \cup \mathcal{M}_\gamma$, for which most of the standard doubly robust estimators would become inconsistent. This was demonstrated by the numerical results in Section 4. Furthermore, we derive the asymptotic normality of the proposed estimator for the average treatment effect, which remains valid for this general class of semiparametric distributions. This is in contrast to the theory in the literature for doubly robust estimators, whose asymptotic normality

relies on the correct specification of either $\mathcal{M}_\beta$ or $\mathcal{M}_\gamma$.

Some alternative methods have been proposed in the literature to relax the parametric specification of working models. Wang et al. (2010) proposed a nonparametric estimator of the outcome regression model in the setting of missing data analysis. On the other hand, Hirano et al. (2003) showed that if the propensity score function is estimated nonparametrically, then the propensity score weighted estimator of the average treatment effect achieves the semiparametric efficiency bounds. However, the nonparametric approach is not feasible for real data analysis when many covariates are available due to the curse of dimensionality. Imai and Ratkovic (2014) introduced covariate balancing propensity score as a method that is robust to mild misspecification of the parametric propensity score model. Several authors (McCaffrey et al. (2004); Ridgeway and McCaffrey (2007); Petersen et al. (2007); Westreich et al. (2010); Lee et al. (2010)) explored machine learning approaches for modeling the propensity score. Although numerical examples suggest promising performance, these works have not studied the asymptotic properties of the resulted average treatment effect estimator. van der Laan Mark and Daniel (2006) proposed targeted maximum likelihood estimators (TMLE) that incorporates the state-of-art of machine learning and uses an ensemble of models. Cross-validation was used to select the best weighted combination of these estimators (Van der Laan et al. (2007)). If the ensemble estimator for the propensity score model or the outcome regression model is consistent for the underlying true model, the TMLE is consistent hence doubly robust. However, if the nuisance parameter estimator is not based on the correctly specified parametric model, but instead on a data-adaptive estimator, the bias of standard TMLE converges to zero at a rate slower than $n^{-1/2}$. van der Laan (2014a) further showed that additional targeting of the estimators of the nuisance parameters can guarantee that the bias of the estimator of the target parameter is of second order and hence asymptotically linear. In practice, this would require double targeting and is more

computationally intensive.

The approach we propose does not rely on parametric specification of the propensity score model or the outcome regression models. It has the advantage of being more robust and is flexible to handle many covariates. Specifically, we relax the commonly imposed parametric assumption on the propensity score model by only assuming the probability of assigning the treatment depends on the p-dimensional covariate vector $\mathbf{X}$ through several linear combinations $\mathbf{B}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{B}$ is a $p \times d$ matrix with $d < p$. We then estimate this conditional probability by employing a nonparametric link function. Note that many works exist in studying how to model the relation between a binary response and many covariates, see for example, Pregibon (1980), Koenker and Yoon (2009), Li et al. (2016). The special case of $d = 1$ yields the single index model and is especially well studied (Hardle et al., 2004). As an intermediate model for the propensity score in the treatment effect estimation, our semiparametric approach for estimating the propensity score is most closely related to the sufficient dimension reduction literature (Cook, 1998) and is of independent interest. Existing methods for estimating the dimension reduction space such as sliced inverse regression (SIR)(Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), directional regression (DR) (Li and Wang, 2007), generalized DR (Li and Dong, 2009; Dong and Li, 2010) have two limitations to be applied to our problem. First, they rely mainly on a linearity condition and/or a constant variance condition, or their generalized form, which may not hold in our problem. Second, they require a reversal of the relation between $\mathbf{X}$ and T to compute expectation of functions of the covariates $\mathbf{X}$ conditional on T, which generates only two different values because T only has two values. This hampers the direct application of these methods. On the other hand, other methods based on nonparametric regression (Xia, 2007) and semiparametric regression (Ma and Zhu, 2012, 2013) exist, but they also need to be adapted instead of directly applied to estimating the propensity score

which concerns binary response.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the multi-index semiparametric estimator of the propensity score function and a robust estimator of the average treatment effect. In Section 2.3, we study the asymptotic properties of the estimators. Simulation studies are conducted and presented in Section 3.4. We illustrate the usefulness of the method in a real data example of analyzing effect of maternal smoking on babies' birth weight in Section 2.5 and conclude the chapter with a brief discussion in Section 2.6. The appendix contains the derivation of the efficient score function and the proof of Theorem 1. The regularity conditions and proofs of Lemmas 1-3 are given in the online supplementary document.

2.2 A robust estimator of the average treatment effect

Notation and setup

We consider the popular setting of a binary treatment T ($T = 1$ for treatment and 0 for control). To evaluate the treatment effect, we adopt the potential or counterfactual outcome framework (Neyman et al., 1990; Rubin, 1974). Let $Y^*(1)$ be the outcome of the subject if s/he received treatment; and $Y^*(0)$ be the outcome if s/he received the non-treatment. Our goal is to estimate the average treatment effect

$$\tau = E\{Y^*(1) - Y^*(0)\}.$$

The complexity of the problem arises because for each individual in the sample, we observe either $Y^*(1)$ or $Y^*(0)$, but not both. The observed outcome is $Y = TY^*(1) + Y^*(0)(1 - T)$, that is, the observed outcome is the potential outcome corresponding to the treatment the subject actually receives, which is often referred to as the consistency assumption in causal inference (Rubin, 1986).

Given data from an observational study $\{Y_i, T_i, X_i\}$, $i = 1, \ldots, n$, where Y_i is the response of the ith subject, T_i is the binary treatment indicator, X_i is a vector of

covariates, we are interested in estimating the average causal effect of the treatment. Direct differencing the sample averages of the treatment and control groups often leads to a biased estimator of τ in observational studies as the two groups often differ in some covariates that are associated with both the treatment and outcome. Let $\pi(X) = P(T = 1|X)$ be the propensity score function and assume that the unconfoundedness assumption is satisfied, that is $\{Y^*(1), Y^*(0)\} \perp T|X$, or the treatment assignment is independent of the potential outcomes given the covariates. Rosenbaum and Rubin (1983) showed that adjusting for the difference in propensity score can completely remove the confounding bias from the difference in covariates.

Hahn (1998) derived the semiparametric efficiency bound for estimating τ. The propensity score can be used in different ways to obtain a consistent estimator for the average treatment effect. Hahn (1998) also proposed an estimator that achieves the semiparametric efficiency bound, but his estimator involves estimating $E(YT|X)$, $E(Y(1-T)|X)$ and $\pi(X)$. Hirano et al. (2003) further showed that a simpler estimator that only estimates $\pi(X)$ nonparametrically can also achieve the semiparametric efficiency bound. However, these nonparametric estimators suffer from the curse of dimensionality in real data analysis even with a moderate amount of covariates such as four covariates.

In practice, existing work on causal inference usually adopts a parametric approach to modeling the propensity score function. For example, logistic models are frequently used to model disease occurrence in case-control studies (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005; Lin and Zeng, 2009; Ma and Carroll, 2016), in missing probability problem (Rubin, 1976; Rubin and Little, 2002), and even in survival models (Efron, 1988). However, the parametric approach is prone to model misspecification and can result in substantial bias.

The crux of our robust estimator of the average treatment effect is to develop a flexible estimator of the propensity score function. Instead of the parametric logistic

regression model for the propensity score function, we assume

$$\text{pr}(T = 1|\mathbf{X} = \mathbf{x}) \quad = \quad \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x})\}}{1 + \exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x})\}}, \tag{2.1}$$

where $\mathbf{X} \in \mathcal{R}^p$, $\mathbf{B} \in {}^{p\times d}$ and η is an arbitrary unspecified function. Note that we use the logit link function here for parameterization purpose to ensure that the depicted probability function takes values between 0 and 1. As the function η is completely unspecified, our model allows the probability of being assigned to the treatment to depend on several linear combinations of X in a nonparametric fashion. In contrast, the popular logistic regression model assumes this probability to depend on one particular linear combination of X in a known parametric fashion.

Flexible estimation of the propensity score

To obtain a more concise form, we rewrite (2.1) equivalently as

$$\text{pr}(T = t \mid \mathbf{X} = \mathbf{x}) = \frac{\exp\{t\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x})\}}{1 + \exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x})\}}. \tag{2.2}$$

The log-likelihood function of $\mathbf{B}$ and η is

$$\sum_{i=1}^{n}(t_i\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i) - \log[(1 + \exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}]).$$

For identifiability of $\mathbf{B}$, we require $\mathbf{B}$ to have the form $\mathbf{B} = (\mathbf{I}_d, \mathbf{B}_l^{\mathrm{T}})^{\mathrm{T}}$, where the upper submatrix $\mathbf{I}_d$ is the $d \times d$ identity matrix while the lower submatrix $\mathbf{B}_l$ is an arbitrary $(p - d) \times d$ matrix. To estimate the semiparametric propensity score function, we need to estimate $\mathbf{B}_l$ and the unknown function η, the former of which contains $p_t = (p - d)d$ free parameters while the latter can be viewed as an infinite dimensional parameter. In the sequel, for notational convenience, for any arbitrary $p\times d$ matrix $\mathbf{B}$, we define the concatenation of the columns contained in the lower $p-d$ rows of $\mathbf{B}$ as $\text{vecl}(\mathbf{B}) = \text{vec}(\mathbf{B}_l) = (\beta_{d+1,1}, ...\beta_{p,1}, ...\beta_{d+1,d}, ...\beta_{p,d})^{\mathrm{T}}$ where "vec" stands for vectorization while "vecl" is the vectorization of the lower part of the original matrix.

Our approach of estimation relies on first deriving the influence function using the geometric technique (Bickel et al., 1993; Tsiatis, 2006). In the Appendix, we derive the efficient score function with respect to $\mathbf{B}$:

$$\begin{aligned} & \mathbf{S}_{\text{eff}}(t_i, \mathbf{x}_i, \mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') \\ = \; & \text{vecl}\left(\{\mathbf{x}_i - E(\mathbf{X}_i \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\} \left[t_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right). \end{aligned} \tag{2.3}$$

We use the Nadaraya-Watson kernel estimator to estimate $E(\mathbf{X}_i \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$, that is,

$$\widehat{E}(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i K_h(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x})}{\sum_{i=1}^{n} K_h(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x})}, \tag{2.4}$$

where h is a bandwidth and K is a multivariate kernel function, $K_h(\cdot) = K(\cdot/h)/h^d$. Neither η nor $\boldsymbol{\eta}'$ is known in real data analysis. To deal with this complexity, in the following we borrow the idea of locally efficient and adaptively efficient estimators in general and especially in Ma and Zhu (2012) and consider two different options, which lead to two different estimators of $\mathbf{B}$.

First, we consider an estimator of $\mathbf{B}$ based on a posited form of η, denoted as η^*, which may not be identical to η. The corresponding derivative is denoted by $\boldsymbol{\eta}^{*\prime}$. This yields the locally efficient score function

$$\begin{aligned} & \mathbf{S}^*_{\text{eff}}(t_i, \mathbf{x}_i, \mathbf{B}, \eta^*, \boldsymbol{\eta}^{*\prime}) \\ = \; & \text{vecl}\left(\{\mathbf{x}_i - E(\mathbf{X}_i \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\} \left[t_i - \frac{\exp\{\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}^{*\prime}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right). \end{aligned} \tag{2.5}$$

Obviously, there are many different choices of η, as long as η^* is a smooth function of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$. For example, when we choose $\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{1}_d$ is a length d vector of ones. Then $\boldsymbol{\eta}^{*\prime}(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d$. The locally efficient estimator of $\mathbf{B}$ solves the estimating equation

$$\sum_{i=1}^{n} \text{vecl}\Big[\{\mathbf{x}_i - \widehat{E}(\mathbf{X}_i \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\} \{ t_i - \frac{\exp(\mathbf{1}_d^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)}{1+\exp(\mathbf{1}_d^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)} \} \mathbf{1}_d^{\mathrm{T}} \Big] = \mathbf{0}. \tag{2.6}$$

We denote this estimator by $\widehat{\mathbf{B}}_1$.

Next, we consider estimating $\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$ and its first derivative nonparametrically to obtain an adaptively efficient estimator of $\mathbf{B}$. We adopt the local linear kernel method to estimate $\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x})$ and its first derivative, which solves

$$\sum_{i=1}^{n}\left[t_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0)\}}{1+\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0)\}}\right] K_h(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0) \;=\; 0 \tag{2.7}$$

$$\sum_{i=1}^{n}\left[t_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0)\}}{1+\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0)\}}\right] (\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0) K_h(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i - \mathbf{B}^{\mathrm{T}}\mathbf{x}_0) \;=\; \mathbf{0}. \tag{2.8}$$

The estimators $\widehat{b}_0$ and $\widehat{\mathbf{b}}_1$ are the estimators of η and $\boldsymbol{\eta}'$ at $\mathbf{B}^{\mathrm{T}}\mathbf{x}_0$, respectively. We can vary $\mathbf{x}_0$ to obtain estimates of the functions at various values. We write the resulting estimators as $\widehat{\eta}(\cdot, \mathbf{B})$ and $\widehat{\boldsymbol{\eta}}'(\cdot, \mathbf{B})$, which can be considered as profiled estimators for η and $\boldsymbol{\eta}'$. We subsequently plug $\widehat{\eta}(\cdot, \mathbf{B})$, $\widehat{\boldsymbol{\eta}}'(\cdot, \mathbf{B})$, $\widehat{E}(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ into (2.3) and solve for $\mathbf{B}$ to obtain the efficient estimator, which we denote by $\widehat{\mathbf{B}}_2$.

Robust estimation of the average treatment effect

To estimate the average treatment effect robustly, we propose to use

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\widehat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\widehat{\pi}(X_i)}\right\}, \tag{2.9}$$

where $\widehat{\pi}(X_i)$ is obtained from the semiparametric model (2.1) and estimated using either of the two options discussed in Section 2.2. Algorithm 1 below depicts the detailed steps of obtaining the estimator $\widehat{\tau}$ when the locally efficient estimator of $\pi(X_i)$ is used (i.e., based on $\widehat{\mathbf{B}}_1$). The algorithm based on $\widehat{\mathbf{B}}_2$ is similar. The above procedure can be considered as an extension of the celebrated Horvitz-Thompson inverse probability weighted estimator (Horvitz and Thompson, 1952), which was originally developed for survey sampling.

The proposed estimator enjoys nice robustness properties. It is more flexible than the parametric propensity score model and hence is less prone to misspecification.

Algorithm 1 Robust estimator of the average treatment effect

Input: $\{Y_i, T_i, X_i\}$, $i = 1, \ldots, n$, where Y_i is the response of the ith subject, T_i is a binary treatment indicator ($T_i = 1$ for treatment and 0 for control), X_i is a vector of covariates.

Output: Estimator $\widehat{\tau}$.

1: Use (2.6) to obtain a local efficient estimator of $\mathbf{B}$, denoted as $\widetilde{\mathbf{B}}$ via, for example, choosing $\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}$.

2: Perform nonparametric estimation of $\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$ and its first derivative $\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$ by implementing (2.7). Write the resulting estimator as $\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \mathbf{B})$ and $\widehat{\boldsymbol{\eta}}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \mathbf{B})$.

3: Perform nonparametric estimation of $E(\mathbf{X}_i \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$. Write the resulting estimator as $\widehat{E}(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)$.

4: Plug $\widehat{\eta}(\cdot, \mathbf{B})$, $\widehat{\boldsymbol{\eta}}'(\cdot, \mathbf{B})$ and $\widehat{E}(\cdot)$ in to $\mathbf{S}_{\text{eff}}$ and solve the estimating equation

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(y_i, \mathbf{x}_i, \mathbf{B}, \widehat{\eta}, \widehat{\boldsymbol{\eta}}', \widehat{E}) = \mathbf{0},$$

using $\widetilde{\mathbf{B}}$ as starting value, to obtain the efficient estimator $\widehat{\mathbf{B}}$.

5: Repeat Step 2 to obtain the final estimator of $\eta(\cdot)$ and form $\widehat{\pi}(X_i) = 1 - 1/[1 + \exp\{\widehat{\eta}(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x})\}]$.

6: **return** $\widehat{\tau} = n^{-1}\sum_{i=1}^{n}\left\{\dfrac{T_i Y_i}{\widehat{\pi}(X_i)} - \dfrac{(1-T_i)Y_i}{1-\widehat{\pi}(X_i)}\right\}$.

Furthermore, it does not rely on the outcome regression models. One can further pursue a double robust estimator by augmenting the estimator we propose. It could further improve estimation efficiency at the price of more complex modeling and/or computation. The estimator can accommodate a large number of covariates. Note that although nonparametric smoothing is used to estimate $\pi(X_i)$, the smoothing is implemented with respect to $\mathbf{B}^{\mathrm{T}}\mathbf{x}$. Under the dimension reduction assumption, it is often sufficient to consider a small d in practice; our estimator does not face the kind of curse of dimensionality that prevents the practical implementation of the estimators in Hahn (1998) and Hirano et al. (2003). Furthermore, we allow the covariate $\mathbf{X}$ to include both continuous and discrete or categorical variables without imposing any distributional assumptions on the covariate.

Remark 1. *A technical detail involved in the nonparametric step of the above pro-*

cedure is bandwidth selection. Through extensive numerical experimentation, we find that the $\boldsymbol{B}$ *estimation procedure is quite insensitive to the bandwidth, while inference precision could be affected by the bandwidth. Thus, guided by the theoretical properties, we recommend to simply set the bandwidth to be* $var(\|X_i\|_2)n^{-1/5}$ *throughout the estimation of* $\boldsymbol{B}$*, and use a leave-one-out cross-validation procedure to obtain the smoothing parameter* h *in estimating* η *after fixing* $\widehat{\boldsymbol{B}}$*. The same bandwidth then can be used in the inference procedure.*

2.3 Asymptotic Properties

In this section, we study the asymptotic properties of the estimators for the propensity score function for the robust estimator of the average treatment effect. The regularity conditions that are needed for the theoretical development are given in the Appendix.

First, we study the asymptotic properties of $\widehat{\mathbf{B}}_1$, the nonparametric estimators of $\eta, \boldsymbol{\eta}'$ and $\widehat{\mathbf{B}}_2$ discussed in Section 2.2. The results are summarized in Lemmas 1-2 below. The proofs are relegated to the Appendix.

Lemma 1. *Let* $\widehat{\boldsymbol{B}}_1$ *be the estimator defined in Section 2.2. Under the regularity conditions (C1)-(C6),* $\widehat{\boldsymbol{B}}_1$ *is locally efficient. As* $n \to \infty$*,*

$$\sqrt{n}\{vecl(\widehat{\boldsymbol{B}}_1) - vecl(\boldsymbol{B})\} \to N\{\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{G}(\boldsymbol{A}^{-1})^{T}\}$$

in distribution, where

$$\begin{aligned}
\boldsymbol{A} &= E\left\{\frac{\partial}{\partial(vecl\,\boldsymbol{B})^{T}} vecl\left(\{\boldsymbol{X}_i - E(\boldsymbol{X}_i \mid \boldsymbol{B}^{T}\boldsymbol{X}_i)\}\left[T_i - \frac{\exp\{\eta^*(\boldsymbol{B}^{T}\boldsymbol{X}_i)\}}{1+\exp\{\eta^*(\boldsymbol{B}^{T}\boldsymbol{X}_i)\}}\right]\right.\right. \\
&\quad \left.\left. \boldsymbol{\eta}^{*\prime}(\boldsymbol{B}^{T}\boldsymbol{x}_i)^{T}\right)\right\}, \\
\boldsymbol{G} &= E\left\{vecl\left(\left\{\boldsymbol{X}_i - E(\boldsymbol{X}_i \mid \boldsymbol{B}^{T}\boldsymbol{X}_i)\right\}\left[T_i - \frac{\exp\{\eta(\boldsymbol{B}^{T}\boldsymbol{X}_i)\}}{1+\exp\{\eta(\boldsymbol{B}^{T}\boldsymbol{X}_i)\}}\right]\boldsymbol{\eta}'(\boldsymbol{B}^{T}\boldsymbol{X}_i)^{T}\right)^{\otimes 2}\right\}.
\end{aligned}$$

Here and throughout the chapter, $\boldsymbol{a}^{\otimes 2} \equiv \boldsymbol{a}\boldsymbol{a}^{T}$*.*

Lemma 2. *Assume the regularity conditions (C1)-(C4) and (C7)-(C8) hold. The local linear kernel estimators of* $\widehat{\eta}(\boldsymbol{B}^T\boldsymbol{x})$ *and* $\widehat{\boldsymbol{\eta}}'(\boldsymbol{B}^T\boldsymbol{x})$ *defined in Section 2.2 satisfy*

$$E\{\widehat{\eta}(\boldsymbol{B}^T\boldsymbol{x})\} - \eta(\boldsymbol{B}^T\boldsymbol{x}) = O(h^m), \;\; E\{\widehat{\boldsymbol{\eta}}'(\boldsymbol{B}^T\boldsymbol{x})\} - \boldsymbol{\eta}'(\boldsymbol{B}^T\boldsymbol{x}) = O(h^m),$$
$$var\{\widehat{\eta}(\boldsymbol{B}^T\boldsymbol{x})\} = O_p\{(nh^d)^{-1}\}, \quad var\{\widehat{\boldsymbol{\eta}}'(\boldsymbol{B}^T\boldsymbol{x})\} = O_p\{(nh^{d+2})^{-1}\}.$$

Furthermore, $\widehat{\boldsymbol{B}}_2$ *defined in Section 2.2 is efficient and satisfies*

$$\sqrt{n}\{vecl(\widehat{\boldsymbol{B}}_2) - vecl(\boldsymbol{B}_2)\} \to N(\boldsymbol{0}, \boldsymbol{V}^{-1})$$

in distribution as $n \to \infty$*, where*

$$\begin{aligned}\boldsymbol{V} &= E\{\boldsymbol{S}_{\mathit{eff}}(T_i, \boldsymbol{X}_i, \boldsymbol{B}^T\boldsymbol{X}_i, \eta, \boldsymbol{\eta}', E)^{\otimes 2}\} \\ &= E\left\{ vecl\left(\left\{\boldsymbol{X}_i - E(\boldsymbol{X}_i \mid \boldsymbol{B}^T\boldsymbol{X}_i)\right\}\left[T_i - \frac{\exp\{\eta(\boldsymbol{B}^T\boldsymbol{X}_i)\}}{1+\exp\{\eta(\boldsymbol{B}^T\boldsymbol{X}_i)\}}\right]\boldsymbol{\eta}'(\boldsymbol{B}^T\boldsymbol{X}_i)^T\right)^{\otimes 2}\right\}.\end{aligned}$$

We provide the asymptotic property of the average treatment estimator $\widehat{\tau}$ defined in Section 2.2, where the propensity is based on the dimension reduction estimation. We adopt two standard assumptions in causal inference. Assume the treatment allocation is independent of the potential treatment outcome given the covariates. Assume further that the probability of treatment is bounded away from 0 and 1.

Theorem 1. *Under the regularity conditions (C1)-(C8), when* $n \to \infty$ *the estimator* $\widehat{\tau}$ *from (2.9) based on* $\widehat{\mathcal{B}}_2$ *satisfies*

$$\sqrt{n}(\widehat{\tau} - \tau) \to N(0, \sigma^2)$$

in distribution, where $\sigma^2 = \sigma^2_{\mathit{eff}} + \boldsymbol{a}^T E(\boldsymbol{S}_{\mathit{eff}}\boldsymbol{S}_{\mathit{eff}}^T)^{-1}\boldsymbol{a}$*, with*

$$\begin{aligned}\sigma^2_{\mathit{eff}} &= var\left[\left\{\frac{T_iY_i}{\pi(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi(\boldsymbol{X}_i)} - \tau\right\} - \left\{\frac{Y_i^*(1)}{\pi(\boldsymbol{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\boldsymbol{X}_i)}\right\}\{T_i - \pi(\boldsymbol{X}_i)\}\right], \\ \boldsymbol{a} &= E\Big(\left[Y_{i1}\{1-\pi(\boldsymbol{X}_i)\} + Y_i^*(0)\pi(\boldsymbol{X}_i)\right]\boldsymbol{\eta}'(\boldsymbol{B}^T\boldsymbol{X}_i) \otimes \boldsymbol{X}_{iL}\Big).\end{aligned}$$

Remark 2. *In the above asymptotic variance expression,* σ^2_{eff} *is the optimal estimation variance (Hahn, 1998; Hirano et al., 2003). The additional term is the price*

we pay when we use a dimension reduction procedure to estimate π instead of doing it fully nonparametrically. In other words, our estimator is in general not efficient. Whether the propensity score is completely known or completely unknown, the efficiency bound in estimating the average treatment effect is the same. In our context, the propensity score is partially known, in that we know it has the dimension reduction structure. Thus, the efficiency bound in estimating the treatment effect should be in between the completely known and completely unknown cases, and hence is also the same as that given in Hahn (1998). Hirano et al. (2003) shows that an inverse probability weighted estimator using the nonparametrically estimated propensity score achieves the optimal efficiency in estimating the treatment effect, regardless whether the true propensity score is known or not. In fact, they show that even if the true propensity score is known, it should not be used otherwise an efficiency loss will occur. However, estimating the propensity score nonparametrically is often infeasible in practice, especially when there are many covariates. Thus, a natural compromise is to adopt a dimension reduction assumption to facilitate the propensity score estimation, which provides a trade-off between efficiency and practical applicability.

2.4 Monte Carlo studies

A simulation study on estimating the propensity score function

We first conduct a simulation study to investigate the performance of the flexible semiparametric estimators proposed in Section 2.2 for the propensity score.

We generate the vector of covariates $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)^T$ as follows. The covariates X_1 and X_2 are generated from independent standard normal distributions. We let $X_3 = 0.2X_1 + 0.2(X_2 + 2.0)^2$, $X_4 = 0.1 + 0.2(X_1 + X_2) + 0.3(X_1 + 1)^2$, and generate X_5 and X_6 independently from Bernoulli distribution with success proba-

bility $\exp(X_3)/\{1+\exp(X_3)\}$ and $\exp(X_4)/\{1+\exp(X_4)\}$, respectively. In (3.1), we consider the following two different functional forms:

- Setting (1): $\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \sin(\mathbf{B}^{\mathrm{T}}\mathbf{x})$,
  where $d = 1$ and $\mathbf{B} = (1.0, -1.2, 0.8, -1.7, -1.5, 0.5)^{\mathrm{T}}$.
- Setting (2): $\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \sin(\mathbf{B}_1^{\mathrm{T}}\mathbf{x}) + \sin(\mathbf{B}_2^{\mathrm{T}}\mathbf{x})$, where $d = 2$,
  $\mathbf{B}_1 = (1.0, 0.0, 1.2, 0.8, -1.2, 0.8)^{\mathrm{T}}$ and $\mathbf{B}_2 = (0.0, 1.0, 1.3, 0.7, 1.1, -0.7)^{\mathrm{T}}$.

For comparison purposes, we implement the oracle estimator and compare with our proposed semiparametric estimators $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{B}}_2$. The oracle estimator assumes the functional form of η in (3.1) is known, although $E(\mathbf{x}|\mathbf{B}^{\mathrm{T}}\mathbf{x})$ is still estimated through the kernel regression in (2.4). Even though the oracle estimator is unrealistic, it provides a benchmark since it is the best performance one could expect to obtain. The local estimator $\widehat{\mathbf{B}}_1$ replaces η with a mis-specified function in the estimation procedure and estimates $E(\mathbf{x}|\mathbf{B}^{\mathrm{T}}\mathbf{x})$ nonparametrically. We posit the models $\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \sin(\mathbf{B}^{\mathrm{T}}\mathbf{x} + 0.8) - 0.3$ and $\eta^*(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = \sin(\mathbf{B}_1^{\mathrm{T}}\mathbf{x} + 0.5) + \cos(\mathbf{B}_2^{\mathrm{T}}\mathbf{x} - 0.5)$ for setting (1) and (2), respectively. The efficient estimator $\widehat{\mathbf{B}}_2$ does not use any posited model for η. Instead, we estimate $E(\mathbf{x}|\mathbf{B}^{\mathrm{T}}\mathbf{x})$, η and $\boldsymbol{\eta}'$ through nonparametric regression, i.e. we followed the algorithm described in Section 2.2. The efficient estimator $\widehat{\mathbf{B}}_2$ is more computationally involved since it solves estimating equations to obtain the nonparametric components η and $\boldsymbol{\eta}'$ at n locations inside the search for $\mathbf{B}$ which does not have a closed form. To alleviate the computational burden, we performed the nonparametric estimation on a set of grid points and then performed a linear interpolation for $d = 1$ and a bilinear interpolation for $d = 2$ to obtain the values at each $\widehat{\mathbf{B}}^{(k)\mathrm{T}}\mathbf{x}_i$, where $\widehat{\mathbf{B}}^{(k)}$ represents the kth iteration of the estimated $\widehat{\mathbf{B}}$ during solving the estimating equation in Step 4 of the algorithm in Section 2.2.

We repeat each experiment 1000 times with sample size $n = 500$ and 1000, re-

Table 2.1: The median and the sample standard errors (std) for various estimates, and the inference results, respectively, the average of the estimated standard deviation ($\widehat{\text{std}}$) and the coverage of the estimated 95% confidence interval (CI) of the oracle estimator and the efficient estimator of $\mathbf{B}$ in simulated example 1.

		$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_4$	$\mathbf{B}_5$	$\mathbf{B}_6$
	True	-1.2	0.8	-1.7	-1.5	0.5
Oracle	median	-1.2000	0.7760	-1.6932	-1.5000	0.4964
n=500	$\widehat{\text{std}}$	0.3044	0.3885	0.2019	0.3854	0.3117
	std	0.3406	0.4116	0.2300	0.3800	0.3670
	CI	0.9320	0.9230	0.9220	0.9610	0.9630
Local	median	-1.0224	0.6503	-1.7137	-1.4016	0.4694
n=500	$\widehat{\text{std}}$	0.2897	0.3431	0.2411	0.5357	0.3581
	std	0.3726	0.4450	0.3736	0.5194	0.4415
	CI	0.8680	0.8830	0.8660	0.9440	0.9300
Efficient	median	-1.2155	0.8105	-1.6986	-1.5037	0.5070
n=500	$\widehat{\text{std}}$	0.5674	0.7080	0.3036	0.5353	0.4337
	std	0.4735	0.4813	0.4129	0.5211	0.5074
	CI	0.9750	0.9860	0.8850	0.9540	0.9440
Oracle	median	-1.1879	0.8133	-1.6843	-1.5061	0.5000
n=1000	$\widehat{\text{std}}$	0.2106	0.2444	0.1435	0.2684	0.2097
	std	0.2405	0.2906	0.1506	0.2924	0.2234
	CI	0.9400	0.9310	0.9400	0.9510	0.9640
Local	median	-1.1802	0.7920	-1.6926	-1.3853	0.4710
n=1000	$\widehat{\text{std}}$	0.2369	0.2463	0.1419	0.2748	0.2196
	std	0.2720	0.2755	0.1874	0.2931	0.2698
	CI	0.9240	0.9430	0.9430	0.9210	0.9450
Efficient	median	-1.1936	0.8030	-1.6999	-1.4953	0.4966
n=1000	$\widehat{\text{std}}$	0.3963	0.3656	0.1712	0.3716	0.2364
	std	0.2561	0.2337	0.1724	0.3168	0.2165
	CI	0.9590	0.9720	0.9400	0.9320	0.9520

Table 2.2: The median and the sample standard errors (std) for various estimates, and the inference results, respectively, the median of the estimated standard deviation ($\widehat{\text{std}}$) and the coverage of the estimated 95% confidence interval (CI) of the oracle estimator and the efficient estimator, of $\mathbf{B}$ in simulated example 2.

		$\mathbf{B}_{13}$	$\mathbf{B}_{14}$	$\mathbf{B}_{15}$	$\mathbf{B}_{16}$	$\mathbf{B}_{23}$	$\mathbf{B}_{24}$	$\mathbf{B}_{25}$	$\mathbf{B}_{26}$
	True	1.2	0.8	-1.2	0.8	1.3	0.7	1.1	-0.7
Oracle	median	1.1874	0.8112	-1.1817	0.8318	1.3251	0.7152	1.0779	-0.7113
n=500	$\widehat{\text{std}}$	0.2085	0.2057	0.3807	0.3622	0.2215	0.2251	0.3949	0.3704
	std	0.2703	0.2861	0.4262	0.4070	0.2873	0.2871	0.4411	0.4085
	CI	0.9380	0.9260	0.9570	0.9610	0.9290	0.9230	0.9690	0.9580
Local	median	1.1939	0.7960	-1.1061	0.8663	1.3070	0.6214	1.2427	-0.7372
n=500	$\widehat{\text{std}}$	0.3259	0.3271	0.5747	0.5526	0.4377	0.4376	0.8213	0.7297
	std	0.3440	0.3748	0.5479	0.5553	0.5138	0.4981	0.7111	0.6871
	CI	0.9270	0.9210	0.9610	0.9700	0.9110	0.9670	0.9490	0.9390
Efficient	median	1.2292	0.8759	-1.2214	0.8315	1.3566	0.7002	1.0998	-0.7723
n=500	$\widehat{\text{std}}$	0.7113	0.6808	0.6997	0.7027	0.6757	0.5555	0.6938	0.6622
	std	0.6104	0.4356	0.4836	0.4129	0.5529	0.5078	0.5195	0.4764
	CI	0.9200	0.9700	0.9770	0.9930	0.9540	0.9260	0.9630	0.9690
Oracle	median	1.1928	0.8154	-1.2070	0.8194	1.3053	0.7098	1.0877	-0.6931
n=1000	$\widehat{\text{std}}$	0.1460	0.1423	0.2620	0.2458	0.1437	0.1447	0.2629	0.2435
	std	0.1742	0.1710	0.2852	0.2700	0.1690	0.1647	0.2778	0.2591
	CI	0.9610	0.9480	0.9560	0.9610	0.9420	0.9540	0.9540	0.9650
Local	median	1.2028	0.7792	-1.0987	0.8109	1.3363	0.6012	1.3007	-0.7161
n=1000	$\widehat{\text{std}}$	0.2224	0.1970	0.3610	0.3381	0.2816	0.2865	0.6493	0.5385
	std	0.2551	0.2402	0.4031	0.3784	0.2226	0.3547	0.5111	0.4654
	CI	0.9450	0.9440	0.9470	0.9550	0.9490	0.9670	0.9620	0.9550
Efficient	median	1.2208	0.8606	-1.2053	0.8055	1.3637	0.7079	1.0827	-0.7109
n=1000	$\widehat{\text{std}}$	0.4734	0.4541	0.3217	0.3032	0.4532	0.3743	0.4572	0.3026
	std	0.4529	0.3142	0.3351	0.2927	0.4261	0.2521	0.3789	0.2716
	CI	0.9230	0.9730	0.9380	0.9450	0.9310	0.9780	0.9570	0.9660

spectively. The results are summarized in Table 2.1 for setting (1) and Table 2.2 for setting (2). From Table 2.1 we observe that the performance of both $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{B}}_2$ is close to that of the oracle estimator. All estimators have very small bias for both sample sizes. The results in the table also provide the median of the estimated standard errors using the results in Lemma 1 and Lemma 2 and the empirical coverage probability of the 95% confidence intervals. These results indicate that the asymptotic normal approximation is accurate for the sample sizes. We observe similar performance in Table 2.2. The standard errors of the $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{B}}_2$ become smaller as the sample size grows and the confidence interval coverage probabilities become closer to the nominal level.

A simulation study on estimating the average treatment effect

We generate the potential outcomes as follows: $Y^*(1) = |X_1 X_5| + \sin\left(\sum_{j=1}^{6} X_j\right)$ and $Y^*(0) = 0$. We generate the treatment indicator following (3.1) while considering the two different functional forms for $\eta(\cdot)$ as specified in Section 4.1 .

We compare estimating the average treatment effect using formula (2.9) but different methods to estimate the propensity function: "True" (use the true form of the propensity score function with known parameters), "Oracle" (use the true form of the propensity score function but estimate the unknown parameters), "Logistic" (use a logistic regression model to estimate the propensity score function), $\hat{\tau}_1$ (proposed semiparametric estimator with $\mathbf{B}$ being estimated by $\widehat{\mathbf{B}}_1$), and $\hat{\tau}_2$ (proposed semiparametric estimator with $\mathbf{B}$ being estimated by $\widehat{\mathbf{B}}_2$). The boxplots of the estimated average treatment effect (based on 1000 simulation runs) using these five methods are displayed in Figure 2.1. We observe that the "Logistic" method based on a misspecified propensity score function has serious bias; while the performance of the proposed semiparametric estimators are close to that of "True" and "Oracle". Note that while the "Oracle" estimator is the gold standard in terms of $\mathcal{B}$ and propensity score estimation in Section 2.4, it is unclear that it should yield the best treatment effect estimation here. Hence we included it as a "standard result" for completeness.

Additional Simulations

We further compare our semiparametric approach with Tan's improved methods (Tan, 2006, 2010), targeted maximum likelihood estimation (TMLE) (van der Laan Mark and Daniel, 2006) and the biased reduced double robust (BRdr) estimator proposed by Vermeulen and Vansteelandt (2015). Because Tan's method requires implementing a regression model on treatment outcome $Y^*(1), Y^*(0)$ separately, we

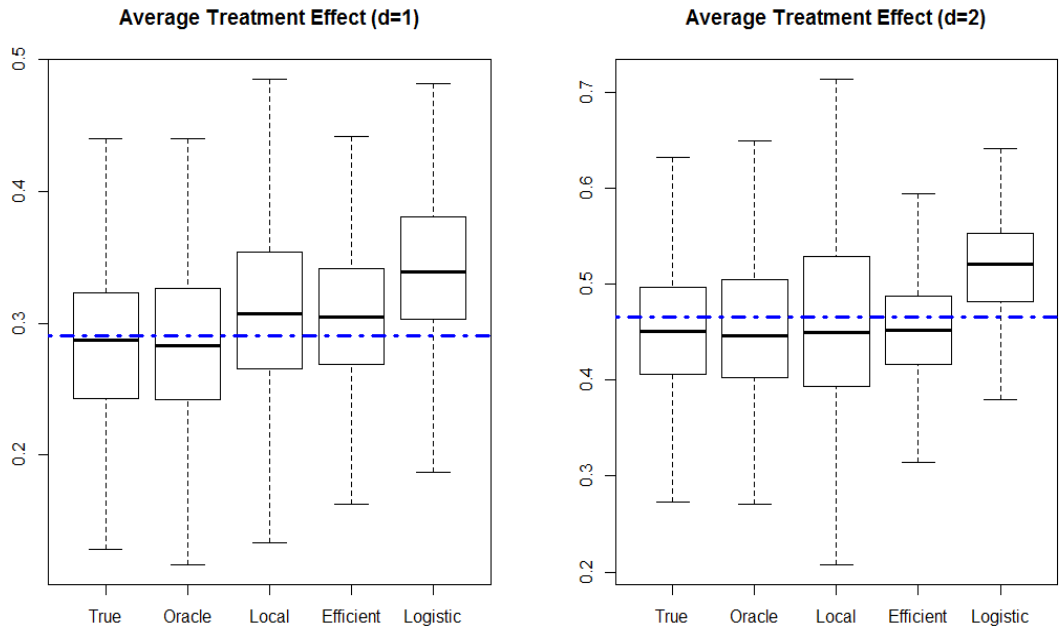

Figure 2.1: Average treatment effect in example 1 (left) and example 2 (right). The blue dash line is the true average treatment effect.

slightly modified $Y^*(0)$ to follow $N(0,1)$ in order to implement the method. We summarize the average treatment effect in Figure (2.2) and Figure (2.3).

We then consider the case where the true outcome model is indeed a linear model. Specifically, we set $Y = X_1 + X_2 + TX_3 + X_4 + 13.5X_5 + X_6 + \epsilon, \epsilon \sim Normal(0, 5^2)$ when $d = 1$ and let $Y = X_1 + X2 + X_3 + X_4 + 3.5X_5 + TX_6 + \epsilon, \epsilon \sim Normal(0, 5^2)$ when $d = 2$. We compare the average treatment effect estimates in Figure (2.4) and Figure (2.5), respectively

Finally, we examine the efficient and locally efficient estimator on the data generated following Kang and Schafer (2007). Specifically, we generated $(Z_1, Z_2, Z_3, Z_4)^{\mathrm{T}}$ from Normal$(\mathbf{0}, \mathbf{I}_4)$ and then form $x_1 = \exp(z_1/2), x_2 = z_2/\{1 + \exp(z_1)\}, x_3 = (z_1 z_3/25 + 0.6)^3, x_4 = (z_2 + z_4 + 20)^2/400$. The outcome model is generated as $y = 210 + 27.4z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + \epsilon$, where $\epsilon \sim N(0,1)$ and the true propensity function is $\pi = \text{expit}(-z_1 + 0.5z_2 - 0.25z_3 - 0.1z_4)$. We use the ob-

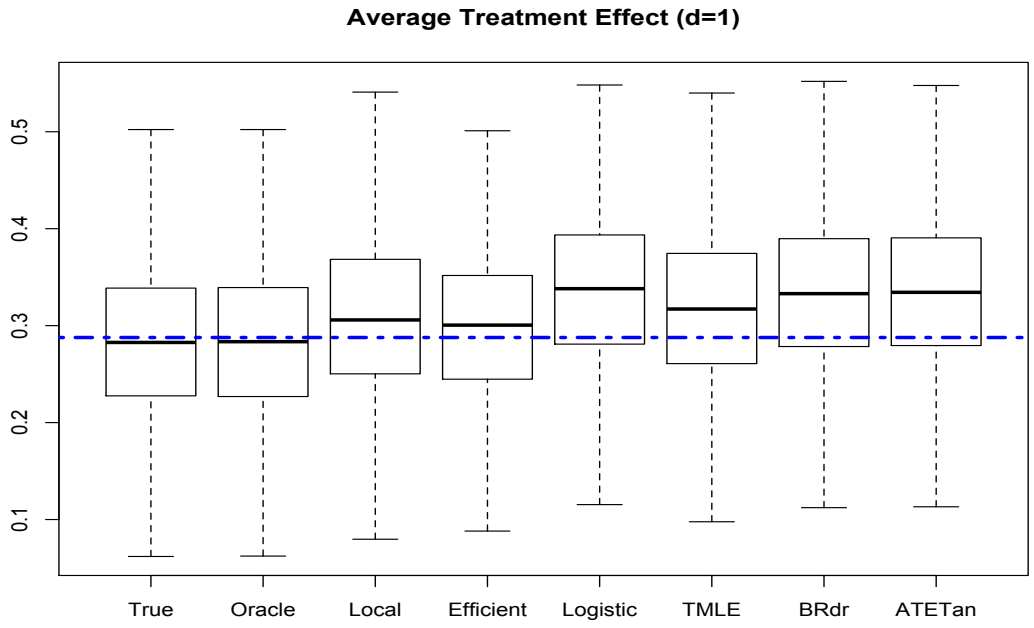

Figure 2.2: Average treatment effect in example 1. The blue dashed line is the true average treatment effect.

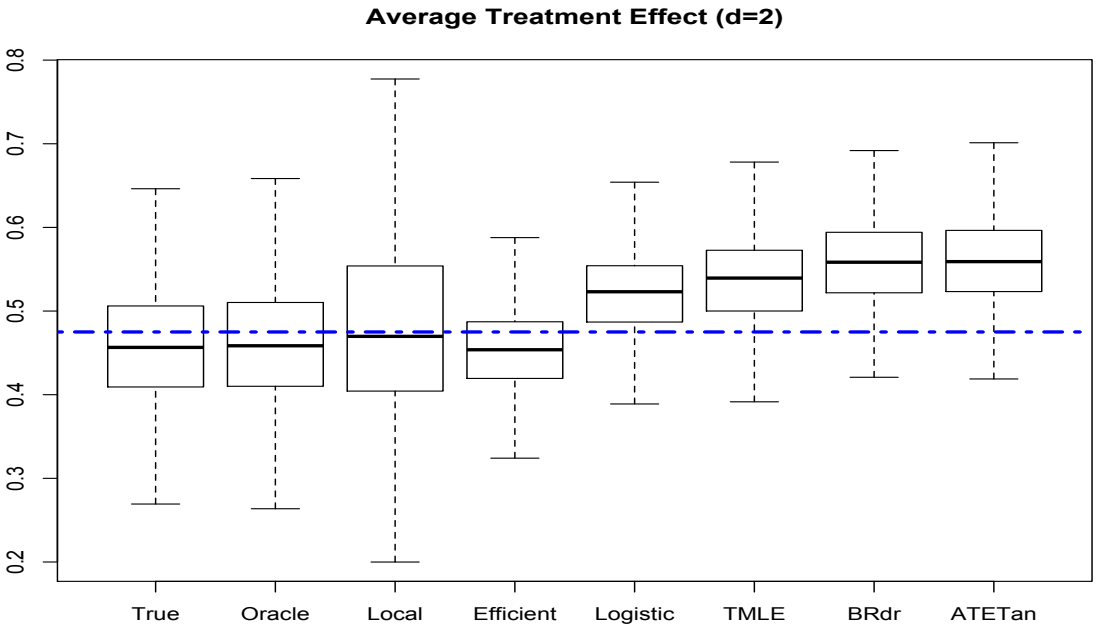

Figure 2.3: Average treatment effect in example 2. The blue dash line is the true average treatment effect.

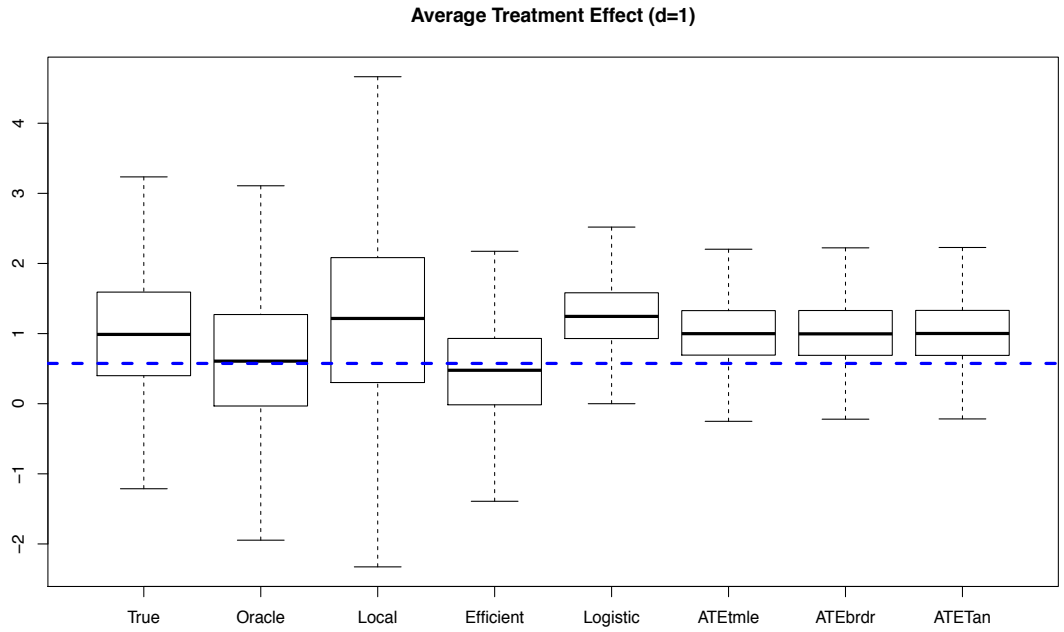

Figure 2.4: Average treatment effect in example 1, where the outcome is $Y = X_1 + X_2 + TX_3 + X_4 + 13.5X_5 + X_6 + \epsilon, \epsilon \sim Normal(0, 5^2)$. The blue dash line is the true average treatment effect.

servable data (Y_i, T_i, X_i) for $i = 1, 2, \cdots, n$ to estimate the propensity score $\widehat{\pi}_i$ for $i = 1, 2, \cdots, n$, then calculate the average treatment effect $\widehat{\tau}$. The performance of the average treatment effect can be found in Figure (2.6), where "True" refers to the average treatment effect calculated from an inverse probability weighted method where the true weight is used. Both the locally efficient and efficient estimators yield reasonable results in comparison with other methods, regardless of whether $d = 1$ or $d = 2$.

2.5 A real data example

We next apply the proposed semiparametric methods to analyze the average effect of maternal smoking on babies' birth weight. The Low Birth Weight data constitute observations from mothers in Pennsylvania, USA and contain birth information on

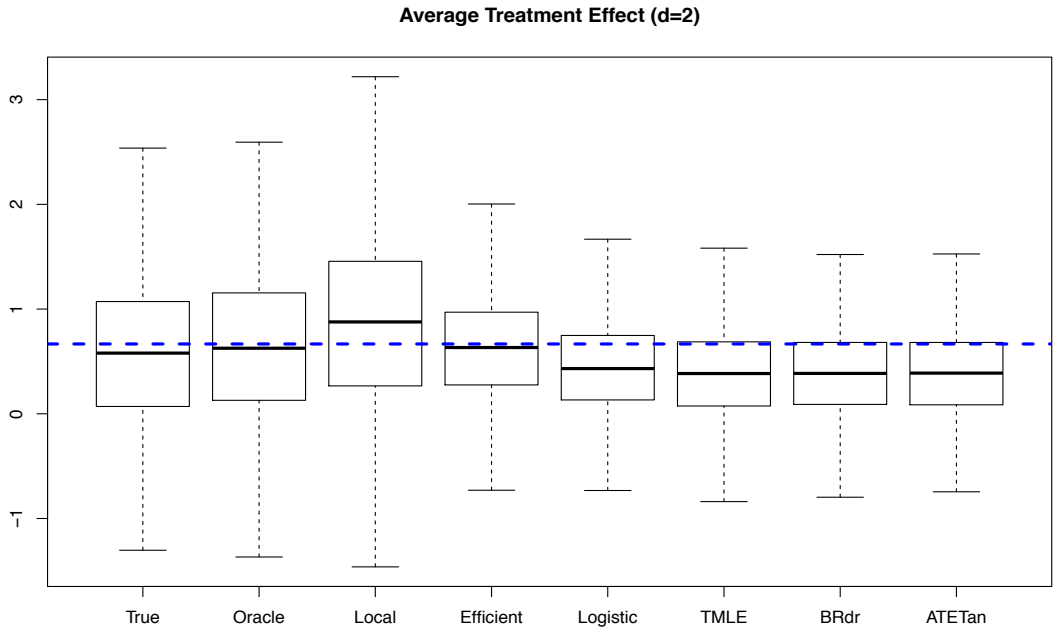

Figure 2.5: Average treatment effect in example 2, where the outcome is $Y = X_1 + X2 + X_3 + X_4 + 3.5X_5 + TX_6 + \epsilon, \epsilon \sim \text{N}ormal(0, 5^2)$. The blue dash line is the true average treatment effect.

4642 infants (Cattaneo, 2010). This dataset was originally used by Almond et al. (2005) and is now available at (http://www.stata-press.com/data/r13/cattaneo2.dta). The outcome of interest Y is infant birth weight measured in grams. The binary variable T denotes the mother's smoking status (1 = smoking, 0 = nonsmoking). The covariates include mother's age, mother's marital status, an indicator variable for alcohol consumption during pregnancy, an indicator for whether there was a previous birth where the newborn died, mother's education, father's education, number of prenatal care visits, mother's race, indicator of first born baby and months since last birth (monthslb).

Based on data from the 4642 infants, the naive average weight difference of the two groups of babies belonging to smoking and non-smoking mothers yields -275.25 grams. Considering that this naive result is not necessarily a valid estimator of the

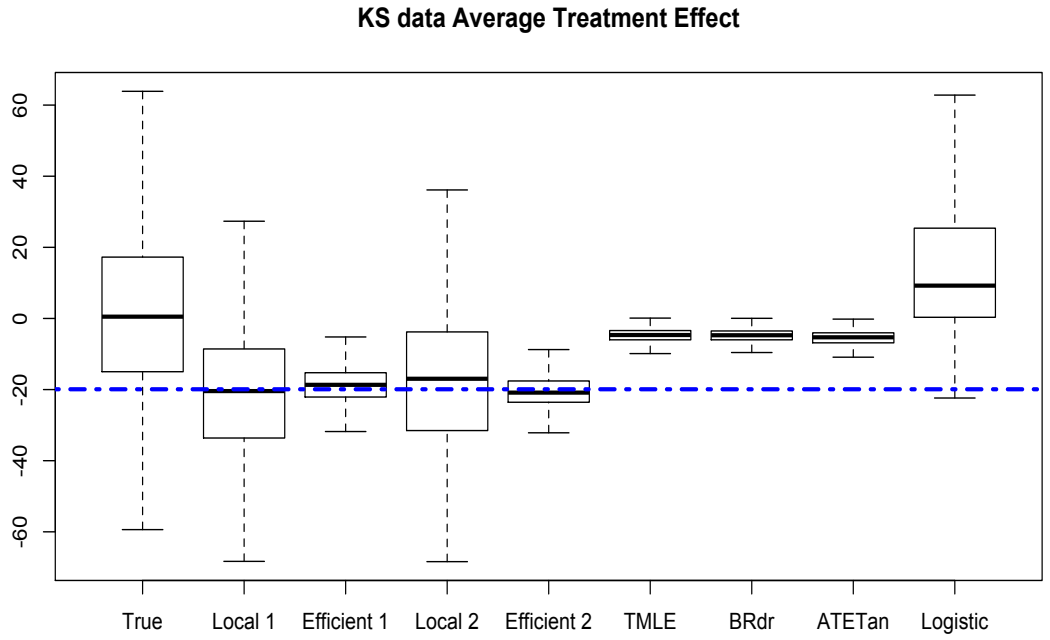

Figure 2.6: Average treatment effect on Kang and Schafer data. Local 1 and Efficient 1 are for $d = 1$. Local 2 and Efficient 2 are for $d = 2$. The blue dash line is the true average treatment effect.

causal result of smoking on birth weight, we next studied the proposed estimators. Specifically, we compare three estimators of average treatment effect discussed in the last section: "Logistic", $\hat{\tau}_1$ and $\hat{\tau}_2$. The estimated propensity score functions are summarized in Table 2.3. The estimated average treatment difference corresponding to "Logistic" , $\hat{\tau}_1$ and $\hat{\tau}_2$ are -352.08, -295.77 and -306.32 grams, respectively. In addition, we compare the average causal effect with Tan's improved method, TMLE and BRdr. The results indicate that maternal smoking has a negative impact on babies' birth weight. The estimate average treatment differences are summarized in Table 2.4 along with the mean and standard deviation from 1000 bootstrap samples for each method. The bootstrap average treatment effect from the seven approaches can be found in Figure (2.7). Note that the estimator using propensity score estimated by logistic regression is substantially different from $\hat{\tau}_1$ and $\hat{\tau}_2$. This suggests logistic

Table 2.3: Low Birth Weight data Example.

	Naive			Local			Efficient		
	Est	Std	P-value	Est	Std	P-value	Est	Std	P-value
(Intercept)	0.9848	0.2631	0.0002						
age	1.0021	0.3607	0.0055						
mmarried	-0.9480	0.1030	0.0000	-0.8020	0.2187	0.0002	-1.6922	0.3537	0.0000
alcohol	1.5886	0.1844	0.0000	1.5021	0.4156	0.0003	2.7014	0.4683	0.0000
deadkids	0.3893	0.0909	0.0000	0.4070	0.1232	0.0010	0.4980	0.1554	0.0014
medu	-0.0964	0.0190	0.0000	-0.0675	0.0281	0.0164	-0.2066	0.0571	0.0003
fedu	-0.0426	0.0118	0.0003	-0.0499	0.0182	0.0061	-0.1067	0.0369	0.0038
nprenatal	-0.0299	0.0111	0.0071	-0.0346	0.0141	0.0143	-0.0513	0.0221	0.0204
monthslb	0.0062	0.0015	0.0000	0.0062	0.0019	0.0012	0.0097	0.0028	0.0007
mrace	0.6888	0.1184	0.0000	0.7607	0.2093	0.0003	1.1446	0.2421	0.0000
fbaby	-0.2574	0.1059	0.0150	-0.2728	0.1181	0.0209	-0.3799	0.1881	0.0435

Table 2.4: Average treatment difference in the Low Birth Weight data. Bootstrap mean (BS mean) and Bootstrap std (BS std). Bootstrap sample $B = 1000$.

	Naive	Efficient	Local	Logistic	TMLE	BRdr	Tan
Estimate	-275.25	-295.77	-306.32	-352.08	-219.96	-228.89	-230.57
BS mean	-275.10	-292.85	-304.69	-352.11	-219.69	-229.33	-231.34
BS std	21.36	38.62	54.50	46.78	29.50	29.34	27.66

regression may not provide an adequate model for the propensity score function.

2.6 Conclusion and discussions

In this chapter, we propose a semiparametric approach to estimating the average treatment effect. The approach is less prone to propensity score model misspecification compared to the logistic regression based inverse probability weighted estimators, which have dominant roles in causal inference. A parametric propensity score model (e.g., logistic regression model) is certainly a lot more informative than a semiparametric model such as the dimension reduction model we propose, but it also bears a greater risk of being misspecified. If the parametric propensity score model is misspecified, then the resulting estimation of the average treatment effect is inconsistent. Furthermore, the semiparametric estimator does not rely on specification of the outcome regression model, and hence is attractive when a reliable outcome regression model is hard to obtain and/or compute, such as when studying treatment effects on complex diseases. We note that if suitable outcome models are obtainable, then

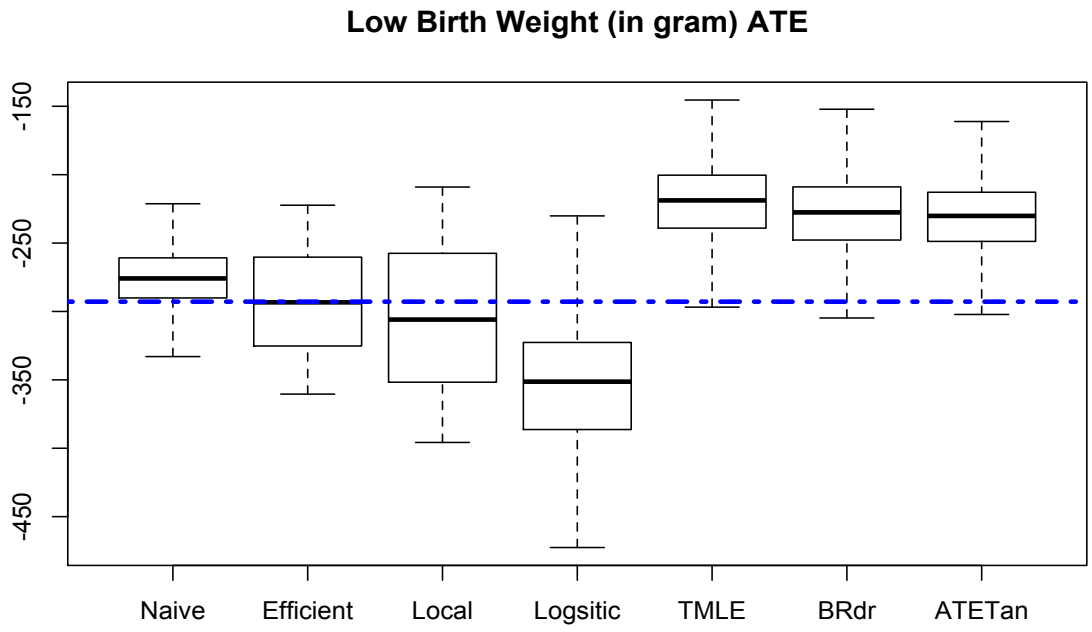

Figure 2.7: Bootstrap Average Treatment Effect. The blue dash line is the mean of the average treatment effect calculated from the efficient estimation procedure.

further extending our method to a doubly robust estimator could bring an additional efficiency gain.

It is of interest to investigate whether a dimension reduction propensity score model will always lead to more efficient treatment effect estimation than a parametric one, in the case that both models are correct. However, we find that is not true in general. The relation can go either way, and it depends on the specific models. We summarize the results in Lemma 3 in the supplementary materials.

Not able to find any definitive relation between the dimension reduction model and a general parametric model, we further investigate the situation of nested models. For the sake of comparing two models that are both correct, this certainly makes much sense. To this end, the model will be the same as in (3.1), except that now η is a known function. Unfortunately, even for this case, as shown in the Appendix, there

is no definitive relation we can claim. Thus, even when the parametric model is a submodel of the dimension reduction model, there is no definitive relation between the two estimators of the average treatment effect based on the two models. Our intuition is that not only the model makes a difference, but also the specific estimator used in the propensity score model has a role to play. The overall picture is unclear and is potentially very complex; much work is needed to fully understand these relations and can lead to interesting research results.

Finally, even though our initial intention is to overcome the potential issue of misspecification of both the propensity score model and the outcome regression model through employing a more relaxed modeling strategy of the former and giving up modeling of the latter, and subsequently proposing inverse property weighting, double robust estimator can be used in combination with our method to further gain efficiency. As it is well known in the original form of the double robust estimator, in combination with the semiparametric propensity score model, when the treatment response is modeled correctly, the method will be more efficient than our method. If the treatment response is modeled incorrectly, depending on how "wrong" the model is, the method could be less efficient than our method. However, if the method of Tan (2010) is adopted, in combination with the semiparametric propensity score model, one can always obtain a more efficient estimator than our method, regardless whether the treatment response is modeled correctly or not. Thus, to achieve improved efficiency, one can strive to propose a "good" model for the treatment response, and further perform additional computation to obtain the correlation adjustment required in Tan (2010).

Chapter 3

Estimation and Inference of Error-Prone Covariate Effect in the Presence of Confounding Variables

3.1 Introduction

Estimating and testing the effect of a covariate of interest while accommodating many other covariates is an important problem in statistical practice. The t-test and the analysis of variance are widely used to evaluate the covariate effect when the covariate of interest is binary or categorical and no confounders are present. When the covariate of interest is not necessarily binary or categorical, evaluating the covariate effect has been studied extensively in the context of linear model, partially linear model (Heckman (1986), Hardle et al. (2000), Ma et al. (2006)) and partially linear single-index model (Carroll et al. (1997), Yu and Ruppert (2002), Li et al. (2011), Ma and Zhu (2012)), as long as both the covariate of interest and the confounders are measured precisely. In this work, we intend to generalize the partially linear single-index model to a larger class where the link function is not restricted to be linear, and we further consider measurement error issues.

When the covariate of interest is measured with error, to evaluate its effect precisely we must reduce the bias caused by measurement error and adjust for the confounding effects simultaneously. This is an interesting yet very challenging problem. To partially address this problem, Carroll et al. (2006) assumed the confounding

effects are linear, and Liang et al. (1999) and Ma and Carroll (2006) assumed the confounders are in fact univariate. These assumptions restrict the usefulness of their methods. To the best of our knowledge, how to assess the covariate effect subject to measurement error while taking into account possibly nonlinear confounding effects still remains an open and difficult problem in the literature.

Estimating and testing the effect of a covariate of interest in the presence of possibly nonlinear confounding effects has many applications in a variety of scientific fields such as econometrics, biology, policy making, etc. Consider the Framingham Heart Study (http://www.framinghamheartstudy.org/) as a typical example. It is common knowledge that high systolic blood pressure (SBP) is directly linked to the occurrence of coronary heart disease (Y). To quantify the effect is however not necessarily straightforward. One difficulty is that SBP can vary significantly from time to time, hence a clinically meaningful covariate is the long term average of SBP ($\widetilde{X}$), which is unfortunately impossible to measure precisely. A widely used practice is to use the average of several measured SBP values ($\widetilde{W}$) during a reasonably long time course as a substitute. Thus, long term average SBP is a variable measured with error. Another difficulty comes from the presence of possibly nonlinear confounding effects ($\mathbf{Z}$) for heart disease, such as smoking status, family history, ethnicity, BMI, lung capacity, age and other laboratory variables. Because these effects are not of medical interest while their connection to the heart disease occurrence might be complex, a suitable modeling strategy is to use an unspecified function to summarize their possibly nonlinear effect. Difficulty with such modeling strategy naturally arises when the dimension of $\mathbf{Z}$ is more than one, since it is well known that nonparametrically estimating a function of multivariate confounding variables suffers from the curse of dimensionality. To tackle this issue, we follow the single index modeling strategy and assume that the combined effect of the covariates in $\mathbf{Z}$ is manifested through a linear combination $\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}$, where $\widetilde{\boldsymbol{\gamma}}$ is a length p vector. For identifiability, we assume that $\mathbf{Z}$

contains at least one continuous variable, the first component of $\widetilde{\boldsymbol{\gamma}}$ is one, and we use $\boldsymbol{\gamma}$ to denote the vector of the last $p-1$ components. Let H be the logistic distribution function. In this Framingham data example, we assume that, given $\widetilde{X}$ and $\mathbf{Z}$, the probability of the occurrence of the coronary heart disease (Y) admits a model of the form

$$\begin{aligned} \mathrm{pr}(Y = 1 \mid \widetilde{X}, \mathbf{Z}) &= H\{\widetilde{X}\beta + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\}, \\ \log(\widetilde{W} - 50) &= \log(\widetilde{X} - 50) + U. \end{aligned}$$

Here we adopt the general assumption that after the transformation from the raw systolic blood pressure, the relation between $W \equiv \log(\widetilde{W} - 50)$ and $X \equiv \log(\widetilde{X} - 50)$ is additive with a normal measurement error, i.e. $U \sim N(0, \sigma_u^2)$, and we assume the error is nondifferential. This relation is verified by Carroll et al. (2006, chapter 6).

The above model can be viewed as a special case of the following general semiparametric measurement error model. To be specific, we write the general probability density/mass function of the response variable Y, for example disease status, conditional on the covariate set $(X, \mathbf{S}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ as

$$g\{y, x, \mathbf{s}, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}), \boldsymbol{\beta}\}, \tag{3.1}$$

where X is an error-prone covariate whose effect on Y is of central research interest, $\mathbf{Z}$, $\mathbf{S}$ contain additional covariates that may be related to Y and may be confounded with X. We model part of these confounders ($\mathbf{S}$) parametrically, such as the categorical variables, and part of these confounders ($\mathbf{Z}$) nonparametrically through an unspecified smooth function θ. Both $\mathbf{S}$ and $\mathbf{Z}$ are measured precisely. In model (3.1), g is a known conditional probability density/mass function, θ is an unspecified smooth function, $\widetilde{\boldsymbol{\gamma}} = (1, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\gamma}$ is an unknown length $p-1$ vector, and $\boldsymbol{\beta}$ is an unknown parameter. In this notation, the example above can be written as $g\{y, x, \mathbf{s}, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}), \boldsymbol{\beta}\} = \exp[y\{\widetilde{x}\beta + \theta(\widetilde{\boldsymbol{\gamma}}\mathbf{z})\}]/[1 + \exp\{\widetilde{x}\beta + \theta(\widetilde{\boldsymbol{\gamma}}\mathbf{z})\}]$. In our context, we assume the covariate X is of our primary interest but is unobservable. Instead, we

observe its erroneous version W, where the relation between W and X is specified, i.e. $f_{W|X}(w \mid x)$ is a known model. In practice, the specification of $f_{W|X}(w \mid x)$ is usually obtained through validation data, instruments or repeated measurements. We treat $\theta(\cdot)$ as an infinite dimensional nuisance parameter. We further make the surrogacy assumption that W and Y are independent given $X, \mathbf{S}, \mathbf{Z}$. The primary interest is in $\boldsymbol{\beta}$, which describes the effect of X on Y. In many applications, $\boldsymbol{\beta}$ enters the model as multiplication coefficient of a linear function of the covariates, such as through $\beta_1 X + \beta_2 S$.

Model (3.1) is an extension of the generalized single index model proposed by Cui et al. (2011) in which neither X nor $\mathbf{S}$ is present. In addition, Tsiatis and Ma (2004) studied a simpler version of model (3.1) where $\mathbf{Z}$ does not appear, and Ma and Carroll (2006) considered a simpler version of model (3.1) where Z is univariate. The generalization to multivariate $\mathbf{Z}$ in model (3.1) is important in practice since it accomodates more realistic applications; see, for example, the Framingham Heart Study in Section 3.5. In particular, model (3.1) allows us to handle the possible nonlinearity of the confounding variables through the unspecified function θ, while the single index structure $\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}$ facilitates nonparametric modeling. Nevertheless, the extension also poses several challenging technical and computational problems. Indeed, when the index vector appears inside an unknown function, its estimation is more complex and interaction between the estimation of the indices and the function has to be taken into account. The variability in estimating these quantities further affects the estimation quality of the parameter of interest. Overall, the three sets of parameters, namely the parameter of interest, the index vector and the unknown smooth function link together intrinsically, which complicates the estimation procedure, the computational treatment and the theoretical development. Compared with the case when the index vector does not appear, such additional complexity can be viewed as a price paid to overcome the curse of dimensionality.

We design a general methodology for the semiparametric measurement error model (3.1), and introduce a bias-correction approach to construct a class of locally efficient estimators. This bias-correction approach is motivated by the projected score idea in semiparametrics (Tsiatis and Ma (2004)) and does not have to resort to a deconvolution method or to correctly specify a distributional model for the error-prone covariate of interest. We further generalize the bias-correction approach to estimating $\widetilde{\boldsymbol{\gamma}}$ in model (3.1), which is a component that does not appear in the models considered in Tsiatis and Ma (2004) or Ma and Carroll (2006). In their studies, $\mathbf{Z}$ is either absent or univariate, hence the issue of estimating $\widetilde{\boldsymbol{\gamma}}$ does not occur. In the presence of multivariate $\mathbf{Z}$, the conditional density of X given $\mathbf{S}$ and $\mathbf{Z}$, denoted $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, is required in implementing the bias-correction approach. However, with a multivariate $\mathbf{Z}$, regardless whether $\mathbf{S}$ is discrete or continuous, estimating $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$ is a thorny issue even if X were observed due to the curse of dimensionality. To alleviate the difficulty in estimating $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, a working model is adopted. If this working model happens to be the underlying true one, the resultant estimator is semiparametrically efficient, whereas if this working model is unfortunately misspecified, then the resultant estimator is still root-n consistent and asymptotically normal. In other words, the resultant estimator is locally efficient. To put the bias-correction approach into practice, we suggest a profiling algorithm for estimating $\boldsymbol{\beta}$.

The article is organized as the following. In Section 3.2 we introduce the bias-correction approach for estimating $\boldsymbol{\beta}$ in the semiparametric measurement error model (3.1). The asymptotic properties of the resultant estimators are given in Section 3.3. We report several simulation studies in Section 3.4 and revisit the Framingham data in Section 3.5. This chapter is concluded with a brief discussion in Section 3.6. All technical details are given in an Appendix.

3.2 Estimation

In this section we discuss estimation of the covariate effect at the sample level. Write the observation as $(y_i, w_i, \mathbf{s}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$. We propose to estimate the effect of the covariate of interest as well as other nuisance parameters through solving the estimating equations derived from the semiparametric log-likelihood.

The surrogacy assumption and the model specification in Section 3.1 directly lead to the semiparametric log-likelihood, subject to an additive term that does not involve the parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta$,

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f_{X|\mathbf{S},\mathbf{Z}}) = \sum_{i=1}^{n} \log \int g\{y_i, x, \mathbf{s}_i, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i), \boldsymbol{\beta}\} f_{W|X}(w_i \mid x) f_{X|\mathbf{S},\mathbf{Z}}(x \mid \mathbf{s}_i, \mathbf{z}_i) dx.$$

Recall that $\boldsymbol{\gamma}$ is defined in Section 1 as a vector of the free parameters in $\widetilde{\boldsymbol{\gamma}}$. Here $f_{X|\mathbf{S},\mathbf{Z}}$ and $f_{W|X}$ represent the probability density function of X conditional on $(\mathbf{S}, \mathbf{Z})$ and the probability density function of W conditional on X respectively. If both θ and $f_{X|\mathbf{S},\mathbf{Z}}$ had been known, the simple maximum likelihood estimator (MLE) would have provided a most natural estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Let

$$\begin{aligned} & \mathbf{S}_{\beta}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f_{X|\mathbf{S},\mathbf{Z}}) \\ = & \frac{\partial \log \int g\{y, x, \mathbf{s}, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}), \boldsymbol{\beta}\} f_{W|X}(w \mid x) f_{X|\mathbf{S},\mathbf{Z}}(x \mid \mathbf{s}, \mathbf{z}) dx}{\partial \boldsymbol{\beta}}, \\ & \mathbf{S}_{\gamma}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f_{X|\mathbf{S},\mathbf{Z}}) \\ = & \frac{\partial \log \int g\{y, x, \mathbf{s}, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}), \boldsymbol{\beta}\} f_{W|X}(w \mid x) f_{X|\mathbf{S},\mathbf{Z}}(x \mid \mathbf{s}, \mathbf{z}) dx}{\partial \boldsymbol{\gamma}} \end{aligned}$$

be the score functions with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, then we could modify the MLE through localization to handle the issue caused by the unknown functional form of θ. Specifically, let us adopt a local parametric model $\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) = \nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}; \boldsymbol{\alpha})$. For example, the most widely used local polynomial model in Fan and Gijbels (1996) can be used as $\nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}; \boldsymbol{\alpha})$. Here $\boldsymbol{\alpha}$ depends on $\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}$, but we suppress the dependence of $\boldsymbol{\alpha}$ on $\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}$ for notational clarity. Then we could estimate θ together with $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, through

iteratively solving

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}_{\beta}\{w_i, \mathbf{s}_i, \mathbf{z}_i, y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i), f_{X|\mathbf{S},\mathbf{Z}}(x_i \mid \mathbf{s}_i, \mathbf{z}_i)\} \\ \mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}_{\gamma}\{w_i, \mathbf{s}_i, \mathbf{z}_i, y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i), f_{X|\mathbf{S},\mathbf{Z}}(x_i \mid \mathbf{s}_i, \mathbf{z}_i)\}\end{aligned}$$

to obtain $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$, and

$$\mathbf{0} = \sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{S}_{\alpha}(w_i, \mathbf{s}_i, \mathbf{z}_i, y_i; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{\alpha}, f_{X|\mathbf{S},\mathbf{Z}})$$

at $\mathbf{z}_0$ to obtain $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = \nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0; \widehat{\boldsymbol{\alpha}})$ for $\mathbf{z}_0 = \mathbf{z}_1, \ldots, \mathbf{z}_n$. Here, $K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z} - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = h^{-1}K\{(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma} - \mathbf{z}_0^{\mathrm{T}}\boldsymbol{\gamma})/h\}$, K is a kernel function and h is a bandwidth. In the above display, $\mathbf{S}_{\alpha}$ is defined analogously as $\mathbf{S}_{\beta}$ except that $\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ is replaced by $\nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}; \boldsymbol{\alpha})$ and the derivative is with respect to $\boldsymbol{\alpha}$, i.e.

$$\begin{aligned}&\mathbf{S}_{\alpha}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, f_{X|\mathbf{S},\mathbf{Z}}) \\ =\ &\frac{\partial \log \int g\{y, x, \mathbf{s}, \nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}; \boldsymbol{\alpha}), \boldsymbol{\beta}\} f_{W|X}(w \mid x) f_{X|\mathbf{S},\mathbf{Z}}(x \mid \mathbf{s}, \mathbf{z}) dx}{\partial \boldsymbol{\alpha}}.\end{aligned}$$

The above idea would have worked if we knew how to actually calculate the score functions. However, without an explicit form of $f_{X|\mathbf{S},\mathbf{Z}}$, the calculation of the score vectors is not an easy task. A natural approach is to estimate $f_{X|\mathbf{S},\mathbf{Z}}$ and then use the estimated version to obtain the corresponding estimated score functions. This is not entirely out of the question, especially when $f_{W|X}(w, x)$ happens to describe an additive independent error model, i.e. when $f_{W|\mathbf{X}}(w, x) = f_U(w-x)$. In this case, from the relation $f_{W|\mathbf{S},\mathbf{Z}}(w, \mathbf{s}, \mathbf{z}) = \int f_U(w-x) f_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z}) dx$, we can use the Fourier transform to obtain $\mathcal{F}_w(t, \mathbf{s}, \mathbf{z}) = \mathcal{F}_u(t)\mathcal{F}_x(t, \mathbf{s}, \mathbf{z})$, where $\mathcal{F}_w(t, \mathbf{s}, \mathbf{z}) = \int f_{W|\mathbf{S},\mathbf{Z}}(w, \mathbf{s}, \mathbf{z}) e^{-2\pi i t w} dw$, $\mathcal{F}_u(t) = \int f_U(u) e^{-2\pi i t u} du$ and $\mathcal{F}_x(t, \mathbf{s}, \mathbf{z}) = \int f_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z}) e^{-2\pi i t x} dx$. Thus, if we can estimate $f_{W|\mathbf{S},\mathbf{Z}}(w, \mathbf{s}, \mathbf{z})$ nonparametrically, then we can obtain an estimated version of $\mathcal{F}_w(t, \mathbf{s}, z)$ and an estimated version of $\mathcal{F}_x(t, \mathbf{s}, \mathbf{z}) = \mathcal{F}_w(t, \mathbf{s}, \mathbf{z})/\mathcal{F}_u(t)$. Performing an inverse Fourier transform on $\mathcal{F}_x(t, \mathbf{s}, \mathbf{z})$ would then yield an estimate of $f_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z})$.

The above analysis reveals some hidden obstacles in estimating $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$. First of all, the deconvolution procedure is only applicable when the measurement error is additive and independent of X. When the measurement error model $f_{W|X}(w,x)$ goes beyond this structure, it is unclear how to recover $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$. Second, the procedure requires estimating $f_{W|\mathbf{S},\mathbf{Z}}(w,\mathbf{s},\mathbf{z})$ nonparametrically. However, when the dimension of $(\mathbf{s},\mathbf{z})$ is moderate or high, in other words, the confounding variables are multivariate, this is again a problem suffering from the curse of dimensionality and is not practically feasible in finite samples. Finally, even when the dimension of $(\mathbf{s},\mathbf{z})$ is sufficiently low and the deconvolution procedure can be carried out in practice, the resulting estimate of $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$ has very slow convergence rate (Carroll and Hall (1998), Fan (1991)), hence using the estimated $\widehat{f}_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$ may yield very different results from using the true $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, which is required in the original score function calculation.

Due to these inherent difficulties involved with estimating $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, we decide not to pursue this route. Instead, we take a somewhat counter-intuitive approach. Instead of striving to obtain an approximation of $f_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, we propose to simply guess a model $f^*_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$, which may or may not reflect the true conditional density function, and calculate the score functions $\mathbf{S}_\beta,\mathbf{S}_\gamma,\mathbf{S}_\alpha$ under this guessed model. Of course, this simple replacement of the true score functions with the guessed version is not guaranteed to yield consistent estimation of $\boldsymbol{\beta},\boldsymbol{\gamma}$ and θ. To correct the possible

bias, we form

$$\begin{aligned}
&\mathbf{L}_\beta(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) \\
&\quad = \mathbf{S}_\beta(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) - E^*\{\mathbf{a}_\beta(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta) \mid w,\mathbf{s},\mathbf{z},y\}, \\
&\mathbf{L}_\gamma(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) \\
&\quad = \mathbf{S}_\gamma(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) - E^*\{\mathbf{a}_\gamma(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta) \mid w,\mathbf{s},\mathbf{z},y\}, \qquad (3.2) \\
&\mathbf{L}_\alpha(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha},f^*_{X|\mathbf{S},\mathbf{Z}}) \\
&\quad = \mathbf{S}_\alpha(w,\mathbf{s},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha},f^*_{X|\mathbf{S},\mathbf{Z}}) - E^*\{\mathbf{a}_\alpha(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}) \mid w,\mathbf{s},\mathbf{z},y\},
\end{aligned}$$

where $\mathbf{a}_\beta, \mathbf{a}_\gamma, \mathbf{a}_\alpha$ are functions of $(X, \mathbf{S}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ that satisfy

$$\begin{aligned}
& E\{\mathbf{S}_\beta(W,\mathbf{s},\mathbf{z},Y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) \mid x,\mathbf{s},\mathbf{z}\} \\
= \;& E[E^*\{\mathbf{a}_\beta(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta)|W,\mathbf{s},\mathbf{z},Y\} \mid x,\mathbf{s},\mathbf{z}], \\
& E\{\mathbf{S}_\gamma(W,\mathbf{s},\mathbf{z},Y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) \mid x,\mathbf{s},\mathbf{z}\} \qquad (3.3) \\
= \;& E[E^*\{\mathbf{a}_\gamma(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta)|W,\mathbf{s},\mathbf{z},Y\} \mid x,\mathbf{s},\mathbf{z}], \\
& E\{\mathbf{S}_\alpha(W,\mathbf{s},\mathbf{z},Y;\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha},f^*_{X|\mathbf{S},\mathbf{Z}}) \mid x,\mathbf{s},\mathbf{z}\} \\
= \;& E[E^*\{\mathbf{a}_\alpha(X,\mathbf{s},\mathbf{z};\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha})|W,\mathbf{s},\mathbf{z},Y\} \mid x,\mathbf{s},\mathbf{z}],
\end{aligned}$$

and E^* represents expectation calculated using $f^*_{X|\mathbf{S},\mathbf{Z}}(x,\mathbf{s},\mathbf{z})$. $E(\mathbf{a}_\beta \mid w,\mathbf{s},\mathbf{z},y)$, $E(\mathbf{a}_\gamma \mid w,\mathbf{s},\mathbf{z},y)$ and $E(\mathbf{a}_\alpha \mid w,\mathbf{s},\mathbf{z},y)$ are respectively the projections of the score vectors $\mathbf{S}_\beta$, $\mathbf{S}_\gamma$ and $\mathbf{S}_\alpha$ onto the tangent space Λ described in Appendix B.1, and has an no explicit form except in some special cases. We give one such special example at the end of this section. It is easy to see that the definition of $\mathbf{a}_\beta, \mathbf{a}_\gamma, \mathbf{a}_\alpha$ in (3.3) guarantees the consistency of $\mathbf{L}_\beta$,$\mathbf{L}_\gamma$, and $\mathbf{L}_\alpha$ automatically, whether or not $f^*_{X|\mathbf{S},\mathbf{Z}}$ reflects the truth. We then use $\mathbf{L}_\beta$, $\mathbf{L}_\gamma$ and $\mathbf{L}_\alpha$ to replace $\mathbf{S}_\beta, \mathbf{S}_\gamma, \mathbf{S}_\alpha$ in the iterative procedure described above to estimate $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and θ. That is, we solve

$$\begin{aligned}
\mathbf{0} &= \sum_{i=1}^n \mathbf{L}_\beta(w_i,\mathbf{s}_i,\mathbf{z}_i,y_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}), \\
\mathbf{0} &= \sum_{i=1}^n \mathbf{L}_\gamma(w_i,\mathbf{s}_i,\mathbf{z}_i,y_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{X|\mathbf{S},\mathbf{Z}}) \qquad (3.4)
\end{aligned}$$

to estimate $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and solve

$$\mathbf{0} \;=\; \sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_\alpha(w_i, \mathbf{s}_i, \mathbf{z}_i, y_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, f^*_{X|\mathbf{S},\mathbf{Z}}) \tag{3.5}$$

at $\mathbf{z}_0 = \mathbf{z}_1, \ldots, \mathbf{z}_n$ to obtain $\widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = \nu(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0; \widehat{\boldsymbol{\alpha}})$. Because different $\mathbf{z}_0$ yields different $\boldsymbol{\alpha}$, hence we could have used a more precise notation $\boldsymbol{\alpha}(\mathbf{z}_0)$ in (3.5). We suppressed the dependence of $\boldsymbol{\alpha}$ on $\mathbf{z}_0$ for notational brevity. The estimation procedure can be either iteratively solving (3.4) and (3.5) (backfitting), or using (3.5) to obtain $\widehat{\theta}$ as a function of $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and then using (3.4) to solve for $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ (profiling). In the following, we carry out all the procedures using the profiling approach.

The bias correction through forming $\mathbf{L}_\beta$ etc. is rooted in the projected score idea in semiparametrics (Bickel et al. (1993), Tsiatis and Ma (2004), Tsiatis (2006)). Given any function, say $\mathbf{S}_\beta$, we can calculate its residual after projecting it onto the nuisance tangent space associated with the model. The projection of $(\mathbf{S}_\beta^{\mathrm{T}}, \mathbf{S}_\gamma^{\mathrm{T}}, \mathbf{S}_\alpha^{\mathrm{T}})^{\mathrm{T}}$ indeed would have been $(\mathbf{L}_\beta^{\mathrm{T}}, \mathbf{L}_\gamma^{\mathrm{T}}, \mathbf{L}_\alpha^{\mathrm{T}})^{\mathrm{T}}$, if we had used $f_{X|\mathbf{S},\mathbf{Z}}$ throughout all the calculations. We defer the detail of this calculation in Appendix B.1. However, due to the lack of knowledge on $f_{X|\mathbf{S},\mathbf{Z}}$, we are forced to perform all the calculations using a proposed $f^*_{X|\mathbf{S},\mathbf{Z}}$. The fortunate fact is that even using the possibly misspecified conditional density, $(\mathbf{L}_\beta^{\mathrm{T}}, \mathbf{L}_\gamma^{\mathrm{T}}, \mathbf{L}_\alpha^{\mathrm{T}})^{\mathrm{T}}$ still has mean zero because this property is enforced by its very construction reflected on the definitions of $\mathbf{a}_\beta, \mathbf{a}_\gamma, \mathbf{a}_\alpha$ in (3.3). It is worth mentioning that if $f_{X|\mathbf{S},\mathbf{Z}}$ happens to be the truth, then $\mathbf{S}_\beta, \mathbf{S}_\gamma, \mathbf{S}_\alpha$ are indeed the score functions. Thus, as the orthogonal projection of the score functions, $\mathbf{L}_\beta$,$\mathbf{L}_\gamma$ and $\mathbf{L}_\alpha$ are the efficient score functions. Hence the resulting estimator is not only consistent, but also efficient.

To further illustrate the estimator, we now investigate the partially linear single index model with normal measurement error. We will show that in this special case, many quantities simplify and a set of explicit estimating equations can be obtained.

Consider an alternative form of Model (3.1) in this case, where $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) + \epsilon$, ϵ follows a normal distribution with mean zero, known constant variance

σ^2 and is independent of $\mathbf{X}$. We adopt an additive normal measurement error $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ follows a normal distribution with mean zero and known constant covariance matrix $\boldsymbol{\Sigma}$ and is independent of $\mathbf{X}$. For estimating $\theta(\cdot)$, we adopt the familiar local linear form $\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) = \alpha_0 + \alpha_1\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}$.

Define $\boldsymbol{\Delta} = \mathbf{W} + Y\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2$. Following Stefanski and Carroll (1987), the forms of $\mathbf{L}_\beta$ is

$$\mathbf{L}_\beta(\mathbf{w}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f^*_{\mathbf{X}|\mathbf{Z}}) = \left\{ y - \frac{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right\} E^*(\mathbf{X}|\boldsymbol{\delta}),$$

where E^* is computed under the model $f^*_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}, \mathbf{z})$. Using similar derivation, we can further obtain

$$\begin{aligned}
\mathbf{L}_\gamma(\mathbf{w}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f^*_{\mathbf{X}|\mathbf{Z}}) &= \left\{ y - \frac{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right\} \alpha_1 \mathbf{z}_{-1}, \\
\mathbf{L}_\alpha(\mathbf{w}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f^*_{\mathbf{X}|\mathbf{Z}}) &= \left\{ y - \frac{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right\} (1, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})^{\mathrm{T}}.
\end{aligned}$$

Then the estimation can be carried out through jointly solving

$$\begin{aligned}
\mathbf{0} &= \sum_{i=1}^{n} \left\{ y_i - \frac{\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right\} E^*(\mathbf{X}_i|\boldsymbol{\delta}_i), \\
\mathbf{0} &= \sum_{i=1}^{n} \left\{ y_i - \frac{\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right\} \theta'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{z}_{-1,i}
\end{aligned}$$

to estimate $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and

$$\mathbf{0} = \sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) \left(y_i - \frac{\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{\beta} + \alpha_0 + \alpha_1\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i}{1 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}/\sigma^2} \right) (1, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{\mathrm{T}}$$

at $\mathbf{z}_0 = \mathbf{z}_1, \ldots, \mathbf{z}_n$ to estimate $\widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = \widehat{\alpha}_0 + \widehat{\alpha}_1\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0$.

Similar calculations can also be made regarding the Poisson model

$$Y \sim \text{Poisson}[\exp\{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\}].$$

In this case, $\mathbf{L}_\beta$ takes the form

$$\mathbf{L}_\beta(\mathbf{w}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, f^*_{\mathbf{X}|\mathbf{Z}}) = a(\mathbf{w}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta) E^*(\mathbf{X}|\boldsymbol{\delta}),$$

where

$$a(\mathbf{w},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)=y-\frac{\sum_{y=0}^{\infty}y\exp[\{\boldsymbol{\delta}^{\mathrm{T}}\beta+\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}y-y^2\beta^{\mathrm{T}}\boldsymbol{\Sigma}\beta/2-\log(y!)]}{\sum_{y=0}^{\infty}\exp[\{\boldsymbol{\delta}^{\mathrm{T}}\beta+\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}y-y^2\beta^{\mathrm{T}}\boldsymbol{\Sigma}\beta/2-\log(y!)]},$$

E^* is computed under the model $f^*_{\mathbf{X}|\mathbf{Z}}(\mathbf{x},\mathbf{z})$. Using similar derivation, we can further obtain

$$\begin{aligned}\mathbf{L}_\gamma(\mathbf{w},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{\mathbf{X}|\mathbf{Z}}) &= a(\mathbf{w},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)\alpha_1\mathbf{z}_{-1},\\ \mathbf{L}_\alpha(\mathbf{w},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta,f^*_{\mathbf{X}|\mathbf{Z}}) &= a(\mathbf{w},\mathbf{z},y;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)(1,\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})^{\mathrm{T}}.\end{aligned}$$

Then the estimation can be carried out through jointly solving

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n}a(\mathbf{w}_i,\mathbf{z}_i,y_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)E^*(\mathbf{X}_i|\boldsymbol{\delta}_i),\\ \mathbf{0} &= \sum_{i=1}^{n}a(\mathbf{w}_i,\mathbf{z}_i,y_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)\theta'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{z}_{-1,i}\end{aligned}\tag{3.6}$$

to estimate $\boldsymbol{\beta},\boldsymbol{\gamma}$ and

$$\mathbf{0}=\sum_{i=1}^{n}a(\mathbf{w}_i,\mathbf{z}_i,y_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta)(1,\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{\mathrm{T}}$$

at $\mathbf{z}_0=\mathbf{z}_1,\ldots,\mathbf{z}_n$ to estimate $\widehat{\theta}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)=\widehat{\alpha}_0+\widehat{\alpha}_1\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0$.

3.3 Asymptotic properties and inference

In this section we show that the estimated covariate effect is asymptotically normal in Theorem 2 and locally efficient in Theorem 3. A by-product of the asymptotic normality property is that it facilitates testing if the estimated covariate effect is statistically significant.

Viewing $\theta(\cdot)$ as a one dimensional parameter, we have $\mathbf{L}_\alpha=L_\theta\boldsymbol{\theta}_\alpha$, where L_θ is obtained the same way as $\mathbf{L}_\alpha$ by replacing $\boldsymbol{\alpha}$ with θ, and $\boldsymbol{\theta}_\alpha$ is the partial derivative of $\theta(\cdot,\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$. Let $\boldsymbol{\theta}_{\alpha\alpha}=\partial\boldsymbol{\theta}_\alpha/\partial\boldsymbol{\alpha}^{\mathrm{T}}$. Let $\mathbf{L}_{\beta\beta}$, $\mathbf{L}_{\beta\gamma}$, $\mathbf{L}_{\beta\alpha}$ and $\mathbf{L}_{\beta\theta}$ be the partial derivative of $\mathbf{L}_\beta$ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ respectively. Similarly define $\mathbf{L}_{\gamma\beta}$, $\mathbf{L}_{\gamma\gamma}$, $\mathbf{L}_{\gamma\alpha}$, $\mathbf{L}_{\gamma\theta}$, $\mathbf{L}_{\alpha\beta}$, $\mathbf{L}_{\alpha\gamma}$, $\mathbf{L}_{\alpha\alpha}$ and $\mathbf{L}_{\alpha\theta}$. Let $\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})=E(L_{\theta\theta}\mid\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})$,

$\mathbf{U}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) = E\{(\mathbf{L}_{\beta\theta}^{\mathrm{T}} \quad \mathbf{L}_{\gamma\theta}^{\mathrm{T}})^{\mathrm{T}} \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}\}\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})^{-1}$, and $\boldsymbol{\theta}_\beta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) = -\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})^{-1}E(\mathbf{L}_{\theta\beta} \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})$, $\boldsymbol{\theta}_\gamma(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) = -\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})^{-1}E(\mathbf{L}_{\theta\gamma} \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})$. Define

$$\begin{aligned}\mathbf{A} &= E\left(\begin{bmatrix} \mathbf{L}_{\beta\beta}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} & \mathbf{L}_{\beta\gamma}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} \\ \mathbf{L}_{\gamma\beta}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} & \mathbf{L}_{\gamma\gamma}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} \end{bmatrix}\right) \\ &\quad + E\left(\begin{bmatrix} \mathbf{L}_{\beta\theta}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} \\ \mathbf{L}_{\gamma\theta}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\} \end{bmatrix}\{\boldsymbol{\theta}_\beta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) \quad \boldsymbol{\theta}_\gamma(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\}\right). \quad (3.7)\end{aligned}$$

Theorem 2. *Under the regularity conditions listed in Appendix B.4, we have the expansion*

$$\begin{aligned}&-\boldsymbol{A}n^{1/2}\begin{pmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\end{pmatrix} \\ =\ & n^{-1/2}\sum_{i=1}^{n}\left(\begin{bmatrix}\boldsymbol{L}_\beta\{Y_i,W_i,\boldsymbol{S}_i,\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\\ \boldsymbol{L}_\gamma\{Y_i,W_i,\boldsymbol{S}_i,\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\end{bmatrix}\right. \\ &\left.-\boldsymbol{U}(\widetilde{\boldsymbol{\gamma}}^{T}\boldsymbol{Z}_i)L_\theta\{Y_i,W_i,\boldsymbol{S}_i,\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\right)+o_p(1).\end{aligned}$$

Consequently, when $n\to\infty$,

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\to N\{\boldsymbol{0},(\boldsymbol{I}_\beta \quad \boldsymbol{0})\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{T})^{-1}(\boldsymbol{I}_\beta \quad \boldsymbol{0})^{T}\}$$

in distribution. Here, $\boldsymbol{I}_\beta$ *is the identity matrix with dimension being the length of* $\boldsymbol{\beta}$, *and*

$$\boldsymbol{B}=cov\left(\begin{bmatrix}\boldsymbol{L}_\beta\{Y,W,\boldsymbol{S},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\\ \boldsymbol{L}_\gamma\{Y,W,\boldsymbol{S},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\end{bmatrix}-\boldsymbol{U}(\widetilde{\boldsymbol{\gamma}}^{T}\boldsymbol{Z})L_\theta\{Y,W,\boldsymbol{S},\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\gamma},\theta(\cdot)\}\right) \quad (3.8)$$

Theorem 3. *If the conjectured model* $f^*_{X|\boldsymbol{S},\boldsymbol{Z}}(x \mid \boldsymbol{s},\boldsymbol{z})$ *is correct, the subsequent estimator* $\widehat{\boldsymbol{\beta}}$ *has the additional property that it is semiparametric efficient.*

The proofs of Theorems 2 and 3 are given in the Appendix B.3 and B.4.

In practice, the matrices $\mathbf{A}$ and $\mathbf{B}$ can be estimated through their sample versions, while $\Omega, \mathbf{U}, \boldsymbol{\theta}_\beta$ and $\boldsymbol{\theta}_\gamma$ need to be estimated via their corresponding nonparametric regression.

Knowing the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ allows us to perform various tests. Specifically, we can test the covariate effect described as $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{M}$ and $\mathbf{c}$ are the corresponding matrices or vectors used to describe the particular test of interest. As an example, we have the following Chi-square test result.

Theorem 4. *Under H_0, the test statistic*

$$T = n(\boldsymbol{M}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^T\{\boldsymbol{M}(\boldsymbol{I}_\beta \quad \boldsymbol{0})\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{A}}^{-1})^T(\boldsymbol{I}_\beta \quad \boldsymbol{0})^T\boldsymbol{M}^T\}^{-1}(\boldsymbol{M}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})$$

follows a chi-square distribution with degrees of freedom d_M, where d_M is the number of rows in $\boldsymbol{M}$.

We provide the proof of Theorem 4 in Appendix B.5.

3.4 Simulation

We perform four simulation studies to examine the finite sample performance of the proposed method.

In the first set of simulation studies, the response variable Y is binary with $Y = 0$ or 1, with the true g function of the form

$$g\{y, x, \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}), \boldsymbol{\beta}\} = \frac{\exp[y\{\beta_1 x + \beta_2 x^2 + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}]}{1 + \exp\{\beta_1 x + \beta_2 x^2 + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}}.$$

Thus, the parameter of interest $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$ consists of two components. The function $\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) = \cos(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})/2 - 1$.

Our first simulation is a relatively simple one, where the covariate vector $\mathbf{Z}$ has dimension $p = 2$. This yields a total of three parameters in addition to the univariate nonparametric function θ and the unknown distribution of X. In simulations 2 and 3, we increase the dimension of the covariate vector $\mathbf{Z}$ to three and four respectively, which yield four and five parameters in addition to the two unknown functions. In all the simulations, the covariate X and the measurement errors are generated from

normal distributions, and the covariate vector $\mathbf{Z}$ is generated from uniform distributions.

To compare the performance of various estimators, we implemented a naive estimator, two versions of the regression calibration estimators and two versions of the semiparametric estimators. In the naive estimator, the presence of measurement error is simply ignored and a profile likelihood estimation procedure is implemented to estimate the parameter $\boldsymbol{\beta}$. In the regression calibration procedures, we first calculate $X^* = E(X \mid W)$ and $X^{*2} = E(X^2 \mid W)$ then treat X^* and X^{*2} as X and X^2, and perform the profile likelihood estimation under the error-free model. In calculating $E(X \mid W)$ and $E(X^2 \mid W)$, we experimented with two situations, where we used two different working distributions of X, respectively normal and uniform. This corresponds to the true and misspecified distributional assumption on X. Finally, we also implemented the proposed semiparametric estimator, with the same working distributions of X. The estimation and inference results of all five estimators are given in Tables 3.1-3.3 respectively, corresponding to the three simulation studies. All the results are based on 1,000 simulated data sets with sample size 500. To see how the estimation procedure behaves with increasing dimension of $\mathbf{Z}$, we also experimented with $p > 4$. In our observation, with all other aspects of the simulation fixed, the procedure performs well until $p = 10$, when we started to see significant biases. Throughout the numerical analysis, we used the bandwidth $h = 3\mathrm{s}d(w)n^{-1/3}$, where $\mathrm{s}d(w)$ is the sample standard deviation of w. We also experimented with the bandwidth $h = 1.5\mathrm{s}d(w)n^{-1/3}$ and $h = 4.5\mathrm{s}d(w)n^{-1/3}$, the results appear insensitive to the bandwidth changes so are omitted.

The common observation across all simulations is that the naive estimator and the two regression calibration estimators tend to produce larger biases while the semiparametric estimators, whether performed under the true or misspecified working model of the distribution of X, have much smaller biases. The relatively large biases

of the naive and regression calibration estimators directly lead to invalid inference results, reflected in the terrible empirical coverage of the 95% confidence intervals. On the contrary, the semiparametric estimators not only yield very small biases, it also provides a close match between the sample standard deviations and their corresponding asymptotic versions. This leads to reasonable approximation of the empirical coverage of the 95% confidence intervals to the nominal level. It is worth pointing out that although we implemented an efficient estimator through adopting the working model for X as normal, and a non-efficient estimator through using uniform as the working model for X, the estimation variability of the two estimators are very close. In other words, the method appears to have certain robustness to the working model, in that in addition to retaining consistency as our theory has promised, it also seems to remain efficient regardless of the working model. The latter property is not within our expectation and whether this is a universal phenomenon with theoretical explanation deserves further investigation.

To further illustrate the generality of the results derived in this chapter, we perform a fourth set of simulation studies concerning a Poisson model. We generate the counting response variable Y with mean $\exp\{\beta x + \theta(\widetilde{\gamma}^{\mathrm{T}}\mathbf{z})\}$ and generate X from $N(0, 1.1^2)$. We set $\beta = 1.1$, $\theta(\widetilde{\gamma}^{\mathrm{T}}\mathbf{z}) = -0.4\cos(2.75\widetilde{\gamma}^{\mathrm{T}}\mathbf{z} - 1.0)$ and allow substantial meansurement error $\sigma_u = 0.8$. Following (3.6), we directly posit $E^*(X \mid \delta) = \delta^2$ and $E^*(X \mid \delta) = \delta\sin(\delta)$ for $E(X \mid \delta)$. We experimented with various dimension of $\mathbf{z}$ from 2 to 11 where $\mathbf{z}$ contain both continuous and discrete. Simulations results are summarized in Table 3.4. The consistency of our estimator, regardless if the posited models are correct or not, as well as the superioty of our method in contrast with the comparison methods are clear from these results.

3.5 Framingham heart study

We use our new methodology to analyse data from the Framingham Heart Study described in Section 1. The data set contains 1,126 male subjects. We use the occurrence of coronary heart disease as the response variable (Y), and systolic blood pressure, after subtracting 50 and taking logarithm transformation, as the covariate measured with error (W), see Carroll et al. (2006) who used this transformation, so that $W = X + U$, where X is the transformed true systolic blood pressure. We included age, the logarithm of 1 + the number of cigarettes smoked per day as reported by the subject and metropolitan relative weight as confounders $\mathbf{Z}$, with age chosen to be the leading component in $\mathbf{Z}$. Metropolitan relative weight is defined as the percentage of desirable weight (the ratio of actual weight to desirable weight times 100). Desirable weight was derived from the 1959 Metropolitan Life Insurance Company tables (Company (1959)) by taking the midpoint of the weight range for the medium build at a specified height, see also Hubert et al. (1983).

We fit the model with systolic blood pressure in its original scale. With $H(\cdot)$ being the logistic distribution function, the final model is

$$\begin{aligned} \mathrm{pr}(Y = 1 \mid X, \mathbf{Z}) &= H\left[\{\exp(X) + 50\}\,\beta + \theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\right], \\ W &= X + U. \end{aligned}$$

Using the available repeated measurements of W, we obtained the measurement error standard deviation to be 0.0745, and the Kolmogorov-Smirnov test for the normality of U yields a p-value of 0.701. We also include the qq-plot of the errors in Figure 3.1, which exhibits a linear pattern. Thus, we assume U has the centered normal distribution with standard deviation 0.0745.

The semiparametric analysis of the Framingham data, as well as the results from naive estimator and regression calibration estimators are given in Table 3.5. Not unexpectedly given the context, all results confirm the significance of the systolic

blood pressure as a risk factor for heart disease. In addition, the two estimates from the two semiparametric methods, conducted under a normal and a uniform working model for the distribution of X respectively, are very close. The naive estimator is attenuated towards zero by approximately 25%. Neither the effects from number of cigarettes smoked nor metropolitan weight is statistically significant. We also plot the estimated $\theta(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ as a function of $\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}$, as well as the 95% pointwise confidence bands in Figure 3.2 from both semiparametric methods, and we can see a general trend of increasing risk with increasing age.

3.6 Discussion

We have developed both estimation and inference tools to analyse covariate effect when the covariate under study is measured with error and also subject to confounding effects. The method is completely general, reflected in the generality of the main regression model. Specifically, we allow arbitrary regression relation between the response variable and the covariate under study, and we do not require a specific parametric model strategy for the confounding effects. Our procedure does not require any model assumption on the unobservable covariate of interest, and the framework can allow arbitrary measurement error structure. Under the special situation, when the regression model has a generalized partially linear form, and the measurement error is normal additive, great simplification occurs (Ma and Tsiatis (2006)) and the estimation procedure degenerates to a backfitted or profiled version of the estimator given in Stefanski and Carroll (1987).

We would like to point out that to solve the estimating equations, one could choose to use backfitting or profiling procedures. In our construction of the estimator, these are only two ways of solving the estimating equations jointly. Upon convergence, the solutions from backfitting and profiling are identical. They are both roots of the estimating equations. This is very different from using backfitting versus profiling before

estimating equations are derived, where using profiling or backfitting could result in different sets of estimating equations and hence both the theoretical and empirical performance can be different. The latter issue is well studied in Van Keilegom and Carroll (2007). Likewise, the nonparametric estimation of $\theta(\cdot)$ can also be carried out via splines, wavelets, etc., and research along these lines are certainly needed.

Table 3.1: Results of Simulation 1 with $p = 2$. The true parameter values, the estimates ($\widehat{\mu}$), the sample standard errors ("sd"), the mean of the estimated standard errors ($\widehat{\text{sd}}$) and the 95% confidence interval of five different estimators are reported. The five estimators are the naive estimator ("Naive"), the regression calibration estimators with two working distributions of X ("RC-nor" and "RC-Uni") and the semiparametric estimators with two working distributions of X ("Semi-nor" and "Semi-Uni").

	Naive		RC-nor		RC-Uni		Semi-nor		Semi-Uni	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
true	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
mean	0.5180	0.5665	0.6745	0.6725	1.0637	0.8451	0.7055	0.7122	0.7201	0.7190
sd	0.1803	0.0923	0.2115	0.1098	0.2704	0.1337	0.2546	0.1377	0.2429	0.1328
$\widehat{\text{sd}}$	0.1747	0.0888	0.2048	0.1054	0.2615	0.1285	0.2505	0.1346	0.2380	0.1301
95%CI	79.5%	62.2%	93.4%	93.3%	71.2%	79.8%	94.7%	94.4%	94.5%	95.3%

Table 3.2: Results of Simulation 2 with $p = 3$. The true parameter values, the estimates ($\widehat{\mu}$), the sample standard errors ("sd"), the mean of the estimated standard errors ($\widehat{\text{sd}}$) and the 95% confidence interval of five different estimators are reported.

	Naive		RC-nor		RC-Uni		Semi-nor		Semi-Uni	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
true	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
mean	0.5075	0.5637	0.6623	0.6694	1.0489	0.8406	0.6915	0.7084	0.7050	0.7151
sd	0.1768	0.0902	0.2072	0.1074	0.2692	0.1317	0.2430	0.1323	0.2328	0.1275
$\widehat{\text{sd}}$	0.1736	0.0882	0.2035	0.1048	0.2613	0.1285	0.2540	0.1370	0.2414	0.1344
95%CI	78.7%	62.8%	94.3%	93.4%	72.4%	81.1%	95.8%	95.6%	96.0%	96.0%

Table 3.3: Results of Simulation 3 with $p = 4$. The true parameter values, the estimates ($\widehat{\mu}$), the sample standard errors ("sd"), the mean of the estimated standard errors ($\widehat{\text{sd}}$) and the 95% confidence interval of five different estimators are reported.

	Naive		RC-nor		RC-Uni		Semi-nor		Semi-Uni	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
true	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
mean	0.5029	0.5606	0.6561	0.6658	1.0440	0.8377	0.6856	0.7043	0.6999	0.7111
sd	0.1828	0.0944	0.2150	0.1123	0.2748	0.1361	0.2572	0.1417	0.2481	0.1378
$\widehat{\text{sd}}$	0.1735	0.0885	0.2033	0.1052	0.2615	0.1283	0.2619	0.1390	0.2459	0.1352
95%CI	77.4%	60.9%	92.4%	91.5%	73.6%	80.6%	94.6%	94.0%	94.9%	94.7%

Figure 3.1: QQ-plot of the measurement errors in Framingham data analysis.

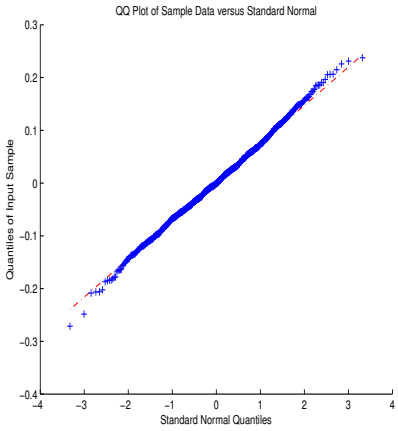

Figure 3.2: The estimated $\theta(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ as a function of $(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ in Framingham data analysis. Vertical axis stands for $\theta(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ and horizontal axis stands for $(\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$. In the left panels, $\widehat{\boldsymbol{\gamma}}$ is obtained with a normal working model on X and in the right panels $\widehat{\boldsymbol{\gamma}}$ is obtained with uniform working model on X. The plots in the lower panels contain the 95% confidence bands.

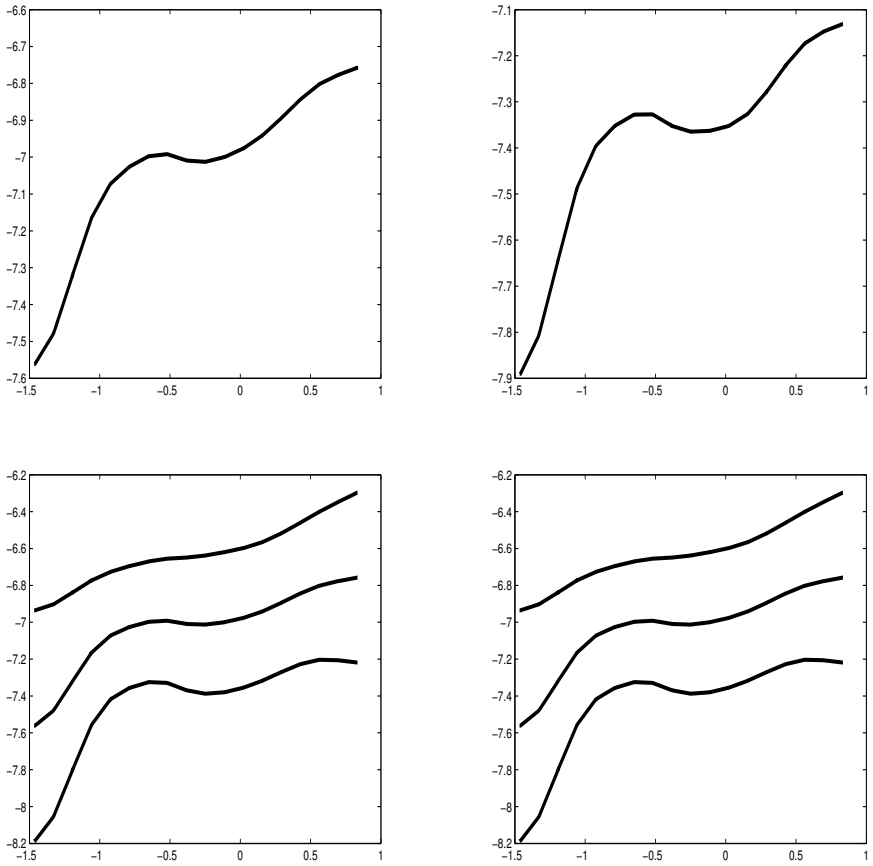

Table 3.4: Results of simulation 4 with $p = 2$ to 11. The true parameter is $\beta = 1.1$. The the estimates("est"), the sample standard errors ("sd"), the mean of the estimated standard errors ($\widehat{\text{sd}}$) and the 95% confidence interval of six different estimators are reported. The six estimators are the naive estimator ("Naive"), the regression calibration estimators with two working distributions of X ("RC-Nor" and "RC-Uni") , the oracle estimator ("Oracle"), and the local estimators with two posited forms of $E(X \mid \delta)$ ("Local 1" and "Local 2").

		Naive	RC-Nor	RC-Uni	Oracle	Local 1	Local 2
$p = 2$	est	0.7726	1.1147	1.1982	1.1128	1.0924	1.1074
	sd	0.1627	0.1900	0.2255	0.1161	0.1102	0.1063
	$\widehat{\text{sd}}$	0.0782	0.1204	0.1493	0.1660	0.1739	0.1820
	CI	0.2200	0.7620	0.7160	0.9560	0.9400	0.9380
$p = 3$	est	0.8092	1.1622	1.2482	1.1042	1.1090	1.1019
	sd	0.1201	0.1110	0.1506	0.1599	0.1382	0.0929
	$\widehat{\text{sd}}$	0.0848	0.1225	0.1579	0.1702	0.1683	0.1320
	CI	0.3080	0.9280	0.8160	0.9520	0.9560	0.9580
$p = 4$	est	0.8192	1.1847	1.2730	1.1046	1.1062	1.0907
	sd	0.1098	0.1149	0.1420	0.1275	0.0877	0.1138
	$\widehat{\text{sd}}$	0.0785	0.1122	0.1341	0.1550	0.1397	0.1296
	CI	0.2620	0.8920	0.7400	0.9740	0.9740	0.9400
$p = 5$	est	0.7294	1.1492	1.2206	1.1032	1.1021	1.1042
	sd	0.1409	0.1703	0.2023	0.1585	0.0755	0.0978
	$\widehat{\text{sd}}$	0.0822	0.1421	0.1598	0.1898	0.1432	0.1587
	CI	0.1760	0.8460	0.7780	0.9580	0.9600	0.9640
$p = 6$	est	0.8267	1.1703	1.2609	1.1221	1.0801	1.1021
	sd	0.1253	0.1248	0.1579	0.1863	0.1422	0.0975
	$\widehat{\text{sd}}$	0.0813	0.1272	0.1527	0.1682	0.1682	0.1993
	CI	0.3140	0.8760	0.7680	0.9620	0.9120	0.9640
$p = 7$	est	0.8076	1.1592	1.2426	1.1280	1.0983	1.0999
	sd	0.1332	0.1496	0.1821	0.2176	0.1263	0.0892
	$\widehat{\text{sd}}$	0.0826	0.1258	0.1478	0.1945	0.1962	0.2280
	CI	0.2900	0.8340	0.7660	0.9540	0.9360	0.9780
$p = 8$	est	0.8098	1.1658	1.2534	1.1256	1.0982	1.0989
	sd	0.1218	0.1362	0.1742	0.2059	0.1394	0.0886
	$\widehat{\text{sd}}$	0.0891	0.1292	0.1453	0.1776	0.1934	0.1906
	CI	0.2920	0.8400	0.7440	0.9560	0.9420	0.9780
$p = 9$	est	0.8079	1.1527	1.2310	1.1329	1.0935	1.1023
	sd	0.1385	0.1491	0.1804	0.2305	0.1514	0.0729
	$\widehat{\text{sd}}$	0.0854	0.1246	0.1464	0.1960	0.1913	0.2243
	CI	0.3120	0.8260	0.7740	0.9580	0.9220	0.9800
$p = 10$	est	0.8009	1.1568	1.2425	1.1224	1.1044	1.0988
	sd	0.1291	0.1506	0.1777	0.1880	0.1542	0.0812
	$\widehat{\text{sd}}$	0.0845	0.1262	0.1478	0.1980	0.2065	0.2181
	CI	0.2860	0.8560	0.7800	0.9620	0.9420	0.9820
$p = 11$	est	0.7958	1.1460	1.2314	1.1213	1.0915	1.1019
	sd	0.1355	0.1501	0.1779	0.2106	0.1364	0.0919
	$\widehat{\text{sd}}$	0.0859	0.1306	0.1455	0.2101	0.1975	0.2476
	CI	0.2900	0.8680	0.7740	0.9580	0.9160	0.9820

Table 3.5: Results of Framingham data analysis. The estimates ($\widehat{\mu}$) and the associated standard errors of five different estimators are reported. All values are multiplied by 100. In the table, $\widehat{\beta}_1$ is the regression coefficient for systolic blood pressure, $\widehat{\gamma}_1$ is the coefficient for transformed number of cigarettes smoked per day and $\widehat{\gamma}_2$ is the coefficient for metropolitan weight.

	$\widehat{\beta}_1$	$\mathrm{s}d(\widehat{\beta}_1)$	$\widehat{\gamma}_1$	$\mathrm{s}d(\widehat{\gamma}_1)$	$\widehat{\gamma}_2$	$\mathrm{s}d(\widehat{\gamma}_2)$
Naive	3.58	0.49	0.09	6.00	0.31	4.79
RC-Nor	4.22	0.60	0.11	5.99	0.25	4.77
RC-Uni	3.73	0.58	0.29	5.81	0.73	4.85
Semi-Nor	4.39	0.77	0.10	3.90	0.24	2.17
Semi-Uni	4.61	0.80	0.10	4.62	0.25	2.12

Chapter 4

Locally Efficient Semiparametric Estimator for Poisson Models with Measurement Error

4.1 Introduction

In regression analysis, it is common that some covariates cannot be measured precisely or directly, thus resulting in the measurement error models. The presence of measurement errors causes biased and inconsistent parameter estimates which leads to erroneous conclusions to various degrees in statistical inferences. As naive methods ignoring measurement errors results in biased estimation or misleading inferences, a large amount of papers and several books have been dedicated to correct such bias. The study on linear measurement error model dates back to Bickel and Ritov (1987) and a comprehensive study on linear models can be found in Fuller (1987). Carroll et al. (1995) extended the measurement error model framework to non-linear cases. Many further research works were devoted to various general measurement error models. For example, Tsiatis and Ma (2004) provided a class of locally efficient estimators for arbitrary parametric regression measurement error models. Ma and Tsiatis (2006) provided a closed form solution for generalized linear models and used it to handle heteroscedastic measurement errors. Ma and Carroll (2006) further extended the work to generalized regression model which contains a nonparametric component. Apanasovich et al. (2009) derived the limiting distribution of SIMEX in semiparametric problems, when the variable X subject to measurement error is modeled parametrically, nonparametrically or a combination of both. Stefanski et al.

(2013) proposed a measurement error model based approach to variable selection with application in nonparametric classification which results in a new kernel-based classifier with LASSO-like shrinkage and variable-selection properties. Zhang et al. (2014) investigated the sample property in efficient variable selection problems and developed a semiparametric profile least-square based estimation procedure to estimate the parameters in partial linear single index models. Measurement error problems have attracted researchers from other scientific research fields as well. For example, Hsiao (1989), Horowitz and Markatou (1996), Dynan (1996) and Parker and Preston (2005) discussed methods in measurement error problems that take advantages of longitudinal data and time series structures. Chesher (1991), Chesher et al. (2002) and Chesher and Schluter (2002) proposed a small noise approximation to assess the effect of measurement errors. Hu and Schennach (2010) introduced a semiparametric sieve estimator for nonclassical measurement error models. See Chen et al. (2007) for a review on recent advances in measurement error models for applied researchers.

Count data analysis has attracted considerable research interest and a large number of inference methods have been proposed in the literature. One popular approach is to utilize a Poisson regression model. However, when some of the covariates cannot be measured directly or correctly, ignoring the measurement error in estimation will suffer severe bias like in the familiar linear regression models. Huang (2014) proposed a trend-constrained corrected score approach for loglinear model for Poisson mean which requires the compactness of the parameter space. Such an approach requires almost surely negative definitive in the first derivative of the local trend of the corrected score as well as almost surely negative derivative in the corrected profile score. As far as we know, very limited work exists to handle Poisson model where some covariates are measured with errors. Even though Stefanski and Carroll (1987) discussed count response models in their applications without actual implementation. Liu et al. (2017) implemented a special case for the Poisson model with covariate er-

rors where simplifications occur to generate closed form estimating equations. The general difficulties lie in both the theoretical challenge and computational complexity involved with count response data. Following the notation of Carroll et al. (1995), let us write Z as the predictor variables that can be measured precisely and write X as those that cannot. Instead of X, we observe its erroneous version W, where the relations between W and X is specified, i.e. $f_{W|X}(w \mid x)$ is a known model. For example, $W = X + U$, where U is a random measurement error. Let Y be the observable count response variable. In this chapter, we develop methodology to correct the bias caused by covariate measurement error in handling the count response variable Y.

The rest of the chapter is organized as follow. In Section 4.2, we investigate five interrelated Poisson models and present our local semiparametric efficient estimation approach for each model. In Section 4.3, we report numerical experimentation results for each model discuss in Section 4.2. Real data application can be found in Section 4.4. We conclude with a brief discussion in Section 4.5.

4.2 Models and Methods

In this section, we study five related Poisson models where the main covariate of interest is measured with error. We start with the simplest case, and then progress to more general and complex cases.

Linear Poisson Model

We first consider a linear Poisson model with a normal additive measurement error, where the variance of the error term is Ω. The relationship between the response variable Y and the covariate $\mathbf{X}$ is

$$Y \mid \mathbf{X} \sim \text{Poisson}(e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}}), \mathbf{W} = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim \text{Normal}(\mathbf{0}, \Omega).$$

Of course, $\mathbf{X}$ cannot be measured correctly. The data we observe is $(Y_i, \mathbf{W}_i)$ for $i = 1, 2, \cdots, n$. Parameters of interest is $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. The specific form of the model is

$$\begin{aligned} p(y \mid \mathbf{x}; \boldsymbol{\theta}) &= \exp\left\{(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})y - e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}} - \log(y!)\right\}, \\ p(w \mid \mathbf{x}) &= \frac{(2\pi)^{-\frac{r}{2}}}{\mid \Omega \mid^{\frac{1}{2}}} \exp\left\{-(\mathbf{w} - \mathbf{x})^{\mathrm{T}}\Omega^{-1}\frac{\mathbf{w} - \mathbf{x}}{2}\right\}, \end{aligned}$$

where r is the dimension of $\boldsymbol{\beta}$. Following Stefanski and Carroll (1987), a complete and sufficient "statistic" is given by $\delta(\mathbf{w}, y; \boldsymbol{\theta}) = \mathbf{w} + y\Omega\boldsymbol{\beta}$. Using a change of variable to replace $\mathbf{w}$ with δ, the Jacobian is $J(\delta, y; \boldsymbol{\theta}) = 1$. In addition, $\frac{d\delta}{d\alpha} = 0$, $\frac{d\delta}{d\boldsymbol{\beta}} = y\Omega$. Thus, we obtain

$$p(y \mid \delta; \boldsymbol{\theta}) = \exp[\xi y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!) - \log\{s(\xi, \boldsymbol{\beta})\}],$$

where $\xi = \alpha + \delta^{\mathrm{T}}\boldsymbol{\beta}$ and $s(\xi, \boldsymbol{\beta}) = \sum_{y=0}^{\infty} \exp\{\xi y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}$. Further, the efficient score function for α is

$$\begin{aligned} S_{\mathrm{eff},\alpha}(y, \mathbf{w}; \boldsymbol{\theta}) &= \frac{\partial}{\partial\alpha}\log\, p(y \mid \delta; \boldsymbol{\theta}) \\ &= y - E(Y \mid \delta) \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} E(Y \mid \delta) &= \frac{d}{d\xi}\log\{s(\xi, \boldsymbol{\beta})\} \\ &= \frac{\sum_{y=0}^{\infty} y \exp\{(\alpha + \delta^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}}{\sum_{y=0}^{\infty} \exp\{(\alpha + \delta^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}}. \end{aligned}$$

Similarly, the efficient score for $\boldsymbol{\beta}$ up to a scalar is given by

$$\mathbf{S}_{\mathrm{eff},\boldsymbol{\beta}}(y, \mathbf{w}; \boldsymbol{\theta}) = \{y - E(Y \mid \delta)\}E(\mathbf{X} \mid \delta).$$

Then the efficient estimator can be obtained through implementing the estimating equations

$$\sum_{i=1}^{n} S_{\mathrm{eff},\alpha}(y_i, \mathbf{w}_i; \boldsymbol{\theta}) = 0 \quad \text{and} \quad \sum_{i=1}^{n} \mathbf{S}_{\mathrm{eff},\boldsymbol{\beta}}(y_i, \mathbf{w}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{4.2}$$

However, in order to compute $E(\mathbf{X} \mid \delta)$, we need to know the distribution model for $\mathbf{X}$. Since we do not know the true form of $E(\mathbf{X} \mid \delta)$, we utilize kernel regression to estimate $\widehat{E}(\mathbf{X} \mid \delta)$ when a validation data set with some $\mathbf{X}$ observations are available. The estimator from such nonparametric approach will serve as a benchmark. Otherwise, when no validation data is available, we directly propose a functional form for $E(\mathbf{X} \mid \delta)$. Regardless of the functional form, the resulting estimator is always consistent. We name the resulting estimator the local efficient estimator. The algorithm is given below.

Step 1. Propose a functional model $\eta^*(\delta)$ for $E(\mathbf{X} \mid \delta)$.

Step 2. Apply a standard profiling method and solve the estimating equations

$$\begin{aligned} 0 &= \sum_{i=1}^{n} S_{\text{eff},\alpha}(y_i, \mathbf{w}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \left[y_i - \frac{\sum_{y=0}^{\infty} y \exp\{(\alpha + \delta_i^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}}{\sum_{y=0}^{\infty} \exp\{(\alpha + \delta_i^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}} \right] \\ \mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}_{\text{eff},\boldsymbol{\beta}}(y_i, \mathbf{w}_i; \boldsymbol{\theta}, \eta^*) \\ &= \sum_{i=1}^{n} \left[y_i - \frac{\sum_{y=0}^{\infty} y \exp\{(\alpha + \delta_i^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}}{\sum_{y=0}^{\infty} \exp\{(\alpha + \delta_i^{\mathrm{T}}\boldsymbol{\beta})y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}} \right] \eta^*(\delta_i), \end{aligned}$$

denote the local efficient estimator of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$.

Linear Poisson Model with a Nonparametric Component

We extend the linear Poisson measurement error model to the following partially linear Poisson case

$$Y \mid (\mathbf{X}, Z) \sim \text{Poisson}\left(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + g(z)}\right), \mathbf{W} = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim \text{Normal}(\mathbf{0}, \Omega), \tag{4.3}$$

where $g(z)$ is an unknown smooth function of z. Similarly, $\mathbf{X}$ could not be measured precisely. Instead we observe $(Y_i, \mathbf{W}_i, Z_i)$ for $i = 1, 2, \cdots, n$. The parameters are

$\boldsymbol{\theta} = (g(z), \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, where $g(z)$ is considered as a nuisance parameter. Our interest lies on $\boldsymbol{\beta}$. Specifically, the model is of the form

$$\begin{aligned} p(y \mid \mathbf{x}, z; \boldsymbol{\theta}) &= \exp\left[\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + g(z)\}y - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}+g(z)} - \log(y!)\right] \\ p(\mathbf{w} \mid \mathbf{x}) &= \frac{(2\pi)^{-\frac{r}{2}}}{\mid \Omega \mid^{\frac{1}{2}}} \exp\left\{-(\mathbf{w}-\mathbf{x})^{\mathrm{T}}\Omega^{-1}\frac{\mathbf{w}-\mathbf{x}}{2}\right\} \end{aligned}$$

where r is the dimension of $\boldsymbol{\beta}$. Even though there is a nonparametric component in the Poisson mean, it is still linear in terms of $\mathbf{X}$ and hence the technique in Stefanski and Carroll (1987) can be employed. To this end, the complete and sufficient "statistic" for this model is $\delta(w, y; \boldsymbol{\beta}) = \mathbf{w} + y\Omega\boldsymbol{\beta}$. Similarly the Jacobian is $J(\delta, z, y; \boldsymbol{\beta}) = 1$ and $\frac{d\delta}{dg(z)} = 0$, $\frac{d\delta}{d\beta} = y\Omega$. We obtain

$$p(y \mid \delta, z; \boldsymbol{\beta}) = \exp\left[\xi y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!) - \log\{s(\xi, \boldsymbol{\beta})\}\right]$$

where $\xi = \delta^{\mathrm{T}}\boldsymbol{\beta} + g(z)$ and $s(\xi, \boldsymbol{\beta}) = \sum_{y=0}^{\infty}\exp\{\xi y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)\}$. The efficient score function for $g(z)$ is

$$\begin{aligned} S_{\text{eff},g}(y, \mathbf{w}, z; \boldsymbol{\beta}) &= \frac{\partial}{\partial g(z)}\log\, p(y \mid \delta, z; \boldsymbol{\beta}) \\ &= y - E\{Y \mid \delta, z, g(z)\} \end{aligned} \tag{4.4}$$

where

$$\begin{aligned} E\{Y \mid \delta, z, g(z)\} &= \frac{d}{d\xi}\log\{s(\xi, \boldsymbol{\beta})\} \\ &= \frac{\sum_{y=0}^{\infty} y\exp[\{\delta^{\mathrm{T}}\boldsymbol{\beta} + g(z)\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]}{\sum_{y=0}^{\infty}\exp[\{\delta^{\mathrm{T}}\boldsymbol{\beta} + g(z)\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]}. \end{aligned}$$

Similarly, the efficient score for $\boldsymbol{\beta}$ up to a scalar is given by

$$\mathbf{S}_{\text{eff},\beta}(y, \mathbf{w}, z; \boldsymbol{\beta}) = \{y - E(Y \mid \delta, z)\}E(\mathbf{X} \mid \delta, z)$$

Then the efficient estimator can be obtained through implementing the estimating equation

$$\sum_{i=1}^{n}\mathbf{S}_{\text{eff},\beta}(y_i, \mathbf{w}_i, z_i; \boldsymbol{\beta}) = \mathbf{0} \tag{4.5}$$

In order to implement the above estimating equations, we need to know the quantity of $E(\mathbf{X} \mid \delta, z)$. As in the linear case, we impose some distribution on $\mathbf{X}$ or Z to facilitate the computation, or even simpler, we directly propose a functional form $\eta^*(\delta, z)$ for $E(\mathbf{X} \mid \delta, z)$. In addition, we also perform a kernel regression to estimate $E(\mathbf{X} \mid \delta, z)$, denoted as $\widehat{E}(\mathbf{X} \mid \delta, z)$ when validation data is available. Of course, the estimator from such nonparametric approach serves only as a benchmark.

For implementation, we still need to handle the unknown function $g(z)$. To do this, we localize part of the estimating equation using kernel weights, and solve

$$\sum_{i=1}^{n} \left(y_i - \frac{\sum_{y=0}^{\infty} y \exp[\{\delta_i^{\mathrm{T}}\boldsymbol{\beta} + g(z_0)\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]}{\sum_{y=0}^{\infty} \exp[\{\delta_i^{\mathrm{T}}\boldsymbol{\beta} + g(z_0)\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]} \right) K_h(z_i - z_0) = 0 \quad (4.6)$$

to obtain $\widehat{g}(z_0, \boldsymbol{\beta})$. Here $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function with bandwidth h. Then we adopt a standard profiling method and implement the estimating equation for $\boldsymbol{\beta}$

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}_{\mathrm{eff},\boldsymbol{\beta}}(y_i, \mathbf{w}_i, \widehat{g}(z_i, \boldsymbol{\beta}); \boldsymbol{\beta}, \eta^*) \\ &= \sum_{i=1}^{n} \left(y_i - \frac{\sum_{y=0}^{\infty} y \exp[\{\delta_i^{\mathrm{T}}\boldsymbol{\beta} + \widehat{g}(z_i, \boldsymbol{\beta})\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]}{\sum_{y=0}^{\infty} \exp[\{\delta_i^{\mathrm{T}}\boldsymbol{\beta} + \widehat{g}(z_i, \boldsymbol{\beta})\}y - \frac{1}{2}y^2\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \log(y!)]} \right) \eta^*(\delta_i, z_i), \end{aligned}$$

where $\widehat{g}(z_i, \boldsymbol{\beta})$ is the estimator of $g(z_0, \boldsymbol{\beta})$ obtained from (4.6) evaluated at $z_0 = z_i$ for $i = 1, 2, \ldots, n$. We denote the resulting local efficient estimator of $\boldsymbol{\beta}^{\mathrm{T}}$ as $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}$.

Nonlinear Poisson Model

To be even more flexible in the modeling, we now consider measurement error models with a nonlinear component $f(\mathbf{x}, \boldsymbol{\beta})$ in the Poisson mean. The model is

$$Y \mid X \sim \text{Poisson}\left(e^{f(\mathbf{x},\boldsymbol{\beta})}\right), \mathbf{W} = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim \text{Normal}(\mathbf{0}, \Omega), \quad (4.7)$$

while the data we observe is $(Y_i, \mathbf{W}_i), i = 1, 2, \cdots, n$. Here $f(\mathbf{x}, \boldsymbol{\beta})$ can be any nonlinear function of $\mathbf{x}$. For example, a polynomial form $f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$,

where the parameter of interest is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$. Specifically, in this case the model can be written as

$$\begin{aligned} p(y \mid \mathbf{x}; \boldsymbol{\beta}) &= \exp\left\{ f(\mathbf{x}, \boldsymbol{\beta}) y - e^{f(\mathbf{x}, \boldsymbol{\beta})} - \log(y!) \right\} \\ p(\mathbf{w} \mid \mathbf{x}) &= \frac{(2\pi)^{-\frac{r}{2}}}{\mid \Omega \mid^{\frac{1}{2}}} \exp\left\{ -(\mathbf{w} - \mathbf{x})^{\mathrm{T}} \Omega^{-1} \frac{\mathbf{w} - \mathbf{x}}{2} \right\} \end{aligned}$$

Assume Y and $\mathbf{W}$ are independent conditional on $\mathbf{X}$. The probability density of the full data is given by

$$p(y, \mathbf{w}, \mathbf{x}) = p(y, \mathbf{w}, \mid \mathbf{x}) p(\mathbf{x}) = p(y \mid \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}) \eta(\mathbf{x})$$

Where the form of $\eta(\mathbf{x})$ is unknown. We posit some model $\eta^*(\mathbf{x})$ for $\eta(\mathbf{x})$, then the observed-data score vector is given by

$$\mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}) = \frac{\int \mathbf{S}^F_{\boldsymbol{\beta}}(y, \mathbf{x}) p(y \mid \mathbf{x}) p(\mathbf{w} \mid \mathbf{x}) \eta^*(\mathbf{x}) d\mu(\mathbf{x})}{\int p(y \mid \mathbf{x}) p(\mathbf{w} \mid \mathbf{x}) \eta^*(\mathbf{x}) d\mu(\mathbf{x})},$$

where $\mathbf{S}^F_{\boldsymbol{\beta}}(y, \mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log p(y \mid \mathbf{x}, \boldsymbol{\beta})$. In the example mentioned above, this leads to

$$\begin{aligned} S^F_{\beta_0}(y, x) &= y - \exp(\beta_0 + \beta_1 x + \beta_2 x^2), \\ S^F_{\beta_1}(y, x) &= x\{y - \exp(\beta_0 + \beta_1 x + \beta_2 x^2)\}, \\ S^F_{\beta_2}(y, x) &= x^2\{y - \exp(\beta_0 + \beta_1 x + \beta_2 x^2)\}, \end{aligned}$$

and the joint distribution of Y and W conditional on x is

$$\begin{aligned} & p(y \mid x) p(w \mid x) \\ = \; & (2\pi\sigma_u^2)^{-\frac{1}{2}} \exp\left\{ (\beta_0 + \beta_1 x + \beta_2 x^2) y - e^{\beta_0 + \beta_1 x + \beta_2 x^2} - \log(y!) - \frac{(w - x)^2}{2} \right\}. \end{aligned}$$

In order to implement the efficient score

$$\mathbf{S}^*_{\mathrm{eff}, \boldsymbol{\beta}}(y, \mathbf{w}) = \mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}) - E^*\{\mathbf{a}(\mathbf{X}) \mid y, \mathbf{w}\},$$

by Theorem 1 of Tsiatis and Ma (2004), we need to solve for $\mathbf{a}(\mathbf{X})$ which satisfies

$$E\{\mathbf{S}^*_{\boldsymbol{\beta}}(Y, \mathbf{W}) \mid \mathbf{x}\} = E[E^*\{\mathbf{a}(\mathbf{X}) \mid Y, \mathbf{W}\} \mid \mathbf{x}].$$

We consider approximating $\eta^*(\mathbf{x})$ as $\eta^*(\mathbf{x}) \approx \sum_{j=1}^m \eta^*(\mathbf{x}_j) I(\mathbf{x} = \mathbf{x}_j)$. For example, when we propose a uniform model, $\eta^*(\mathbf{x}) \approx \sum_{j=1}^m I(\mathbf{x} = \mathbf{x}_j)$, when we propose a standard normal model, $\eta^*(\mathbf{x}) \approx \sum_{j=1}^m \phi(\mathbf{x}_j) I(\mathbf{x} = \mathbf{x}_j)$, where $\phi(\mathbf{x}_j)$ is the standard normal pdf. Therefore,

$$\begin{aligned} & E[E^*\{\mathbf{a}(\mathbf{X}) \mid Y, \mathbf{W}\} \mid \mathbf{X} = \mathbf{x}_i] \\ = & \int \frac{\sum_{j=1}^m \mathbf{a}_j p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)} p(y \mid \mathbf{x}_i) p(\mathbf{w} \mid \mathbf{x}_i) d\mu(y) d\mu(\mathbf{w}) \end{aligned}$$

where $\mathbf{a}_j = \mathbf{a}(\mathbf{x}_j)$. Also

$$\begin{aligned} & E\{\mathbf{S}^*_\beta(Y, W) \mid \mathbf{X} = \mathbf{x}_i\} \qquad (4.8) \\ = & \int \frac{\sum_{j=1}^m \mathbf{S}^F_\beta(y, \mathbf{x}_j) p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)} p(y \mid \mathbf{x}_i) p(\mathbf{w} \mid \mathbf{x}_i) d\mu(y) d\mu(\mathbf{w}) \end{aligned}$$

Consequently, the solutions to the integral equations reduce to the linear equations $\mathbf{A}\mathbf{a}^{\mathrm{T}} = \mathbf{b}^{\mathrm{T}}$ where the solutions $\mathbf{a}$ is the $q \times m$ matrix $\{\mathbf{a}_1, \cdots, \mathbf{a}_m\}$, $\mathbf{b}$ is a $q \times m$ matrix whose i^{th} column is $E\{\mathbf{S}^*_\beta(Y, \mathbf{W}) \mid \mathbf{x} = \mathbf{x}_i\}$ defined in (4.8), and $\mathbf{A}$ is a $m \times m$ matrix whose (i, j) element is given by

$$A_{ij} = \int \frac{p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)} p(y \mid \mathbf{x}_i) p(\mathbf{w} \mid \mathbf{x}_i) d\mu(y) d\mu(\mathbf{w}).$$

Hence the efficient score is given by

$$\mathbf{S}^*_{\text{eff}}(y, \mathbf{w}) = \frac{\sum_{j=1}^m \{\mathbf{S}^F_\beta(y, \mathbf{x}_j) - \mathbf{a}_j\} p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*(\mathbf{x}_j)}. \qquad (4.9)$$

We then solve $\sum_{i=1}^n \mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}(y_i, \mathbf{w}_i) = \mathbf{0}$ to obtain $\widehat{\boldsymbol{\beta}}$.

Nonlinear Poisson with a Nonparametric Component Model

In additional to the nonlinear structure discussed in the above section, there could be some other pertinent factor contributing to the outcome in an unknown fashion. In this section, we consider this more general situation by including a nonparametric function $g(z)$, where Z is a variable whose contribution to the response is left

unspecified. Specifically, we consider the model

$$Y \mid (\mathbf{X}, Z) \sim \text{Poisson}[e^{f\{\mathbf{x},\boldsymbol{\beta},g(z)\}}]. \tag{4.10}$$

Similar as before, we let $\mathbf{W} = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim \text{Normal}(\mathbf{0}, \Omega)$, and the observations are $(Y_i, \mathbf{W}_i, Z_i)$ where $i = 1, 2, \cdots, n$. For example, if $f\{x, \boldsymbol{\beta}, g(z)\} = g(z) + \beta_1 x + \beta_2 x^2$, then the model can be written as

$$\begin{aligned} p(y \mid x, z) &= \exp\left[\{g(z) + \beta_1 x + \beta_2 x^2\}y - e^{g(z)+\beta_1 x+\beta_2 x^2} - \log(y!)\right], \\ p(w \mid x) &= (2\pi\sigma_u^2)^{-\frac{1}{2}} \exp\left\{-\frac{(w-x)^2}{2\sigma_u^2}\right\}. \end{aligned}$$

Assume Y and $\mathbf{W}$ are independent conditional on $(\mathbf{X}, Z)$, then the probability density of the full data becomes $p(y, \mathbf{w}, \mathbf{x}, z; \boldsymbol{\beta}, g, \eta_1, \eta_2) = p\{y \mid z, \mathbf{x}; \boldsymbol{\beta}, g(z)\}p(\mathbf{w} \mid \mathbf{x}, z)\eta_1(\mathbf{x} \mid z)\eta_2(z)$, where the conditional probability density function of $\mathbf{X}$ given Z, denoted by $\eta_1(\mathbf{x} \mid z)$, and the density function of Z, denoted $\eta_2(z)$, are both unknown. The parameters in this model are $\boldsymbol{\theta} = (\boldsymbol{\beta}, g, \eta_1, \eta_2)$ while our interest lies solely in $\boldsymbol{\beta}$. Let us denote the expectations computed under the correct model and the posited model $\eta_1^*(\mathbf{x} \mid z)$ for $\mathbf{X}$ given Z by $E(\cdot)$ and $E^*(\cdot)$, respectively. Conceptually, we treat g as if it is an unknown constant, then following (4.9), we derive the locally efficient estimating function for $\boldsymbol{\beta}$ to be

$$\mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}(y, \mathbf{w}, z, \boldsymbol{\beta}, g, \eta_1) = \mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}, z) - E^*\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, z) \mid (y, \mathbf{w}, z)\}, \tag{4.11}$$

where $\mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}, z)$ is the observed data score vector w.r.t $\boldsymbol{\beta}$. Note that η_2 drops out of the derivation so it does not play a role. $\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, z)$ in (4.11) satisfies

$$E\{\mathbf{S}^*_{\boldsymbol{\beta}}(Y, \mathbf{W}, z) \mid \mathbf{x}, z\} = E[E^*\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, z) \mid Y, \mathbf{W}, z\} \mid \mathbf{x}, z].$$

Unlike η_2, the nuisance function $g(\cdot)$ carries all the way through the implementation for $\boldsymbol{\beta}$ in (4.11). Therefore, even though $g(z)$ is not of our primary interest, we are obliged to estimatie it. We propose to estimate $g(z)$ at $z = z_0$ via

$$\sum_{i=1}^{n} S^*_{\text{eff},g}(y_i, \mathbf{w}_i, z_i, \boldsymbol{\beta}, g(z_0, \mathbf{x}, \boldsymbol{\beta}), \eta_1^*)K_h(z_i - z_0) = 0, \tag{4.12}$$

where $S^*_{\text{eff},g}(y, \mathbf{w}, z, \boldsymbol{\beta}, g, \eta_1) = S^*_g(y, \mathbf{w}, z) - E^*\{\alpha_g(\mathbf{X}, z) \mid (y, \mathbf{w}, z)\}$, $S^*_g(y, \mathbf{w}, z)$ is the score vector w.r.t. the unknown function $g(z)$, i.e.,

$$S^*_g(y, \mathbf{w}, z) = \frac{\int S^F_g(y, \mathbf{x}, z) p(y \mid \mathbf{x}, z) p(\mathbf{w} \mid \mathbf{x}) \eta^*_1(\mathbf{x} \mid z) d\mu(\mathbf{x})}{\int p(y \mid \mathbf{x}, z) p(\mathbf{w} \mid \mathbf{x}) \eta^*_1(\mathbf{x} \mid z) d\mu(\mathbf{x})},$$

and $\alpha_g(\mathbf{X}, z)$ in (4.12) satisfies

$$E\{S^*_g(Y, \mathbf{W}, z) \mid \mathbf{x}, z\} = E[E^*\{\alpha_g(\mathbf{X}, z) \mid Y, \mathbf{W}, z\} \mid \mathbf{x}, z].$$

When the Poisson mean model is $f\{x, \boldsymbol{\beta}, g(z)\} = g(z) + \beta_1 x + \beta_2 x^2$, then $S^F_g(y, x, z) = y - \exp\{g(z) + \beta_1 x + \beta_2 x^2\}$. Again, we consider approximating $\eta_1(\mathbf{x} \mid z)$ as $\eta^*_1(\mathbf{x} \mid z) \approx \sum_{j=1}^m c_j(z) I(\mathbf{x} = \mathbf{x}_j)$ where $\sum_{j=1}^m c_j(z) = 1$ for all z over the support of Z. For example, when we propose a same uniform model for all z, $\eta^*_1(\mathbf{x} \mid z) \propto \sum_{j=1}^m I(\mathbf{x} = \mathbf{x}_j)$, when we propose a normal model with variance σ^2 and independent of Z, $\eta^*_1(\mathbf{x} \mid z) \propto \sum_{j=1}^m \phi(\mathbf{x}_j) I(\mathbf{x} = \mathbf{x}_j)$, where $\phi(\cdot)$ is the multivariate standard normal pdf. Then

$$\begin{aligned} & E[E^*\{\alpha_g(\mathbf{X}, z) \mid Y, \mathbf{W}, z\} \mid \mathbf{X} = \mathbf{x}_i, z] \\ = & \int \frac{\sum_{j=1}^m \alpha_g(\mathbf{x}_j, z) p(y \mid \mathbf{x}_j, z) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*_1(\mathbf{x}_j \mid z)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j, z) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*_1(\mathbf{x}_j \mid z)} p(y \mid \mathbf{x}_i, z) p(\mathbf{w} \mid \mathbf{x}_i) d\mu(y) d\mu(\mathbf{w}) \end{aligned}$$

and

$$\begin{aligned} & E\{S^*_g(Y, \mathbf{W}, z) \mid \mathbf{X} = \mathbf{x}_i, z\} \\ = & \int \frac{\sum_{j=1}^m S^F_g(y, \mathbf{x}_j, z) p(y \mid \mathbf{x}_j, z) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*_1(\mathbf{x}_j \mid z)}{\sum_{j=1}^m p(y \mid \mathbf{x}_j, z) p(\mathbf{w} \mid \mathbf{x}_j) \eta^*_1(\mathbf{x}_j \mid z)} p(y \mid \mathbf{x}_i, z) p(\mathbf{w} \mid \mathbf{x}_i) d\mu(y) d\mu(\mathbf{w}). \end{aligned}$$

We adopt a standard profile likelihood approaches to solve the estimating equation (4.12) to obtain $\widehat{g}(z_0, \boldsymbol{\beta})$ and then solve (4.11) to obtain the locally efficient estimator, denoted as $\widehat{\boldsymbol{\beta}}$.

Nonlinear Partial Index Poisson Model

Often times, the pertinent factor $\mathbf{Z}$ is of high dimension, say $p + 1$, thus models in the Nonlinear Poisson with a Nonparametric Component Model Section are not

feasible to use. The linear model of $\mathbf{Z}$ is a popular choice, although somewhat restrictive. We summarize the effect of $\mathbf{Z}$ using an index and consider a data adaptive model $E(Y \mid \mathbf{x}, \mathbf{z}) = \exp[f\{\mathbf{x}, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}]$, where $g(\cdot)$ is unspecified and can be any smooth function. Therefore, we arrive at a flexible semiparametric model that overcomes the curse of dimensionality. We write the association between the outcome Y and covariate $(\mathbf{X}, \mathbf{Z})$ and the measurement error structure as

$$Y \mid (\mathbf{X}, \mathbf{Z}) \sim \text{Poisson}[e^{f\{\mathbf{x}, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\}}], \mathbf{W} = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim \text{Normal}(\mathbf{0}, \Omega) \tag{4.13}$$

where $g(\cdot)$ is an unknown function. For identification purpose, without loss of generality, assume $\widetilde{\boldsymbol{\gamma}} = (1, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a p-dimensional unknown vector that can be estimated from the data. In this model, we observe data $(Y_i, \mathbf{W}_i, \mathbf{Z}_i)$ for $i = 1, 2, \cdots, n$. For example, if $f\{x, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\} = g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2$, then the specific form of the nonlinear single index model becomes

$$\begin{aligned} p(y \mid x, \mathbf{z}) &= \exp\left[\{g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2\}y - e^{g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2} - \log(y!)\right] \\ p(w \mid x) &= (2\pi\sigma_u^2)^{-\frac{1}{2}} \exp\left\{-\frac{(w-x)^2}{2\sigma_u^2}\right\}. \end{aligned}$$

We assume Y and $\mathbf{W}$ are independent given the covariates $(\mathbf{X}, \mathbf{Z})$. Then we write the probability density of the observed data as $p(y, \mathbf{w}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1, \eta_2) = \int p(y \mid \mathbf{z}, \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\gamma}, g) p(\mathbf{w} \mid \mathbf{x}, \mathbf{z}) \eta_1(\mathbf{x} \mid \mathbf{z}) \eta_2(\mathbf{z}) d\mu(\mathbf{x})$. In addition to the single index structure function $g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$, the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ which denoted by $\eta_1(\mathbf{x} \mid \mathbf{z})$ and the marginal density function of $\mathbf{Z}$, denoted $\eta_2(\mathbf{z})$ are also unspecified. The parameter space is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1, \eta_2)$ while our interest only lies in $\boldsymbol{\beta}$ which associates with the error-prone covariate $\mathbf{X}$. Following the general idea of Liu et al. (2017) and following the same spirit in the previous sections, we employ a working version of $\eta_1(\mathbf{x} \mid \mathbf{z})$, denote as $\eta_1^*(\mathbf{x} \mid \mathbf{z})$, and propose a semiparametric approach to estimate the effect of the covariate of interest as well as the nuisance parameters $g, \boldsymbol{\gamma}$ through solving estimating equations. At any nonconstant given function $g(\cdot)$, we estimate $\boldsymbol{\beta}$

and $\boldsymbol{\gamma}$ through solving the efficient estimating equations

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}\{y_i, \mathbf{w}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*(\mathbf{x}_i \mid \mathbf{z}_i)\} \\ \mathbf{0} &= \sum_{i=1}^{n} \mathbf{S}^*_{\text{eff},\boldsymbol{\gamma}}\{y_i, \mathbf{w}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*(\mathbf{x}_i \mid \mathbf{z}_i)\}, \end{aligned} \tag{4.14}$$

where $\mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}, \mathbf{S}^*_{\text{eff},\boldsymbol{\gamma}}$ are the residuals of the orthogonal projections of the score vectors $\mathbf{S}^*_{\boldsymbol{\beta}}$, $\mathbf{S}^*_{\boldsymbol{\gamma}}$ onto the nuisance tangent space with respect to η_1 (Tsiatis (2006)). The construction of $\mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}, \mathbf{S}^*_{\text{eff},\boldsymbol{\gamma}}$ is essentially identical to that in the Nonlinear Poisson with a Nonparametric Component Model Section by viewing $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ as one parameter of interest, and is detailed later. We denote the estimators as $\widehat{\boldsymbol{\beta}}(g)$ and $\widehat{\boldsymbol{\gamma}}(g)$. Of course, the function g is unknown, and the construction of $\mathbf{S}^*_{\text{eff},\boldsymbol{\gamma}}$ also relies on g', the derivative of g, so we adopt a local linear kernel estimator of $g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0), g'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)$ by implementing

$$\mathbf{0} = \sum_{i=1}^{n} \mathbf{S}^*_{\text{eff},g}\{y_i, \mathbf{w}_i, \mathbf{z}_i, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \mathbf{x}, \boldsymbol{\beta}), \eta_1^*(\mathbf{x}_i \mid \mathbf{z}_i)\} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) \tag{4.15}$$

at $\mathbf{z}_0 = \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$ to estimate $\widehat{g}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = c_0$ and $g'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) = c_1$ where $K_h(\cdot)$ is defined as before, c_0, c_1 are explained below, and $\mathbf{S}^*_{\text{eff},g}$ is the efficient score function derives from the posited model of η_1, η_1^*. Specifically, The efficient score functions for $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and g, g' can be written as

$$\begin{aligned}\mathbf{S}^*_{\text{eff},\boldsymbol{\beta}}(y, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) &= \mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) - E^*\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}, \\ \mathbf{S}^*_{\text{eff},\boldsymbol{\gamma}}(y, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) &= \mathbf{S}^*_{\boldsymbol{\gamma}}(y, \mathbf{w}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) - E^*\{\mathbf{a}_{\boldsymbol{\gamma}}(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}, \\ \mathbf{S}^*_{\text{eff},g}(y, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) &= \mathbf{S}^*_{g}(y, \mathbf{w}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, g, \eta_1^*) - E^*\{\mathbf{a}_{g}(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}. \end{aligned}$$

where $\mathbf{S}^*_{\boldsymbol{\beta}}(\cdot)$ and $\mathbf{S}^*_{\boldsymbol{\gamma}}(\cdot)$ and $\mathbf{S}^*_{g}(\cdot)$ are the score vectors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively, i.e.,

$$\begin{aligned}\mathbf{S}^*_{\boldsymbol{\beta}}(y, \mathbf{w}, \mathbf{z}) &= \frac{\int \mathbf{S}^F_{\boldsymbol{\beta}}(y, \mathbf{x}, \mathbf{z}) p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})}{\int p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})} \\ \mathbf{S}^*_{\boldsymbol{\gamma}}(y, \mathbf{w}, \mathbf{z}) &= \frac{\int \mathbf{S}^F_{\boldsymbol{\gamma}}(y, \mathbf{x}, \mathbf{z}) p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})}{\int p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})} \\ \mathbf{S}^*_{g}(y, \mathbf{w}, \mathbf{z}) &= \frac{\int \mathbf{S}^F_{g}(y, \mathbf{x}, \mathbf{z}) p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})}{\int p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{w} \mid \mathbf{x}) \eta_1^*(\mathbf{x} \mid \mathbf{z}) d\mu(\mathbf{x})}. \end{aligned}$$

In the example $f\{x, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\} = g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2$, we have

$$\begin{aligned}
\mathbf{S}^F_{\boldsymbol{\beta}}(y, x, \mathbf{z}) &= \begin{pmatrix} x[y - \exp\{g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2\}] \\ \\ x^2[y - \exp\{g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2\}] \end{pmatrix} \\
\mathbf{S}^F_{\boldsymbol{\gamma}}(y, x, \mathbf{z}) &= (\mathbf{0}\ \mathbf{I}_p)\mathbf{z}\left[y - \exp\{g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2\}\right] g'(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}),
\end{aligned}$$

where g' is the first partial derivative of g. Also, at $\mathbf{z}_0$,

$$\mathbf{S}^F_g(y, x, \mathbf{z}) = \left[y - \exp\{c_0 + c_1(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z} - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) + \beta_1 x + \beta_2 x^2\}\right](1, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z} - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)^{\mathrm{T}}.$$

$E^*\{\mathbf{a}_\beta(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}$, $E^*\{\mathbf{a}_{\boldsymbol{\gamma}}(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}$ and $E^*\{\mathbf{a}_g(\mathbf{X}, \mathbf{z}) \mid (y, \mathbf{w}, \mathbf{z})\}$ are respectively the projections of the score vectors $\mathbf{S}^*_{\boldsymbol{\beta}}$, $\mathbf{S}^*_{\boldsymbol{\gamma}}$ and $\mathbf{S}^*_g$ onto the tangent space (Tsiatis, 2006), which satisfy

$$\begin{aligned}
E\{\mathbf{S}^*_{\boldsymbol{\beta}}(Y, \mathbf{W}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}\} &= E[E^*\{\mathbf{a}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{z}) \mid Y, \mathbf{W}, \mathbf{z}\} \mid \mathbf{x}, \mathbf{z}] \\
E\{\mathbf{S}^*_{\boldsymbol{\gamma}}(Y, \mathbf{W}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}\} &= E[E^*\{\mathbf{a}_{\boldsymbol{\gamma}}(\mathbf{X}, \mathbf{z}) \mid Y, \mathbf{W}, \mathbf{z}\} \mid \mathbf{x}, \mathbf{z}] \\
E\{\mathbf{S}^*_g(Y, \mathbf{W}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}\} &= E[E^*\{\mathbf{a}_g(\mathbf{X}, \mathbf{z}) \mid Y, \mathbf{W}, \mathbf{z}\} \mid \mathbf{x}, \mathbf{z}].
\end{aligned}$$

For $f\{x, \boldsymbol{\beta}, g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})\} = g(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}) + \beta_1 x + \beta_2 x^2$, we write the conditional expectations as

$$\begin{aligned}
& E\{\mathbf{S}^*_{\boldsymbol{\beta}}(Y, W, \mathbf{z}_i) \mid x = x_i, \mathbf{z}_i\} \\
= & \int \frac{\sum_{j=1}^m \mathbf{S}^F_{\boldsymbol{\beta}}(y, x_j, \mathbf{z}_i) p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)}{\sum_{j=1}^m p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)} p(y \mid x_i, \mathbf{z}_i) p(w \mid x_i) d\mu(y) d\mu(w), \\
& E\{\mathbf{S}^*_{\boldsymbol{\gamma}}(Y, W, \mathbf{z}_i) \mid x = x_i, \mathbf{z}_i\} \\
= & \int \frac{\sum_{j=1}^m \mathbf{S}^F_{\boldsymbol{\gamma}}(y, x_j, \mathbf{z}_i) p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)}{\sum_{j=1}^m p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)} p(y \mid x_i, \mathbf{z}_i) p(w \mid x_i) d\mu(y) d\mu(w), \\
& E\{\mathbf{S}^*_g(Y, W, \mathbf{z}_i) \mid x = x_i, \mathbf{z}_i\} \\
= & \int \frac{\sum_{j=1}^m \mathbf{S}^F_g(y, x_j, \mathbf{z}_i) p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)}{\sum_{j=1}^m p(y \mid x_j, \mathbf{z}_i) p(w \mid x_j) \eta^*_1(x_j \mid \mathbf{z}_i)} p(y \mid x_i, \mathbf{z}_i) p(w \mid x_i) d\mu(y) d\mu(w).
\end{aligned}$$

Similarly, we follow the discretization technique that are described in the Nonlinear Poisson Model Section and the Nonlinear Poisson with a Nonparametric Component Model Section to solve for $\mathbf{a}_{\boldsymbol{\beta}}$, $\mathbf{a}_{\boldsymbol{\gamma}}$ and $\mathbf{a}_g$. We iteratively solve (4.15) at the $\mathbf{z}_0 = \mathbf{z}_i$ for $i = 1, 2, \cdots, n$ and (4.14) until convergence for $\widehat{g}$ and $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$.

We want to point out that when the Poisson mean is the exponential of a partially linear single index model, then a complete and sufficient "statistic" exists. In such case, we do not have to go through the approximation algorithm or discretization technique in order to implement the efficient score functions. Instead, we can borrow the idea from the Linear Poisson Model with a Nonparametric Component Section and combine with the method we discuss in this section to solve for $\boldsymbol{\beta}$. This simplifies to a special case discussed in Liu et al. (2017).

4.3 Simulation Studies

In this section, we illustrate five scenarios which are discussed in the previous section via simulations. For each case, we report the Naive estimator which ignores the measurement error. Although it is unrealistic, we provide an oracle estimator where we have the true distribution of $X \mid \delta$ or $\mathbf{X} \mid \mathbf{Z}$ to serve as a benchmark. We provide two locally efficient estimators along with two regression calibration estimators for comparison.

Linear Poisson Model

We generated Y which is related to a single covariate X, $X \sim$ Normal $(0, 1)$, through a linear Poisson regression model. Specifically, $E(Y \mid x) = \exp(\alpha + \beta x)$. We further generated $W = X + U$ where U is independent of Y and X, and has a standard normal distribution. We set $\alpha = -0.4, \beta = 1.1$. We conducted 500 simulations, each with sample size $n = 150$. To conduct the semiparametric estimator, we posited two different functional forms for $E(X \mid \delta)$, one is δ and the other is $4\sin(\delta/20)$. We compared the performance of the semiparametric estimators with the naive Poisson regression which simply ignores the measurement error, as well as regression calibration approach in Table 4.1. From these results, we can see clearly that ignoring the measurement error results in severe bias in both α and β estimation.

Although regression calibration can correct such bias in the estimation of β, it does not perform well for the α estimation even when we calculated $E(X \mid W)$ under the correct distribution of $\mathbf{X}$. As can be seen from Table (4.1), the semiparametric method provides consistent and more efficient resultin estimating both α and β.

Linear Poisson Model with a Nonparametric Component

We generate Y, X and W as in Section (4.3) and set $\beta = 1.1$. In addition, we generate $Z \sim$ Uniform $(0, 1)$ and let $g(z) = -0.4\cos(3.2z)$. Such data generating process results in a complete and sufficient "statistic" to be $\delta(w, y; \beta) = w + 1.1y$ which is free of z. Since the conditional expectation of X given the complete and sufficient "statistic" remains unknown to us, we adopt a nonparametric kernel method to estimate the quantity $E(X \mid \delta)$ and the concomitant $\widehat{\beta}$ is served as a benchmark. We conducted 500 simulations, each with sample size $n = 150$. For a comparison purpose, we retain the posited forms of $\eta^*(\delta)$ for the locally efficient estimation procedure. Results are summarized in Table (4.2). It is not surprised to see that the naive estimator significantly departs from the true value and of course the coverage is as low as 17%. When the mean model involves a nonparametric component, regression calibration is not efficient in estimating the parameter in the linear term. It is not difficulty to observe that, regardless of what model we posited for $\eta(\delta)$, our semiparametric estimator is consistent and efficient.

Nonlinear Poisson Model

In this case, we generate the counting measure Y related to unobservable X with a polynomial form, $E(Y \mid x) = \exp(\beta_0 + \beta_1 x + \beta_2 x^2)$. We utilize the same data generating process as stated in the linear Poisson case for X while allowing a substantial amount of error $\sigma_\epsilon = 0.65$. This results in observing $W \sim N(X, 0.65^2)$. In the estimating procedure, it is unavoidable to solve a double integration for a in

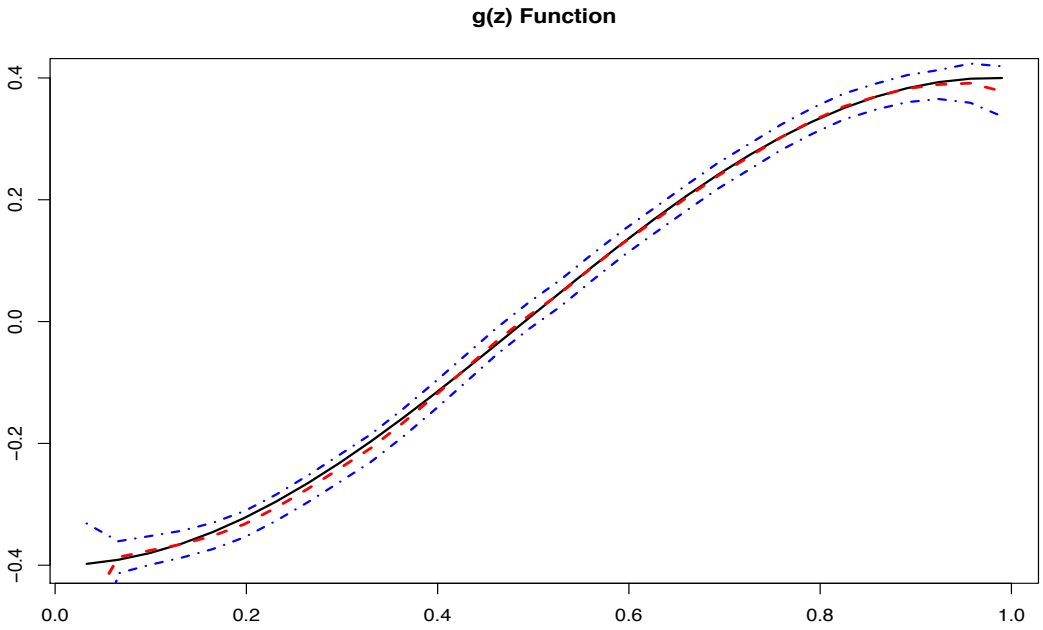

Figure 4.1: Nonparametric kernel estimation of $g(z)$ in linear Poisson model with a nonparametric component. $g(z) = -0.4\cos(3.2z)$

the efficient score function. We adopt Hermite Gaussian quadrature method due to the concern of accuracy. The true values of $(\beta_0, \beta_1, \beta_2)$ are set to be $(1.0, 0.7, -0.2)$. In order to implement the estimating equations, we need to know the probability distribution of X. To examine the robustness of the semiparametrc method, we consider two extremes where one correctly specified the distribution of X, $N(0, 1)$ while the other X is uniformly weighted in $[0, 1.5]$. We also provide a exponential weight in the range of $[-3, 3]$. Results for 500 simulations with each sample size 100 are summarized in Table (4.3). It is straightforward to see that the semiparametrc approach is much more capable to handle the measurement error in a nonlinear Poisson model.

Nonlinear Poisson with a Nonparametric Component Model

For nonlinear Poisson mean model with a nonparametric component, we generate the counting measure Y related to unobservable X and observable Z with a polynomial form, $E(Y \mid x, z) = \exp\{g(z) + \beta_1 x + \beta_2 x^2\}$. We set $g(z) = 0.3\{\sin(3.3z - 0.05) + 1.57\} - 1$. We utilize the same data generating process as the linear Poisson model with a Nonparametric component case for X, Z and allow a substantial amount of error $\sigma_\epsilon = 0.65$, i.e., $W \sim N(X, 0.65^2)$ and $Z \sim \text{Uniform}(0, 1)$. To compare performance, we set the true values of (β_1, β_2) to be $(0.7, -0.2)$. To implement (4.11) and (4.12) for $\boldsymbol{\beta}$ and g, we need the probability distribution of $X \mid Z$. Since X and Z are independently generated, we plug in the correctly specified distribution of X as well as two extreme weights, e.g., a $\text{Uniform}(0, 1.5)$ and an exponential distribution over the range of -3 to 3. In the estimating procedure, in addition to solve a double integration for a in the efficient score functions, we have to deal with the nonparametric term $g(z)$. To handle the unknown function $g(z)$, we adopt a standard profiling method in the estimation procedure and then estimate $g(z)$ and (β_1, β_2) iteratively. The entire estimating procedure requires high computation demand because we need to solve double integrations n times at n points and then iteratively estimate (β_1, β_2) using the updated $g(z)$. Results for 500 simulations with each sample size 150 are summarized in Table (4.4).

Nonlinear Partial Index Poisson Model

To mimic real world scenario, we inspect the numerical performance under the situation that the observable covariate $\mathbf{Z}$ contains both continuous and categorical data. We generate $Z_1 \sim \text{Binominal}(0.5)$, $Z_2 \sim \text{Uniform}(0, 0.5)$ and $Z_3 \sim \text{Normal}(0.3, 0.1^2)$. The true parameters inside the single index structure $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}$ is $(1.0, -0.9, -0.9)$. We set $g(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}) = -0.4\cos(2.75\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z} - 1.0)$, therefore $g\prime(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}) = 1.1\sin(2.75\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z} - 1.0)$. We generate $X \sim N(0, 1.1^2)$, set $\beta = 1.1$ and $\sigma_\epsilon = 0.8$. The Poisson mean becomes

$E(Y \mid x, \mathbf{z}) = 1.1x - 0.4\cos\{2.75(1.0z_1 - 0.9z_2 - 0.9z_3) - 1.0\}$. Similarly, we implement the "Oracle" estimator where the conditional expectation of X given δ is estimated nonparametrically. The two forms we posited for $\eta(\delta)$ are $\eta^*(\delta) = \delta^2$ and $\eta^*(\delta) = \delta\sin(\delta)$. Results are summarized in Table (4.5).

4.4 Empirical Applications

Cigarette Consumption and Mortality

The data is obtained from STATLIB-DASL (http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html), hosted by Carnegie Mellon University. It contains a measure of the number of cigarettes smoked per capita along with the death rates per thouusand population from lung and other cancers, for 43 states and the District of Columbia in 1960. We eliminated two cases where cigarette consumption are beyond than 40 for this analysis. The cigarette consumption is obviously an estimate quantity rather than an exact value. We fit a Poisson regression model with log-link relating the lung cancer rate (Y) to cigarette consumption (W), by allowing for measurement error in cigarette consumption. We predict the lung cancer death rate using a bias-correction linear Poisson model propose in this chapter, regression calibration, SIMEX and the modified estimating equation (MEE) method Buonaccorsi (2010). Comparison plots are summarize in Figure (4.2). We compare the mean square predicted error in Table (4.6).

Stroke Recovery in Underserved Populations

The Stroke Recovery in Underserved Populations study was conducted in 2005-2006 by National Institute on Aging (NIA). The survey fellowed 1216 patients from 11 rehabilitation facilities at the time they were admitted to and discharge from the rehabilitation facility, 80-189 days and 365-425 days after discharge. The study

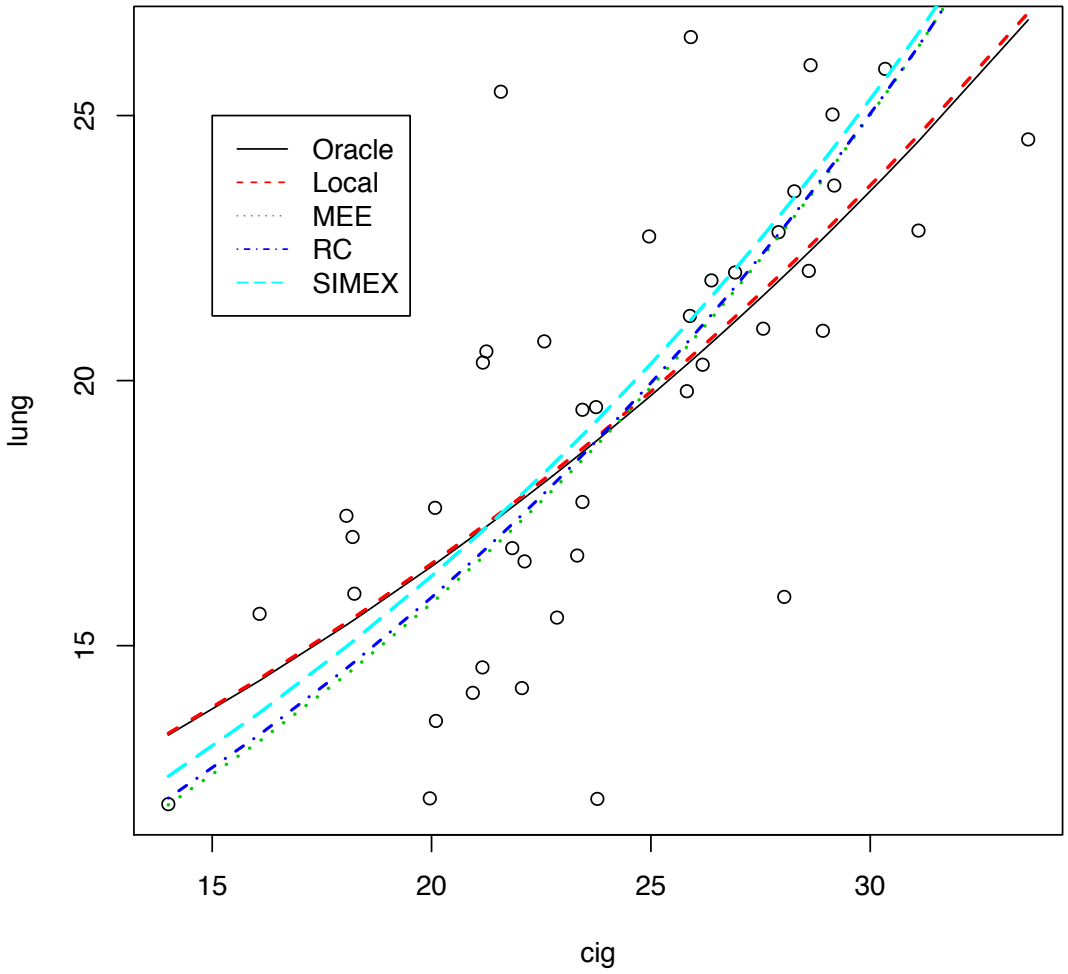

Figure 4.2: Cigeratte comsumption and lung cancer

aims to exam how positive emotion such as joy, gratitude, love, and social networks independently and interactively contribute to recovery of functional status after stroke within two underserved groups. Data was collected via face-to-face interview or phone interview. It contains 216 variables which include demographics information, stroke symptoms, functional recovery, emotional well-being, etc..

We fit a Poisson model with a quadratic term in the mean by taking the measurement error of PAIN level into account. Results are summarized in Table (4.7).

4.5 Discussion

We have explored a class of Poisson models where the major covariate of interest can not be accessed correctly. The models and methods we study in this chapter cover almost all possible situations in counting response data analyses. We have constructed a locally efficient semiparemetric estimator for a general class of Poisson models with measurement errors, in which there exists an infinite-dimensional nuisance function. Rather than taking the route of estimating $E(\mathbf{X} \mid \boldsymbol{\delta})$ or $f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x} \mid \mathbf{z})$ nonparatetrically, we infuse the idea of combining a parametric model estimator and a local kernel estimator via profiling. Our approaches do not require any model assumption on the unobservable covariate of interest. The resulting profiling-based estimator retains the consistence and semiparatric efficiency locally. Although we implemented through a profiling technique, backfitting should yield similar results as Liu et al. (2017) pointed out in their paper.

Table 4.1: Case 1:"Oracle" estimate $E(X \mid \delta)$ nonparametrically;"Local 1" used a posited $\eta^*(\delta) = \delta$; "Local 2" used a posited $\eta^*(\delta) = 4\sin(\frac{\delta}{20})$. RC Normal is regression calibration where calculate $E(\mathbf{X} \mid \mathbf{W})$ under a normal distribution. RC Uniform is regression calibration where calculate $E(\mathbf{X} \mid \mathbf{W})$ under a uniform distribution. The truth is $(\alpha, \beta) = (-0.4, 1.1)$

	Oracle	Local 1	Local 2	Naive	RC Normal	RC Uniform
$\widehat{\alpha}$	-0.3806	-0.3770	-0.3763	-0.0944	-0.1091	-0.1140
emp.sd	0.2169	0.2257	0.2294	0.1214	0.1304	0.1265
est.sd	0.1981	0.1992	0.2063	0.1214	0.1204	0.1235
95% CI	0.9420	0.9340	0.9460	0.2920	0.3400	0.3660
$\widehat{\beta}$	1.0999	1.1012	1.1011	0.5365	1.0730	1.1313
emp.sd	0.0951	0.1011	0.1037	0.0935	0.1871	0.1999
est.sd	0.0817	0.0835	0.0873	0.0758	0.1515	0.1787
95% CI	0.9520	0.9480	0.9660	0.0020	0.8560	0.9320

Table 4.2: Case 2:"Oracle" estimate $E(X \mid \delta)$ nonparametrically;"Local 1" used a posited $\eta^*(\delta) = \delta$; "Local 2" used a posited $\eta^*(\delta) = 4\sin(\frac{\delta}{20})$. The truth is $\beta = 1.1$

	Oracle	Local 1	Local 2	Naive	RC Normal	RC Uniform
$\widehat{\beta}$	1.1027	1.1014	1.1016	0.6309	1.2619	1.3313
emp.sd	0.0402	0.0337	0.0343	0.1598	0.0799	0.1727
est.sd	0.0366	0.0363	0.0366	0.1598	0.0717	0.0925
95% CI	0.9400	0.9620	0.9600	0.1680	0.4020	0.3160

Table 4.3: Case 3:"Oracle" used the true normal weight for X;"Local 1" used a uniform weight of X;"Local 2" used a exponential weight of X

	Oracle	Local 1	Local 2	Naive	RC Normal	RC Uniform
$\widehat{\beta}_0$	1.0137	1.0144	1.0131	1.0099	1.0669	1.1052
emp.sd	0.0797	0.0806	0.0812	0.0711	0.0823	0.1174
est.sd	0.0832	0.0838	0.0830	0.0683	0.0787	0.1147
95% CI	0.9600	0.9680	0.9640	0.9320	0.8600	0.8580
$\widehat{\beta}_1$	0.6935	0.6888	0.6936	0.4503	0.6406	0.6051
emp.sd	0.0959	0.0971	0.0974	0.0648	0.0922	0.0833
est.sd	0.1143	0.1131	0.1148	0.0618	0.0879	0.0813
95% CI	0.9740	0.9600	0.9700	0.0540	0.8800	0.7620
$\widehat{\beta}_2$	-0.2087	-0.2081	-0.2087	-0.0949	-0.1920	-0.2253
emp.sd	0.0728	0.0756	0.0747	0.0359	0.0727	0.1122
est.sd	0.0786	0.0791	0.0779	0.0329	0.0665	0.1114
95% CI	0.9540	0.9660	0.9620	0.1680	0.9140	0.9480

Table 4.4: Case 4:"Oracle" used the true normal weight for X;"Local 1" used a uniform weight of X;"Local 2" used a exponential weight of X.

	Oracle	Local 1	Local 2	Naive	RC Normal	RC Uniform
$\widehat{\beta}_1$	0.7250	0.7166	0.7051	0.6402	0.7775	0.6797
emp.sd	0.3229	0.3124	0.3067	0.2187	0.2540	0.2038
est.sd	0.2736	0.2615	0.2743	0.9751	0.2734	0.2171
95% CI	0.9400	0.9280	0.9560	0.9780	0.9440	0.9480
$\widehat{\beta}_2$	-0.2113	-0.2036	-0.2045	-0.1763	-0.2192	-0.1939
emp.sd	0.1834	0.1937	0.1764	0.0825	0.0767	0.0723
est.sd	0.1689	0.1606	0.1612	0.3198	0.1714	0.1631
95% CI	0.9760	0.9760	0.9840	0.9900	0.9980	0.9940

Table 4.5: Case 5:"Oracle" estimate $E(X \mid \delta)$ nonparametrically;"Local 1" used a posited $\eta^*(\delta) = \delta^2$; "Local 2" used a posited $\eta^*(\delta) = \delta \sin(\delta)$. The truth is $\beta = 1.1$. Dimensions of Z is 3.

	Oracle	Local 1	Local 2	Naive	RC Normal	RC Uniform
$\widehat{\beta}_0$	1.1042	1.1090	1.1019	0.8092	1.1622	1.2482
emp.sd	0.1599	0.1382	0.0929	0.1201	0.1110	0.1506
est.sd	0.1702	0.1683	0.1320	0.0848	0.1225	0.1579
95% CI	0.9520	0.9560	0.9580	0.3080	0.9280	0.8160

Table 4.6: Mean square prediction error on lung cancer death rate.

	Oracle	Local	MEE	RC	SIMEX
$MSPE$	27.4413	27.5851	33.8242	33.5601	33.3963

Table 4.7: Stroke data. PAIN_F2 is (PAIN_F)^2. LL is the 95% lower confidence limit, UL is the 95% upper confidence limit. Naive Poisson ignored the measurement error

		Estimate	Std.	p-value	LL	UL
Naive Poisson	(Intercept)	-1.3074	0.0710	0.0000	-1.3118	-1.3031
	PAIN_F	0.3126	0.0488	0.0000	0.3096	0.3156
	PAIN_F2	-0.0183	0.0061	0.0028	-0.0187	-0.0180
Exponential weight	(Intercept)	-1.3069	0.1005	0.0000	-1.3131	-1.3007
	PAIN_F	0.1318	0.0458	0.0037	0.1290	0.1346
	PAIN_F2	-0.0068	0.0028	0.0139	-0.0070	-0.0067
Uniform weight	(Intercept)	-1.3881	0.1124	0.0000	-1.3950	-1.3812
	PAIN_F	0.1101	0.0421	0.0093	0.1075	0.1126
	PAIN_F2	-0.0056	0.0027	0.0394	-0.0057	-0.0054

Chapter 5

Conclusion

In this dissertation, we have been focused on efficient semiparametric estimation and inference with applications on high dimension data problem and measurement error models. We propose flexible semiparametric mothod to model the propensity score function in the inverse probability weighted approach to evaluate causal effect. In the measurement error data framework, where the main covariate of interest can not be accessed correctly, we introduce a semiparametric bias-correction approach to estimate the effect of the covariate of interest in the presence of many other confounding covariate. We further extend our work to a class of Poisson measurement error models and provide bias-reduction solutions which yield locally efficient estimators.

Prior to this dissertation, numerous works have been done to improve the consistency in estimating the treatment effect, for example, Tan (2006), Tan (2010), van der Laan (2014b), Vermeulen and Vansteelandt (2015),

Vermeulen and Vansteelandt (2016) and among others. However, these methods either require model specification or intensive computation or a hybrid of both. The approach we propose does not reply on model specification of the propensity score or the outcome regression models. In the meanwhile, our method is more robust in estimation and very flexible to handle high dimensional covariate. We have provide rigors mathematical proofs and lots of numerical results to compare the performance with other famous approaches. Future work involves using the augmented inverse probability weighted (AIPW) estimator in estimating the average treatment effect. Under the causal inference framework, we also are interested in the average treatment

effect for the treated, which is defined as $E\{Y^*(1) - Y^*(0) \mid T = 1\}$.

Motivated by the Framingham Heart Study, we investigate the effect of a covariate of interest in the presence of possibly nonlinear confounding effects. In Chapter 3, we design a general methodology for the semiparametric measurement error model. We construct a class of locally efficient estimators which correct potential bias. We show that the semiparametric bias-correct estimator is root-n consistent, asymptotically normal and locally efficient. Through various simulation studies which account for increasing dimension of variable $\mathbf{Z}$, we demonstrated that our semiparametric methods result in much smaller biases, comparing to two regression calibration estimators and a naive estimator which neglect the error in X. The tools we developed for both estimation and inference is completely general, reflected in the generality of the main regression model. We illustrate the generality of the results via extensive simulation studies of a Poisson model for where $\mathbf{Z}$ has the dimension from 2 to 11. The locally semiparametric efficient estimator we propose is flexible, on the other hand, avoid the curse of dimensionality.

Finally, stemmed from the studies in general measurement error models, we investigate five interrelated Poisson models in Chapter 4. We gradually stretch the model flexibility from a linear Poisson mean model to a model that is nonlinear with a partial index structure in the mean of a counting measure. We integrate the results of Stefanski and Carroll (1987), Ma and Tsiatis (2006), Tsiatis and Ma (2004), Ma and Carroll (2006) and Liu et al. (2017) to provide a class of constructive locally efficient semiparametric estimators for a wide range of Poisson mean models with functional measurement errors. To the best of our knowledge, the estimation procedure developed in Chapter 4 is the first to give a locally efficient estimator without the specification in the probability density function $\eta(\mathbf{x})$ or the conditional covariate distribution $\eta(\mathbf{x} \mid \mathbf{z})$ for a Poisson model.

In conclusion, semiparametric methodology plays an important role in many sta-

tistical modeling in the real world. It is to our advantage to compare the flexibility in modeling and the efficiency in estimation of a class of semiparametric estimator in order to make a determination as to which model is most preferable for the type of data being analyzed.

Bibliography

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. Quarterly Journal of Economics, 120:1031–1083.

Apanasovich, T. V., Carroll, R. J., and Maity, A. (2009). Simex and standard error estimation in semiparametric measurement error models. Electron J Stat, **3**:318–348.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). Efficient and Adaptive Estimation for Semiparametric Models. The Johns Hopkins University Press, Baltimore, MD.

Bickel, P. J. and Ritov, Y. (1987). Efficient estimation in the errors in variables model. The Annals of Statistics, **15**:513–540.

Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods, and Applications. Chapman and Hall/CRC, London.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika, page asp033.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data.

Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(3):571–584.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. Journal of the American Statistical Association, **92**:477–489.

Carroll, R. J. and Hall, P. (1998). Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Association, **83**:1184–1186.

Carroll, R. J., Ruppert, D., A., S. L., and Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective. CRC Press, Boca Raton.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (1995). Measurement Error in Nonlinear Modeals. Chapman and Hall/CRC, London.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. Journal of Econometrics, 155:138–154.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. Biometrika, 92:399–418.

Chen, X., Hong, H., and Nekipelov, D. (2007). Measurement error models. Manuscript.

Chesher, A. (1991). The effect of measurement error. Biometrika, **78**:451–462.

Chesher, A., Dumangane, M., and Smith, R. J. (2002). Duration response measurement error. Journal of Econometrics, **111**:169–194.

Chesher, A. and Schluter, C. (2002). Welfare measurement and measurement error,. Review of Economic Studies, **69**:357–378.

Company, M. L. I. (1959). New weight standards for men and women. Statistical Bulletin of the Metropolitan Life Insurance Company.

Cook, D. R. (1998). Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York.

Cook, D. R. and Weisberg, S. (1991). Discussion of âĂIJsliced inverse regression for dimension reduction. Journal of the American Statistical Association, **86**:28–33.

Cui, X., Hardle, W., and Zhu, L. (2011). The efm approach for single-index models. Annals of Statistics, **39**:1658–1688.

De Luna, X., Waernbaum, I., and Richardson, T. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika, **98**:861–875.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order moments. Biometrics, **97**:279–294.

Dynan, K. E. (1996). Habit formation in consumer preferences: Evidence from panel data. Review of Economic Studies, **90**:391–406.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. Journal of the American Statistical Association, 83:414–425.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. Annals of Statistics, **19**:1257–1272.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

Fuller, W. A. (1987). Measurement Error Models. Wiley-Interscience, New Jersey.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica, 66:315–331.

Hardle, W., Liang, H., and Gao, J. (2000). Partially Linear Models. Physica-Verlag, Heidelberg.

Hardle, W., Werwatz, A., Müller, M., and Sperlich, S. (2004). Nonparametric and Semiparametric Models. Springer, New York.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. Journal of the Royal Statistical Society: Series B, **48**:244–8.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71:1161–1189.

Horowitz, J. L. and Markatou, M. (1996). Semiparametric estimation of regression models for panel data. Review of Economic Studies, **63**:145–168.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260):663–685.

Hsiao, C. (1989). Identification and estimation of dichotomous latent variables models using panel data. Review of Economic Studies, **58**:717–731.

Hu, Y. and Schennach, S. M. (2010). Instrumental variable treatment of nonclassical measurement error models. Econometrica, **76**:195–216.

Huang, Y. (2014). Corrected score with sizable covariate measurement error: Pathology and remedy. Statistica Sinica, **24**:357–374.

Hubert, H. B., Feinleib, M., McNamara, P. M., and Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the framingham heart study. Circulation, **67**:968–977.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science, pages 523–539.

Koenker, R. and Yoon, J. (2009). Parametric links for binary choice models: A fisherian-bayesian colloquy. Journal of Econometrics, 152:120–130.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3):337–346.

Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. The Annuals of Statistics, **37**:1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. Journal of the American Statistical Association, **102**:997–1008.

Li, D., Wang, X., Lin, L., and Dey, D. K. (2016). Flexible link functions in nonparametric binary regression with gaussian process priors. Biometrics, 72:707–719.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, **86**:316–327.

Li, L., Zhu, L., and Zhu, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. Journal of the Royal Statistical Society: Series B, **73**:59–80.

Liang, H., Hardle, W., and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. The Annals of Statistics, **27**:1519–1535.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. Genetic Epidemiology, 33:256–265.

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

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. Journal of the American Statistical Association, **101**.

Ma, Y. and Carroll, R. J. (2016). Semiparametric estimation in the secondary analysis of case–control studies. Journal of the Royal Statistical Society, Series B, 78:127–151.

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

Ma, Y. and Tsiatis, A. A. (2006). On closed form semiparametric estimators for measurement error models. Statistica Sinica, **16**:183–193.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. Journal of the American Statistical Association, **107**:168–1798.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. The Annuals of Statistics, **41**:250–268.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods, 9(4):403.

Newey, W. K. (1990). Semiparametric efficiency bounds. Journal of Applied Econometrics, **5**:99–135.

Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments: essay on principles, section 9. Statistical Scince, 5:465–480.

Parker, J. and Preston, B. (2005). Precautionary savings and consumption fluctuations. American Economic Review, **95**:1119–1144.

Petersen, M. L., Wang, Y., Van Der Laan, M. J., Guzman, D., Riley, E., and Bangsberg, D. R. (2007). Pillbox organizers are associated with improved adherence to hiv antiretroviral therapy and viral suppression: a marginal structural model analysis. Clinical Infectious Diseases, 45(7):908–915.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. Journal of the Royal Statistical Society, Series C, 29:15–23.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika, 66:403–411.

Ridgeway, G. and McCaffrey, D. F. (2007). Comment: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Sciences, 22:540–543.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the bickel and kwon article, âĂIJinference for semiparametric models: Some questions and an answer,âĂİ. Statistica Sinica, 11:920–936.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41–55.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. Biometrika, page ass013.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. Journal of educational Psychology, 66:688–701.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63:581–592.

Rubin, D. B. (1986). Which ifs have causal answers. Journal of the American Statistical Association, 81:961–962.

Rubin, D. B. and Little, R. A. (2002). Statistical analysis with missing data (2nd ed.). Wiley, New York.

Rubin, D. B. and van der Laan, M. J. (2007). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. The International Journal of Biostatistics, 4(1):Article–5.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. Biometrika, **74**:703–16.

Stefanski, L. A., Wu, Y., and White, K. (2013). Variable selection in nonparametric classification via measurement error model selection likelihoods. Journal of the American Statistical Association, **109**:574–589.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association, 101(476):1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika, 97(3):661–682.

Tsiatis, A. (2006). Semiparametric Theory and Missing Data. Springer, New York.

Tsiatis, A. A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. Biometrika, **91**:835–848.

van der Laan, M. J. (2014a). Targeted estimation of nuisance parameters to obtain valid statistical inference. The international journal of biostatistics, 10(1):29–57.

van der Laan, M. J. (2014b). Targeted estimation of nuisance parameters to obtain valid statistical inference. International Journal of Biostatistics, 10:29–57.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).

Van der Laan, M. J. and Rose, S. (2011). Targeted learning.

van der Laan Mark, J. and Daniel, R. (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1):1–40.

Van Keilegom, I. and Carroll, R. J. (2007). Backfitting versus profiling in general criterion functions. Statistica Sinica, **17**:797–816.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. Statistical methods in medical research, 21(1):7–30.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. Journal of the American Statistical Association, 110(511):1024–1036.

Vermeulen, K. and Vansteelandt, S. (2016). Data-adaptive bias-reduced doubly robust estimation. The international journal of biostatistics, 12(1):253–282.

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

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. Journal of clinical epidemiology, 63(8):826–833.

Xia, Y. C. (2007). A constructive approach to the estimation of dimension reduction directions. Annals of Statistics, 35:2654–2690.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, **97**:1042–1054.

Zhang, J., Wang, X., Yu, Y., and Gai, Y. (2014). Estimation and variable selection in partial linear single index models with error-prone linear covariates. A Journal of Theoretical and Applied Statistics, **48**:1048–1070.

Appendix A

Supplement to Chapter 2

A.1 Derivation of the efficient score function

Taking derivative with respect to $\mathbf{B}$ of the logarithm of the probability density function function, it is easy to verify that the score function with respect to $\mathbf{B}$ is

$$\mathbf{S}_{\mathbf{B}}(T_i, \mathbf{x}_i, \mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') = \mathrm{vecl}\left(\mathbf{x}_i\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right).$$

The efficient score is the residual after projecting the score vector with respect to $\mathbf{B}$ onto the nuisance tangent space Λ (Tsiatis, 2006). The nuisance tangent space, denoted Λ, is the mean-squared closure of all nuisance tangent spaces of all parametric submodels. We can verify that

$$\Lambda = \left(\left[T - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X})\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X})\}}\right]\mathbf{a}(\mathbf{B}^{\mathrm{T}}\mathbf{X}):\ \forall \mathbf{a}(\mathbf{B}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)\times d}\right)$$

We then obtain its orthogonal complement

$$\begin{aligned}\Lambda^{\perp} \;=\; & \Bigg[\mathbf{f}(Y,\mathbf{X}):\ \forall \mathbf{f} \in \mathcal{R}^{(p-d)d}\ s.t.\ E\{\mathbf{f}(1,\mathbf{X}) \mid T=1, \mathbf{B}^{\mathrm{T}}\mathbf{X}\}\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X})\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X})\}} \\ & = F\{\mathbf{f}(0,\mathbf{X}) \mid T=0, \mathbf{B}^{\mathrm{T}}\mathbf{X}\}\Bigg].\end{aligned}$$

We now write

$$\begin{aligned} & \mathbf{S}_{\mathbf{B}}(T_i, \mathbf{x}_i, \mathbf{B}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') \\ = \;& \mathrm{vecl}\left(\mathbf{x}_i\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right) \\ = \;& \mathrm{vecl}\left(E(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right) \\ & +\mathrm{vecl}\left(\mathbf{x} - E(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right).\end{aligned}$$

We can readily verify that

$$\operatorname{vecl}\left(E(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right) \in \Lambda$$

and

$$\operatorname{vecl}\left(\mathbf{x} - E(\mathbf{X} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\left[T_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right) \in \Lambda^{\perp},$$

hence this yields the desired result.

A.2 Proof of Theorem 1

From (2.9), we write

$$\begin{aligned}
& n^{1/2}(\widehat{\tau}-\tau) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_iT_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau\right\} \\
& + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_iT_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{Y_iT_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\widehat{\pi}(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)}\right\} \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_iT_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau\right\} \\
& + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}\{\pi(\mathbf{X}_i)-\widehat{\pi}(\mathbf{X}_i)\} - \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\{\widehat{\pi}(\mathbf{X}_i)-\pi(\mathbf{X}_i)\}\right] \\
& + O_p\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\widehat{\pi}(\mathbf{X}_i)-\pi(\mathbf{X}_i)\}^2\right] \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_iT_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau\right\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right] \\
& \times\{\widehat{\pi}(\mathbf{X}_i)-\pi(\mathbf{X}_i)\} + O_p(n^{1/2}h^{2m} + n^{-1/2}h^{-d}).
\end{aligned}$$

Now

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\{\widehat{\pi}(\mathbf{X}_i)-\pi(\mathbf{X}_i)\} \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left[\frac{\exp\{\widehat{\eta}(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right] \\
= \;& T_1 + T_2 + T_2,
\end{aligned}$$

where

$$
\begin{aligned}
T_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left[\frac{\exp\{\eta(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right],\\
T_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left[\frac{\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right],\\
T_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left[\frac{\exp\{\widehat{\eta}(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\\
&\quad -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\\
&\quad \times\left[\frac{\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right].
\end{aligned}
$$

It is easy to see that

$$
\begin{aligned}
& T_3\\
=\ & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\\
& \times\frac{\partial}{\partial\mathbf{B}}\left[\frac{\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\Bigg|_{\mathbf{B}=\mathbf{B}^*}(\widehat{\mathbf{B}}-\mathbf{B})\\
=\ & \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\Big(\frac{\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\widehat{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)}{[1+\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}]^2}\\
& -\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\eta'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)}{[1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}]^2}\Big)^{\mathrm{T}}\Bigg|_{\mathbf{B}=\mathbf{B}^*}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\sqrt{n}\mathrm{vecl}(\widehat{\mathbf{B}}-\mathbf{B})\\
=\ & o_p(1),
\end{aligned}
$$

where the last equality is because $\sqrt{n}\mathrm{vecl}(\widehat{\mathbf{B}}-\mathbf{B})=O_p(1)$ based on Lemma 2, and because of the consistency of $\widehat{\eta},\widehat{\eta}'$ established in Lemma 2.

It is also easy to see that

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\eta'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}}{[1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}]^2}\\
&\quad\otimes\mathbf{X}_{iL}^{\mathrm{T}}\sqrt{n}\mathrm{vecl}(\widehat{\mathbf{B}}-\mathbf{B})+o_p(1)\\
&= E\left(\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\pi(\mathbf{X}_i)\{1-\pi(\mathbf{X}_i)\}\eta'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\right)\\
&\quad\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1)\\
&= E\left(\left[Y_i^*(1)\{1-\pi(\mathbf{X}_i)\}+Y_i^*(0)\pi(\mathbf{X}_i)\right]\eta'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\right)\\
&\quad\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1)\\
&= \mathbf{a}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1),
\end{aligned}$$

where $Y_i^*(1)$ and $Y_i^*(0)$ are potential outcomes under treatment and no treatment respectively, and we used the independence assumption between potential outcomes and treatment in the second last equality.

We now analyze T_2. To this end, with the same notation as in the proof of Lemma 2 in the online supplement,

$$\begin{aligned}
T_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left[\frac{\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\left\{\widehat{H}(\mathbf{t}_i)-H(\mathbf{t}_i)\right\}.
\end{aligned}$$

Here again we used the independence assumption in the last equality. We consider $\widehat{H}(\mathbf{t}_i)$ as the direct kernel estimator of $H(\mathbf{t}_i)$, i.e. $\widehat{H}(\mathbf{t}_i) = \{\sum_{j=1}^{n}K_h(\mathbf{t}_j -$

$\mathbf{t}_i)Y_j\}/\{\sum_{j=1}^n K_h(\mathbf{t}_j - \mathbf{t}_i)\}$. We further obtain

$$
\begin{aligned}
T_2 &= \frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{n^{-1}\sum_{k=1}^n K_h(\mathbf{t}_k-\mathbf{t}_i)}-H(\mathbf{t}_i)\right\}\\
&= \frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{f(\mathbf{t}_i)}\right.\\
&\quad \left.\left\{1-\frac{n^{-1}\sum_{k=1}^n K_h(\mathbf{t}_k-\mathbf{t}_i)-f(\mathbf{t}_i)}{f(\mathbf{t}_i)}\right\}-H(\mathbf{t}_i)\right]+O_p(n^{1/2}h^{2m}+n^{-1/2}h^{-d})\\
&= \frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{f(\mathbf{t}_i)}-H(\mathbf{t}_i)\right\}\\
&\quad -\frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}H(\mathbf{t}_i)\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)-f(\mathbf{t}_i)}{f(\mathbf{t}_i)}\right\}+o_p(1)\\
&= \frac{1}{n^{3/2}}\sum_{i=1}^n\sum_{j=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]+o_p(1)\\
&= n^{-1/2}\sum_{i=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}E\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\mid\mathbf{t}_i,T_i\right]\\
&\quad +n^{-1/2}\sum_{j=1}^n E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]\mid\mathbf{t}_j,Y_j\right)\\
&\quad -n^{1/2}E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]\right)+o_p(1)\\
&= n^{-1/2}\sum_{j=1}^n E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]\mid\mathbf{t}_j,Y_j\right)\\
&\quad +o_p(1)\\
&= n^{-1/2}\sum_{i=1}^n\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\{T_i-H(\mathbf{t}_i)\}+o_p(1)\\
&= n^{-1/2}\sum_{i=1}^n\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i-\pi(\mathbf{X}_i)\}+o_p(1).
\end{aligned}
$$

Combining the above results regarding T_1, T_2 and T_3, we obtain

$$
\begin{aligned}
&n^{1/2}(\widehat{\tau}-\tau)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\left\{\frac{Y_iT_i}{\pi(\mathbf{X}_i)}-\frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)}-\tau\right\}-\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i-\pi(\mathbf{X}_i)\}\right]\\
&\quad -\mathbf{a}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^n E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1). \qquad \text{(A.1)}
\end{aligned}
$$

Comparing with the results in Hirano et al. (2003), it is now clear that the component in (A.1) is the efficient influence function, while the remaining component in the

expansion of $n^{1/2}(\widehat{\tau}-\tau)$ is the difference between the influence functions of our estimator and the efficient estimator, hence is orthogonal to the efficient influence function. In fact the orthogonality is also easily checked by direct calculation. □

A.3 Statement of Lemma 3

Lemma 3. *Assume the treatment allocation is independent of the potential treatment outcome given the covariates. Assume further that the probability of treatment is bounded away from 0 and 1. Assume a parametric model* $\pi(\boldsymbol{X}_i, \boldsymbol{\gamma})$ *with true parameter value* $\boldsymbol{\gamma}_0$. *Then when* $n \to \infty$, *the estimator* $\widehat{\tau}$ *from (2.9) satisfies* $\sqrt{n}(\widehat{\tau}-\tau) \to N(0,\sigma^2)$, *where* $\sigma^2 = \sigma^2_{\mathit{eff}} + E(B_i^2)$ *where* σ^2_{eff} *is the same as in Theorem 1, and* $B_i = \left\{\frac{Y_i^*(1)}{\pi(\boldsymbol{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\boldsymbol{X}_i)}\right\}\{T_i - \pi(\boldsymbol{X}_i)\} - E\left(\left[\frac{Y_i^*(1)}{\pi(\boldsymbol{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\boldsymbol{X}_i)}\right]\frac{\partial\pi(\boldsymbol{X}_i,\gamma)}{\partial\gamma_0}\right)\phi(\boldsymbol{X}_i, T_i)$, *where* $\phi(\boldsymbol{X}_i, T_i)$ *is the influence function of* $\widehat{\boldsymbol{\gamma}}$.

A.4 Comparing average treatment effect estimators for nested propensity models

When η is a known function, the efficient score function for $\mathbf{B}$ is

$$\begin{aligned}
\widetilde{\mathbf{S}}_{\text{eff}}(y_i, \mathbf{x}_i, \mathbf{B}^{\mathrm{T}}\mathbf{x}_i) &= \operatorname{vecl}\left(\mathbf{x}_i\left[y_i - \frac{\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right) \\
&= \operatorname{vecl}\left[\mathbf{x}_i\left\{y_i - \pi(\mathbf{X}_i)\right\}\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right] \\
&= \left\{y_i - \pi(\mathbf{X}_i)\right\}\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{x}_i) \otimes \mathbf{x}_{iL},
\end{aligned}$$

and the efficient influence function is $E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}}$. Using the results in Lemma 3, we have

$$\begin{aligned}
B_i &= \left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i - \pi(\mathbf{X}_i)\} \\
&\quad -E\left(\left[\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right]\pi(\mathbf{X}_i)\{1-\pi(\mathbf{X}_i)\}\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{X}_{iL}^{\mathrm{T}}\right) \\
&\quad \times E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}} \\
&= \left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i - \pi(\mathbf{X}_i)\} \\
&\quad -E\left(\left[Y_i^*(1)\{1-\pi(\mathbf{X}_i)\} + Y_i^*(0)\pi(\mathbf{X}_i)\right]\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{X}_{iL}^{\mathrm{T}}\right) E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}} \\
&= \left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i - \pi(\mathbf{X}_i)\} - \mathbf{a}^{\mathrm{T}}E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \\
&\quad \times \{T_i - \pi(\mathbf{X}_i)\},
\end{aligned}$$

Now let

$$\begin{aligned}
C_i &\equiv B_i + \mathbf{a}^{\mathrm{T}}E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\text{eff}}(\mathbf{X}_i, T_i) \\
&= \left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i - \pi(\mathbf{X}_i)\} - \mathbf{a}^{\mathrm{T}}E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \\
&\quad \times \{T_i - \pi(\mathbf{X}_i)\} + \mathbf{a}^{\mathrm{T}}E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\} \\
&\quad \times \{T_i - \pi(\mathbf{X}_i)\} \\
&= \left[\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\} - \mathbf{a}^{\mathrm{T}}E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\}\right. \\
&\quad \left. +\mathbf{a}^{\mathrm{T}}E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\right]\{T_i - \pi(\mathbf{X}_i)\}.
\end{aligned}$$

Now, following the previous notation to let $\mathbf{t}_i = \mathbf{B}^{\mathrm{T}}\mathbf{X}_i$, and $H(\mathbf{t}_i) = \pi(\mathbf{X}_i)$,

$$
\begin{aligned}
& E\{C_i \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i, T_i)\} \\
= \; & E\left(\left[\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\} - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\mathrm{eff}}\widetilde{\mathbf{S}}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\}\right. \right. \\
& \left. + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\right] \{T_i - \pi(\mathbf{X}_i)\}^2 \\
& \left. \times \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}\right) \\
= \; & E\left(\left[\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)} + \frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\} - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\mathrm{eff}}\widetilde{\mathbf{S}}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\{\boldsymbol{\eta}'(\mathbf{t}_i) \otimes \mathbf{X}_{iL}\}\right. \right. \\
& \left. + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\}\right] H(\mathbf{t}_i)\{1 - H(\mathbf{t}_i)\} \\
& \left. \times \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\}\right) \\
= \; & E\left(\left[\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)} + \frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\} - \mathbf{a}^{\mathrm{T}}\{E(\widetilde{\mathbf{S}}_{\mathrm{eff}}\widetilde{\mathbf{S}}_{\mathrm{eff}}^{\mathrm{T}})^{-1} - {}^{\mathrm{T}}\, E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\}\boldsymbol{\eta}'(\mathbf{t}_i)\right. \right. \\
& \left. \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\}\right] \times H(\mathbf{t}_i)\{1 - H(\mathbf{t}_i)\}\, \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\boldsymbol{\eta}'(\mathbf{t}_i) \\
& \left. \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\}\right),
\end{aligned}
$$

which is not necessarily zero. Thus, there is no definitive relation we can say even when the parametric model is a submodel of the dimension reduction model.

Appendix B

Supplement to Chapter 3

B.1 Calculation of the Projection of $(\mathbf{S}_\beta^{\mathrm{T}}, \mathbf{S}_\gamma^{\mathrm{T}}, \mathbf{S}_\alpha^{\mathrm{T}})^{\mathrm{T}}$

Replacing $\theta(\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z})$ with $\nu(\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}; \boldsymbol{\alpha})$, the conditional model of Y on $(\mathbf{X}, \mathbf{S}, \mathbf{Z})$ in (3.1) is a fully parametric model. Following Tsiatis and Ma (2004), we know that the nuisance tangent space Λ and its orthogonal complement $\Lambda^\perp$ are respectively

$$\begin{aligned} \Lambda &= [E\{\mathbf{f}(X, \mathbf{s}, \mathbf{z}) \mid w, \mathbf{s}, \mathbf{z}, y\} : E\{\mathbf{f}(X, \mathbf{s}, \mathbf{z}) \mid \mathbf{s}, \mathbf{z}\} = \mathbf{0}], \\ \Lambda^\perp &= [\mathbf{f}(w, \mathbf{s}, \mathbf{z}, y) : E\{\mathbf{f}(W, \mathbf{s}, \mathbf{z}, Y) \mid x, \mathbf{s}, \mathbf{z}\} = \mathbf{0}]. \end{aligned}$$

We can then easily verify from the definition of $\mathbf{L}_\beta, \mathbf{L}_\gamma, \mathbf{L}_\alpha$ that

$$\begin{pmatrix} \mathbf{L}_\beta^{\mathrm{T}}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, f_{X|\mathbf{S},\mathbf{Z}}) \\ \mathbf{L}_\gamma^{\mathrm{T}}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, f_{X|\mathbf{S},\mathbf{Z}}) \\ \mathbf{L}_\alpha^{\mathrm{T}}(w, \mathbf{s}, \mathbf{z}, y; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, f_{X|\mathbf{S},\mathbf{Z}}) \end{pmatrix}$$ is an element of $\Lambda^\perp$ and

$$\begin{pmatrix} E\{\mathbf{a}_\beta^{\mathrm{T}}(X, \mathbf{s}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \mid w, \mathbf{s}, \mathbf{z}, y\} \\ E\{\mathbf{a}_\gamma^{\mathrm{T}}(X, \mathbf{s}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \mid w, \mathbf{s}, \mathbf{z}, y\} \\ E\{\mathbf{a}_\alpha^{\mathrm{T}}(X, \mathbf{s}, \mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \mid w, \mathbf{s}, \mathbf{z}, y\} \end{pmatrix}$$ is an element of Λ. Equivalently, the projection of $(\mathbf{S}_\beta^{\mathrm{T}}, \mathbf{S}_\gamma^{\mathrm{T}}, \mathbf{S}_\alpha^{\mathrm{T}})^{\mathrm{T}}$ is indeed $(\mathbf{L}_\beta^{\mathrm{T}}, \mathbf{L}_\gamma^{\mathrm{T}}, \mathbf{L}_\alpha^{\mathrm{T}})^{\mathrm{T}}$.

B.2 List of Regularity Conditions

1. The function $\theta(\cdot)$ is twice differentiable and its second derivative is Lipschitz-continuous.

2. The density function of $\mathbf{Z}$ has a compact support and is positive on the support.

3. The matrix $\mathbf{A}$ and $\mathbf{B}$ defined in (3.7) and (3.8) are non-singular and their elements are bounded away from infinity.

4. The kernel function $K(\cdot)$ has compact support, is bounded on its support, and satisfies $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx > 0$.

5. The bandwidth $h = O(n^{-r})$ for $1/8 < r < 1/2$.

Condition 1 is a standard smoothness requirement on $\theta(\cdot)$ required for general nonparametric smoothing methods. Condition 2 requires the distribution of $\mathbf{Z}$ to have some properties to avoid technical issues such as dividing by zero. This requirement can be slightly relaxed at the price of more tedious technical treatment. Condition 3 ensures that the estimators of the parameters do not degenerate. Condition 4 rquires the kernel function to be the usual compactly supported second order kenel. Condition 5 states the bandwidth reuirement and illustrates that the method does not require under smoothing.

B.3 Proof of Theorem 2

For notational simplicity, we define $\boldsymbol{\zeta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{L}_\zeta = (\mathbf{L}_\beta^{\mathrm{T}}, \mathbf{L}_\gamma^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{L}_{\zeta\zeta} = \partial \mathbf{L}_\zeta / \partial \boldsymbol{\zeta}^{\mathrm{T}}$, $\mathbf{L}_{\zeta\alpha} = \partial \mathbf{L}_\zeta / \partial \boldsymbol{\alpha}^{\mathrm{T}}$. Let $\boldsymbol{\theta}_\zeta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) = \{\boldsymbol{\theta}_\beta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) \quad \boldsymbol{\theta}_\gamma(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\}$. When solving for $\boldsymbol{\alpha}$ in (3.5), we have

$$\mathbf{0} \;=\; n^{-1/2}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_\alpha\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}$$

at any $\boldsymbol{\zeta}$, therefore

$$\begin{aligned}
\mathbf{0} \;=\;& n^{-1}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_{\alpha\zeta}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
&+n^{-1}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_{\alpha\alpha}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}\partial\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}^{\mathrm{T}} \\
&+n^{-1}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_{\alpha}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
&\times\left[\mathbf{0}_\beta^{\mathrm{T}} \;\; \frac{K'\{\mathbf{e}_1^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z}_0)/h\}(\mathbf{z}_i - \mathbf{z}_0)^{\mathrm{T}}}{K\{\mathbf{e}_1^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z}_0)/h\}h} \;\; \cdots \;\; \frac{K'\{\mathbf{e}_d^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z}_0)/h\}(\mathbf{z}_i - \mathbf{z}_0)^{\mathrm{T}}}{K\{\mathbf{e}_d^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z}_0)/h\}h}\right],
\end{aligned}$$

where $\mathbf{0}_\beta$ is a zero vector with the same length as $\boldsymbol{\beta}$. Note also that

$$\begin{aligned}
& \mathbf{L}_{\alpha\alpha}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
=\;& \mathbf{L}_{\theta\theta}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
& +\mathbf{L}_{\theta}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_{\alpha\alpha}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}.
\end{aligned}$$

Taking into account that $E\{\mathbf{L}_\theta(Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\zeta},\boldsymbol{\alpha}) \mid \mathbf{z}\} = \mathbf{0}$, this yields

$$
\begin{aligned}
& \boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}} \\
= \; & -\{\boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha})\boldsymbol{\theta}_\alpha(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha})\}^{-1} \left[E\{\mathbf{L}_{\theta\theta}(Y_i, W_i, \mathbf{S}_i, \mathbf{Z}_i) \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i = \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0\}\right]^{-1} \\
& \boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha}) E\{\mathbf{L}_{\alpha\zeta}(Y_i, W_i, \mathbf{S}_i, \mathbf{Z}_i) \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i = \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0\} + o_p(1) \\
= \; & -\{\boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha})\boldsymbol{\theta}_\alpha(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha})\}^{-1} \Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)^{-1} \\
& \boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha})\boldsymbol{\theta}_\alpha(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0, \boldsymbol{\alpha}) E\{\mathbf{L}_{\theta\zeta}(Y_i, W_i, \mathbf{S}_i, \mathbf{Z}_i) \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0\} + o_p(1) \\
= \; & \boldsymbol{\theta}_\zeta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) + o_p(1).
\end{aligned}
$$

Now we expand (3.4) and obtain

$$
\begin{aligned}
\mathbf{0} \; = \; & n^{-1/2}\sum_{i=1}^{n} \mathbf{L}_\zeta\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
& + n^{-1}\sum_{i=1}^{n} \Big[\mathbf{L}_{\zeta\zeta}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} + \mathbf{L}_{\zeta\theta}\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} \\
& \times \frac{\partial \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}^{\mathrm{T}}}\Big] n^{1/2}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(1) \\
= \; & n^{-1/2}\sum_{i=1}^{n} \mathbf{L}_\zeta\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} + \mathbf{A} n^{1/2}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + o_p(1) \\
= \; & \mathbf{A} n^{1/2}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) + n^{-1/2}\sum_{i=1}^{n} \mathbf{L}_\zeta\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \theta(\cdot)\} \qquad \text{(A.1)} \\
& + n^{-1/2}\sum_{i=1}^{n} [\mathbf{L}_\zeta\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\} - \mathbf{L}_\zeta\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \theta(\cdot)\}] + o_p(1).
\end{aligned}
$$

From (3.5), we also have

$$
\begin{aligned}
\mathbf{0} \; = \; & n^{-1/2}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_\alpha\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& + n^{-1/2}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\mathbf{L}_{\alpha\theta}[y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \theta\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \boldsymbol{\alpha}(\boldsymbol{\zeta})\}]\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i; \boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& \times \{\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta}) - \boldsymbol{\alpha}(\boldsymbol{\zeta})\} + o_p(1) \\
= \; & n^{-1/2}\sum_{i=1}^{n} K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0) < \mathbf{L}_\alpha\{y_i, w_i, \mathbf{s}_i, \mathbf{z}_i; \boldsymbol{\zeta}, \boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& + n^{1/2} E\left(\mathbf{L}_{\theta\theta}[Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}; \boldsymbol{\alpha}(\boldsymbol{\zeta})\}] \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0\right) \\
& \times \boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0; \boldsymbol{\alpha}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0; \boldsymbol{\alpha}(\boldsymbol{\zeta})\} f_{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_0)\{\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta}) - \boldsymbol{\alpha}(\boldsymbol{\zeta})\} + o_p(1),
\end{aligned}
$$

hence

$$
\begin{aligned}
& \boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}n^{1/2}\{\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})-\boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
= \; & -\left[E\left(\mathbf{L}_{\theta\theta}\{Y,W,\mathbf{S},\mathbf{Z};\boldsymbol{\zeta},\theta\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}\}\mid\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i\right)\right]^{-1}[\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}]^{-1} \\
& n^{-1/2}\sum_{j=1}^n f_{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{-1}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j-\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{L}_\alpha\{y_j,w_j,\mathbf{s}_j,\mathbf{z}_j;\boldsymbol{\zeta},\boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& +o_p(1) \\
= \; & -\Omega^{-1}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)f_{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}}^{-1}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)[\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}]^{-1}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& n^{-1/2}\sum_{j=1}^n K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j-\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{L}_\alpha\{y_j,w_j,\mathbf{s}_j,\mathbf{z}_j;\boldsymbol{\zeta},\boldsymbol{\alpha}(\boldsymbol{\zeta})\}+o_p(1).
\end{aligned}
$$

Incorporating the above, we have

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^n\left(\mathbf{L}_\zeta[y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}]-\mathbf{L}_\zeta\{y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\}\right) \\
= \; & n^{-1/2}\sum_{i=1}^n\mathbf{L}_{\zeta\theta}\{y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\{\widehat{\boldsymbol{\alpha}}(\boldsymbol{\zeta})-\boldsymbol{\alpha}(\boldsymbol{\zeta})\}+o_p(1) \\
= \; & -n^{-1/2}\sum_{i=1}^n\mathbf{L}_{\zeta\theta}\{y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\}\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{-1} \\
& \times[\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}]^{-1}f_{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{-1}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& \times\left[n^{-1}\sum_{j=1}^n K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j-\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{L}_\alpha\{y_j,w_j,\mathbf{s}_j,\mathbf{z}_j;\boldsymbol{\zeta},\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\right]+o_p(1) \\
= \; & -n^{-1/2}\sum_{j=1}^n n^{-1}\sum_{i=1}^n K_h(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j-\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\mathbf{L}_{\zeta\theta}\{y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)\}\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{-1} \\
& \times f_{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)^{-1}[\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}\boldsymbol{\theta}_\alpha\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\}]^{-1}\boldsymbol{\theta}_\alpha^{\mathrm{T}}\{\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i;\boldsymbol{\alpha}(\boldsymbol{\zeta})\} \\
& \times\mathbf{L}_\alpha\{y_j,w_j,\mathbf{s}_j,\mathbf{z}_j;\boldsymbol{\zeta},\boldsymbol{\alpha}(\boldsymbol{\zeta})\}+o_p(1) \\
= \; & -n^{-1/2}\sum_{j=1}^n E(\mathbf{L}_{\zeta\theta}\mid\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j)\Omega(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j)^{-1}\{\boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j)\boldsymbol{\theta}_\alpha(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j)\}^{-1} \\
& \times\boldsymbol{\theta}_\alpha^{\mathrm{T}}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_j;\boldsymbol{\alpha})\mathbf{L}_\alpha\{y_j,w_j,\mathbf{s}_j,\mathbf{z}_j;\boldsymbol{\zeta},\boldsymbol{\alpha}(\boldsymbol{\zeta})\}+o_p(1) \\
= \; & -n^{-1/2}\sum_{i=1}^n\mathbf{U}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{z}_i)L_\theta\{y_i,w_i,\mathbf{s}_i,\mathbf{z}_i;\boldsymbol{\zeta},\theta(\cdot)\}+o_p(1).
\end{aligned}
$$

Plugging the above into (A.1), we obtain the expansion in Theorem 2. The subsequent results in Theorem 2 are easy to obtain hence their proofs are omitted. □

B.4 Proof of Theorem 3

The asymptotic expansion in Theorem 2 indicates that $\mathbf{A}^{-1}\mathbf{S}_{\text{eff}}\{Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta(\cdot)\}$ is an influence function (Newey 1989), where

$$\mathbf{S}_{\text{eff}}\{Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta(\cdot)\} \equiv \mathbf{L}_{\zeta}\{Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta(\cdot)\} - \mathbf{U}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})L_{\theta}\{Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta(\cdot)\}.$$

To show the efficiency, we only need to show that when $f^{*}_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z}) = f_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z})$, $\mathbf{S}_{\text{eff}}$ is the residual of the orthogonal projection of $\mathbf{S}_{\zeta}$ onto the nuisance tangent space, denoted Λ. Following Tsiatis and Ma (2004), the nuisance tangent space with respect to $f_{X|\mathbf{S},\mathbf{Z}}(x, \mathbf{s}, \mathbf{z})$ is

$$\Lambda_f = [E\{\mathbf{a}(X, \mathbf{S}, \mathbf{Z}) \mid Y, W, \mathbf{S}, \mathbf{Z}\} : E(\mathbf{a}) = \mathbf{0}],$$

and $\mathbf{L}_{\zeta}$ is the orthogonal projection of $\mathbf{S}_{\zeta}$ onto $\Lambda_f^{\perp}$, the orthogonal complement of Λ_f. Taking derivative of $l^{*}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, y, w, \mathbf{s}, \mathbf{z})$ with respect to $\boldsymbol{\alpha}$ and considering all possible $\boldsymbol{\alpha}$, we obtain the nuisance tangent space with respect to $\theta(\cdot)$ as

$$\Lambda_{\theta} = \{S_{\theta}(Y, W, \mathbf{S}, \mathbf{Z})\mathbf{a}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\}.$$

Thus, the nuisance tangent space is $\Lambda = \Lambda_f + \Lambda_{\theta}$. Defining

$$\widetilde{\Lambda}_{\theta} = \{L_{\theta}(Y, W, \mathbf{S}, \mathbf{Z})\mathbf{a}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\} = \left([S_{\theta}(Y, W, \mathbf{S}, \mathbf{Z}) - E\{a_{\theta}(X, \mathbf{S}, \mathbf{Z}) \mid Y, W, \mathbf{S}, \mathbf{Z}\}]\mathbf{a}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})\right),$$

where a_{θ} satisfies

$$E\{S_{\theta}(Y, W, \mathbf{S}, \mathbf{Z}) \mid X, \mathbf{S}, \mathbf{Z}\} = E[E\{a_{\theta}(X, \mathbf{S}, \mathbf{Z}) \mid Y, W, \mathbf{S}, \mathbf{Z}\} \mid X, \mathbf{S}, \mathbf{Z}],$$

and

$$L_{\theta}(Y, W, \mathbf{S}, \mathbf{Z}) = S_{\theta}(Y, W, \mathbf{S}, \mathbf{Z}) - E\{a_{\theta}(X, \mathbf{S}, \mathbf{Z}) \mid Y, W, \mathbf{S}, \mathbf{Z}\},$$

then $\Lambda = \Lambda_f \oplus \widetilde{\Lambda}_{\theta}$. Subsequently, the orthogonal complement of Λ is

$$\Lambda^{\perp} = \{\mathbf{b}(Y, W, \mathbf{S}, \mathbf{Z}) : E(\mathbf{b} \mid \mathbf{X}, \mathbf{S}, \mathbf{Z}) = E(\mathbf{b}S_{\theta} \mid \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}) = \mathbf{0}\}.$$

It is easy to see that $\mathbf{U}(\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z})L_{\theta}\{Y, W, \mathbf{S}, \mathbf{Z}; \boldsymbol{\zeta}, \theta(\cdot)\} \in \widetilde{\Lambda}_{\theta} \cap \Lambda_f^{\perp}$. On the other hand, we already have $\mathbf{S}_{\text{eff}} \in \Lambda^{\perp} = \widetilde{\Lambda}_{\theta}^{\perp} \cap \Lambda_f^{\perp}$. Thus, $\mathbf{S}_{\text{eff}}$ is the orthogonal projection of $\mathbf{L}_{\zeta}$ on $\Lambda^{\perp}$, hence equivalently, the orthogonal projection of $\mathbf{S}_{\zeta}$ on $\Lambda^{\perp}$. This proves the efficiency result. □

B.5 Proof of Theorem 4

Following the results in Theorem 2, under H_0, $n^{1/2}(\mathbf{M}\widehat{\boldsymbol{\beta}} - \mathbf{c})$ follows a normal distribution with mean zero and variance-covariance matrix $\mathbf{M}(\mathbf{I}_{\beta}, \mathbf{0})\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1^{\mathrm{T}}}(\mathbf{I}_{\beta}, \mathbf{0})^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}$ asymptotically. Consequently, T given in Theorem 4 has an asymptotic Chi-square distribution with d_M degrees of freedom. □