SEMIPARAMETRIC EFFICIENT ESTIMATORS IN PRIMARY AND SECONDARY ANALYSIS OF CASE-CONTROL STUDIES

A Thesis

by

LIANG LIANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond Carroll
Committee Members,	Jeffrey Hart
	Qi Li
	Mohsen Pourahmadi
Head of Department,	Valen Johnson

May 2017

Major Subject: Statistics

ABSTRACT

As a cost-efficient alternative to cohort design, case-control design is widely used in epidemiological studies. The primary analysis of the case-control studies focuses on the relationship between disease status and the potential risk factors, while the secondary analysis lies in analyzing the interrelationship between risk factors. The dissertation considers three semiparametric models arose in primary and ki secondary analysis of case-control studies and develops novel semiparametric estimators with great estimation efficiency.

We first investigate a special primary analysis problem, the gene-environment interaction model under independence assumption. While all existing approaches that exploit gene-environment independence assumption rely on a rare disease assumption or/and a distributional assumption on the genetic variable, we allow the disease rate and the distributions of the genetic and environmental variables in the underlying source population to be unknown. Under such a flexible semiparametric model, we derive the semiparametric efficient estimator and show that it outperformed the prospective logistic regression, the standard approach in primary analysis, through various numerical illustrations.

In the secondary conditional mean regression model, we analyze the interrelationship between covariates while only a conditional mean model is specified. Due to the unknown error distribution and the case-control nature of the data, semiparametric efficient estimation requires multivariate nonparametric regression on various quantities, which meets the curse of dimensionality as the dimension of covariates increases. We bypass this problem by devising a dimension reduction approach. The resulting estimator is robust against the misspecification of the regression error distribution and it shows great efficiency gain over several existing methods.

Lastly, we consider a secondary conditional quantile regression problem, which is a more preferable model in epidemiology when high or low values in the population are associated with high risks. Under a semiparametric framework that allows the covariates distribution to be nonparametric, we derive a class of consistent semiparametric estimators and spot the efficient member. The resulting estimator dominates the weighted estimating equation approach, the only published approach on secondary quantile regression, both theoretically and numerically.

DEDICATION

I dedicate this to my mother and father for their endless support.

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my advisors, Dr. Yanyuan Ma and Dr. Raymond J. Carroll for their invaluable guidance and persistent support, without which I would not have been able to complete this dissertation. Dr. Ma nourishes me with her patience, enthusiasm, optimism, and profound statistical knowledge. She has spent an enormous amount of time on teaching me how to do research, from little details like composing emails and polishing codes, to big pictures involving thinking independently and writing research papers. I cannot be more grateful for having her company during my doctoral training, especially when I got lost in my research. Dr. Carroll, as one of the most successful and experienced researchers in statistics, offers me generous criticisms and suggestions on both research and life, and helps me avoid making useless effort. I am especially thankful to him for providing unlimited opportunities of touching with new research directions and excellent statisticians.

I thank my committee members, Dr. Jeffery Hart and Dr. Mohsen Pourahmadi for their insightful advice and continuous help on this dissertation and defense, as well as their support on my career development; and Dr. Qi Li for serving on my committee.

My sincere thanks also goes to all faculty members here at Department of Statistics, Texas A&M University for their willingness to encourage and help me when needed. Particularly, I owe a great many thanks to Dr. Michael Longnecker for his endless support on teaching, career development, administrative issues and many more. I would also like to thank my colleagues, Alex Asher, Shahina Rahman, Ya Su, and Tianying Wang for valuable discussions and suggestions.

Lastly, I would like to thank my parents and boyfriend. They always respect my decisions and support me spiritually. Their love and encouragement push me to move forward.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Raymond Carroll [advisor], Jeffery Hart and Mohsen Pourahmadi of the Department of Statistics and Professor Qi Li of the Department of Economics. In addition, Professor Yanyuan Ma from the Department of Statistics, Pennsylvania State University also.

The data set analyzed in Section 2 was provided by Dr. Neal Freedman from the National Cancer Institute, while the data sets analyzed in Section 3 and 4 were provided by Professor Raymond Carroll.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a graduate assistantship from Texas A&M University and a grant from National Cancer Institute with Grant Numbers R37-CA057030 and U01-CA057030.

Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Texas A&M University or the National Cancer Institute.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 General Semiparametric Framework

Parametric regression, such as generalized linear models, attains great parsimony and efficiency but only works for well-specified model, while nonparametric regression maintains substantial model flexibility at the price of exploded computational cost and reduced efficiency. In contrast, semiparametric regression is an integration of parametric and nonparametric regression, where the model parameter can be split into a finite-dimensional parameter of interest, which represents simple but essential features that may be extracted from complex data sets, and an infinite-dimensional nuisance parameter, which denotes the intricate but unimportant details that can be neglected. It allows us to achieve a balance between estimation efficiency and model flexibility by modeling the finite-dimensional parameter parametrically and the infinite-dimensional nuisance parameter nonparametrically. A common semiparametric model is linear regression with unspecified error distribution, where the interest lies in estimating the regression coefficients with the error distribution being treated as a nuisance parameter. When implemented properly, semiparametric regression can be a prominent tool in solving complex problems arised from epidemiology, psycology, and various other scientific fields.

Suppose $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are independent, and identically distributed (i.i.d.) random variables taken from a collection of distributions $\{\mathcal{P}_\omega : \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, indexed by a parameter $\boldsymbol{\omega}$. In a general semiparametric model, $\boldsymbol{\omega}$ consists of a $p-$dimensional parameter $\boldsymbol{\theta}$ and an infinite-dimensional nuisance parameter $\boldsymbol{\eta}$, i.e., $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\eta})$. Semiparametric theories aim at finding consistent and ideally efficient estimator for $\boldsymbol{\theta}$. In this dissertation, we restrict ourselves to regular semiparametric estimators, namely regular asymptotically linear (RAL) estimators first introduced by Newey (1990). Each RAL estimator $\widehat{\boldsymbol{\theta}}$ can be uniquely identified by an influence function $\phi(\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\eta})$, i.e., a mean-zero p-dimensional random function with finite and positive-definite variance, through

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} \phi(\mathbf{X}_i) + o_p(1).$$

Subsequently, the asymptotic property of $\widehat{\boldsymbol{\theta}}$ can be uniquely characterized by its influence function. Specifically,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow \text{Normal}[\mathbf{0}, \text{cov}\{\phi(\mathbf{X})\}].$$

Hence, the problem of finding the efficient semiparametric estimator of $\boldsymbol{\theta}$ is equivalent to determing the influence function with smallest variance.

Bickel et al. (1993) and Tsiatis (2007) formulate the above semiparametric problem using a geometric approach, considering a Hilbert space $\mathcal{H}$ that consists of of all p-dimensional measurable functions with mean zero and finite variance and defining the inner product of two arbitrary functions in $\mathcal{H}$ to be their covariance. The Hilbert space $\mathcal{H}$ can be decomposed into a so-called nuisance tangent space Λ and its the orthogonal complement $\Lambda^{\perp}$. All valid influence functions must fall into the orthogonal space $\Lambda^{\perp}$ and the most efficient influence function can be obtained by projecting the score function, defined as the derivative of the log-likelihood with respect to $\boldsymbol{\theta}$, onto $\Lambda^{\perp}$. The score function, if scaled properly, is a valid influence function. Projecting the score function onto $\Lambda^{\perp}$ is conceptually a procedure of removing the greatest posssible variability due to the nuisance parameter $\boldsymbol{\eta}$. Although the general methodology of this geometric approach is standard and intuitive, the derivation of the nuisance tangent space Λ and the projection of score function onto the orthogonal space $\Lambda^{\perp}$ can be highly mathematically involved depending on the specific problems.

This dissertation is dedicated to develop semiparametric (locally) efficient estimators in primary and secondary analysis of case-control studies. Due to its special sampling scheme, case-control samples are not i.i.d. samples taken from the underlying source population. Thus the direct application of the above geometric approach is prohibited. To conquer this problem, we adopt the hypothetical population framework by Ma (2010). Such a hypothetical population has the same disease to non-disease ratio as the case-control sample, and it allows us to treat case-control samples as i.i.d. samples taken from this hypothetical population. The validity of the hypothetical population is discussed thoroughly in Ma (2010).

1.2 Primary and Secondary Analysis in Case-Control Studies

Case-control studies are popular tools in investigating risk factors associated with various uncommon diseases, such as cancer and myocardial infarction. Typically, a population-based case-control study employs a random sample of cases (diseased subjects) and a separate random sample of controls (non-diseased subjects). It also collects covariate information on the exposure of interest and other risk factors. As a result, the case-control samples are no longer representative samples of the underlying source populaiton. The primary task of case-control studies lies in understanding the relationship between disease status and covariates, usually via a prospective logistic regression analysis, which gives an efficient estimator of all parameters except the intercept, under the conditions that the disease rate is unknown and no parametric model for the predictors is available in the underlying source population (Prentice and Pyke, 1979).

In Section 2, we consider a special primary model, i.e., the gene-environment interaction model under gene-environment independence assumption. With the independence assumption on genetic and environmental variables, the prospective logistic regression is still consistent but no more efficient. All other existing approaches that exploit gene-environment independence assumption rely on a rare disease assumption or/and a distributional assumption on the genetic variables such as the genetic variable is discrete and takes finitely many values. Although the resulting estimators display appealing performances in terms of efficiency, they are not suitable for more general settings. We relax both assumptions. We construct a semiparametric estimator in case-control studies exploiting gene-environment independence, while the distributions of genetic susceptibility and environmental exposures are both unspecified and the disease rate is assumed unknown and is not required to be close to zero. The resulting estimator is semiparametric efficient and its superiority over prospective logistic regression, the usual analysis in case-control studies, is demonstrated in various numerical illustrations.

Recently, there has been considerable interest in using case-control data for a *secondary analysis*, namely examining the interrelationship between covariates, say Y and $\mathbf{X}$. As the case-control data is not a random sample from the underlying source population, which we refer to as *true population* throughout the paper, the relationship between covariates Y and $\mathbf{X}$ in the secondary analysis under the case-control context can be very different from the relationship in the true population. Hence, simply regressing Y on $\mathbf{X}$ and ignoring the case-control sampling scheme can be grossly misleading.

In Section 3, we examine the secondary analysis problem when multiple covariates are available, while only a regression mean model between $\mathbf{X}$ and Y is specified. Despite the completely parametric modeling of the regression mean function, the case-control nature of the data requires special treatment and semiparametric efficient estimation generates various nonparametric estimation problems with multivariate covariates. We devise a dimension reduction approach that fits with the specified primary and secondary models in the original problem setting, and use reweighting to adjust for the case-control nature of the data, even when the disease rate in the source population is unknown. The resulting estimator is both locally efficient and robust against the misspecification of the regression error distribution, which can be heteroscedastic as well as non-Gaussian. We demonstrate the advantage of our method over several existing methods, both analytically and numerically.

In Section 4, we consider the secondary analysis where the association between $\mathbf{X}$ and Y is specified by a conditional quantile regression model, becuase quantiel regression is often preferable in epidemiology, especially when the interest lies in studying high or low values of a population. We approach the secondary quantile regression problem from a semiparametric perspective, allowing the covariates distribution to be completely unspecified. The problem is identifiable excluding a few special cases. We derive a class of consistent semiparametric estimators and spot the efficient member. The implementation and the asymptotic properties of the resulting estimator are discussed in detail. Simulation results and a real data analysis are provide to show satistifactory performance.

Section 5 is a summary of this dissertation, in which we discussed the advantages and limitations of the proposed semiparametric approaches as well as the potential future works worth exploration.

2. SEMIPARAMETRIC EFFICIENT ESTIMATION IN GENE-ENVIRONMENT INTERACTION MODEL UNDER INDEPENDENCE ASSUMPTION

2.1 Introduction

The etiology of most complex diseases, such as cancers and cardiovascular diseases, is the joint effect of genetic susceptibility and environmental or non-genetic exposures, as well as their interactions. Even subtle differences in genetic factors between people, when exposed to the same environmental factors, can lead to dramatically different responses. One common example is that sunlight exposure results in higher risk of developing skin cancer among fair-skinned individuals than people with dark skin (Hunter, 2005; Ottman, 1996). Studying gene-environment interactions is thus of great importance to understand disease mechanisms and develop new treatments and prevention strategies.

The case-control study design is commonly used to investigate the intricate interplay of genetic susceptibility and environment effects. It is cost-efficient and convenient to implement compared to a cohort study, especially when dealing with relatively rare diseases (Chatterjee et al., 2009). Instead of taking a random sample from the underlying source population, the case-control design randomly draws a fixed number of cases (diseased subjects) and a comparable number of controls (non-diseased subjects) from the respective case and control subpopulations. Genetic and environmental factors are then measured and recorded for these sampled subjects in a retrospective fashion. The standard approach for the analysis of such a case-control study is prospective logistic regression, which ignores the underlying retrospective nature of the case-control design. Cornfield (1956) showed the equivalence of prospective and retrospective odds ratios, which validates the prospective approach. Prentice and Pyke (1979) further showed that prospective logistic regression analysis gives an efficient estimator, in the sense that it yields the maximum likelihood estimates of the odds ratio parameters under a semiparametric model that allows an *arbitrary* covariate distribution.

Despite of this, prospective logistic regression treatment in a case control study can still require a large sample size to obtain adequate statistical power for detecting gene-environment interactions or testing other hypotheses of interest. As a consequence, epidemiological researchers often exploit the potential efficiency gain from further assuming certain parametric or semiparametric structures for the covariate distribution. For example, in practice, a common assumption is that genetic susceptibility and environmental exposure are independent in the underlying source population (Piegorsch et al., 1994), possibly given strata. Under such a model, prospective logistic regression analysis is still valid but may not be efficient because it ignores gene-environment independence.

A growing number of articles have been published in the last two decades, proposing analytical methods that exploit gene-environment independence assumption (Chatterjee and Carroll, 2005; Gauderman et al., 2013; Han et al., 2015; Ma, 2010; Murcray et al., 2009; Piegorsch et al., 1994). Piegorsch et al. (1994) showed that under gene-environment independence and a rare disease assumption, the multiplicative interaction odds-ratio parameter can be estimated by cases alone and the resulting estimator is more precise than the estimator from traditional prospective logistic regression analysis using both cases and controls. However, the misuse of a rare disease assumption in analyzing diseases with moderate prevalence or diseases with small marginal probability in the source population but high risk for certain combination of genetic and environmental exposures can lead to considerable bias in the estimation. Noting this fact, Chatterjee and Carroll (2005) developed a semiparametric maximum likelihood estimator employing the gene-environment independence assumption but not requiring any rare-disease assumption. Their approach leaves the distribution of the environmental exposures totally unspecified but restricts genetic susceptibility to have a discrete distribution that takes values in a finite and fixed set. Ma (2010) proposed a semiparametric efficient estimator in the same setting as Chatterjee and Carroll (2005) except the distribution of genetic susceptibility is allowed to be either discrete or continuous with a finite-dimensional parameter. The key ingredient of this approach is to construct a hypothetical population with infinite population size and a disease to non-disease ratio of n_1/n_0, where n_1 and n_0 are the numbers of cases and controls in the case-control sample. Section 2 of Ma (2010) showed that the case-control sample can be viewed as a size $n = n_0 + n_1$ random sample of independent and identically distributed observations from this hypothetical population, and hence classical semiparametric analysis is applicable. The validity and usefulness of such a hypothetical population was established in Ma (2010).

In this section, we consider a more general setting which keeps the gene-environment independence assumption, while further allowing an unknown disease rate and completely nonparametric distributions for both the genetic susceptibility and the environmental exposure. Under such a model setting, we adopt the hypothetical population framework of Ma (2010) and derive the semiparametric efficient estimator by employing a semiparametric approach, which links the efficient estimator with the efficient score function. Throughout our work, the underlying source population is referred to as the true population to emphasize the difference between the underlying source population and the hypothetical population. The inherent connection between the two populations allows us to transport parameter estimation and inference results derived in the hypothetical population directly to those in the true population, see Theorem 1. Although general semiparametric theory applies in the hypothetical population framework, computing the efficient estimator in this context is technically challenging because the efficient score does not have an explicit form and must be solved from an integral equation. We adopt a simple numerical approach to solve the integral equation by discretizing the distribution of the genetic susceptibility when it is continuous. The resulting estimator, when properly implemented, is asymptotically linear with optimal efficiency.

The rest of the section is organized as follows. The specific model and the hypothetical population framework are presented in Section 2.2, with the corresponding identifiability conditions provided in Appendix A.1. In Section 2.3, we formulate the problem by using a conventional semiparametric approach. The analytic expression of our semiparametric efficient estimator as well as its detailed implementation are discussed in this section. Section 2.4 illustrates the asymptotic properties of the resulting estimator. Several simulation studies are conducted in Section 2.5 to demonstrate the numerical performance of our semiparametric efficient estimator compared with prospective logistic regression. A real data analysis is provided in Section 2.6, followed with a brief discussion in Section 2.7. Technical details and proofs are given in Appendix A.

2.2 Model And Framework

Let D denote the binary indicator of disease status, with $D = 0$ representing the absence and $D = 1$ the presence of a disease. Let the genetic susceptibility be G and the environmental exposures X. Assume that the prospective risk given the covariates (G, X) follows a logistic model

$$\begin{aligned} \text{pr}(D = d \mid G = g, X = x) &= f^{\text{true}}_{D|G,X}(d, g, x) = H(d, g, x, \boldsymbol{\theta}) \\ &= \frac{\exp[d\{\alpha + m(g, x, \boldsymbol{\beta})\}]}{1 + \exp\{\alpha + m(g, x, \boldsymbol{\beta})\}}, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \alpha)^{\mathrm{T}}$ and $m(\cdot)$ is a function known up to the parameter $\boldsymbol{\beta}$. Here and throughout the text, the superscript "true" is used to emphasize that those quantities are related to the true source population. In addition, in the true population, G and X are assumed to be independent so that the joint probability density/mass function of G, X can be written as $f^{\text{true}}_{G,X}(g, x) = f^{\text{true}}_{G}(g) f^{\text{true}}_{X}(x) = \eta_1(g)\eta_2(x)$. Here, for notational simplicity, we write $\{f_G^{\rm true}(g), f_X^{\rm true}(x)\}$ as $\{\eta_1(g), \eta_2(x)\}$. The problem stated above is identifiable in the case-control study under mild conditions, which are given in Appendix A.1, along with the proof of identifiability.

The hypothetical population/tilted study joint density/mass function of (D, G, X) is

$$\begin{aligned} f_{D,G,X}(d,g,x,\boldsymbol{\theta},\eta_1,\eta_2) &= (n_d/n)f_{G,X|D}(d,g,x) = (n_d/n)f_{G,X|D}^{\rm true}(d,g,x) \\ &= \frac{n_d}{n}\frac{f_G^{\rm true}(g)f_X^{\rm true}(x)f_{D|G,X}^{\rm true}(d,g,x,\boldsymbol{\theta})}{\int f_G^{\rm true}(g)f_X^{\rm true}(x)f_{D|G,X}^{\rm true}(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g)}. \\ &= \frac{n_d\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})}{n\int \eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g)} \\ &= \frac{n_d}{n\pi_d}\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta}), \end{aligned} \tag{2.2}$$

where

$$\pi_d = \int \eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g). \tag{2.3}$$

We consider $\eta(\cdot) = \{\eta_1(\cdot), \eta_2(\cdot)\}$ as the infinite-dimensional nuisance parameter. The approach of Ma (2010) views this as a semiparametric problem, to be solved using techniques explained in Bickel et al. (1993) and Tsiatis (2007). Here, the concept of hypothetical population and the corresponding tilted likelihood is used as a vehicle to allow us to transport the semiparametric tools for direct application. It enables us to construct consistent estimators without having to concern about the non-random sample issue in case-control study. Because the non-random sampling issue is already taken into account when we formulate the tilted likelihood, the resulting estimator is indeed automatically consistent under the original case-control sampling framework, that is, if the case-control sample size grows to infinity while retaining the relative sample proportion of n_1/n_0, the estimator will converge to the true parameter value. We formally write out this result in

Theorem 1.

<u>Theorem</u> 1. *Assume (d_i, g_i, x_i), $i = 1, \ldots, n$, is a case-control sample with n_1 cases, n_0 controls, and with disease model (2.1) and independence of X and G. Assume $\widetilde{d}_i, \widetilde{g}_i, \widetilde{x}_i$, $i = 1, \ldots, n$, is a random sample of iid observations with size n from model (2.2). Then, if $\widehat{\boldsymbol{\theta}}\{(\widetilde{d}_1, \widetilde{g}_1, \widetilde{x}_1), \ldots, (\widetilde{d}_n, \widetilde{g}_n, \widetilde{x}_n)\}$ is a root-n consistent regular asymptotically linear estimator of $\boldsymbol{\theta}$ and satisfies $E[\widehat{\boldsymbol{\theta}}\{(\widetilde{d}_1, \widetilde{g}_1, \widetilde{x}_1), \ldots, (\widetilde{d}_n, \widetilde{g}_n, \widetilde{x}_n)\} \mid D] - \boldsymbol{\theta} = o_p(n^{-1/2})$, then so is $\widehat{\boldsymbol{\theta}}\{(d_1, g_1, x_1), \ldots, (d_1, g_1, x_1)\}$.*

Theorem 1 essentially says that if we can develop a root-n consistent estimator based on a random sample from model (2.2), then we can simply apply this estimation procedure to the case-control sample and we will still get a root-n consistent estimator. The proof of the Theorem 1 is the entire content of Section 2 of Ma (2010).

We take advantage of this property to generate an estimation procedure, which we will then show consistently estimates the parameters when using the case-control data. In particular, the procedure is not dependent on the hypothetical population/tilted study formalism.

2.3 Analytic Derivations: Efficient Score and Algorithm

The outline of the semiparametric approach is to first construct a Hilbert space $\mathcal{H}$, consisting of all measurable functions with mean zero and finite variance. We next decompose $\mathcal{H}$ into nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$. The efficient estimator can then be obtained by solving $0 = \sum_{i=1}^{N} \mathbf{S}_{\text{eff}}(D_i, G_i, X_i; \boldsymbol{\theta})$, where $\mathbf{S}_{\text{eff}}$ is the projection of the score function $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$, and thus $\mathbf{S}_{\text{eff}}$ is called efficient score function.

Careful calculation shows that the score function under the hypothetical population (2.2) takes the form

$$\mathbf{S}_{\boldsymbol{\theta}}(d, g, x) = \mathbf{S}(d, g, x) - E(\mathbf{S} \mid d),$$

where $\mathbf{S} = \{d - H(1, g, x, \boldsymbol{\theta})\}\{\mathbf{m}'_{\boldsymbol{\beta}}(g, x, \boldsymbol{\beta})^{\mathrm{T}}, 1\}^{\mathrm{T}}$ and $\mathbf{m}'_{\boldsymbol{\beta}}(g, x, \boldsymbol{\theta}) \equiv \partial m(g, x, \boldsymbol{\theta})/\partial \boldsymbol{\beta}$. Let p denote the dimension of $\boldsymbol{\theta}$. The final form of the Λ and $\Lambda^{\perp}$ are listed below with the detailed derivation provided in Appendix A.2. Specifically,

$$\begin{aligned}
\Lambda &= [\mathbf{a}_1(G) + \mathbf{a}_2(X) - E\{\mathbf{a}_1(G) + \mathbf{a}_2(X) \mid D\} : \text{ for all } \mathbf{a}_1(G), \mathbf{a}_2(X)], \\
\Lambda^{\perp} &= [\mathbf{f}(D, G, X) : E(\mathbf{f} \mid G) = E\{E(\mathbf{f} \mid D) \mid G\}, \\
&\qquad E(\mathbf{f} \mid X) = E\{E(\mathbf{f} \mid D) \mid X\}, E(\mathbf{f}) = \mathbf{0}].
\end{aligned}$$

Define $\mathbf{S}_x(x) = E(\mathbf{S}_{\boldsymbol{\theta}} \mid x) = E(\mathbf{S} \mid x) - E\{E(\mathbf{S} \mid D) \mid x\}$ and $\mathbf{S}_g(g) = E(\mathbf{S}_{\boldsymbol{\theta}} \mid g) = E(\mathbf{S} \mid g) - E\{E(\mathbf{S} \mid D) \mid g\}$. Projecting the score function onto $\Lambda^{\perp}$ shows that

$$\mathbf{S}_{\text{eff}}(d, g, x) = \mathbf{S}(d, g, x) - \mathbf{a}(g) - \mathbf{b}(x) - E\{\mathbf{S}(d, G, X) \mid d\} + E\{\mathbf{a}(G) + \mathbf{b}(X) \mid d\},$$

where

$$\begin{aligned}
E\{\mathbf{a}(G) \mid x\} + \mathbf{b}(x) - E\{E(\mathbf{a} + \mathbf{b} \mid D) \mid x\} &= \mathbf{S}_x(x), && (2.4) \\
\mathbf{a}(g) + E\{\mathbf{b}(X) \mid g\} - E\{E(\mathbf{a} + \mathbf{b} \mid D) \mid g\} &= \mathbf{S}_g(g). && (2.5)
\end{aligned}$$

It is easy to check that $E\{\mathbf{S}_{\text{eff}}(d, G_i, X_i) \mid d\} = \mathbf{0}$.

In order to obtain the efficient score function, we need to solve $\mathbf{a}$ and $\mathbf{b}$ from the integral equations (2.4) and (2.5). The existence of the solution is automatically guaranteed by the identifiability of the problem, whereas the uniqueness is not. However, it is shown in Appendix A.3 that $\mathbf{a}$ and $\mathbf{b}$ are unique up to constant shifts. Thus, (2.4) and (2.5) have a unique solution under the constraints $E(\mathbf{a}) = E(\mathbf{b}) = \mathbf{0}$. It is further proved in Appendix A.4 that, under the mean zero constraint, (2.4) and (2.5) have an equivalent expression, which is given by equations (A.1), (A.2), and (A.3) in Appendix. Such an equivalent expression allows us to separate $\mathbf{a}$ and $\mathbf{b}$ by introducing an intermediate variable $\mathbf{u}_0 = E(\mathbf{a} + \mathbf{b} \mid D = 0)$. However, there is no explicit expression for $\mathbf{a}$ and $\mathbf{b}$. We still need to solve the integral equation (A.1). In Appendix A.5, we propose an approximation to its solution in the spirit of Tsiatis and Ma (2004), by discretizing X if X is continuous.

The detailed algorithm for constructing the efficient score function and computing the efficient estimator for $\boldsymbol{\theta}$ is given in Algorithm 1 below.

2.4 Distribution Theory

It is not surprising that the semiparametric estimator described in Algorithm 1 is asymptotically normal with a parametric convergence rate and optimal efficiency as it is formed by estimating all conditional expectations in the efficient score nonparametrically. The asymptotic properties of our estimator are described in Theorem 2 under regularity conditions C1-C2 listed below. The proof is provided in Appendix A.6.

C1 The univariate kernel function $K(\cdot)$ has support $(-1, 1)$ and satisfies $\int K(u)udu = 0$, $\int K(u)u^2du < \infty$. The bandwidth h satisfies $nh^2 \to \infty$ and $nh^8 \to 0$.

C2 Any discrete covariate has finitely many levels. Any continuous covariate has compact support and its density function is twice continuously differentiable.

Theorem 2. *Under the regularity conditions C1 and C2, the estimator $\widehat{\boldsymbol{\theta}}$ obtained from solving the estimating equation $\mathbf{0} = \sum_{i=1}^{N} \widehat{\mathbf{S}}_{\text{eff}}(D_i, G_i, X_i, \widehat{\boldsymbol{\theta}})$ is asymptotically normal with optimal efficiency, i.e., $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to Normal\{0, var(\mathbf{S}_{\text{eff}})^{-1}\}$, and is semiparametric efficient.*

2.5 Simulation Study

We performed simulations to understand the finite sample performance of the semiparametric efficient estimator described in Section 2.3 and demonstrate its superiority to

Algorithm 1: Computing the Efficient Estimator

1. Estimate $f_{X|D=d}(\cdot)$, the conditional density/mass function of X given disease status $D = d$, by nonparametric estimation among the data with $D_i = d$ for $d = 0, 1$. Denote the result by $\widehat{f}_{X|D}(\cdot)$.

2. Estimate $f_{G|D=d}(\cdot)$, the conditional density/mass function of G given disease status $D = d$, by nonparametric estimation among the data with $D_i = d$ for $d = 0, 1$. Denote the result by $\widehat{f}_{G|D}(\cdot)$.

3. Define $\widehat{\eta}_1(g, \pi_0) = \pi_0 \widehat{f}_{G|D=0}(g) + (1 - \pi_0)\widehat{f}_{G|D=1}(g),\ \widehat{\eta}_2(x, \pi_0) = \pi_0 \widehat{f}_{X|D=0}(x) + (1 - \pi_0)\widehat{f}_{X|D=1}(x)$, what we call a weighted nonparametric density/mass function estimate, being weighted by the (estimated) population probabilities.

4. When (π_0, π_1) is unknown, estimate them by solving the integral equation
$$\pi_0 = \int H(0, g, x)\widehat{\eta}_1(g, \pi_0)\widehat{\eta}_2(x, \pi_0) d\mu(g) d\mu(x),$$
and setting $\widehat{\pi}_1 = 1 - \widehat{\pi}_0,\ \widehat{\eta}_1(g) = \widehat{\eta}_1(g, \widehat{\pi}_0),\ \widehat{\eta}_2(x) = \widehat{\eta}_2(x, \widehat{\pi}_0)$.

5. Follow the method described in Appendix A.5 to obtain the solution of the integral equations (2.4) and (2.5), with result $\widehat{\mathbf{a}}, \widehat{\mathbf{b}}$, and approximate $E(\widehat{\mathbf{a}} + \widehat{\mathbf{b}} \mid D)$ using nonparametric density estimates $\widehat{f}_{X|D}(\cdot)$ and $\widehat{f}_{G|D}(\cdot)$, with result $\widehat{E}(\widehat{\mathbf{a}} + \widehat{\mathbf{b}} \mid D)$.

6. Form $\widehat{\mathbf{S}}_{\text{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}) = \widehat{\mathbf{S}}_{\boldsymbol{\theta}}(D_i, G_i, X_i) - \widehat{\mathbf{a}}(G_i) - \widehat{\mathbf{b}}(X_i) + \widehat{E}\{\widehat{\mathbf{a}}(G_i) + \widehat{\mathbf{b}}(X_i) \mid D_i\}$, and estimate $\boldsymbol{\theta}$ by solving the estimating equation
$$\sum_{i=1}^{n} \widehat{\mathbf{S}}_{\text{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}) = \mathbf{0}.$$

It is critical that we estimate $E\{\widehat{\mathbf{a}}(G_i) + \widehat{\mathbf{b}}(X_i) \mid D_i\}$ and $E(\mathbf{S} \mid D_i)$ involved in Steps 5 and 6 using $\widehat{f}_{X|D}(\cdot)$ and $\widehat{f}_{G|D}(\cdot)$ described in Steps 1 and 2 of the above algorithm, instead of simply taking a sample version of the expectations. This ensures that all the conditional expectations are computed using the same kind of approximation and the gene-environment independence assumption is fully employed.

prospective logistic regression method under the gene-environment independent model. Two scenarios are considered: (a) $\text{pr}(D = 1) = 0.045$ and (b) $\text{pr}(D = 1) = 0.10$, corresponding to cases with a relatively rare disease rate and a common disease rate, respectively. In each scenario, we generated X from the standard normal distribution $\text{Normal}(0, 1)$ or the gamma distribution with mean 20 and variance 20 $\text{Gamma}(20, 1)$, while the distribution of G is one of the following: (i) Bernoulli with success probability 0.6, where for example $G = 1$ or $G = 0$ corresponds to the presence or absence of a genetic mutation, and (ii) $\text{Normal}(0, 1)$, which can be used to model gene expression levels or continuous traits, such as height and skin color, that are controlled by several genes. Given G and X, we generated disease status D from the logistic regression model $\text{logit}\{\text{pr}(D = 1 \mid G, X)\} = \alpha + \beta_1 G + \beta_2 X + \beta_3 GX$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (0.76, 0.36, -0.63)$ for both settings with normal X, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (3.577, 0.080, -0.141)$ for both settings with gamma X. We varied the intercept β_0 in different simulations to get the desired disease rate. Specifically speaking, in the case of $X = \text{Normal}(0, 1)$, we set $\alpha = -3.61$ and -3.465 for binary G and normal G respectively to achieve a disease rate of 4.5%, and we set $\alpha = -2.74$ and -2.538 for binary G and normal G respectively to achieve a disease rate of 10%. In the case of $X = \text{Gamma}(20, 1)$, we set $\alpha = -5.220$ and -5.086 for binary G and normal G respectively to achieve a disease rate of 4.5%, and we set $\alpha = -4.352$ and -4.158 for binary G and normal G respectively to achieve a disease rate of 10%. For each setting, we simulated 1,000 data sets, each with $n_1 = 1,000$ cases and $n_0 = 1,000$ controls. In the computation of the weighted nonparametric density/mass function estimates defined in Algorithm 1, we used the asymptotically justified bandwidth $h = cn^{-1/5}$, where $c \in [0.4, 1.2]$, and the results were insensitive to the choice of c.

The results are summarized in Tables 2.1-2.4. For 4.5% disease prevalence and normally distributed X (Table 2.1), it is clear that prospective logistic regression and our semiparametric efficient estimator are both consistent, while the semiparametric estimator has smaller variance. Specifically, the semiparametric efficient estimator has a mean squared error efficiency gain as large as 57% (the interaction term between G and X) for binary G, and 46% (the interaction term between G and X) for the normal G. For 4.5% disease prevalence and gamma X (Table 2.3), when G follows a Bernoulli distribution, our semiparametric efficient estimator has a mean squared error efficiency gain between 31% (the main effect of X) and 56% (the interaction term between G and X); when G is normal, the corresponding efficiency gain of the interaction term is 44%.

The results for the 10% disease rate case (Table 2.2 and 2.4) are similar. Both approaches are asymptotically valid, with our approach being superior to prospective logistic regression in the sense that our semiparametric efficient estimator has better mean squared error (MSE) results.

		Binary G, Normal X			Normal G, Normal X		
	$\boldsymbol{\beta}$	0.76	0.36	-0.63	0.76	0.36	-0.63
Logistic	mean	0.761	0.363	-0.635	0.762	0.363	-0.634
	se	0.101	0.088	0.103	0.055	0.053	0.056
	est se	0.101	0.084	0.101	0.056	0.054	0.055
	95%	0.952	0.939	0.942	0.950	0.954	0.942
Semi	mean	0.761	0.360	-0.630	0.761	0.362	-0.627
	se	0.101	0.077	0.082	0.054	0.051	0.046
	est se	0.100	0.073	0.079	0.053	0.051	0.041
	95%	0.953	0.939	0.941	0.949	0.953	0.921
	MSE Eff	1.003	1.325	1.566	1.068	1.112	1.457

Table 2.1: Simulation studies based upon 1,000 simulated case-control samples taken from a population with a disease rate of approximately 4.5%, and independent genetic and environmental variables, under the logistic model with gene-environment interaction. The results for binary G ~Bernoulli(0.6) and $X \sim \text{Normal}(0, 1)$ is displayed on the left whereas the results for $G \sim \text{Normal}(0, 1)$ and $X \sim \text{Normal}(0, 1)$ is on the right. Each replicate contains $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary G, Normal X			Normal G, Normal X		
	$\boldsymbol{\beta}$	0.76	0.36	-0.63	0.76	0.36	-0.63
Logistic	mean	0.762	0.365	-0.638	0.762	0.363	-0.633
	se	0.102	0.084	0.100	0.056	0.051	0.057
	est se	0.100	0.083	0.100	0.056	0.053	0.057
	95%	0.943	0.952	0.955	0.957	0.960	0.952
Semi	mean	0.762	0.359	-0.628	0.761	0.363	-0.629
	se	0.102	0.077	0.087	0.055	0.050	0.053
	est se	0.100	0.074	0.081	0.055	0.052	0.050
	95%	0.944	0.932	0.936	0.953	0.960	0.934
	MSE Eff	1.004	1.180	1.325	1.032	1.065	1.145

Table 2.2: Simulation results from 1,000 simulated case-control samples taken from a population with a disease rate of approximately 10%, and independent genetic and environmental variables, under the logistic model with gene-environment interaction. The results for binary G $\sim$Bernoulli(0.6) and $X \sim$ Normal$(0,1)$ is displayed on the left whereas the results for $G \sim$ Normal$(0,1)$ and $X \sim$ Normal$(0,1)$ is on the right. Each replicate contains $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary G, Gamma X			Normal G, Gamma X		
	$\boldsymbol{\beta}$	3.577	0.080	-0.141	3.577	0.080	-0.141
Logistic	mean	3.599	0.081	-0.142	3.592	0.080	-0.141
	se	0.456	0.018	0.022	0.269	0.012	0.012
	est se	0.462	0.018	0.022	0.259	0.012	0.012
	95%	0.957	0.953	0.949	0.937	0.950	0.942
Semi	mean	3.586	0.080	-0.141	3.569	0.080	-0.140
	se	0.375	0.016	0.018	0.230	0.011	0.010
	est se	0.369	0.016	0.017	0.202	0.011	0.009
	95%	0.950	0.949	0.942	0.914	0.940	0.919
	MSE Eff	1.484	1.305	1.559	1.372	1.059	1.437

Table 2.3: Simulation results from 1,000 simulated case-control samples taken from a population with a disease rate of approximately 4.5%, and independent genetic $G \sim$ Bernoulli(0.6) and environmental $X \sim$ Gamma(20,1) variables, under the logistic model with gene-environment interaction. Each replicate contains $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary G, Gamma X			Normal G, Gamma X		
	$\boldsymbol{\beta}$	3.577	0.080	-0.141	3.577	0.080	-0.141
Logistic	mean	3.589	0.081	-0.141	3.600	0.081	-0.142
	se	0.459	0.018	0.022	0.274	0.012	0.013
	est se	0.460	0.018	0.022	0.269	0.012	0.012
	95%	0.949	0.950	0.947	0.950	0.934	0.944
Semi	mean	3.565	0.080	-0.140	3.590	0.081	-0.142
	se	0.394	0.016	0.019	0.268	0.012	0.012
	est se	0.381	0.016	0.018	0.247	0.011	0.011
	95%	0.945	0.953	0.938	0.934	0.937	0.930
	MSE Eff	1.360	1.240	1.406	1.048	1.031	1.061

Table 2.4: Simulation results from 1,000 simulated case-control samples taken from a population with a disease rate of approximately 10%, and independent genetic G $\sim$Bernoulli(0.6) and environmental X $\sim$Gamma(20,1) variables, under the logistic model with gene-environment interaction. Each replicate contains $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

2.6 Example

Prostate cancer is a heterogeneous disease resulting from the complex interplay of genetic susceptibility and environmental exposures. It is the second leading cause of cancer death among men in the US (Siegel and Jemal, 2015). Prostate cells (both primary and cancer cells) were demonstrated to have 1α-OHase activity, whereas 1α-OHase is the enzyme responsible for converting [25(OH)D], the major circulating form of vitamin D that reflects both dietary and sunlight exposures, into 1,23-dihydroxy-vitamin D [1,25(OH)2D], the most active form of this vitamin that can induce cell-cycle regulation, apoptosis and differentiation in prostate cancer cells via the vitamin D receptor (VDR). Thus, (a) [25(OH)D] is hypothesized to have an anticancer effect, and (b) an important question is whether its relationship with the risk of developing prostate cancer is modified by genetic polymorphisms in the VDR gene.

In this section, we implemented our methodology in a case-control study of prostate cancer, using the same data set analyzed differently but in a different context by Chen et al. (2009), see that reference for details about the study. Specifically, our analysis is based on a polygenic risk score, a single risk factor incorporating information from susceptibility SNPs, whereas Chen et al. (2009) focused on haplotypes. The data consist of $n_1 = 690$ cases and $n_0 = 717$ controls randomly selected from the screening arm of a large population-based cohort study, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) at the National Cancer Institute. The PLCO cohort study recruited a total of 76,685 men aged 55 to 74 at 10 screening centers between November 1993 and July 2001, then randomly assigned 38,340 of them to the screening arm and the rest to the non-screening arm. In a 10 year follow-up period, in the study population, the cumulative incidence rate for prostate cancer in the screening arm was 108.4 per 10,000 person-years (Andriole et al., 2012). Apart from case-control status, [25(OH)D] level (nmol/L) and genotype data on 19 single-nucleotide polymorphisms (SNPS) are available for each subject involved in the case-control study. According to Chen et al. (2009), these polymorphisms, our G, are unlikely to affect the [25(OH)D] level, our X, as the VDR gene plays a "downstream" role in the vitamin-D pathway. In other words, the gene-environment independence assumption is likely to be valid in this application. Detailed information about the design can be found in Andriole et al. (2012), Hayes et al. (2000), and Prorok et al. (2000).

One difficulty in investigating the genetic modification of the VDR gene to [25(OH)D] on the risk of prostate cancer is that the VDR gene contains multiple underlying susceptibility SNPs, where each individual SNP may only confer a small component of overall risk. In fact, running a logistic regression of case-control status on each of the 19 SNPs shows only three SNPs have p-values ≤ 0.10. Recently, it has been recognized that the polygenic risk score has the potential of improving risk prediction for some common diseases (Aly et al., 2011; Chatterjee et al., 2016; Dudbridge, 2013; Evans et al., 2009; Fuchsberger et al., 2016; Purcell et al., 2009). Therefore, we created a polygenic risk score for the prostate cancer data by weighting those 19 SNPs, where the weights are the effect sizes of separate logistic regressions applied to each SNP.

The results of prospective logistic regression and our semiparametric approach based on 1,000 bootstrap samples are given in Table 2.5. The two sets of estimates are fairly consistent as expected. However, our semiparametric efficient estimator has smaller standard errors than does the prospective logistic regression, in accordance with theory and our simulations. This leads to a substantial difference in inference for the interaction between the polygenic risk score and the [25(OH)D] level. Specifically, both prospective logistic regression and our semiparametric efficient method show that the main effects of both the polygenic risk score and the [25(OH)D] level is statistically significant and positive. That is, if ignoring the interaction, men with higher polygenic risk scores or/and higher [25(OH)D] levels tend to have higher risk of developing prostate cancer.

		β_G	β_X	β_{GX}
Logistic	Estimates	0.169	0.123	-0.101
	se, bootstrap	0.056	0.056	0.054
	est se, asymptotic	0.055	0.055	0.055
	p-value, bootstrap	0.002	0.028	0.064
	p-value, asymptotic	0.002	0.024	0.066
Semi	Estimates	0.168	0.124	-0.110
	se, bootstrap	0.056	0.056	0.049
	est se, asymptotic	0.055	0.054	0.042
	p-value, bootstrap	0.003	0.027	0.026
	p-value, asymptotic	0.002	0.021	0.009

Table 2.5: Analysis of the case-control study on prostate cancer, containing $n_1 = 690$ cases and $n_0 = 717$ controls. Two approaches were implemented, (1) “Logistic” is ordinary logistic regression, and (2) “Semi” is our semiparametric efficient estimator. Displayed are the estimates, bootstrap standard error (“se, bootstrap”), mean estimated asymptotic standard error (“est se, asymptotic”), bootstrap p-value (“p-value, bootstrap”), and asymptotic p-value (“p-value, asymptotic”) of the coefficients for the standardized polygenic risk score (G), [25(OH)D] level (X), and the interaction between them (GX).

Importantly, the estimates of the interaction parameter from the prospective logistic regression is not significant at the 5% level. However, our approach shows significant evidence of interaction, i.e., the effects of [25(OH)D] level on prostate cancer risk differ depending on the polygenic risk score.

In addition, our approach provides an estimated disease rate in the population of 10.6%, whereas the disease rate in the PLCO cohort study is 10.8% per person-year. This validation of our methodology suggests an additional use to which it can be applied.

2.7 Discussion

We have developed a semiparametric efficient estimator in case-control studies for the gene-environment independent model, where the distributions of genetic susceptibility and environmental exposure are allowed to be arbitrary and the disease rate is assumed completely unknown. We showed that despite of these weak assumptions, the problem is identifiable in most cases. The proposed estimator is derived under the so called hypothetical population framework, which enables us to view the case-control sample as a random sample from a hypothetical distribution and thus facilitates the application of a conventional semiparametric approach. Such an estimator is semiparametric efficient and its superiority over the prospective logistic regression was demonstrated in various simulations. The general methodology of our approach can be extended to parametric models other than the logistic model, such as the probit model, and it can be used to consider assumptions other than gene-environment independence, such as Hardy-Weinberg equilibrium, as long as the resulting model is identifiable.

To handle the nuisance parameters in the estimation procedure, nonparametric density/mass function estimation are used. When the dimensions of genetic susceptibility or environmental exposures increase, such nonparametric estimation suffers from the curse of dimensionality. In such cases, dimension reduction techniques might be needed to maintain model flexibility as well as ensure computation feasibility. This will be pursued in future work.

3. DIMENSION REDUCTION AND ESTIMATION IN THE SECONDARY ANALYSIS OF CASE-CONTROL STUDIES

3.1 Introduction

The analysis on gene-environment independence model discussed in Section 3 is a special case of primary analysis, where the covariates can be separated into independent genetic and environmental (non-genetic) variables. Generally, covariates in primary analysis can be arbitrary risk factors potentially associated with the disease of interest and the relationship between covariates is not necessary to be independence. For example, in Section 3.6, we describe a case-control study involving breast cancer and its well known risk factors including mammographic density and age at first live birth. We discover a statistically significant effect of age at first live birth on mammographic density. Particulary, women with a relatively late age at first live birth often have a lower mammographic density.

The study that examine the interrelationship between covariates is known as *secondary analysis*. Its main difficulty lies in the fact that the case-control data is not a random sample from the underlying source population. In fact the case-control samples are taken separately from the case subpopulation and the control subpopulation. As a consequence, the relationship between covariates Y and $\mathbf{X}$ in the secondary analysis under the case-control context can be very different from the relationship in the true population. Hence, simply regressing Y on $\mathbf{X}$ and ignoring the case-control sampling scheme can be grossly misleading.

A simple approach to secondary analysis is using only controls if the disease rate is rare, say less than 1%. This type of approach is widely used, because if the disease rate is $< 1\%$, the controls make up more than 99% of the population, and analysis of them is close to that of the entire population. However, this approach can have relatively low efficiency because it ignores the information carried by the cases. A more efficient approach is to adopt a semiparametric framework, assuming a parametric distribution for Y given $\mathbf{X}$, e.g., linear regression with normally distributed and homoscedastic regression errors, as well as known or rare disease rate (Jiang et al., 2006; Li et al., 2010; Lin and Zeng, 2009; Tchetgen, 2014; Wei et al., 2013). This approach improves estimation efficiency compared with the controls only method because both cases and controls are taken into account.

However, the disease rate in the source population being sampled is often unknown and some diseases may not be so rare as less than 1%, so that the controls-only analysis can have considerable bias. This prompted Ma and Carroll (2016) to propose a further improved approach, which does not require a known or a rare disease assumption, and also, unlike the papers referenced above, does not assume normality or homoscedasticity of the regression error. In fact, they only specify a mean model to describe the relationship between covariates. Their semiparametric estimator involves positing density functions for $\mathbf{X}$ and Y given $\mathbf{X}$ that may or may not be true. The resulting estimator is (a) consistent and asymptotically normally distributed even if the posited functions are incorrectly specified; and (b) it is efficient if the posited functions are correctly specified. An estimator with the properties (a) and (b) will be called locally efficient throughout this article.

Because the approach of Ma and Carroll (2016) was developed by adopting a hypothetical population concept and viewing case-control samples as independent and identically distributed observations sampled from the hypothetical population, they need to link the quantities in the hypothetical population to the ones in the true population. As a consequence, several additional conditional distributions arise in the likelihood formulation, including quantities conditional on the covariates. This leads to the need to perform several nonparametric regressions on the covariates in their estimator. When the covariate dimension increases, such nonparametric regressions inevitably suffer from the curse of dimensionality.

In this section, we work in the hypothetical population framework and handle the potential dimensionality problem using a dimension reduction modeling approach. We assume several quantities of interest depend on the covariates $\mathbf{X}$ only through linear combinations of $\mathbf{X}$ and/or known functions of $\mathbf{X}$. This allows us to avoid multivariate nonparametric regression. However, because of the inherent relation between the covariates assumed in the original true population, the dimension reduction structure is not completely arbitrary. Instead, it is subject to various constraints, which makes the problem different from the classical dimension reduction modeling and estimation. Taking these various special features into consideration, we construct asymptotically consistent estimators for the regression parameters in the true population model. These estimators have a parametric convergence rate and are robust to the misspecification of the conditional distribution of Y given $\mathbf{X}$.

The rest of the section is organized as follows. In Section 3.2, we introduce the model and notation we use. We propose a locally efficient estimator using the single index model in Section 3.3. We derive the asymptotic properties of the resulting estimators in Section 3.4. In Section 3.5, we report simulation results, compare our method with the control only method and the semiparametric efficient method that assumes normality and homoscedasticity on the error (Lin and Zeng, 2009). We analyze a mammographic density data set with our approach in Section 3.6, and follow with a brief conclusion in Section 3.7. Technical details and proofs are provided in an Appendix.

3.2 Methodology

3.2.1 Background

Let D be disease status, where $D = 1$ denotes a case and $D = 0$ denotes a control. Also let $(\mathbf{X}^{\mathrm{T}}, Y)^{\mathrm{T}}$ be a $(p+1) \times 1$ vector of covariates, where $\mathbf{X}$ is a p-dimensional vector and Y is a scalar. We assume that both $\mathbf{X}$ and Y are continuous and they are related to disease status D via a logistic regression model

$$\begin{aligned} \text{pr}(D = d|\mathbf{X} = \mathbf{x}, Y = y) &= f_{D|X,Y}^{\text{true}}(d, \mathbf{x}, y) = H(d, \mathbf{x}, y, \boldsymbol{\alpha}) \\ &= \frac{\exp\{d(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)\}}{1 + \exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\alpha} = (\alpha_c, \boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}}$.

As mentioned before, the goal of secondary analysis is to investigate the relationship between $\mathbf{X}$ and Y in the source population, which we assume is of the form

$$Y = m(\mathbf{X}, \boldsymbol{\beta}) + \epsilon, \tag{3.2}$$

where $m(\cdot)$ is a smooth function known up to a parameter $\boldsymbol{\beta}$. The error term ϵ satisfies $E_{\text{true}}(\epsilon|\mathbf{X}) = 0$, but no other assumptions about ϵ are made, especially normality or homoscedasticity or independence from $\mathbf{X}$. Under mild conditions, the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ defined in (3.1) and (3.2) are identifiable (Ma and Carroll, 2016).

3.2.2 Hypothetical Population Model Framework and Efficient Estimator

From model (3.2), the conditional distribution of Y given $\mathbf{X}$ and the marginal distribution of $\mathbf{X}$ with respect to the true population are

$$f_{Y|\mathbf{X}}^{\text{true}}(y, \mathbf{x}, \boldsymbol{\beta}) = \eta_2\{y - m(\mathbf{x}, \boldsymbol{\beta}), \mathbf{x}\} = f_{\epsilon|\mathbf{X}}^{\text{true}}(\epsilon, \mathbf{x}), \tag{3.3}$$

$$f_{\mathbf{X}}^{\text{true}}(\mathbf{x}) = \eta_1(\mathbf{x}). \tag{3.4}$$

Here η_2 is an unknown probability density function with mean 0, which is free of the unknown parameters $\boldsymbol{\beta}$, ϵ is the error term defined in (3.2), i.e., $\epsilon = Y - m(\mathbf{X}, \boldsymbol{\beta})$ and η_1 is another probability density function which is also unknown. The superscript "true" emphasizes that the probability densities in (3.3) - (3.4) are defined under the true population.

Suppose we draw a case-control sample with N_1 cases and N_0 controls. Because of the sampling design, classical large-sample asymptotic theory does not work here. The idea of a hypothetical population is to construct a hypothetical population with infinite sample size and a fixed ratio of cases to controls, N_1/N_0, then treat the case-control sample as a random sample from the hypothetical population with sample size $N = N_0 + N_1$ (Ma, 2010). The explicit form of the joint density of $(\mathbf{X}, Y, D)$ in such a hypothetical population is

$$\begin{aligned} f_{\mathbf{X},Y,D}(\mathbf{x}, y, d) &= (N_d/N) f^{\text{true}}_{X,Y|D}(\mathbf{x}, y, d) \\ &= \frac{N_d}{N} \frac{\eta_1(\mathbf{x})\eta_2(\epsilon, \mathbf{x}) H(d, \mathbf{x}, y, \boldsymbol{\alpha})}{\int \eta_1(\mathbf{x})\eta_2(\epsilon, \mathbf{x}) H(d, \mathbf{x}, y, \boldsymbol{\alpha}) d\mu(\mathbf{x}) d\mu(y)}. \end{aligned}$$

Here we use the fact that the distribution of $(\mathbf{X}, Y)$ conditional on the disease status D in the hypothetical population and in the true population are identical, which links the distributions in these two populations.

Ma and Carroll (2016) derived the semiparametric efficient score function corresponding to the above hypothetical population, $\mathbf{S}_{\text{eff}}(\mathbf{X}_i, Y_i, D_i) = \{\mathbf{S}(\mathbf{X}_i, Y_i, D_i) - \mathbf{g}\{Y_i - m(\mathbf{X}_i, \boldsymbol{\beta}), \mathbf{X}_i)\} - (1 - D_i)\mathbf{v}_0 - D_i\mathbf{v}_1$. The resulting efficient estimating equation is

$$\textstyle\sum_{i=1}^{N} \{\mathbf{S}(\mathbf{X}_i, Y_i, D_i) - \mathbf{g}\{Y_i - m(\mathbf{X}_i, \boldsymbol{\beta}), \mathbf{X}_i)\} - (1 - D_i)\mathbf{v}_0 - D_i\mathbf{v}_1 = \mathbf{0}, \qquad (3.5)$$

where

$$\mathbf{S}(\mathbf{x}, y, d, \boldsymbol{\theta}) = \begin{Bmatrix} \partial \text{log}\{H(d, \mathbf{x}, y, \boldsymbol{\alpha})\}/\partial \boldsymbol{\alpha} \\ \partial \text{log}\{\eta_2(\epsilon, \mathbf{x})\}/\partial \boldsymbol{\beta} \end{Bmatrix}. \qquad (3.6)$$

Although as a function, η_2 does not depend on $\boldsymbol{\beta}$, its first argument ϵ contains $\boldsymbol{\beta}$. Other quantities used in (3.5) are defined in (3.7).

$$\begin{aligned}
&\pi_0 \equiv p_D^{\text{true}}(0) = \textstyle\int \eta_1(\mathbf{x})\eta_2(\epsilon,\mathbf{x})H(0,\mathbf{x},y)d\mu(\mathbf{x})d\mu(y);\\
&\pi_1 \equiv p_D^{\text{true}}(1) = \textstyle\int \eta_1(\mathbf{x})\eta_2(\epsilon,\mathbf{x})H(1,\mathbf{x},y)d\mu(\mathbf{x})d\mu(y);\\
&b_0 \equiv E\{f_{D|\mathbf{X},Y}(1,\mathbf{X},y) \mid D=0\}; b_1 \equiv E\{f_{D|\mathbf{X},Y}(0,\mathbf{X},y) \mid D=1\};\\
&\boldsymbol{\mu}_s(\mathbf{x},y) \equiv E(\mathbf{S} \mid \epsilon, \mathbf{X}=\mathbf{x}); \mathbf{c}_0 \equiv E(\mathbf{S} \mid D=0) - E\{\boldsymbol{\mu}_s(\mathbf{X},Y) \mid D=0\};\\
&\mathbf{c}_1 \equiv E(\mathbf{S} \mid D=1) - E\{\boldsymbol{\mu}_s(\mathbf{X},Y) \mid D=1\};\\
&\kappa(\mathbf{x},y) \equiv \left[\textstyle\sum_{d=0}^{1}\{N_d H(d,\mathbf{x},y)\}/(N\pi_d)\right]^{-1};\\
&t_1(\mathbf{X}) \equiv \left[E_{\text{true}}\left\{\epsilon^2\kappa(\mathbf{X},Y) \mid \mathbf{X}\right\}\right]^{-1};\\
&\mathbf{t}_2(\mathbf{X}) \equiv E_{\text{true}}\left\{\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y) \mid \mathbf{X}\right\} - (\mathbf{c}_0/b_0)E_{\text{true}}\left\{\epsilon f_{D|\mathbf{X},Y}(0,\mathbf{X},Y) \mid \mathbf{X}\right\};\\
&t_3(\mathbf{X}) \equiv -b_0^{-1}E_{\text{true}}\left\{\epsilon f_{D|\mathbf{X},Y}(0,\mathbf{X},Y) \mid \mathbf{X}\right\}; \mathbf{a}(\mathbf{x}) \equiv t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x}) + t_3(\mathbf{x})\mathbf{u}_0\};\\
&\mathbf{u}_0 \equiv \left(1 - E\left[\epsilon t_1(\mathbf{X})t_3(\mathbf{X})\kappa(\mathbf{X},Y) \mid D=0\right]\right)^{-1} E\left[\epsilon t_1(\mathbf{X})\mathbf{t}_2(\mathbf{X})\kappa(\mathbf{X},Y) \mid D=0\right];\\
&\mathbf{u}_1 \equiv -(N_0/N_1)\mathbf{u}_0; \mathbf{v}_0 \equiv (\pi_1/b_0)(\mathbf{u}_0+\mathbf{c}_0); \mathbf{v}_1 \equiv -(\pi_0/b_0)(\mathbf{u}_0+\mathbf{c}_0);\\
&\mathbf{g}(\epsilon,\mathbf{x}) \equiv \boldsymbol{\mu}_s(\mathbf{x},y) - \epsilon\mathbf{a}(\mathbf{x})\kappa(\mathbf{x},y) - \mathbf{v}_0 f_{D|\mathbf{X},Y}(0,\mathbf{x},y) - \mathbf{v}_1 f_{D|\mathbf{X},Y}(1,\mathbf{x},y).
\end{aligned} \tag{3.7}$$

3.3 Approach via Dimension Reduction

3.3.1 Background

The estimating equation (3.5) contains three expectations conditional on covariates $\mathbf{X}$, i.e., $E_{\text{true}}\left\{\epsilon^2\kappa(\mathbf{X},Y) \mid \mathbf{X}\right\}$, $E_{\text{true}}\left\{\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y) \mid \mathbf{X}\right\}$ and $E_{\text{true}}\left\{\epsilon f_{D|\mathbf{X},Y}(0,\mathbf{X},Y) \mid \mathbf{X}\right\}$, which need to be estimated nonparametrically. However, such estimation may be extremely hard when the covariates $\mathbf{X}$ are multivariate. To bypass the potential curse of dimensionality problem caused by the multivariate nature of $\mathbf{X}$, we use a dimension reduction modeling strategy, i.e., we assume all three quantities in the conditional expectations depend on $\mathbf{X}$ only through several linear combinations $\mathbf{X}^{\mathrm{T}}\boldsymbol{\gamma}$ or several linear combinations of functions of $\mathbf{X}$. Under such a dimension reduction structure, we can construct nonparametric regression estimators for high dimensional covariates $\mathbf{X}$ in a way similar to the univariate case with desired bias and MSE order, hence facilitating the estimation procedure via solving the estimating equation (3.5).

Let $f_0(\mathbf{X},Y,\boldsymbol{\alpha}) = f_{D|\mathbf{X},Y}(0,\mathbf{X},Y)$. All three functions $\kappa(\mathbf{x},y)$, $\boldsymbol{\mu}_s(\mathbf{x},y)$ and $f_0(\mathbf{x},y)$

depend on $\pi_d = \pi_d(\boldsymbol{\alpha})$. To emphasize this, we replace π_d with $\pi_d(\widetilde{\boldsymbol{\alpha}})$ in those three functions and we use the notation $\kappa(\mathbf{x}, y, \widetilde{\boldsymbol{\alpha}}), \boldsymbol{\mu}_s(\mathbf{x}, y, \widetilde{\boldsymbol{\alpha}}), f_0(\mathbf{x}, y, \widetilde{\boldsymbol{\alpha}})$ to distinguish them from the ones using the true parameter value $\boldsymbol{\alpha}$. In addition, we define $\epsilon(\mathbf{X}, Y, \widetilde{\boldsymbol{\beta}}) = Y - m(\mathbf{X}, \widetilde{\boldsymbol{\beta}})$ to distinguish it from the true $\epsilon = Y - m(\mathbf{X}, \boldsymbol{\beta})$.

There are two cases that need to be considered, namely that (i) $m(\cdot)$ defined in (3.2) is a linear function of $\mathbf{X}$; and (ii) that $m(\cdot)$ is not a linear function of $\mathbf{X}$. In case (i), we set $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} = \mathbf{X}$, while in case (ii), we set $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} = \{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \widetilde{\boldsymbol{\beta}})\}^{\mathrm{T}}$.

Then our dimension reduction models are

$$\begin{aligned} E_{\text{true}}\{\epsilon^2(\mathbf{X}, Y, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widetilde{\boldsymbol{\alpha}}) \mid \mathbf{X}\} &= \zeta_1(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\alpha}}), && (3.8)\\ E_{\text{true}}\{\epsilon(\mathbf{X}, Y, \widetilde{\boldsymbol{\beta}})\boldsymbol{\mu}_s(\mathbf{X}, Y, \widetilde{\boldsymbol{\alpha}}) \mid \mathbf{X}\} &= \boldsymbol{\zeta}_2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_2, \mathbf{X}, \widetilde{\boldsymbol{\alpha}}), && (3.9)\\ E_{\text{true}}\{\epsilon(\mathbf{X}, Y, \widetilde{\boldsymbol{\beta}})f_0(\mathbf{X}, Y, \widetilde{\boldsymbol{\alpha}}) \mid \mathbf{X}\} &= \zeta_3(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_3, \widetilde{\boldsymbol{\alpha}}), && (3.10) \end{aligned}$$

for $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$ that are in a neighborhood of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Here $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}$ is a finite dimensional vector, each element of which is a function of $\mathbf{X}$. The subscript $\widetilde{\boldsymbol{\beta}}$ indicates $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}$ may depend on the unknown parameter $\widetilde{\boldsymbol{\beta}}$. The three indices $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ are vectors or matrices that have the same row size as the length of $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}$ and with ℓ columns. The lower square blocks of all three matrices $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ are set to be identity to ensure identifiability. Throughout the text, we use the notation $\boldsymbol{\gamma}_{-1}$ to denote the submatrix of $\boldsymbol{\gamma}$ without the lower square block for any matrix $\boldsymbol{\gamma}$. $\zeta_1(\cdot), \boldsymbol{\zeta}_2(\cdot), \zeta_3(\cdot)$ are three unknown functions. Strictly speaking, model (3.9) is not a standard dimension reduction model. However, in Appendix B.1, we describe its actual form, which in general consists of three different standard dimension reduction models.

3.3.2 Data Generating Mechanisms for Which (3.8)-(3.10) are Valid

The dimension reduction models (3.8)-(3.10) are used here only as working models to facilitate the estimation procedure. We do not intend to include these models as part of our original model assumptions and thereby take these structures into account to further improve estimation efficiency.

There are at least two simple and important data generating mechanisms for which (3.8)-(3.10) hold: (a) when ϵ is independent of $\mathbf{X}$; and (b) when, as in equation (1) of Lian et al. (2015), $\epsilon = v(\mathbf{X}^{\mathrm{T}}\omega)\epsilon^*$, where $\mathbf{v}(\cdot)$ is an unknown smooth function and ϵ^* is independent of $\mathbf{X}$ with mean 0 and variance 1. More generally, we have the following result, proved in Appendix B.4, and including the two special cases given above..

Proposition 1. Suppose $\epsilon = Q(\mathbf{X}^{\mathrm{T}}\omega, \epsilon^*)$, where $Q(\cdot)$ is an arbitrary smooth function and ϵ^* is independent of $\mathbf{X}$. Then the dimension reduction models (3.8)-(3.10) hold.

3.3.3 Estimation

As stated in Section 3.3.2, models (3.8)-(3.10) can often be used as working models to facilitate the multivariate nonparametric regression. Therefore, in the rest of the derivation, we use the general model (3.8)-(3.10) without specifying the particular form of $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}$. Of course, we need to estimate $\boldsymbol{\gamma}_j$ and $\zeta_j(\cdot)$ for $j = 1, 2, 3$. To resolve the issue of estimating conditional expectations in the true population while we only have a random sample from the hypothetical population, the key point is to recognize the connection between the two populations and to adjust the case-control data in the context of conditional expectations via

$$E_{\text{true}}\{h(D, \mathbf{X}, Y)\} = \textstyle\sum_{d=0}^{1} \pi_d E\{h(D, \mathbf{X}, Y) \mid D = d\},$$

where $h(\cdot)$ is any function such that $h(D, \mathbf{X}, Y)$ has finite mean. Hence we can simply weight cases by π_1/N_1 and controls by π_0/N_0 and this will give us the $\zeta_j(\cdot)$'s. Take $\zeta_1(\cdot)$ as an example. A valid estimating equation for $\zeta_1(\cdot)$ is

$$0 = \textstyle\sum_{d=0}^{1}(\pi_d/N_d)\sum_{i=1}^{N} I(D_i = d)\{\epsilon_i^2 \kappa(\mathbf{X}_i, Y_i) - \zeta_1(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1)\} K_h(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_1), \tag{3.11}$$

since

$$\begin{aligned} &E[\textstyle\sum_{d=0}^{1}(\pi_d/N_d)\sum_{i=1}^{N} I(D_i = d)\{\epsilon_i^2 \kappa(\mathbf{X}_i, Y_i) - \zeta_1(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1)\} K_h(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_1)] \\ &\quad = \textstyle\sum_{d=0}^{1}\pi_d E_{\mathrm{true}}[\{\epsilon^2 \kappa(\mathbf{X}, Y) - \zeta_1(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1)\} K_h(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_1) | D = d] \\ &\quad = E_{\mathrm{true}}[\{\epsilon^2 \kappa(\mathbf{X}, Y) - \zeta_1(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1)\} K_h(\mathbf{z}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_1)] = 0. \end{aligned}$$

Here $K_h(\mathbf{u}) = \prod_{i=1}^{\ell} K(u_i/h)/h^{\ell}$ for $\mathbf{u} = (u_1, \cdots, u_\ell)^{\mathrm{T}}$ for any ℓ-dimensional vector $\mathbf{u}$.

Of course π_d is not known. Thus, to implement the idea stated in (3.11), we need an estimator of $\pi_d = \pi_d(\boldsymbol{\alpha})$. As an equation for π,

$$E\left[\frac{H(0, \mathbf{X}, Y, \boldsymbol{\alpha})}{(N_0/N)H(0, \mathbf{X}, Y, \boldsymbol{\alpha}) + (N_1/N)H(1, \mathbf{X}, Y, \boldsymbol{\alpha})\{\pi/(1-\pi)\}}\right] = 1 \tag{3.12}$$

has a solution $\pi = \pi_0(\boldsymbol{\alpha})$. It is the unique solution as long as pr$\{H(0, \mathbf{X}, Y, \boldsymbol{\alpha}) > 0\} > 0$, since $\pi/(1-\pi)$ is strictly increasing, ranging from 0 to ∞. Based on (3.12), we can construct a root-N consistent estimator of π_0 and plug it into (3.11). The resulting estimators of the $\zeta_1(\cdot)$ have the same bias and mean squared error order as the usual nonparametric estimator. The proof is provided in **Supplementary Material** B.5.

For simplicity, one may use the same index in (3.8)-(3.10), i.e. assuming $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = \boldsymbol{\gamma}$. As before, we restrict the lower square block of $\boldsymbol{\gamma}$ to be identity. We provide detailed estimation procedures and algorithms for both cases, with the algorithm for different indices in Appendix B.2.1 and that for the same index in Appendix B.2.2.

Remark 1. It is worth pointing out that the estimation of π via (3.12) originates from

$$\begin{aligned}\pi_0 &= \int H(0,\mathbf{X},Y,\boldsymbol{\alpha}) f^{\text{true}}_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}) f^{\text{true}}_{\mathbf{X}}(\mathbf{x}) d\mu(\mathbf{x}) d\mu(y) \\ &= \int \frac{H(0,\mathbf{X},Y,\boldsymbol{\alpha})}{\sum_d N_d/(N\pi_d) H(d,\mathbf{X},Y,\boldsymbol{\alpha})} \\ &\qquad \times {\textstyle\sum_d} N_d/(N\pi_d) H(d,\mathbf{X},Y,\boldsymbol{\alpha}) f^{\text{true}}_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\beta}) f^{\text{true}}_{\mathbf{X}}(\mathbf{x}) d\mu(\mathbf{x}) d\mu(y) \\ &= \int \frac{H(0,\mathbf{X},Y,\boldsymbol{\alpha})}{\sum_d N_d/(N\pi_d) H(d,\mathbf{X},Y,\boldsymbol{\alpha})} f_{\mathbf{X},Y}(y,\mathbf{x},\boldsymbol{\beta}) d\mu(\mathbf{x}) d\mu(y).\end{aligned}$$

Thus, the estimator takes into account the difference between the hypothetical population and the population from which the case-control sample is drawn, and thus leads to a consistent estimator of π_0.

3.3.4 Estimation Algorithm Using Different Indices

The estimating equation in (3.5) relies on the unknown probability density function η_2. Here, we use a posited model η_2^*, which is not necessarily the truth, to calculate the efficient score and other related quantities. The resulting estimating function is denoted by $\mathbf{S}^*_{\text{eff}}$. We will show that the resulting estimator is still consistent, and it is efficient if the posited model η_2^* is the correct one.

The main difficulty in calculating $\mathbf{S}^*_{\text{eff}}$ lies in approximating functions $\mathbf{g}, \mathbf{v}_0, \mathbf{v}_1$, because they depend on three expectations conditional on covariates $\mathbf{X}$, which need to be estimated nonparametrically. We bypass this difficulty via the dimension reduction strategy described in Section 3.3.1-3.3.3. A sketch of the algorithm is the following.

1. Posit a model for $\eta_2(\epsilon, \mathbf{x})$ which has mean zero. Under this posited model, calculate S^* from (3.6).

2. Solve $\widehat{\pi}_0(\boldsymbol{\alpha}) = \sum_{i=1}^N H(0,\mathbf{X}_i,Y_i,\boldsymbol{\alpha})[N_0 H(0,\mathbf{X}_i,Y_i,\boldsymbol{\alpha})/\widehat{\pi}_0(\boldsymbol{\alpha}) + N_1 H(1,\mathbf{X}_i,Y_i,$

$\boldsymbol{\alpha})/\{1-\widehat{\pi}_0(\boldsymbol{\alpha})\}]^{-1}$, and set $\widehat{\pi}_1(\boldsymbol{\alpha}) = 1-\widehat{\pi}_0(\boldsymbol{\alpha})$.

3. Estimate the indices $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ and the corresponding functions $\zeta_1, \boldsymbol{\zeta}_2, \zeta_3$ defined in (3.8)-(3.10) respectively by following the procedure in Section 3.3.3.

4. Plug the estimation from Step 3 into the expression of functions $\mathbf{g}, \mathbf{v}_0$ and v_1 in (3.7) to get $\widehat{\mathbf{g}}, \widehat{\mathbf{v}}_0$ and $\widehat{\mathbf{v}}_1$.

5. Form $\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i) = \mathbf{S}^*_i - \widehat{\mathbf{g}}_i - \widehat{\mathbf{v}}_{D_i}$ and solve the corresponding estimating equation.

For convenience, we adopt $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = \boldsymbol{\gamma}$ in all the simulations, where the lower square block of $\boldsymbol{\gamma}$ is set to be identity to ensure identifiability. The algorithm in this simplified case is identical to the one described above except step 3. The detailed algorithms for cases using different indices and using a common index are given in Appendix B.2.

3.4 Distribution Theory

We now establish the asymptotic distribution theory of our estimators, stated as Theorem 3 below, with necessary regularity conditions C1-C11 listed in Appendix B.3. The proof of Theorem 3 is detailed and lengthy and is thus sketched in the **Supplementary Material** Section B.5. While Theorem 3 holds for both the estimator using different indices and the estimator using a common index, we only provide the proof and regularity conditions for the algorithm with different indices. One can easily adapt the conditions and proof to the case of a common index.

Under the regularity conditions C1-C11 listed in Appendix B.3, the following theorem holds. The proof is in the **Supplementary Material** Section B.5.

Theorem 3. *Define* $\mathbf{A} = E\left\{\partial \mathbf{S}^*_{\text{eff}}(D, \mathbf{X}, Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}\right\}$ *and* $\mathbf{B} = cov\left\{\mathbf{S}^*_{\text{eff}}(D, \mathbf{X}, Y, \boldsymbol{\theta})\right\}$. *The* *estimator $\widehat{\theta}$ obtained from solving the estimating equation*

$$\sum_{i=1}^{N} \widehat{\mathbf{S}}^{*}_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0} \tag{3.13}$$

satisfies $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to$ Normal$\{0, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$ and $\widehat{\boldsymbol{\theta}}$ is locally efficient, see the definition of locally efficient in Section 3.3.1.

3.5 Simulations

3.5.1 Setup

We performed a series of simulations to understand the behaviour of our method and compare it to competitors. The simulations displayed in this section are for the case that the regression errors ϵ are Gaussian or centered Gamma, both homoscedastic and heteroscedastic.

In these simulations, we considered different disease rates, different dimensions and distributions for $\mathbf{X}$ and different error variance structures. The results indicate that our methods have small bias and good coverage probability in all the cases we examined. Here, due to space limitations, we only list the results for two typical scenarios, where the first one is homoscedastic and the second one is heteroscedastic. In both cases, we chose a balanced design with $N_1 = 1000$ cases and $N_0 = 1000$ controls, set the disease rate to be approximately 4.5% and let $\mathbf{X}$ be exchangeable with $p = \dim(\mathbf{X}) = 4$.

More specifically, we generated $\mathbf{X} = (X_1, \cdots, X_4)^{\mathrm{T}}$ in the following way.

1. Generate $\mathbf{X}^* =$ Normal$(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\Sigma_{i,j})_{1 \le i,j \le 4}$ and $\Sigma_{i,j} = 1$ if $i = j$ and $\Sigma_{i,j} = \rho$ for $|\rho| < 1$ if $1 \le i \ne j \le 4$.

2. Let $\mathbf{X} = \Phi(\mathbf{X}^*) = \{\Phi(\mathbf{X}^*_1), \cdots, \Phi(\mathbf{X}^*_4)\}^{\mathrm{T}}$, where Φ is the cumulative distribution function of a standard normal random variable.

Hence, $\mathbf{X}$ is an exchangeable vector of random variables satisfying $X_i = \text{Uniform}[0,1]$ for $i = 1, \cdots, 4$ and $\text{corr}(X_i, X_j) = \text{corr}(X_k, X_l)$ for all $i \neq j, k \neq l$. In our simulation studies, we used $\rho = 0.2$, which resulted in $\text{corr}(X_i, X_j) \approx 0.191$ for all $1 \leq i \neq j \leq 4$.

As mentioned in the opening paragraph of this section, in this section we display results when the regression errors are Gaussian or Gamma. Specifically, we generated homoscedastic errors ϵ as $\text{Normal}(0,1)$ and we generated heteroscedastic errors ϵ such that $[\epsilon \mid \mathbf{X}] = \text{Normal}\left(0, [1 + \{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\alpha}_1 + \alpha_2\boldsymbol{\beta}_1)\}^2]^{3/2}/4\right)$. In the Gamma case, we generated homoscedastic errors ϵ from a Gamma distribution with shape parameter 0.4, scale parameter 1.8 and then normalized it to have mean 0 and variance 1; we generated heteroscedastic errors ϵ using the same distribution except that ϵ was multiplied by $[1 + \{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\alpha}_1 + \alpha_2\boldsymbol{\beta}_1)\}^2]^{3/4}/2$.

To obtain an approximately 4.5% disease rate in both Gaussian and Gamma cases with both homoscedastic and heteroscedastic errors, we first set $\alpha_c = -3.6, \boldsymbol{\alpha}_1 = (-1.0, 0.3, 0.5, 0.7)^{\mathrm{T}}$ and $\alpha_2 = 0.6$ in the logistic model $\text{pr}(D = 1|\mathbf{X}, Y) = H(\alpha_c + \mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}_1 + Y\alpha_2)$. Then we set the regression model for Y to be linear, i.e., $Y = \beta_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1 + \epsilon$ and let $\beta_0 = -1.1, \boldsymbol{\beta}_1 = (0.5, 1.0, 0.3, 0.5)^{\mathrm{T}}$. For each setting, we generated 1,000 simulated data sets.

We set the posited model η_2^* to be $\text{Normal}(0,1)$ and adopted the estimation algorithm discussed in Section 3.3.4 and Appendix B.2 for the important conditional expectations $E_{\text{true}}\{\epsilon^2\kappa(\mathbf{X}, Y) \mid \mathbf{X}\}$, $E_{\text{true}}\{\epsilon\boldsymbol{\mu}_s(\mathbf{X}, Y) \mid \mathbf{X}\}$ and $E_{\text{true}}\left\{\epsilon f_{D|\mathbf{X},Y}(0, \mathbf{X}, Y) \mid \mathbf{X}\right\}$. In steps (1)-(3) in Appendix B.2 that involves nonparametric calculations, we used the asymptotically justified bandwidth $h = cn_0^{-1/5}$: we found that when $c \in [1, 6]$, the estimation results are very similar.

3.5.2 Results

We contrasted three methods. The first one is ordinary least squares using controls only. The second one is the semiparametric efficient method that assumes the regression error ϵ to be normally distributed with homoscedastic variance and $E(Y \mid \mathbf{X})$ to be linear in $\mathbf{X}$, or equivalently in our notation, $m(\mathbf{X}, \boldsymbol{\beta}) = \beta_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1$ (Lin and Zeng, 2009). This method also requires a rare or known disease rate, which was set to 0.1% in the simulations. The third is our method described in Section 3.3.4, which does not require the rare disease assumption and does not put any restriction on ϵ other than that $E(\epsilon|\mathbf{X}) = 0$.

To implement Lin and Zeng's method, we used their software SPREG provided on http://dlin.web.unc.edu/software/spreg-2/, which adopts the rare disease assumption if the input disease rate is less than 1%. This software was designed to work in a semiparametric framework where it assumes a fully parametric Gaussian model for ϵ but the distribution of $\mathbf{X}$ is nonparametric. However, through multiple attempts we found that their software can only handle the case where components of $\mathbf{X}$ are independent. Thus, before running SPREG, we decorrelated $\mathbf{X}$ by multiplying it by L^{-1}, where L is the Cholesky decomposition of the $\mathrm{cov}(\mathbf{X}) = \Sigma$ satisfying $LL^{\mathrm{T}} = \Sigma$. In the simulations, we used the true covariance matrix Σ to fulfill the restriction of SPREG. However when dealing with the mammographic density data in Section 3.6, the true covariance matrix Σ is unknown. We estimated it using only the controls.

The results are summarized in Tables 3.1-3.2. In the homoscedastic Gaussian scenario (Table 3.1), the approach using only controls ("Control") is asymptotically valid with small bias and near nominal coverage. Lin and Zeng's method ("Param"), which assumes normality and homoscedasticity, has the smallest standard deviation among the three methods since it is efficient if the errors are normal. However, it suffers from slight bias since the true disease rate is 4.5%, larger than 1%. Our method ("Semi"), which assumes neither normality nor rare disease, is superior considering overall performance. It has the smallest bias compared with the other two methods. In addition, its mean-squared error efficiency is from 60.0% to 79.9% greater than using only controls and is comparable to Lin and Zeng's method. In the homoscedastic Gamma case (Table 3.2), Lin and Zeng's methods has considerable bias, under-coverage and loss of mean squared error efficiency.

In the heteroscedastic scenario, for both Gaussian and Gamma errors, both the "Controls" and the "Param" methods suffered from low coverage probabilities while our approach ("Semi") maintains nominal coverage. The approach using only controls is reasonably unbiased in the Gaussian case but suffers from much larger bias in the Gamma case. In both cases, Lin and Zeng's parametric method gives badly biased estimates, low coverage probabilities and low mean squared error efficiency. Taking β_{13}, the third element in $\boldsymbol{\beta}_1$, as an example, while the nominal coverage is 95%, the actual coverage rates are 40.6% and 43.7% in the Gaussian and Gamma case, respectively. Our approach has no larger than 4% bias compared with the truth, which is the best among three methods. It also achieves the best coverage probabilities and smallest mean-squared errors.

Remark 2. We have compared our approach to two methods, the control only method and Lin and Zeng's method. The control only method is simple and quick and can work surprisingly well when the disease is truly rare. Lin and Zeng's method is the gold standard in practice. In fact, there are a number of other works on secondary analysis in the literature, however none of them is applicable in our setting. For example, Jiang et al. (2006) and Li et al. (2010) focused on binary Y, for which a logistic regression model between Y and $\mathbf{X}$ (or Y and $(\mathbf{X}, D)$) was considered. Ma and Carroll (2016) adopted kernel density regression in their estimation procedure, and thus it is not applicable to the cases with multivariate $\mathbf{X}$ due to the curse of dimensionality. Wei et al. (2013) requires the rare disease assumption as well as homoscedastic regression errors, and hence is not applicable in our model setting.

3.6 Analysis of Mammographic Density Data

Here we apply our methodology in a case-control study of breast cancer, where the data were collected from women in the breast cancer detection demonstration project (BCDDP), see Chen et al. (2006) and Chen et al. (2008). The study recruited a total of 284,780 women, starting from January 1, 1973 and ended December 31, 1995. Then in the following five years, follow-up annual screening was performed for each subject. Here the period from 1973-1980 is referred to as the "screening phase" of the study. At the end of the screening phase, the study selected all cases, i.e. women who developed breast cancer, and sampled from the controls. All the selected women were included in a further extended follow-up study from 1980 to 1995. Standard risk factors, including age at menarche, age at first live birth and body mass index, were available in this study. However, we were only able to retrieve mammographic density measurements at baseline in 1973-1975 for $N_1 = 2092$ cases and $N_0 = 3295$ controls.

Mammographic density is a measure of the average of dense tissue percentage in both breasts. Women's breasts consist of fat, breast tissue, nerves, veins, arteries and connective tissue that holds everything in place. Both breast tissue and connective tissue are denser than fat. Previous studies showed that higher mammographic density is a strong risk factor for breast cancer. In addition, age at menarche and age at first live birth are both known to be associated with breast cancer. Women who have their first menstruation before age 12 have a slightly higher chance of developing breast cancer compared with those who have their first period after 14; women who give birth to their first child at a young age tend to have a relatively lower risk of developing breast cancer. Body mass index is another risk factor for breast cancer. Before menopause, being slightly overweight can reduce breast cancer risk. However, there is little existing work discussing the interrelationship between mammographic density, age at menarche, age at first live birth and body mass index. The goal of our analysis is to investigate this interrelationship. Before implementing our method, we used an inverse logistic transformation on mammographic density and rescaled the other three risk factors to [0,1] by subtracting their minimums and dividing by the ranges.

Preliminary analysis based on only the controls data showed that mammographic density is reasonably linear in age at menarche, age at first live birth and body mass index. To check this, we fit both a linear regression model and a quadratic regression model using controls and compared these two models via analysis of variance. The p-value is about .78, which indicates the linear model is preferred over the quadratic model. Hence, we adopted a linear $m(\cdot)$ in the secondary analysis. The diagnostic plots of linear regression are given in Figure 3.1. The left plot is the kernel density estimate of the residuals from a linear fit on the controls, with an overlaid normal density. It shows that the regression error almost follows a normal distribution but with slightly negative skewness. The right plot is the LOWESS smoother of fitted values versus the square roots of absolute values of residuals, which indicates the regression error is homoscedastic.

The results of secondary analysis using only controls, Lin and Zeng's parametric method and our semiparametric approach based on 1000 bootstrap samples are given in Table 3.3. All three methods have fairly consistent results as expected, since the regression error is homoscedastic and close to normal. For all three methods, age at first live birth is highly statistically significant with a positive effect on mammographic density. That is women who gave birth to their first children earlier tend to have a lower mammographic density, and hence obtain some protective effect from developing breast cancer. Both age at menarche and body mass index have negative coefficients, which indicates that having a relatively late first period or being moderately overweight can slightly reduce mammographic density. However, neither of them is statistically significant.

As expected, Lin and Zeng's parametric method has a much smaller bootstrap standard deviation compared with the ordinary least squares using only controls, with an average efficiency of 1.60. Here the efficiency is defined as the square of the ratio of bootstrap standard deviation compared with using only controls. Our semiparametric approach, which assumes neither homoscedasticity nor normality, has almost the same bootstrap standard deviation as Lin and Zeng's method. The bootstrap standard errors of Lin and Zeng's parametric approach for age at menarche, age at first live birth and body mass index are 0.131, 0.106, 0.138, respectively, while that of our semiparametric approach are 0.129, 0.107 and 0.137 respectively. The average efficiency of our approach is 1.63, which is even slightly larger than that of Lin and Zeng's method.

3.7 Discussion

We have extended the work of Ma and Carroll (2016) and have overcome the potential dimensionality issue involved in their nonparametric kernel regression. Multivariate kernel regression is avoided by using dimension reduction modeling ideas. We repeat that our work is not about fitting dimension reduction models per se, but to use them in the secondary analysis of case-control studies. Our method makes no assumptions about the regression errors, and we do not need to make a rare disease assumption or require known disease rate.

The dimension reduction assumptions stated in (3.8)-(3.10) are mild in general, see Proposition 1, and are applicable in many practical situations. An interesting topic for future work would be to consider using regularization to further reduce the dimension of $\mathbf{Z}_\beta$ so as to obtain an even more parsimonious model.

Alternative dimension reduction modeling approaches could exist, although it is not easy to identify them based on our preliminary analysis along this line. For example, generalized additive models do not appear to be suitable in the common regression error structures described in Section 3.3.2. For example, in (3.8),

$$E\{\epsilon^2\kappa(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\} = E(\epsilon^2 G[\{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \boldsymbol{\beta})\}(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}} + \epsilon\alpha_2] \mid \mathbf{X}).$$

where G is a function of the logistic distribution function, i.e., a function of several exponential functions. It is not clear that this can be written as a generalized additive model. Even if it can be done, using such a dimension reduction approach will still require careful exploration and new methodology development because off-the-shelf results on generalized additive models may not apply due to the case-control sampling nature.

Finally, in some cases, it might be possible to posit a parametric form for var$(\epsilon \mid \mathbf{X})$. We believe that our approach can be extended to this case, and would further improve efficiency in estimating $\boldsymbol{\beta}$. This will be pursued in future work.

		Homoscedastic Gaussian error				Heteroscedastic Gaussian error			
	$\boldsymbol{\beta}_1$	0.5	1.0	0.3	0.5	0.5	1.0	0.3	0.5
Controls	mean	0.514	0.977	0.280	0.480	0.543	0.940	0.254	0.433
	s.d.	0.113	0.115	0.114	0.111	0.106	0.103	0.100	0.101
	est. sd	0.114	0.113	0.113	0.113	0.102	0.102	0.102	0.102
	95%	0.957	0.941	0.951	0.947	0.922	0.910	0.937	0.900
Param	mean	0.523	0.970	0.273	0.461	0.264	1.257	0.495	0.781
	s.d.	0.082	0.085	0.087	0.084	0.089	0.083	0.088	0.086
	est. sd	0.083	0.084	0.087	0.087	0.089	0.082	0.088	0.088
	95%	0.948	0.942	0.933	0.932	0.250	0.115	0.406	0.101
	MSE Eff	1.759	1.717	1.618	1.484	0.204	0.196	0.263	0.170
Semi	mean	0.507	0.992	0.292	0.493	0.510	0.986	0.289	0.484
	s.d	0.089	0.088	0.086	0.087	0.102	0.093	0.092	0.098
	est. sd	0.091	0.093	0.091	0.094	0.093	0.095	0.089	0.100
	95%	0.960	0.964	0.961	0.975	0.932	0.957	0.936	0.950
	MSE Eff	1.600	1.755	1.799	1.666	1.240	1.606	1.396	1.490

Table 3.1: Simulation study in Section 3.5 with $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, disease rate of approximately 4.5% and 4-dimensional correlated covariates $\mathbf{X}$ over 1,000 simulated data sets. The results for the homoscedastic normal error model are listed on the left and the results for the heteroscedastic normal error model are listed on the right. The three analyses performed are “Controls”, which is ordinary least squares using only controls, “Param”, which is semiparametric efficient method proposed by Lin and Zeng (2009) assuming normality and homoscedasticity, and “Semi”, which is our new estimator described in Section 3.4. Here, we list the sample mean (“mean”), the sample standard deviation (“s.d.”), the mean estimated standard deviation (“est. sd”) and the coverage for the nominal 95% confidence intervals (“95%”) for all three methods. In addition, we computed the mean squared error efficiency compared to using only controls for the “Param” and “Semi” methods.

		Homoscedastic Gamma error				Heteroscedastic Gamma error			
	$\boldsymbol{\beta}_1$	0.5	1.0	0.3	0.5	0.5	1.0	0.3	0.5
Controls	mean	0.522	0.967	0.277	0.470	0.581	0.902	0.228	0.394
	s.d.	0.102	0.101	0.103	0.099	0.086	0.087	0.084	0.090
	est. sd	0.100	0.100	0.100	0.100	0.087	0.087	0.087	0.087
	95%	0.942	0.939	0.934	0.938	0.858	0.782	0.876	0.751
Param	mean	0.630	0.830	0.165	0.301	0.173	1.393	0.585	0.922
	s.d.	0.135	0.135	0.135	0.135	0.144	0.124	0.127	0.137
	est. sd	0.131	0.134	0.135	0.136	0.138	0.127	0.133	0.130
	95%	0.820	0.750	0.831	0.691	0.368	0.124	0.427	0.105
	MSE Eff	0.307	0.239	0.307	0.186	0.110	0.100	0.125	0.098
Semi	mean	0.502	0.995	0.299	0.501	0.513	0.981	0.291	0.482
	s.d	0.068	0.068	0.067	0.068	0.084	0.081	0.073	0.088
	est. sd	0.066	0.068	0.066	0.069	0.087	0.096	0.085	0.105
	95%	0.948	0.958	0.947	0.955	0.948	0.958	0.953	0.946
	MSE Eff	2.314	2.449	2.528	2.345	1.922	2.463	2.261	2.388

Table 3.2: Simulation study in Section 3.5 with $N_1 = 1,000$ cases and $N_0 = 1,000$ controls, disease rate of approximately 4.5% and 4-dimensional correlated covariates $\mathbf{X}$ over 1,000 simulated data sets. The results for the homoscedastic gamma error model are listed on the left and the results for the heteroscedastic gamma error model are listed on the right. The three analyses performed are "Controls", which is ordinary least squares using only controls, "Param", which is semiparametric efficient method proposed by Lin and Zeng (2009) assuming normality and homoscedasticity, and "Semi", which is our new estimator described in Section 3.4. Here, we list the sample mean ("mean"), the sample standard deviation ("s.d."), the mean estimated standard deviation ("est. sd") and the coverage for the nominal 95% confidence intervals ("95%") for all three methods. In addition, we computed the mean squared error efficiency compared to using only controls for the "Param" and "Semi" methods.

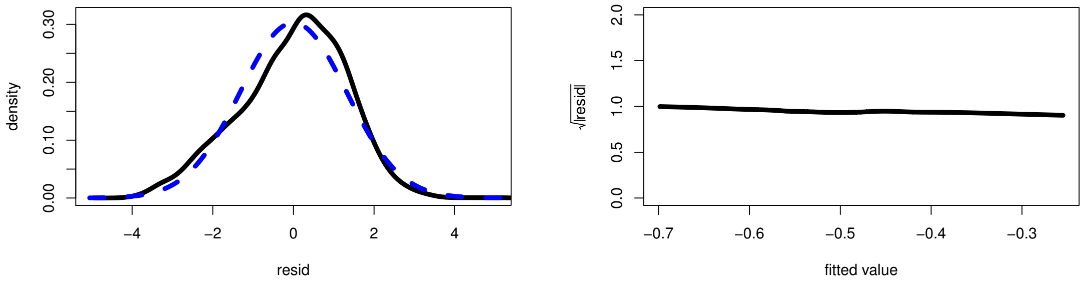

Figure 3.1: Mammographic density data in Section 3.6. The left plot is the kernel density estimate (solid black line) of the residuals from a linear fit on the controls, with an overlaid normal density (dashed blue line). The right plot is the LOWESS smoother of fitted values versus the square roots of absolute values of residuals: the fact that it is flat indicates little heteroscedasticity.

		MENARCHE	1STLB	BMI
Controls	mean	-0.047	0.428	-0.105
	boot. sd	0.164	0.139	0.172
	est. sd	0.165	0.144	0.176
	Lower	-0.371	0.146	-0.449
	Upper	0.277	0.710	0.240
Param	mean	-0.054	0.356	-0.121
	boot. sd	0.131	0.106	0.138
	est. sd	0.127	0.107	0.135
	Lower	-0.302	0.147	-0.385
	Upper	0.195	0.565	0.144
	Eff	1.550	1.710	1.547
Semi	mean	-0.061	0.363	-0.135
	boot. sd	0.129	0.107	0.137
	est. sd	0.130	0.113	0.140
	Lower	-0.315	0.142	-0.410
	Upper	0.194	0.584	0.140
	Eff	1.606	1.698	1.575

Table 3.3: Analyses of the mammographic density data from the breast cancer detection demonstration project (BCDDP) in Section 3.6, which has $N_1 = 2092$ cases and $N_0 = 3295$ controls, using only controls ("Controls"), Lin and Zeng's method ("Param") and our approach ("Semi"). Displayed are the mean estimates of the coefficients for age at menarche (MENARCHE), age at first live birth (1STLB) and body mass index (BMI), their bootstrap standard deviation ("boot. sd"), the mean estimated bootstrap standard deviation ("est. sd") and the lower and upper end values of the 95% confidence intervals ("Lower" and "Upper"). Also displayed is the efficiency ("Eff"), which is the square of the ratio of bootstrap standard deviation to that using only controls.

4. SEMIPARAMETRIC EFFICIENT ESTIMATION IN QUANTILE REGRESSION OF SECONDARY ANALYSIS

4.1 Introduction

The secondary analysis we discussed in Section 3 focus on the association between covariates $\mathbf{X}$ and Y based on the conditional mean function $E(Y \mid \mathbf{X})$. However, in epidemiological studies, it is often of interest to make inference for high or low values in a population distribution. The reason is those high or low values are potentially associated with high risks. For example, people with high body mass index and high blood pressure are at a higher risk of developing diabetes (Wei et al., 2016). In such a case, quantile regression, which provides a complete picture of the relationship between covariates and secondary outcome at any percentile and is robust to the skewed distribution, is preferred over conditional mean regression, which only describes the effect of the covariates on the mean of the secondary outcome and is often sensitive to the misspecification of the regression error distribution.

To our best knowledge, the weighted estimating equation (WEE) approach proposed by Wei et al. (2016) is the first and the only approach on the secondary quantile regression. The main idea of the WEE approach is to construct estimating equations that incorporate both observed and pseudo counter-factual secondary outcomes, where the counter-factual secondary outcomes refer to the pseudo outcomes under alternative disease status. The WEE approach can be further classified into simulated counter-factual outcomes (SICO) approach and kernel smoothing (KS) approach. The SICO approach simulates the counter-factual outcomes directly and assembles the estimating equation, whereas the KS approach replaces the counter-factual part of the estimating equation with its conditional expectation and thus avoids simulating pseudo outcomes. Their simulation result shows SICO approach with 100 replicates has comparable or better performance than KS approach and IPW approach in terms of the mean squared error. Therefore, for comparison presented in this work, we focus on the SICO approach.

The SICO method relies on the following two assumptions: (i) the disease rate in the underlying source population is known, and (ii) the conditional quantile of the secondary outcome is linear in $\mathbf{X}$ for any percentile. However, the disease rate in the source population is not always available. Besides, as shown in the simulation study of Wei et al. (2016), those two assumptions does not help SICO method gain much efficiency compared with the controls only approach.

In this article, we propose to work under the hypothetical population framework (Ma, 2010; Ma and Carroll, 2016). Such a hypothetical population has the same case-to-control ratio as the case-control sample and it allows us to view the case-control sample as a sample of independent identically distributed (i.i.d.) observations taken from this hypothetical population. We further derive a class of semiparametric esimators by imposing a density function of Y given $\mathbf{X}$, which is not necessarily to be the true density. The resulting estimator is consistent. Moreover, it is efficient if the posited density is the truth.

The rest of the section is organized as follows. In Section 4.2, we introduce the secondary quantile regression model and the hypothetical population framework, and provide necessary identifiability conditions. In Section 4.3, we construct the semiparametric efficient estimator through a conventional semiparametric approach. The implementation of the resulting estimator is described in Section 4.4, while its asymptotic properties is discussed in Section 4.5. In Section 4.6, we demonstrate the superiority of our semiparametric efficient estimator over existing approaches via various simulation studies. Section 4.7 illustrates the practical application of our approach through the analysis of a colorectal cancer data set. Section 8 contains a short discussion. Technical details and proofs are given in an Appendix.

4.2 Model and Hypothetical Population Framework

4.2.1 Model

Let disease status be D, with $D = 0$ representing a control and $D = 1$ representing a case. Set $\mathbf{X}$ be the exposures of interest, which can be either continuous, e.g., the first several principle components of the non-genetic variables, or discrete, e.g., SNP information. Other than $\mathbf{X}$, a secondary outcome Y is collected, which can be important biomarkers or characterization of the disease. The disease risk is related to covariates $\mathbf{X}$ and Y through a logistic regression model,

$$\mathrm{pr}^{\mathrm{true}}(D = d \mid \mathbf{X} = \mathbf{x}, Y = y) = H(d, \mathbf{x}, y, \boldsymbol{\alpha}) = \frac{\exp\{d(\alpha_c + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x} + \alpha_2 y)\}}{1 + \exp(\alpha_c + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x} + \alpha_2 y)}. \quad (4.1)$$

Here and throughout the text, the superscript $^{\mathrm{true}}$ is used to denote a model in the underlying source population from which we obtain the case-control sample.

Let $q_{\tau,Y|\mathbf{X}}$ denote the τ^{th} conditional quantile of Y given $\mathbf{X}$. The secondary conditional quantile model that represents the interrelationship between $\mathbf{X}$ and Y in the underlying source population is given by

$$q^{\mathrm{true}}_{\tau,Y|\mathbf{X}}(y, \mathbf{x}, \tau, \boldsymbol{\beta}_\tau) = \beta_{\tau,c} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \quad (4.2)$$

where $\tau \in (0, 1)$. Model (4.2) can be alternatively written as

$$Y = \beta_{\tau,c} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_\tau + \epsilon_\tau,$$

where ϵ_τ has τ^{th} quantile zero, but its distribution is otherwise not specified. That is, $\int u_\tau \eta_2(\epsilon_\tau, \mathbf{x}) d\epsilon_\tau = 0$, where $u_\tau \equiv I(\epsilon_\tau < 0) - \tau$, and $\eta_2(\epsilon_\tau, \mathbf{x})$ denotes the density function of ϵ_τ conditional on $\mathbf{X} = \mathbf{x}$ in the true population, which has a unknown form.

Besides, the distribution of $\mathbf{X}$ is also unspecified.

Suppose we draw a case-control sample with n_1 cases and n_0 controls from the underlying source population model (4.1) - (4.2). That is, we assume that the source population can be split into a disease population and a disease-free population, which we call case subpopulation and control subpopulation, respectively. We then sample n_1 cases randomly from the case subpopulation and n_0 controls randomly from the control subpopulation. Typically, n_1 and n_0 are chosen to be comparable in practice. As a result, the case-to-control ratio in the case-control sample is usually higher than it is in the underlying source population. Subsequently, the association between $\mathbf{X}$ and Y in the case-control sample may differ dramatically from the association in the underlying source population.

To understand this numerically, we set $\boldsymbol{\alpha} = (\alpha_c, \boldsymbol{\alpha}_1, \alpha_2) = (-4.5, 1, 1)$ in the logistic regression model (4.1) and $\tau = 0.5, \boldsymbol{\beta} = (\beta_c, \boldsymbol{\beta}_\tau) = (0, 1)$ in the secondary quantile regression model (4.2). Additionally, we simulate $\mathbf{X}$ from Uniform(0,1) and ϵ_τ from Normal(0,1). The resulting disease rate is about 5%. From this specific setting, we generated 1,000 case-control data sets, each consists of 500 cases and 500 controls. We then ran quantile regression on each of the simulated data sets, ignoring the case-control sampling scheme, and averaged the estimated slope and intercept across all 1,000 simulations. The result is summarized in Figure 4.1. The dashed blue line is the quantile regression relationship between covariate X and secondary outcome Y under the case-control context. It has an intercept of 0.225 and slope of 1.352. The solid black line is the relationship in the underlying source population, which has an intercept of 0 and slope of 1. Clearly, neglecting the retrospective nature of the case-control sample produce bias in both intercept and slope.

As illustrated by the artificial example given above, the main difficulty in secondary analysis is that the case-control sample is not a representative sample of the underlying source population. To conquer this problem, we borrow strength from the hypothetical population framework used in Ma (2010); Ma and Carroll (2016). Such a hypothetical population has the same case-to-control ratio as the case-control sample. It is connected with the true population through the fact that they are identical given the disease status. Ma (2010) proved the first order asymptotic equivalence between the case-control sampling and random sampling in the hypothetical population. Her result permits us to view the case-control sample as a prospective random sample taken from this hypothetical population. The exact form of the hypothetical population for the secondary quantile regression problem we are studying is given in Section 4.2.2.

4.2.2 Hypothetical Population

We consider the case that the disease rate in the true population, $\pi_1 \equiv \text{pr}^{\text{true}}(D = 1)$, is unknown, and it can be rare or common. Let $\pi_0 = 1 - \pi_1$. The goal is to estimate $\boldsymbol{\alpha} = (\alpha_c, \boldsymbol{\alpha}_1^{\text{T}}, \alpha_2)^{\text{T}}$ and $\boldsymbol{\beta} = (\beta_c, \boldsymbol{\beta}_\tau^{\text{T}})^{\text{T}}$. Using the concept of hypothetical population (Ma, 2010), we treat the case-control sample as a random sample from a hypothetical population, where the disease and non-disease ratio is n_1/n_0. For a random observation of the hypothetical population $(\mathbf{X}, Y, D)$, its density function has an explicit form,

$$\begin{aligned}
&f_{\mathbf{X},Y,D}(\mathbf{x}, y, d, \boldsymbol{\beta}_\tau, \boldsymbol{\alpha}, \eta_1, \eta_2) \\
&\quad = f_D(d) f_{\mathbf{X},Y|D}(\mathbf{x}, y, d) = \frac{n_d}{n} f^{\text{true}}_{\mathbf{X},Y|D}(\mathbf{x}, y, d) \\
&\quad = \frac{n_d}{n} \frac{\eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x}) f^{\text{true}}_{D|\mathbf{X},Y}(d, \mathbf{x}, y, \boldsymbol{\alpha})}{\int \eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x}) f^{\text{true}}_{D|\mathbf{X},Y}(d, \mathbf{x}, y, \boldsymbol{\alpha}) d\mu(\mathbf{x})\mu(y)} \\
&\quad = \frac{n_d \eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\text{T}}\boldsymbol{\beta}_\tau, \mathbf{x}) H(d, \mathbf{x}, y, \boldsymbol{\alpha})}{n \int \eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\text{T}}\boldsymbol{\beta}_\tau, \mathbf{x}) H(d, \mathbf{x}, y, \boldsymbol{\alpha}) d\mu(\mathbf{x})\mu(y)},
\end{aligned} \tag{4.3}$$

where $n = n_0 + n_1$, μ denotes a Lebesgue measure for a continuous random variable and a counting measure for a discrete random variable. η_1 and η_2 are the respective probability density/mass function of $\mathbf{X}$ and $Y \mid \mathbf{X}$. Define $\epsilon_\tau = Y - \beta_{\tau,c} - \mathbf{X}^{\text{T}}\boldsymbol{\beta}_\tau$. We have $\eta_1, \eta_2 \geq 0, \int \eta_1(\mathbf{x}) d\mu(\mathbf{x}) = 1, \int \eta_2(\epsilon_\tau, \mathbf{x}) d\mu(\epsilon_\tau) = 1$, and $\int u_\tau \eta_2(\epsilon_\tau, \mathbf{x}) d\mu(\epsilon_\tau) = \tau$. However, both η_1 and η_2 have unknown forms. We use the notation η_1 and η_2 instead of $f_{\mathbf{X}}^{\text{true}}$ and f_Y^{true} to emphasize they are infinite-dimensional nuisance parameters.

Under the hypothetical population framework (4.3), we obtain a locally efficient semiparametric estimator for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using a geometric approach illustrated in Bickel et al. (1993) and Tsiatis (2007). A sketch of the approach is given in Section 4.3 while its technical details are given in Appendix C.4 and C.5.

Remark 3. *It is known that a consistent estimator of $\boldsymbol{\alpha}$ excluding α_c can be obtained via the prospective logistic regression, treating the case-control sample as if it is a random sample. We can then plug in the estimator of $\boldsymbol{\alpha}$ excluding α_c, and focus on estimating $\boldsymbol{\beta}$ and α_c only. Here we choose to treat the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ together instead.*

4.2.3 Identifiability

Before introducing our locally efficient semiparametric estimator, it is useful to study the identifiability of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the hypothetical population. We assume the following conditionds.

Assumption 1. For any $\delta > 0$, there exists $K > 0$ such that $\lim_{x^\ell \to \pm\infty} \text{pr}(\epsilon_\tau < -K \mid \mathbf{x}) < \delta$, where x^ℓ is the ℓ^{th} element of $\mathbf{x}$.

Assumption 2. $\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_\tau \alpha_2 \neq 0$ and $\alpha_2 \neq 0$.

Assumption 1 ensures the left tail of ϵ_τ given $\mathbf{x}$ is not too heavy when an arbitrary element of $\mathbf{x}$ diverges to $\pm\infty$. This is a natural condition to guarantee the mean signal $\beta_c + \mathbf{X}^{\text{T}} \boldsymbol{\beta}_\tau$ can be separated from the noise ϵ_τ. Assumption 2 ensures the logistic regression model (4.1) indeed depends on the value of $\boldsymbol{\beta}$. When it is violated, we can prove $\boldsymbol{\beta}_\tau$ is still identifiable, but β_c and α_c are no longer identifiable. See Appendix C.2 and C.3 for details.

The identifiability result is stated in Proposition 2 below, while its proof is provided in Appendix C.1.

Proposition 2. *Under Assumptions 1-2, the parameter $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_\tau$ are identifiable.*

4.3 Analytical Derivation

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, and $\eta = (\eta_1, \eta_2)$. Define $p = \dim(\boldsymbol{\theta})$. The hypothetical population model (4.3) is a semiparametric model, where $\boldsymbol{\theta}$ is a finite-dimensional parameter we are interested in, and η is an infinite-dimensional nuisance parameter. One general approach to handle such a semiparametric problem is to find a p-dimensional influence function $\phi(\mathbf{X}, Y, D; \boldsymbol{\theta}, \eta)$, i.e., an arbitrary function with mean $E\{\phi(\mathbf{X}, Y, D; \boldsymbol{\theta}, \eta)\} = \mathbf{0}$, and solve the corresponding estimating equation $\sum_{i=1}^{n} \phi(\mathbf{X}_i, Y_i, D_i; \boldsymbol{\theta}, \eta) = \mathbf{0}$. The resulting estimator is a semiparametric estimator with variance var$\{\phi(\mathbf{X}, Y, D; \boldsymbol{\theta}, \eta)\}$. For instance, the score function of the hypothetical population (4.3), $\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S} - E(\mathbf{S} \mid d)$, is a valid influence function that leads to a semiparametric estimator. Here

$$\mathbf{S}(\mathbf{x}, y, d, \boldsymbol{\theta}, \eta_2) = \begin{bmatrix} \partial \log H(d, \mathbf{x}, y, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha} \\ \partial \log \eta_2(y - \beta_c - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x})/\partial \boldsymbol{\beta} \end{bmatrix}, \tag{4.4}$$

with η_2 satisfying $\int \eta_2(\epsilon_\tau, \mathbf{x}) d\epsilon_\tau = 1$ and $E_{\text{true}}(u_\tau \mid \mathbf{x}) = \int u_\tau \eta_2(\epsilon_\tau, \mathbf{x}) d\epsilon_\tau = 0$. In fact, $\mathbf{S}$ is the score function of the underlying source population. Among the class of all semiparametric estimators, the optimal estimator, which is usually referred to as the semiparametric efficient estimator, is the one with smallest variance.

We adopt a geometric approach (Bickel et al., 1993; Tsiatis, 2007) to drive the semiparametric efficient estimator. Specifically, we consider a Hilbert space $\mathcal{H}$ that consists of of all p-dimensional measurable functions with mean zero and finite variance and define the inner product of two arbitrary functions in $\mathcal{H}$ to be their covariance. We then decompose the Hilbert space $\mathcal{H}$ as $\mathcal{H} = \Lambda \oplus \Lambda^{\perp}$, where Λ is the nuisance tangent space and $\Lambda^{\perp}$

is the orthogonal complement of Λ. The semiparametric efficient estimator can be solved from $\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(\mathbf{X}_i, Y_i, D_i; \boldsymbol{\theta}, \eta) = \mathbf{0}$, where $\mathbf{S}_{\text{eff}}$ is the projection of the score function $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$. Consequently, $\mathbf{S}_{\text{eff}}$ is usually called the efficient score function.

Under the hypothetical population model (4.3), the nuisance tangent space has the form

$$\Lambda = [\mathbf{g}(\epsilon_\tau, \mathbf{x}) - E\{\mathbf{g}(\epsilon_\tau, \mathbf{X}) \mid d\} : E_{\text{true}}(\mathbf{g}) = \mathbf{0}, E_{\text{true}}\{u_\tau \mathbf{g}(\epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\} = \mathbf{0}, a.s.],$$

where $\epsilon_\tau = Y - \beta_c - \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_\tau$. Its orthogonal complement is

$$\begin{aligned} \Lambda^{\perp} = & \Big[\mathbf{h}(d, \epsilon_\tau, \mathbf{x}) : E(\mathbf{h}) = \mathbf{0}, E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{x}\} \\ & \times \sum_{d=0}^{1} \frac{n_d H(d, \mathbf{x}, y, \boldsymbol{\alpha})}{n \pi_d} = \mathbf{a}(\mathbf{x}) u_\tau, a.s. \ \forall \mathbf{a} \Big]. \end{aligned}$$

The detailed derivation of Λ and $\Lambda^{\perp}$ is provided in Appendix C.4.

The projection of the score function $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$ is very mathematically involved. Here we list the final form of the efficient score function $\mathbf{S}_{\text{eff}}$, while deferring all the technical details to Appendix C.5. Particularly,

$$\mathbf{S}_{\text{eff}}(\mathbf{X}, Y, D; \boldsymbol{\theta}, \eta) = \mathbf{S}(\mathbf{X}, Y, D; \boldsymbol{\theta}, \eta) - \mathbf{g}(Y - \beta_c - \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}_\tau, \mathbf{X}) - (1 - D)\mathbf{v}_0 - D\mathbf{v}_1,$$

where

$$
\begin{aligned}
&\pi_0 \equiv \text{pr}^{\text{true}}(D=0) = \textstyle\int \eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})H(0, \mathbf{x}, y)d\mu(\mathbf{x})d\mu(y);\\
&\pi_1 \equiv \text{pr}^{\text{true}}(D=1) = \textstyle\int \eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})H(1, \mathbf{x}, y)d\mu(\mathbf{x})d\mu(y);\\
&b_0 \equiv E\{f_{D|\mathbf{X},Y}(1, \mathbf{X}, Y) \mid D=0\}; b_1 \equiv E\{f_{D|\mathbf{X},Y}(0, \mathbf{X}, Y) \mid D=1\};\\
&\mathbf{c}_0 \equiv E(\mathbf{S} \mid D=0) - E\{E(\mathbf{S} \mid \epsilon_\tau, \mathbf{X}) \mid D=0\};\\
&\mathbf{c}_1 \equiv E(\mathbf{S} \mid D=1) - E\{E(\mathbf{S} \mid \epsilon_\tau, \mathbf{X}) \mid D=1\};\\
&\kappa(\mathbf{x}, y) \equiv \left[\textstyle\sum_{d=0}^{1}\{n_d H(d, \mathbf{x}, y)\}/(n\pi_d)\right]^{-1}; t_1(\mathbf{x}) \equiv \left[E_{\text{true}}\left\{u_\tau^2 \kappa(\mathbf{X}, Y) \mid \mathbf{x}\right\}\right]^{-1}; \qquad (4.5)\\
&\mathbf{t}_2(\mathbf{x}) \equiv E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\} - (\mathbf{c}_0/b_0)E_{\text{true}}\left\{u_\tau f_{D|\mathbf{X},Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\};\\
&t_3(\mathbf{x}) \equiv -b_0^{-1}E_{\text{true}}\left\{u_\tau f_{D|\mathbf{X},Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\}; \mathbf{a}(\mathbf{x}) \equiv t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x}) + t_3(\mathbf{x})\mathbf{u}_0\};\\
&\mathbf{u}_0 \equiv \left(1 - E\left[u_\tau t_1(\mathbf{X})t_3(\mathbf{X})\kappa(\mathbf{X}, Y) \mid D=0\right]\right)^{-1} E\left[u_\tau t_1(\mathbf{X})\mathbf{t}_2(\mathbf{X})\kappa(\mathbf{X}, Y) \mid D=0\right];\\
&\mathbf{u}_1 \equiv -(n_0/n_1)\mathbf{u}_0; \mathbf{v}_0 \equiv (\pi_1/b_0)(\mathbf{u}_0 + \mathbf{c}_0); \mathbf{v}_1 \equiv -(\pi_0/b_0)(\mathbf{u}_0 + \mathbf{c}_0);\\
&\mathbf{g}(\epsilon, \mathbf{x}) \equiv E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) - u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, y) - \mathbf{v}_0 f_{D|\mathbf{X},Y}(0, \mathbf{x}, y) - \mathbf{v}_1 f_{D|\mathbf{X},Y}(1, \mathbf{x}, y).
\end{aligned}
$$

The semiparametric efficient estimator is then obtained by solving

$$\textstyle\sum_{i=1}^{N}\{\mathbf{S}(\mathbf{X}_i, Y_i, D_i) - \mathbf{g}(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{X}_i)\} - n_0\mathbf{v}_0 - n_1\mathbf{v}_1 = \mathbf{0}. \qquad (4.6)$$

The estimating equation (4.6) involves the disease and non-disease rate, π_1 and $\pi_0 = 1 - \pi_1$, which are assumed to be unknown. We notice that

$$
\begin{aligned}
\pi_0 &= \int H(0, \mathbf{X}, Y, \boldsymbol{\alpha}) f_{Y|\mathbf{X}}^{\text{true}}(y, \mathbf{x}, \boldsymbol{\beta}) f_{\mathbf{X}}^{\text{true}}(\mathbf{x}) d\mu(\mathbf{x}) d\mu(y)\\
&= \int \frac{H(0, \mathbf{X}, Y, \boldsymbol{\alpha})}{\sum_d N_d/(N\pi_d)H(d, \mathbf{X}, Y, \boldsymbol{\alpha})}\\
&\qquad \times \textstyle\sum_d N_d/(N\pi_d)H(d, \mathbf{X}, Y, \boldsymbol{\alpha}) f_{Y|\mathbf{X}}^{\text{true}}(y, \mathbf{x}, \boldsymbol{\beta}) f_{\mathbf{X}}^{\text{true}}(\mathbf{x}) d\mu(\mathbf{x}) d\mu(y)\\
&= \int \frac{H(0, \mathbf{X}, Y, \boldsymbol{\alpha})}{\sum_d N_d/(N\pi_d)H(d, \mathbf{X}, Y, \boldsymbol{\alpha})} f_{\mathbf{X},Y}(y, \mathbf{x}, \boldsymbol{\beta}) d\mu(\mathbf{x}) d\mu(y)\\
&= E\left[\frac{H(0, \mathbf{X}, Y, \boldsymbol{\alpha})}{n_0 H(0, \mathbf{X}, Y, \boldsymbol{\alpha})/\pi_0 + n_1 H(1, \mathbf{X}, Y, \boldsymbol{\alpha})/(1-\pi_0)}\right]. \qquad (4.7)
\end{aligned}
$$

Moreover, $\pi_0 = \text{pr}^{\text{true}}(D=0)$ is the unique solution to equation (4.7) if $\text{pr}\{H(0, \mathbf{X}, Y, \boldsymbol{\alpha}) > 0\} > 0$. Hence, we can obtain a consistent estimator of π_0 through solving the following estimating equation,

$$\pi_0 = \textstyle\sum_{i=1}^{N} \frac{H(0, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})}{n_0 H(0, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})/\pi_0 + n_1 H(1, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})/(1-\pi_0)}.$$

Denote the resulting estimator by $\widehat{\pi}_0$. We then estimate π_1 by $\widehat{\pi}_1 = 1 - \widehat{\pi}_0$.

4.4 Algorithm

Some other quantities involved in the estimating equation (4.6) depends on the unknown distributions of $\mathbf{X}$ and $Y \mid \mathbf{X}$, i.e., η_1 and η_2. All those quantities can be estimated nonparametrically if we know the exact form of the score function in the underlying source population, i.e., $\mathbf{S}$, which is defined in (4.4). Unfortunately, $\mathbf{S}$ itself also relies on the unknown probability density function η_2. Here, we propose an algorithm based on a posited score function $\mathbf{S}^*$. It is obtained by replacing η_2 with an arbitrary density function with τ^{th} quantile zero in (4.4). The resulting estimator of $\boldsymbol{\theta}$ is locally efficient.

Here is the detailed algorithm.

1. Posit a model for $\eta_2(\epsilon_\tau, \mathbf{x})$ which has τth quantile zero, and calculate (4.4), calling the result $\mathbf{S}^*$.

2. Solve $\widehat{\pi}_0 = \sum_{i=1}^{N} H(0, \mathbf{X}_i, Y_i)\{n_0 H(0, \mathbf{X}_i, Y_i)/\widehat{\pi}_0 + n_1 H(1, \mathbf{X}_i, Y_i)/(1-\widehat{\pi}_0)\}^{-1}$ to obtain $\widehat{\pi}_0$.

3. Set $\widehat{\pi}_1 = 1 - \widehat{\pi}_0$ and

$$\begin{aligned}
\widehat{\kappa}_i &= \widehat{\kappa}(\mathbf{X}_i, Y_i) = \{\textstyle\sum_d n_d H(d, \mathbf{X}_i, Y_i)/(n\widehat{\pi}_d)\}^{-1} \\
\widehat{f}_{0i} &= \widehat{f}_{D|X,Y}(0, \mathbf{X}_i, Y_i) = n_0 H(0, \mathbf{X}_i, Y_i)\widehat{\kappa}_i/(n\widehat{\pi}_0) \\
\widehat{f}_{1i} &= \widehat{f}_{D|X,Y}(1, \mathbf{X}_i, Y_i) = n_1 H(1, \mathbf{X}_i, Y_i)\widehat{\kappa}_i/(n\widehat{\pi}_1) \\
\widehat{\boldsymbol{\mu}}_{si} &= \widehat{E}(\mathbf{S}_i^* \mid \epsilon_{\tau,i}, \mathbf{X}_i) = \textstyle\sum_d n_d H(d, \mathbf{X}_i, Y_i)\mathbf{S}^*(d, \mathbf{X}_i, Y_i)\widehat{\kappa}_i/(n\widehat{\pi}_d) \\
\widehat{b}_0 &= \textstyle\sum_{i=1}^N \widehat{f}_{1i}\widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i} \\
\widehat{b}_1 &= \textstyle\sum_{i=1}^N \widehat{f}_{0i}\widehat{f}_{1i}/\sum_{i=1}^N \widehat{f}_{1i} \\
\widehat{\mathbf{c}}_0 &= \textstyle\sum_{i=1}^N \left\{\mathbf{S}^*(0, \mathbf{X}_i, Y_i) - \widehat{\boldsymbol{\mu}}_{si}\right\} \widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i} \\
\widehat{\mathbf{c}}_1 &= \textstyle\sum_{i=1}^N \left\{\mathbf{S}^*(1, \mathbf{X}_i, Y_i) - \widehat{\boldsymbol{\mu}}_{si}\right\} \widehat{f}_{1i}/\sum_{i=1}^N \widehat{f}_{1i}.
\end{aligned}$$

4. Perform a nonconventional weighted version of the nonparametric regression to form $\widehat{E}_{\text{true}}(u_\tau^2\widehat{\kappa} \mid \mathbf{x}) = \{\sum_d \widehat{\pi}_d/n_d \sum_{i=1}^N I(D_i = d) u_{\tau,i}^2 \widehat{\kappa}_i K_h(\mathbf{X}_i - \mathbf{x})\}/\{\sum_d \widehat{\pi}_d/n_d \sum_{i=1}^N I(D_i = d) K_h(\mathbf{X}_i - \mathbf{x})\}$ and $\widehat{t}_1(\mathbf{x}) = \{\widehat{E}_{\text{true}}(u_\tau^2\widehat{\kappa} \mid \mathbf{x})\}^{-1}$.

5. (a) Perform nonparametric regression using the data $(\mathbf{X}_i, u_{\tau,i}\widehat{\boldsymbol{\mu}}_{si})$ with $D_i = 0$ to obtain $\widehat{E}(u_\tau\widehat{\boldsymbol{\mu}}_s \mid \mathbf{x}, D = 0)$. Similarly, perform nonparametric regression using the data $(\mathbf{X}_i, u_{\tau,i}\widehat{\boldsymbol{\mu}}_{si})$ with $D_i = 1$ to obtain $\widehat{E}(u_\tau\widehat{\boldsymbol{\mu}}_s \mid \mathbf{x}, D = 1)$.

   (b) Form $\widehat{E}_{\text{true}}(u_\tau\widehat{\boldsymbol{\mu}}_s \mid \mathbf{x}) = \sum_d \widehat{\pi}_d \widehat{E}(u_\tau\widehat{\boldsymbol{\mu}}_s \mid \mathbf{x}, D = d)\widehat{f}_{\mathbf{X}|D}(\mathbf{x}, d)/\sum_d \widehat{\pi}_d \widehat{f}_{X|D}(\mathbf{x}, d)$.

6. (a) Perform nonparametric regression using the data $(\mathbf{X}_i, u_{\tau,i}\widehat{f}_{0i})$ with $D_i = 0$ to obtain $\widehat{E}(u_\tau\widehat{f}_0 \mid \mathbf{x}, 0)$. Similarly, perform nonparametric regression using the data $(\mathbf{X}_i, u_{\tau,i}\widehat{f}_{0i})$ with $D_i = 1$ to obtain $\widehat{E}(u_\tau\widehat{f}_0 \mid \mathbf{x}, 1)$.

   (b) Form $\widehat{E}_{\text{true}}(u_\tau\widehat{f}_0 \mid \mathbf{x}) = \sum_d \widehat{\pi}_d \widehat{E}(u_\tau\widehat{f}_0 \mid \mathbf{x}, d)\widehat{f}_{X|D}(\mathbf{x}, d)/\sum_d \widehat{\pi}_d \widehat{f}_{X|D}(\mathbf{x}, d)$.

7. (a) Perform nonparametric regression using the data $(\mathbf{X}_i, u_{\tau,i}\widehat{f}_{1i})$ with $D_i = 0$ to obtain $\widehat{E}(u_\tau\widehat{f}_1 \mid \mathbf{x}, 0)$. Similarly, perform nonparametric regression using the

data $(\mathbf{X}_i, u_{\tau,i}\widehat{f}_{1i})$ with $D_i = 1$ to obtain $\widehat{E}(u_\tau \widehat{f}_1 \mid \mathbf{x}, 1)$.

(b) Form $\widehat{E}_{\text{true}}(u_\tau \widehat{f}_1 \mid \mathbf{x}) = \sum_d \widehat{\pi}_d \widehat{E}(u_\tau \widehat{f}_1 \mid \mathbf{x}, d)\widehat{f}_{X|D}(\mathbf{x}, d)/\sum_d \widehat{\pi}_d \widehat{f}_{X|D}(\mathbf{x}, d)$.

8. (a) Form $\widehat{\mathbf{t}}_2(\mathbf{x}) = \widehat{E}_{\text{true}}(u_\tau \widehat{\boldsymbol{\mu}}_s \mid \mathbf{x}) - (\widehat{\mathbf{c}}_0/\widehat{b}_0)\widehat{E}_{\text{true}}(u_\tau \widehat{f}_0 \mid \mathbf{x})$ and $\widehat{t}_3(\mathbf{x}) = -\widehat{b}_0^{-1} \times \widehat{E}_{\text{true}}(u_\tau \widehat{f}_0 \mid \mathbf{x})$.

(b) Form $\widehat{E}\{u_\tau t_1(\mathbf{x})t_3(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\} = \sum_{i=1}^N u_{\tau,i}\widehat{t}_1(\mathbf{X}_i)\widehat{t}_3(\mathbf{X}_i)\widehat{\kappa}(\mathbf{X}_i, Y_i) \widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i}$, $\widehat{E}\{u_\tau t_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\} = \sum_{i=1}^N u_{\tau,i}\widehat{t}_1(\mathbf{X}_i)\widehat{\mathbf{t}}_2(\mathbf{X}_i) \widehat{\kappa}(\mathbf{X}_i, Y_i)\widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i}$ and $\widehat{\mathbf{u}}_0 = \left[1 - \widehat{E}\left\{u_\tau t_1(\mathbf{x})t_3(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\right\}\right]^{-1} \times \widehat{E}\left\{u_\tau t_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\right\}$.

(c) Form $\widehat{\mathbf{u}}_1 = -(n_0/n_1)\widehat{\mathbf{u}}_0$, $\widehat{\mathbf{v}}_0 = (\widehat{\pi}_1/\widehat{b}_0)(\widehat{\mathbf{u}}_0 + \widehat{\mathbf{c}}_0)$ and $\widehat{\mathbf{v}}_1 = -(\widehat{\pi}_0/\widehat{b}_0)(\widehat{\mathbf{u}}_0 + \widehat{\mathbf{c}}_0)$.

(d) Form $\widehat{\mathbf{a}}(\mathbf{x}) = \widehat{t}_1(\mathbf{x})\{\widehat{\mathbf{t}}_2(\mathbf{x}) + \widehat{t}_3(\mathbf{x})\widehat{\mathbf{u}}_0\}$.

(e) Form $\widehat{\mathbf{g}}_i = \widehat{\boldsymbol{\mu}}_{si} - u_{\tau,i}\widehat{\mathbf{a}}(\mathbf{X}_i)\widehat{\kappa}_i - \widehat{\mathbf{v}}_0\widehat{f}_{0i} - \widehat{\mathbf{v}}_1\widehat{f}_{1i}$.

(f) Form $\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i) = \mathbf{S}^*_i - \widehat{\mathbf{g}}_i - \widehat{\mathbf{v}}_{D_i}$ and solve the corresponding estimating equation.

We point out the algorithm described above is designed for continuous $\mathbf{X}_i$. When $\mathbf{X}_i$ is discrete, one simply replaces the various nonparametric regressions with the corresponding averages associated with the different $\mathbf{x}_i$ values.

4.5 Asymptotics

The asymptotic distribution of the proposed estimator is given in Theorem 4 below, with proof provided in Appendix C.6. We assume the following regularity conditions.

C1: There exists a constant $0 < C < \infty$ such that $\lim_{n\to\infty} n_1/n_2 = C$.

C2: The univariate kernel function is a probability density function with support $(-1, 1)$ and order r, i.e., $\int K(x)x^t dx = 0$ if $1 \le t < r$ and $\int K(x)x^r dx \neq 0$. The d-

dimensional kernel function, still represented with K, is a product of d univariate kernel functions, that is, $K(\mathbf{x}) = \prod_{i=1}^{d} K(x_i)$ for a d-dimensional $\mathbf{x}$.

C3: For $d = 1, 0$, $f_{X|D}(\mathbf{x} \mid D = d), E(u_\tau^2 \kappa \mid \mathbf{X}, D = d), E(u_\tau \mu_s \mid \mathbf{X}, D = d), E(u_\tau f_0 \mid \mathbf{X}, D = d), E(u_\tau f_1 \mid \mathbf{X}, D = d)$ have compact support and have continuous r^{th} derivatives.

C4: The bandwidth $h = n^{-\tau}$ where $1/(2d) > \tau > 1/(4r)$, where d is the dimension of $\mathbf{x}$. This allows the optimal bandwidth $h = O\{n^{-1/(2r+d)}\}$ as long as we choose a kernel of order $2r > d$.

The preceding regularity conditions are typical assumptions to ensure the consistency of the nonparametric estimators built in Section 4.4 and the subsequent semiparametric estimator of $\boldsymbol{\theta}$. Specifically, condition C1 is a general assumption in all case-control studies of the type we are considering. It ensures the number of cases and controls are comparable in the case-control sample. Condition C2 and C4 are standard requirements on kernel function K and bandwidth h. Condition C3 is the smoothness assumption on the functions that are needed to be estimated nonparametrically.

Theorem 4. Under the regularity conditions C1-C4 listed above, the estimator $\widehat{\boldsymbol{\theta}}$ obtained from solving the estimating equation $\sum_{i=1}^{n} \widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \widehat{\boldsymbol{\theta}}) = 0$ satisfies

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \text{Normal}\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{A}^{-1}\right)^{\text{T}}\}$$

when $n \to \infty$, where $\mathbf{A} = E\left\{\partial \mathbf{S}^*_{\text{eff}}(D, \mathbf{X}, Y, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\text{T}}\right\}$ and $\mathbf{B} = \text{cov}\left\{\mathbf{S}^*_{\text{eff}}(D, \mathbf{X}, Y, \boldsymbol{\theta})\right\}$.

4.6 Simulation Study

In this section, we study the finite sample performance of our estimator via various simulations. Specifically, we considered the following 54 simulation settings. First, we consider a relatively rare disease rate of 4.5%, an extremely rare disease rate of 1%, and a common disease rate of 10%. Second, we set the secondary model to be $Y = \beta_0 + \beta_1 X + \epsilon_\tau$, where $\beta_0 = 0.5$ and $\beta_1 = 1$, and generate X from a Uniform(0,1) distribution. In the secondary model, we consider three different quantiles τ = 0.25, 0.5, and 0.75, corresponding to first quartile, median, and third quartile, respectively. In addition, we consider three different regression error distributions. Specifically, we first generate $\epsilon^* \sim F$, and then set the regression error $\epsilon = \{\epsilon^* - F^{-1}(\tau)\}(1 + X^2)^{3/4}/2$, where F is the standard normal distribution Normal$(0, 1)$, or the standardized Gamma distribution with shape parameter 0.4. Third, the logistic regression model was pr$(D = 1|Y, X) = H(\alpha_c + \alpha_1 X + \alpha_2 Y)$, where $\alpha_1 = 1$ and α_2 = 0.50. The intercept α_c is chosen to achieve specific disease rates given above.

For each setting, we simulated 1,000 data sets, each with 1,000 cases and 1,000 controls, and applied the algorithm discussed in Section 4.4. Specifically, we set the posited model for η_2 to be Normal$\{-\Phi^{-1}(\tau), 1\}$, where Φ is the distribution function of the standard normal distribution. It is easy to check the τ^{th} quantile of the posited model is zero, and the second element in S* has a simple and clear form, $\{y - \beta_c - \beta_\tau x + \Phi^{-1}(\tau)\}(1, x)^{\mathrm{T}}$. Note that the posited model for η_2 is misspecified because we simulate the regression error from normal distribution and gamma distribution with heteroscedastic variance. When performing the nonparametric regressions, we use a bandwidth $h = cn_0^{-1/3}$, where c is a constant and n_0 is the number of controls. We have tested the performance of our estimator using different values of c between 0.5 and 1.5 and found out the results are similar. Here we only report the result with $c = 1$.

We contrasted our locally efficient semiparametric approach, refered to as “Semi”, with two methods. The first one is the ordinary quantile regression on only controls, which we refer to as “Controls”. It produces estimators with negligible bias when the disease rate is rare. The second one is the SICO approach by Wei et al. (2016), which forms a weighted estimating equation by combining both observed secondary outcomes and unobserved counter-factual secondary outcomes simulated under an alternative disease status. Because simulating counter-factual secondary outcomes brings extra variability into the SICO estimator, Wei et al. (2016) suggested to stabilize their estimator by replication. Here we use 100 replicates in all the simulations. Besides, the SICO approah requires a known disease rate. We passed the true disease rate, i.e., 1% for the cases with an extremely rare disease rate, 4.5% for the cases with a relatively rare disease rate, and 10% for the cases with a common disease rate, to the SICO approach, and refered to the resulting estimator as “SICO, true”. To check the robustness of SICO approach to the misspecification of the disease rate, we also passed a rare disease rate of 1% to the SICO approach for the cases with a true disease rate of 4.5% and 10%. The resulting estimator is refered to as “SICO, rare”.

The results are summarized in Table 4.1-4.3. We display three key features, i.e., the mean estimates, the standard deviation across the simulation, and the mean squared error efficiency (MSE Eff) of SICO approach and our semiparametric approach relative to using only controls. A notable finding is, our locally efficient semiparametric estimator, which does not assume a known or rare disease rate, shows dominating advantages over the controls only approach and SICO approach in terms of the mean squared error. Besides, the SICO rare approach suffers from inflated bias and variability when the true disease rate is 10%.

Specifically, in the case of an extremely rare disease of 1% (Table 4.1), all three approaches showed asymptotical consistency with small bias. However, the SICO approach using the true disease rate has a maximum MSE Eff of 1.083 and 1.052 for the cases with heteroscedastic normal error and gamma error, respectively. That is, the efficiency gain of SICO approach over the controls only approach is marginal. In contrast, our approach has a minimum MSE Eff of 3.194 for the heteroscedastic normal case, and a minimum MSE

Eff of 2.680 for the heteroscedastic gamma case.

As the disease rate increased to 4.5% and 10% (Table 4.2 - 4.3), the controls only approach showed moderate bias, epecially for the slope β_τ, so is the SICO rare approach. Meanwhile, the SICO true approach and our semiparametric approach remained consistent. Besides, compared to the controls only approach, the SICO rare approach has similar efficiency and the SICO true approach has slightly better efficiency. In contrast, our semiparametric estimator remains two to five times more efficient than the controls only approach. Take the case with 10% disease rate, quantile $\tau = 0.75$, and heteroscedastic gamma regression error as an example and focus on the estimates of β_τ. Both the controls only estimator and the SICO rare estimator have more than 7% bias, while the SICO true estimator and our semiparametric estimator have no more than 2% bias. Moreover, the SICO rare and SICO true estimator have a respective MSE Eff of 1.078 and 1.718, whereas our semiparametric has a MSE Eff of 5.605.

4.7 Real Data Analysis

The consumption of red meat, e.g., beef, is known to be positively associated with colorectal cancer. Although red meat can provide necessary nutritions such as protein, vitamins, and minerals, it can also produce MeIQx, a carcinogenic heterocyclic amine (HCA), if cooked at high temperatures for long duration. Besides, red meat may also be connected with colorectal cancer through other nutrition facts such as saturated fat and cholesterol. Analyzing the relationship between red meat and MeIQx may yield valuable insights about the etiology of colorectal cancer.

In this section, we apply our locally efficient semiparametric approach to a case-control data set of colorectal adenoma, taken from a large population-based cohort study, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). The PLCO cohort study recruited a total of 33,971 participants, 10% of which were cases, i.e., participants developed at least one histologically verified colorectal adenoma (Peters et al., 2003). $n_1 = 640$ cases and $n_0 = 665$ controls were randomly taken from the case subgroup and control subgroup of this cohort study, respectively, and formed a case-control data set.

In our analysis, we set X to be the red meat consumption in grams and Y to be the MeIQx produced during the cooking in nanograms per gram of red meat. Because both X and Y are heavily skewed in their original measurement scales, we transformed X and Y by first adding 1.0 and taking logrithms, and then dividing their respective standard deviations. We fit model (4.1) and (4.2) to this case-control data set with $\tau = 0.5$ using three approaches, i.e., controls only approach, SICO approach, and our semiparamtric approach described in Section 4.4. Although the true disease rate π_1 is known, we feed the SICO approach with an extremely rare disease rate of 1%, a rare disease rate of 4.5%, and a true disease rate of 10%. In contrast, our locally efficient semiparametric approach is applied assuming the disease rate is unknown.

The result is summarized in Table 4.4. All three approaches showed there is a positive association between red meat consumption and the MeIQx, or equivalently, high assumption of red meat leads to high intake of MeIQx. Besides, the SICO estimator with 1% disease rate has marginal improvement in efficiency compared with controls only estimator, in line with expectations. Here efficiency is defined as the ratio of variance relative to the controls only estimator. While the inputted disease rate increases, the efficiency of the SICO approach gets enhanced and it achieves the maximum efficiency at the true disease rate $\pi_1 = 10\%$. In comparison, our semiparametric estimator, which assume the disease rate to be unknown, has the greatest efficiency among all three estimators.

4.8 Discussion

The quantile regression is often preferred in epidemiology, especially when it is of interest to make inference about the high or low values of the population. In this article, we considered the secondary quantile regression problem with minimal model assumptions. Particular, we only specify a linear relationship between covariates at a given quantile in the secondary model, while the covariate distribution is modeled completely nonparametric. We showed that despite of these weak assumptions, the problem is identifiable excluding a few cases. Under those weak assumptions, we developed a class of consistent semiparametric estimators and identified the most efficient member by adopting the hypothetical population framework and viewing the case-control sample as a prospective random sample taken from the hypothetical population (Ma, 2010). The utilization of the hypothetical population permits the application of a conventional semiparametric approach; however, the derivation is highly non-standard and non-trivial. The superiority of our semiparametric estimator is demonstrated both theoretically and numerically.

The implementation of algorithm discussed in Section 4.4 involves several nonparametric regression, which meets the curse of dimensionality when the dimension of covariates increases. One possible future work is employing dimension reduction techniques such as single index model or B-spline in the secondary quantile regression model.

Another compelling direction is to further improve the efficiency of our locally efficient semiparametric estimator by imposing certain parametric structure on the regression error ϵ_τ, say $\epsilon_\tau = \mathbf{X}^{\mathrm{T}}\zeta\epsilon_\tau^*$, where ϵ_τ^* has τ^{th} quantile zero. The general methodology of our approach can definitely be extended here, but the asymptotic property of the resulting estimator would need to be re-established.

			$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		$\boldsymbol{\beta}$	0.5	1.0	0.5	1.0	0.5	1.0
Normal	Controls	mean	0.501	0.988	0.501	0.988	0.501	0.987
		s.d.	0.063	0.119	0.056	0.107	0.063	0.119
	SICO, true	mean	0.502	0.992	0.501	0.993	0.501	0.994
		s.d	0.062	0.116	0.055	0.105	0.061	0.115
		MSE Eff	1.048	1.056	1.043	1.060	1.062	1.083
	Semi	mean	0.496	1.002	0.498	1.003	0.500	1.003
		s.d	0.034	0.055	0.032	0.053	0.032	0.051
		MSE Eff	3.368	4.764	3.194	4.180	3.807	5.520
Gamma	Controls	mean	0.500	1.000	0.502	0.994	0.502	0.990
		s.d.	0.006	0.012	0.021	0.042	0.056	0.114
	SICO, true	mean	0.500	1.000	0.501	0.997	0.502	0.998
		s.d	0.005	0.011	0.021	0.041	0.055	0.111
		MSE Eff	1.042	1.042	1.030	1.046	1.047	1.052
	Semi	mean	0.499	1.000	0.501	1.000	0.504	1.004
		s.d	0.003	0.005	0.013	0.023	0.032	0.056
		MSE Eff	3.828	4.954	2.680	3.288	3.119	4.121

Table 4.1: Simulation study in Section 4.6 with $n_1 = 1,000$ cases and $n_0 = 1,000$ controls, and a disease rate of approximately 1% over 1,000 simulated data sets. The results for the heteroscedastic normal error model are listed on the top and the results for the heteroscedastic gamma error model are listed at the bottom. The three analyses performed are "Controls", which is quantile regression using only controls, "SICO, true", which is simulated counter-factual outcomes approach proposed by Wei et al. (2016) assuming the true disease rate is known, and "Semi", which is our new estimator described in Section 4.4. Here, we list the sample mean ("mean"), the sample standard deviation ("s.d."), and the mean squared error efficiency compared to using only controls ("MSE Eff").

			$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		$\boldsymbol{\beta}$	0.5	1.0	0.5	1.0	0.5	1.0
Normal	Controls	mean	0.499	0.976	0.498	0.974	0.500	0.967
		s.d.	0.064	0.120	0.058	0.106	0.064	0.119
	SICO, true	mean	0.499	0.999	0.499	0.999	0.501	0.996
		s.d	0.061	0.113	0.056	0.101	0.061	0.112
		MSE Eff	1.092	1.170	1.082	1.176	1.111	1.230
	SICO, rare	mean	0.499	0.981	0.498	0.980	0.500	0.974
		s.d	0.062	0.117	0.057	0.105	0.063	0.116
		MSE Eff	1.043	1.067	1.034	1.058	1.042	1.082
	Semi	mean	0.496	1.008	0.497	1.006	0.500	1.002
		s.d	0.036	0.059	0.033	0.053	0.032	0.052
		MSE Eff	3.137	4.245	3.144	4.273	4.046	5.675
Gamma	Controls	mean	0.501	0.999	0.501	0.990	0.500	0.970
		s.d.	0.006	0.011	0.021	0.043	0.053	0.107
	SICO, true	mean	0.501	1.000	0.500	0.999	0.499	1.002
		s.d	0.005	0.011	0.020	0.042	0.051	0.104
		MSE Eff	1.071	1.095	1.065	1.120	1.063	1.163
	SICO, rare	mean	0.501	0.999	0.501	0.992	0.500	0.976
		s.d	0.005	0.011	0.021	0.042	0.052	0.105
		MSE Eff	1.041	1.053	1.034	1.049	1.040	1.073
	Semi	mean	0.499	1.000	0.501	1.000	0.502	1.008
		s.d	0.003	0.005	0.013	0.024	0.029	0.050
		MSE Eff	3.601	4.620	2.549	3.347	3.335	4.789

Table 4.2: Simulation study in Section 4.6 with $n_1 = 1,000$ cases and $n_0 = 1,000$ controls, and a disease rate of approximately 4.5% over 1,000 simulated data sets. The results for the heteroscedastic normal error model are listed on the top and the results for the heteroscedastic gamma error model are listed at the bottom. The four analyses performed are "Controls", which is quantile regression using only controls, "SICO, true", which is simulated counter-factual outcomes approach proposed by Wei et al. (2016) assuming the true disease rate is known, "SICO, rare", which is simulated counter-factual outcomes approach assuming a rare disease rate of 1%, and "Semi", which is our new estimator described in Section 4.4. Here, we list the sample mean ("mean"), the sample standard deviation ("s.d."), and the mean squared error efficiency compared to using only controls ("MSE Eff").

			$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		$\boldsymbol{\beta}$	0.5	1.0	0.5	1.0	0.5	1.0
Normal	Controls	mean	0.501	0.941	0.494	0.948	0.496	0.942
		s.d.	0.062	0.118	0.058	0.108	0.061	0.115
	SICO, true	mean	0.502	0.992	0.497	1.001	0.500	1.000
		s.d	0.058	0.106	0.053	0.096	0.056	0.101
		MSE Eff	1.166	1.538	1.183	1.556	1.205	1.626
	SICO, rare	mean	0.500	0.947	0.494	0.953	0.496	0.948
		s.d	0.061	0.115	0.057	0.106	0.060	0.113
		MSE Eff	1.050	1.084	1.038	1.081	1.038	1.078
	Semi	mean	0.498	1.009	0.499	1.006	0.500	1.007
		s.d	0.033	0.056	0.031	0.048	0.032	0.049
		MSE Eff	3.480	5.523	3.502	6.211	3.696	6.867
Gamma	Controls	mean	0.500	0.997	0.502	0.976	0.502	0.921
		s.d.	0.005	0.010	0.021	0.041	0.054	0.105
	SICO, true	mean	0.500	1.000	0.502	0.992	0.501	0.980
		s.d	0.005	0.010	0.020	0.039	0.051	0.098
		MSE Eff	1.110	1.225	1.105	1.442	1.108	1.718
	SICO, rare	mean	0.500	0.997	0.502	0.978	0.502	0.927
		s.d	0.005	0.010	0.020	0.040	0.053	0.103
		MSE Eff	1.039	1.061	1.035	1.070	1.032	1.078
	Semi	mean	0.499	1.001	0.502	0.999	0.503	1.007
		s.d	0.003	0.005	0.012	0.023	0.030	0.055
		MSE Eff	3.188	4.111	2.774	4.277	3.144	5.605

Table 4.3: Simulation study in Section 4.6 with $n_1 = 1,000$ cases and $n_0 = 1,000$ controls, and a disease rate of approximately 10% over 1,000 simulated data sets. The results for the heteroscedastic normal error model are listed on the top and the results for the heteroscedastic gamma error model are listed at the bottom. The four analyses performed are "Controls", which is quantile regression using only controls, "SICO, true", which is simulated counter-factual outcomes approach proposed by Wei et al. (2016) assuming the true disease rate is known, "SICO, rare", which is simulated counter-factual outcomes approach assuming a rare disease rate of 1%, and "Semi", which is our new estimator described in Section 4.4. Here, we list the sample mean ("mean"), the sample standard deviation ("s.d."), and the mean squared error efficiency compared to using only controls ("MSE Eff").

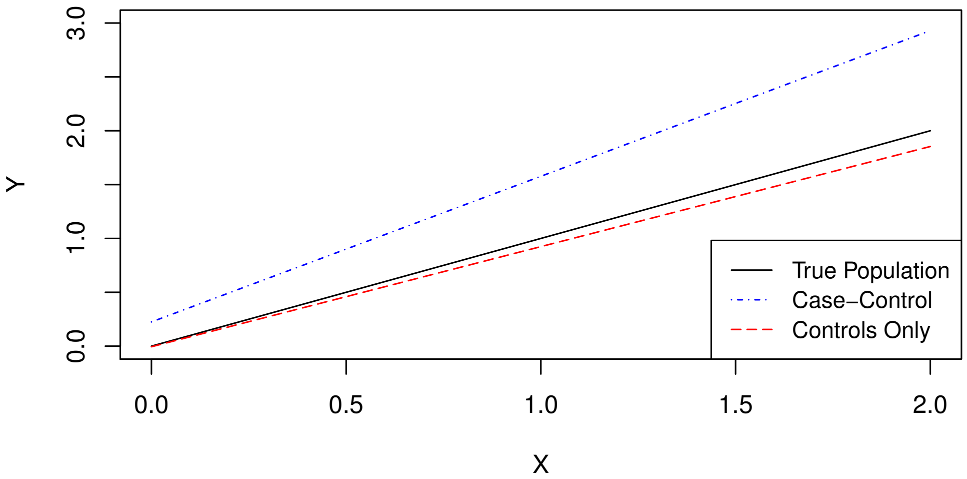

Figure 4.1: Artificial example discussed in Section 4.2.1. The solid black line is the relationship between covariate X and secondary outcome Y in the underlying source population; the dashed blue line is the relationship under the case-control context; the dashed red line is the regression function among only controls.

	β_0			β_X		
	mean	sd	Eff	mean	sd	Eff
Controls	-3.607	0.194	1.000	0.732	0.039	1.000
SICO, 1%	-3.605	0.185	1.098	0.732	0.037	1.087
SICO, 4.5%	-3.595	0.178	1.185	0.730	0.036	1.175
SICO, 10%	-3.583	0.170	1.306	0.728	0.034	1.303
Semi	-3.529	0.159	1.494	0.721	0.031	1.537

Table 4.4: Results of the secondary analysis of the colorectal adenoma data set discussed in Section 4.7 across 1,000 bootstrap samples: “Controls" is the quantile regression on only controls, “SICO" is the simulated counter-factual outcomes approach by Wei et al. (2016), and “Semi" is our locally efficient semiparametric approach. The “SICO" approach is fitted using three different disease rates, 1%, 4.5%, and 10%, while the “Controls" and “Semi" approaches are fitted without specifying the disease rate. Mean (“mean"), sample standard deviation (“sd"), and the square of the ratio of the sample standard deviation compared to controls only approach (“Eff") are reported.

5. CONCLUSION

Semiparametric theory framework is of substantial value in dealing with a vast majority of statistical problems, where the objective is to estimate a finite-dimensional parameter, which is practically important, in the presence of an infinite-dimensional nuisance parameter, which is often complex and of no interest. Three semiparametric models generally used in primary and secondary analysis of case-control studies are considered in this dissertation, i.e., (a) the gene-environment interaction model under independence assumption, (b) the secondary conditional mean regression model, and (c) the secondary conditional quantile regression model. The direct use of the semiparametric theory, which works for i.i.d. samples, in case-control studies, where the samples are biased and taken in a retrospective way, is invalid. We extend the semiparametric framwork to the case-control studies and derive novel semiparametric estimators by adopting a hypothetical population framework and viewing the case-control sample as a prospective random sample generated from the hypothetical population.

Prior to this work, existing approaches exploit gene-environment independence assumption in the gene-environment interaction model either restrict the disease rate to be rare or require distributional assumptions on genetic variables, and hence are not applicable when those assumptions are violated. In this work, we obtain the efficient semiparametric estimator, i.e., a semiparametric estimator that is root-N consistent and asymptotically normal with minimal variance, under a flexible semiparametric model that allows the disease rate to be unknown and the distribution of both genetic and environment variables to be unspecified. Besides, we provide the asymptotic variance of our semiparametric efficient estimator, which can be used for inference. Various simulations with a wide range of disease rates and a group of distinct covariate distributions reveal the eminence of our approach compared with the prospective logistic regression. Moreover, in the real data analysis in Section 2.6, our method detects statistically significant interaction between [25(OH)D] level and VDR gene on the prostate cancer while the prospective logistic regression concludes the interaction term is not significant. Future work entails exploring dimensional reduction approaches, such as single-index modeling and B-spline, to deal with the dimensionality problem arised in nonparametric regression.

In the secondary conditional mean regression model, we devise a dimension reduction technique to bypass the curse of dimensionality involved in the nonparametric regression of the efficient estimation. The dimension reduction assumptions we made are mild in general and are applicable in various practical situations. Our approach maintains the model flexibility as we make no assumption about the regression errors and allow the disease rate to be unknown. Compared to the semiparametric approach by Lin and Zeng (2009) that assumes a known or rare disease rate and requires the regression error to be normally distributed with homoscadistic variance, our approach shows similar performance when those parametric assumptions holds, and it performs considerably better than Lin and Zeng (2009) when those parametric assumptions are violated. An interesting direction for future work would be to further reduce the dimension of the covariates using regularization and to further improve the estimation efficiency by positing a parametric form, e.g., single-index model, for the conditional variance of the regression error.

Finally, we derive the locally consistent semiparametric estimators for the secondary quantile regression model, where only a quantile regression model between covariates is specified and both the disease rate and the covariate distribution are assumed to be unknown. Our method involves positing a density function for Y given $\mathbf{X}$ that may or may not be true. The resulting estimator is consistent for an arbitrary posited density and it is further efficient if the posited model is the truth. Our method outperforms the weighted estimating equation approach (Wei et al., 2016), the only published approach for secondary quantile regression, in terms of efficiency. Similar to the gene-environment independent model, the curse of dimensionality emerges in the estimating process, specifically, the nonparametric regression part. Dimension reduction would be a possible future work. Another appealing topic is exploring the efficiency gain by further assuming a parametric form for the regression error as discussed in the secondary mean regression model.

REFERENCES

Aly, M., Wiklund, F., Xu, J., Isaacs, W. B., Eklund, M., D'Amato, M., Adolfsson, J., and Grönberg, H. (2011). Polygenic risk score improves prostate cancer risk prediction: Results from the stockholm-1 cohort study. *European Urology*, 60, 21–28.

Andriole, G. L., Crawford, E. D., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Isaacs, C., Kvale, P. A., Reding, D. J., et al. (2012). Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: Mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, 104, 125–132.

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

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

Chatterjee, N., Chen, Y.-H., Luo, S., and Carroll, R. J. (2009). Analysis of case-control association studies: SNPs, imputation and haplotypes. *Statistical Science*, 24, 489–502.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17, 392–406.

Chen, J., Ayyagari, R., Chatterjee, N., Pee, D. Y., Schairer, C., Byrne, C., Benichou, J., and Gail, M. H. (2008). Breast cancer relative hazard estimates from case–control and cohort designs with missing data on mammographic density. *Journal of the American Statistical Association*, 103, 976–988.

Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J., and Gail, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute*, 98, 1215–1226.

Chen, Y. H., Chatterjee, N., and Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104, 220–233.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 135–148.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9, e1003348.

Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18, 3525–3531.

Fuchsberger, C., Flannick, J., Teslovich, T. M., et al. (2016). The genetic architecture of type 2 diabetes. *Nature (doi: 10.1038/nature18642)*, .

Gauderman, W. J., Zhang, P., Morrison, J. L., and Lewinger, J. P. (2013). Finding novel genes by testing G$\times$ E interactions in a Genome-Wide Association Study. *Genetic Epidemiology*, 37, 603–613.

Han, S. S., Rosenberg, P. S., Ghosh, A., Landi, M. T., Caporaso, N. E., and Chatterjee, N. (2015). An exposure-weighted score test for genetic associations integrating environmental risk factors. *Biometrics*, 71, 596–605.

Hayes, R. B., Reding, D., Kopp, W., Subar, A. F., Bhat, N., Rothman, N., Caporaso, N., Ziegler, R. G., Johnson, C. C., Weissfeld, J. L., et al. (2000). Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21, 349S–355S.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6, 287–298.

Jiang, Y., Scott, A. J., and Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine*, 25, 1323–1339.

Li, H., Gail, M. H., Berndt, S., and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*, 34, 427–433.

Lian, H., Liang, H., and Carroll, R. J. (2015). Variance function partially linear single-index models. *Journal of the Royal Statistical Society: Series B*, 77, 171–194.

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

Ma, Y. (2010). A semiparametric efficient estimator in case-control studies. *Bernoulli*, 16, 585–603.

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. and Zhu, L. (2012). Efficiency loss caused by linearity condition in dimension reduction. *Biometrika*, 99, 1–13.

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

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169, 219–226.

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

Ottman, R. (1996). Gene–environment interaction: Definitions and study designs. *Preventive Medicine*, 25, 764.

Peters, U., Sinha, R., Chatterjee, N., Subar, A. F., Ziegler, R. G., Kulldorff, M., Bresalier, R., Weissfeld, J. L., Flood, A., Schatzkin, A., et al. (2003). Dietary fibre and colorectal adenoma in a colorectal cancer early detection programme. *The Lancet*, 361, 1491–1495.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*, 13, 153–162.

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

Prorok, P. C., Andriole, G. L., Bresalier, R. S., Buys, S. S., Chia, D., Crawford, E. D., Fogel, R., Gelmann, E. P., Gilbert, F., Hasson, M. A., et al. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21, 273S–309S.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., Ruderfer, D. M., McQuillin, A., Morris, D. W., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460, 748–752.

Siegel, R. and Jemal, A. (2015). *Cancer Facts & Figures 2015*. American Cancer Society Atlanta, Ga, USA.

Tchetgen, E. J. T. (2014). A general regression framework for a secondary outcome in case–control studies. *Biostatistics*, 15, 117–128.

Tsiatis, A. (2007). *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.

Wei, J., Carroll, R. J., Müller, U. U., Van Keilegom, I., and Chatterjee, N. (2013). Robust estimation for homoscedastic regression in the secondary analysis of case–control data. *Journal of the Royal Statistical Society, Series B*, 75, 185–206.

Wei, Y., Song, X., Liu, M., Ionita-Laza, I., and Reibman, J. (2016). Quantile regression in the secondary analysis of case–control data. *Journal of the American Statistical Association*, 111, 344–354.

APPENDIX A

SKETCH OF TECHNICAL ARGUMENTS FOR SECTION 2

A.1 Identifiability

A1 There exists c_x so that when $x \to \mathbf{c}_x$, $m(g, x, \boldsymbol{\beta}) \to \infty$ or $m(g, x, \boldsymbol{\beta}) \to -\infty$ for any g.

A2 There exists g_1 and x_1, x_2 such that $m(g_1, x_1, \beta) \neq m(g_1, x_2, \beta)$.

A3 There exists c_g so that when $g \to c_g$, $m(g, x, \boldsymbol{\beta}) \to \infty$ or $m(g, x, \boldsymbol{\beta}) \to -\infty$ for any x.

A4 There exists x_1 and g_1, g_2 such that $m(g_1, x_1, \beta) \neq m(g_2, x_1, \beta)$.

Proposition 3. *The problem stated in (2.2) is identifiable,*

- *If condition A1 holds, and at least one of the conditions A3 and A4 holds;*
- *or if at least one of the conditions A1 and A2 holds, and condition A3 holds.*

Remark 4. *In practice, a widely used model is the one including main effects and two-way interaction, i.e.,* $m(g, x, \boldsymbol{\beta}) = \alpha + \beta_1 g + \beta_2 x + \beta_3 xg$. *It can be easily verified that if* g *and* x *both have the support on* $\mathbb{R}$ *then this model satisfies conditions A1 and A3 described above and hence is identifiable.*

Remark 5. *Proposition 3 applies in the case where at most one of G and X is discrete. In the case where both G and X are discrete with levels* l_G *and* l_X *respectively, identifiability requires* $l_G l_X \geq 2l_G + 2l_X - 2$ *as a necessary condition. Additional conditions may be needed. Although for a specific model with known* l_G *and* l_X*, it can be easy to derive the sufficient conditions for identifiability, such result is difficult to describe in general.*

Proof of Proposition 3. From Prentice and Pyke (1979), $\boldsymbol{\beta}$ is identifiable. Thus, we aim at establishing the identifiability of η_1, η_2 and α.

We first prove the result under A1 and A3. Assume there are α, η_1, η_2 and $\alpha^*, \eta_1^*, \eta_2^*$, so that

$$\frac{n_d}{n\pi_d}\eta_1(g)\eta_2(x)H(d, g, x, \boldsymbol{\beta}, \alpha) = \frac{n_d}{n\pi_d^*}\eta_1^*(g)\eta_2^*(x)H(d, g, x, \boldsymbol{\beta}, \alpha^*).$$

This yields

$$\begin{aligned}\frac{1}{\pi_1}\eta_1(g)\eta_2(x)H(1, g, x, \boldsymbol{\beta}, \alpha) &= \frac{1}{\pi_1^*}\eta_1^*(g)\eta_2^*(x)H(1, g, x, \boldsymbol{\beta}, \alpha^*),\\ \frac{1}{\pi_0}\eta_1(g)\eta_2(x)H(0, g, x, \boldsymbol{\beta}, \alpha) &= \frac{1}{\pi_0^*}\eta_1^*(g)\eta_2^*(x)H(0, g, x, \boldsymbol{\beta}, \alpha^*).\end{aligned}$$

Taking the ratio of the above two and solving, we obtain $\exp(\alpha^*) = \exp(\alpha)\pi_0\pi_1^*/(\pi_1\pi_0^*)$. This leads to

$$\frac{\eta_2^*(x)}{\eta_2(x)}\frac{\eta_1^*(g)}{\eta_1(g)} = \frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g, x, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g, x, \boldsymbol{\beta})\}}.$$

Under condition A1, letting $x \to \mathbf{c}_x$, we obtain $\eta_1^*(g) = \eta_1(g)$. Similarly, under condition A3, letting $g \to c_g$, we obtain $\eta_2^*(x) = \eta_2(x)$. This in turn leads to $\pi_0^* = \pi_0, \pi_1^* = \pi_1$. Finally, these results lead to $\alpha^* = \alpha$.

We now prove the result under A1 and A4. Under condition A1 alone, the same derivation as before leads to

$$\frac{\eta_2^*(x)}{\eta_2(x)} = \frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g, x, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g, x, \boldsymbol{\beta})\}}.$$

Thus A4 further implies

$$\frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g_1, x_1, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g_1, x_1, \boldsymbol{\beta})\}} = \frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g_2, x_1, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g_2, x_1, \boldsymbol{\beta})\}},$$

or equivalently, $(\pi_0^*/\pi_0 - \pi_1^*/\pi_1)[\exp\{\alpha + m(g_1, x_1, \boldsymbol{\beta})\} - \exp\{\alpha + m(g_2, x_1, \boldsymbol{\beta})\}] = 0$. Hence, $\pi_d^* = \pi_d$ for $d = 0, 1$. As a result, $\alpha^* = \alpha$ and $\eta_2^*(x) = \eta_2(x)$.

The result under A2 and A3 is symmetric to the one under A1 and A4 hence is omitted.

□

The requirements in A1 and A3 are appropriate in the case where G and X are both continuous. The requirements in A1 and A4 are suitable in the case where G is discrete and X is continuous. The requirements in A2 and A3 are suitable in the case where X is discrete and G is continuous.

A.2 Nuisance Tangent Space Λ and its Orthogonal Complement $\Lambda^\perp$

The nuisance tangent space Λ is computed in two steps. First, replacing the nuisance parameter $\eta = (\eta_1, \eta_2)$ with a finite-dimensional parameter, say $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$, and taking the derivative of $\log f_{D,G,X}(d, g, x; \boldsymbol{\beta}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ to get $\mathbf{S}_{\boldsymbol{\gamma}} = (\mathbf{S}_{\boldsymbol{\gamma}_1}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}_2}^{\mathrm{T}})^{\mathrm{T}}$. Second, finding the mean squared closure that contains all such $\mathbf{S}_{\boldsymbol{\gamma}}$, which is Λ.

For any finite-dimensional parameter $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$, we have $\mathbf{S}_{\boldsymbol{\gamma}} = (\mathbf{S}_{\boldsymbol{\gamma}_1}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma}_2}^{\mathrm{T}})^{\mathrm{T}}$,

where

$$\begin{aligned}
\mathbf{S}_{\gamma_1} &= \eta_1(g,\boldsymbol{\gamma}_1)^{-1}\partial\eta_1(g,\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1 - \pi_d^{-1}\int \partial\eta_1(g,\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1\eta_2(x) \\
&\qquad \times H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g) \\
&= \eta_1(g,\boldsymbol{\gamma}_1)^{-1}\partial\eta_1(g,\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1 - E\{\eta_1(g,\boldsymbol{\gamma}_1)^{-1}\partial\eta_1(G,\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1 \mid D\}, \\
\mathbf{S}_{\gamma_2} &= \eta_2(x,\boldsymbol{\gamma}_2)^{-1}\partial\eta_2(x,\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 - \pi_d^{-1}\int \eta_1(g)\partial\eta_2(x,\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 \\
&\qquad \times H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g) \\
&= \eta_2(x,\boldsymbol{\gamma}_2)^{-1}\partial\eta_2(x,\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 - E\{\eta_2(x,\boldsymbol{\gamma}_2)^{-1}\partial\eta_2(X,\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 \mid D\}.
\end{aligned}$$

It is easy to show the nuisance tangent spaces associated with η_1 and η_2 are respectively

$$\begin{aligned}
\Lambda_1 &= \Big[\mathbf{a}(g) - \pi_d^{-1}\int \mathbf{a}(g)\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g) : \\
&\qquad E^{\text{true}}\{\mathbf{a}(G)\} = \mathbf{0}, \mathbf{a}(g) \in \mathcal{R}^p\Big] \\
&= \left[\mathbf{a}(g) - E\{\mathbf{a}(G) \mid d\} : \forall\mathbf{a}(g) \in \mathcal{R}^p\right], \\
\Lambda_2 &= \Big[\mathbf{a}(x) - \pi_d^{-1}\int \mathbf{a}(x)\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g) : \\
&\qquad E^{\text{true}}\{\mathbf{a}(X)\} = \mathbf{0}, \mathbf{a}(x) \in \mathcal{R}^p\Big] \\
&= \left[\mathbf{a}(x) - E\{\mathbf{a}(X) \mid d\} : \forall\mathbf{a}(x) \in \mathcal{R}^p\right].
\end{aligned}$$

Then

$$\begin{aligned}
\Lambda &= \Lambda_1 + \Lambda_2 \\
&= \left[\mathbf{a}_1(g) + \mathbf{a}_2(x) - E\{\mathbf{a}_1(G) + \mathbf{a}_2(X) \mid d\} : \forall\mathbf{a}_1(g), \mathbf{a}_2(x) \in \mathcal{R}^p\right].
\end{aligned}$$

Define $\Lambda_1^{\perp,\text{conj}} = [\mathbf{f}(d,g,x) : E(\mathbf{f}) = \mathbf{0}, E(\mathbf{f} \mid G) = E\{E(\mathbf{f} \mid D) \mid G\}]$. Now consider

$\mathbf{f} \perp \Lambda_1$. Then for any $\mathbf{a}(g) - E\{\mathbf{a}(G) \mid d\} \in \Lambda_1$,

$$\begin{aligned} 0 &= E(\mathbf{f}^{\mathrm{T}}[\mathbf{a}(G) - E\{\mathbf{a}(G) \mid D\}]) \\ &= E\left[\mathbf{f}^{\mathrm{T}}\mathbf{a}(G) - \mathbf{f}^{\mathrm{T}}E\{\mathbf{a}(G) \mid D\}\right] \\ &= E\left[\mathbf{f}^{\mathrm{T}}\mathbf{a}(G) - E(\mathbf{f}^{\mathrm{T}} \mid D)E\{\mathbf{a}(G) \mid D\}\right] \\ &= E\left\{\mathbf{f}^{\mathrm{T}}\mathbf{a}(G) - E(\mathbf{f}^{\mathrm{T}} \mid D)\mathbf{a}(G)\right\} \\ &= E\left[E\{\mathbf{f}^{\mathrm{T}} - E(\mathbf{f}^{\mathrm{T}} \mid D) \mid G\}\mathbf{a}(G)\right]. \end{aligned}$$

Hence, $E\{\mathbf{f} - E(\mathbf{f} \mid D) \mid G\} = \mathbf{0}$ almost surely. Besides, $\Lambda_1^{\perp}$ need to be a subspace of the Hilbert space $\mathcal{H}$, hence $E(\mathbf{f}) = \mathbf{0}$. Thus, we have shown $\Lambda_1^{\perp} \subset \Lambda_1^{\perp,\mathrm{conj}}$. On the other hand, for any $\mathbf{f} \in \Lambda_1^{\perp,\mathrm{conj}}$,

$$\begin{aligned} &E\left[\mathbf{f}^{\mathrm{T}}\mathbf{a}(G) - \mathbf{f}^{\mathrm{T}}E\{\mathbf{a}(G) \mid D\}\right] \\ &\quad = E\left\{\mathbf{f}^{\mathrm{T}}\mathbf{a}(G) - E(\mathbf{f}^{\mathrm{T}} \mid D)\mathbf{a}(G)\right\} \\ &\quad = E\left[E\{\mathbf{f}^{\mathrm{T}} - E(\mathbf{f}^{\mathrm{T}} \mid D) \mid G\}\mathbf{a}(G)\right] \\ &\quad = \mathbf{0}, \end{aligned}$$

hence $\Lambda_1^{\perp,\mathrm{conj}} \subset \Lambda_1^{\perp}$. Thus, we have obtained $\Lambda_1^{\perp} = \Lambda_1^{\perp,\mathrm{conj}}$. Similarly, we can prove

$$\Lambda_2^{\perp} = [\mathbf{f}(d, g, x) : E(\mathbf{f}) = \mathbf{0}, E(\mathbf{f} \mid X) = E\{E(\mathbf{f} \mid D) \mid X\}]$$

Hence,

$$\begin{aligned} \Lambda^{\perp} = {} & [\mathbf{f}(d, g, x) : E(\mathbf{f} \mid G) = E\{E(\mathbf{f} \mid D) \mid G\}, E(\mathbf{f} \mid X) = E\{E(\mathbf{f} \mid D) \mid X\}, \\ & E(\mathbf{f}) = \mathbf{0}, \mathbf{f} \in \mathcal{R}^p]. \end{aligned}$$

A.3 Uniqueness of $\mathbf{a}$ and $\mathbf{b}$ up to Constants

To prove that $\mathbf{a}$ and $\mathbf{b}$ defined in (2.4) - (2.5) are unique up to constant shifts, we consider the following. If there exists $\mathbf{a}_1, \mathbf{a}_2, \mathbf{b}_1, \mathbf{b}_2$ such that

$$\begin{aligned}\mathbf{S}_{\text{eff}}(d,g,x) &= \mathbf{S}(d,g,x) - \mathbf{a}_1(g) - \mathbf{b}_1(x) - E\{\mathbf{S}(d,G,X) \mid d\} \\ &\quad + E\{\mathbf{a}_1(G) + \mathbf{b}_1(X) \mid d\} \\ &= \mathbf{S}(d,g,x) - \mathbf{a}_2(g) - \mathbf{b}_2(x) - E\{\mathbf{S}(d,G,X) \mid d\} \\ &\quad + E\{\mathbf{a}_2(G) + \mathbf{b}_2(X) \mid d\},\end{aligned}$$

then

$$\mathbf{a}_2(g) - \mathbf{a}_1(g) = \mathbf{b}_1(x) - \mathbf{b}_2(x) - E\{\mathbf{a}_1(G) + \mathbf{b}_1(X) \mid d\} + E\{\mathbf{a}_2(G) + \mathbf{b}_2(X) \mid d\}.$$

The left-hand side is a function of g while the right-hand side is a function of x and d. Hence $\mathbf{a}_1(g) - \mathbf{a}_2(g)$ is a constant. Similarly, $\mathbf{b}_1(x) - \mathbf{b}_2(x)$ is also a constant. □

A.4 Equivalent Expression of Equations (2.4) - (2.5) and the Proof Under the Condition $E(\mathbf{a}) = E(\mathbf{b}) = \mathbf{0}$

We claim under the mean zero constraint $E(\mathbf{a}) = E(\mathbf{b}) = \mathbf{0}$, (2.4) and (2.5) are equivalent to (A.1), (A.2), and (A.3) below, namely

$$\begin{aligned}\mathbf{S}_g(g) - E\{\mathbf{S}_x(X) \mid g\} &= \mathbf{a}(g) + \mathbf{u}_0 c_g(g) - E\{E(\mathbf{a} \mid X) \mid g\} \\ &\quad - \mathbf{u}_0 E\{c_x(X) \mid g\}, \qquad (A.1) \\ \mathbf{S}_x(x) &= E(\mathbf{a} \mid x) + \mathbf{b}(x) + \mathbf{u}_0 c_x(x), \qquad (A.2) \\ \mathbf{u}_0 &= E(\mathbf{a} + \mathbf{b} \mid D = 0), \qquad (A.3)\end{aligned}$$

where $c_x(x) = E[\{n_0 - nI(D=0)\}/n_1 \mid x], c_g(g) = E[\{n_0 - nI(D=0)\}/n_1 \mid g]$.

Proof. Suppose $\mathbf{a}$ and $\mathbf{b}$ are the solution of equations (2.4) and (2.5). Let $E(\mathbf{a}+\mathbf{b} \mid D = 0) = \mathbf{u}_0$, $E(\mathbf{a}+\mathbf{b} \mid D=1) = \mathbf{u}_1$. Then (A.3) automatically holds. It is easy to verify that $\mathbf{u}_0 n_0 + \mathbf{u}_1 n_1 = nE(\mathbf{a}+\mathbf{b}) = \mathbf{0}$. Hence (2.4) and (2.5) become

$$\begin{aligned} E(\mathbf{a} \mid x) + \mathbf{b}(x) + \mathbf{u}_0\{(n_0/n_1)f_{D|X}(1,x) - f_{D|X}(0,x)\} &= \mathbf{S}_x(x), \\ \mathbf{a}(g) + E(\mathbf{b} \mid g) + \mathbf{u}_0\{(n_0/n_1)f_{D|G}(1,g) - f_{D|G}(0,g)\} &= \mathbf{S}_g(g). \end{aligned}$$

Further write

$$\begin{aligned} c_x(x) &= (n_0/n_1)f_{D|X}(1,x) - f_{D|X}(0,x) = \{n_0 - nf_{D|X}(0,x)\}/n_1 \\ &= E[\{n_0 - nI(D=0)\}/n_1 \mid x] = E[\{n_0/n - I(D=0)\}/(n_1/n) \mid x], \\ c_g(g) &= (n_0/n_1)f_{D|G}(1,g) - f_{D|G}(0,g) = \{n_0 - nf_{D|G}(0,g)\}/n_1 \\ &= E[\{n_0 - nI(D=0)\}/n_1 \mid g] = E[\{n_0/n - I(D=0)\}/(n_1/n) \mid g]. \end{aligned}$$

Then

$$E(\mathbf{a} \mid x) + \mathbf{b}(x) + \mathbf{u}_0 c_x(x) = \mathbf{S}_x(x), \tag{A.4}$$

$$\mathbf{a}(g) + E(\mathbf{b} \mid g) + \mathbf{u}_0 c_g(g) = \mathbf{S}_g(g). \tag{A.5}$$

Note that (A.4) above is exactly (A.2) defined in Section 2.3. Taking conditional expectation of (A.4) given $G = g$, we obtain

$$E\{E(\mathbf{a} \mid X) \mid g\} + E(\mathbf{b} \mid g) + \mathbf{u}_0 E\{c_x(X) \mid g\} = E\{\mathbf{S}_x(X) \mid g\}.$$

Subtracting the above from (A.5), we obtain (A.1), namely

$$\mathbf{a}(g)+\mathbf{u}_0 c_g(g)-E\{E(\mathbf{a}\mid X)\mid g\}-\mathbf{u}_0 E\{c_x(X)\mid g\}=\mathbf{S}_g(g)-E\{\mathbf{S}_x(X)\mid g\}.$$

From the above derivation, it is clear that any mean zero functions $\mathbf{a}(g), \mathbf{b}(x)$ that solve (2.4) and (2.5) also satisfy (A.1), (A.2), and (A.3). We now prove the other way around, that is any mean zero functions $\mathbf{a}(g), \mathbf{b}(x)$ that satisfy (A.1), (A.2), and (A.3) also satisfy (2.4) and (2.5).

Taking the expectation of (A.2) conditionally on $G=g$ and adding the resulting equation to (A.1), we obtain exactly (A.5). Hence equations (A.2), (A.1) lead to equations (A.2), (A.5).

For preparation, note also that $c_g(g)=(n_0/n_1)f_{D|G}(1,g)-f_{D|G}(0,g)$. Hence under (A.3) and the condition $n_1E(\mathbf{a}+\mathbf{b}\mid D=1)+n_0E(\mathbf{a}+\mathbf{b}\mid D=0)=nE(\mathbf{a}+\mathbf{b})=\mathbf{0}$, we can further write

$$\begin{aligned}
\mathbf{u}_0 c_g(g) &= E(\mathbf{a}+\mathbf{b}\mid D=0)\{(n_0/n_1)f_{D|G}(1,g)-f_{D|G}(0,g)\}\\
&= E(\mathbf{a}+\mathbf{b}\mid D=0)(n_0/n_1)f_{D|G}(1,g)-E(\mathbf{a}+\mathbf{b}\mid D=0)f_{D|G}(0,g)\\
&= -E(\mathbf{a}+\mathbf{b}\mid D=1)f_{D|G}(1,g)-E(\mathbf{a}+\mathbf{b}\mid D=0)f_{D|G}(0,g)\\
&= -E\{E(\mathbf{a}+\mathbf{b}\mid D)\mid g\}.
\end{aligned}$$

Similarly, $\mathbf{u}_0 c_x(x)=-E\{E(\mathbf{a}+\mathbf{b}\mid D)\mid x\}$. From (A.2), we obtain

$$\mathbf{S}_x(x) \;=\; E(\mathbf{a}\mid x)+\mathbf{b}(x)+\mathbf{u}_0 c_x(x)=E(\mathbf{a}\mid x)+\mathbf{b}(x)-E\{E(\mathbf{a}+\mathbf{b}\mid D)\mid x\},$$

which is exactly (2.4). Similarly, from (A.5), we obtain (2.5). □

Equation (A.1) allows us to solve for $\mathbf{a}(g)$ as a function of $\mathbf{u}_0$ and other known quantities, say $\mathbf{a}(g) = \mathbf{F}_a(g, \mathbf{u}_0) - E\{\mathbf{F}_a(G, \mathbf{u}_0)\}$, where $\mathbf{F}_a$ is a function that solves (A.1) which does not need to have mean $\mathbf{0}$. Then we can solve $\mathbf{b}(\cdot)$ from (A.2) as a function of $\mathbf{u}_0$ to obtain

$$\mathbf{b}(x) = \mathbf{S}_x(x) - \mathbf{u}_0 c_x(x) - E\{\mathbf{F}_a(G, \mathbf{u}_0) \mid x\} + E\{\mathbf{F}_a(G, \mathbf{u}_0)\}.$$

Now

$$\begin{aligned} \mathbf{u}_0 &= E\{\mathbf{a}(G) + \mathbf{b}(X) \mid D = 0\} \\ &= E[\mathbf{F}_a(G, \mathbf{u}_0) + \mathbf{S}_x(X) - \mathbf{u}_0 c_x(X) - E\{\mathbf{F}_a(G, \mathbf{u}_0) \mid X\} \mid D = 0], \end{aligned}$$

which allows us to solve for $\mathbf{u}_0$. Having obtained $\mathbf{u}_0$, we can then solve for all other quantities easily. Unfortunately, the integral equation (A.1) does not have an explicit solution. We propose an approximation to its solution in the spirit of Tsiatis and Ma (2004), which is provided in Appendix A.5, by discretizing X if X is continuous.

The efficient score $\mathbf{S}_{\text{eff}}$, especially the procedure of solving for $\mathbf{a}$ and $\mathbf{b}$, contains several expectations conditional on D, G, or X. To get estimations of these conditional expectations, we need density estimators of the nuisance parameter $\eta = (\eta_1, \eta_2)$. If the disease rate π_1 or the non-disease rate $\pi_0 = 1 - \pi_1$ is known, then η can be approximated by

$$\widehat{\eta}_1 = \pi_0 \widehat{f}_{G|D=0} + (1 - \pi_0)\widehat{f}_{G|D=1}, \quad \widehat{\eta}_2 = \pi_0 \widehat{f}_{X|D=0} + (1 - \pi_0)\widehat{f}_{X|D=1},$$

where $\widehat{f}_{G|D=d}$ and $\widehat{f}_{X|D=d}$ are the nonparametric estimators of the conditional density/mass function $f_{G|D=d}$ and $f_{X|D=d}$ respectively for $d = 0, 1$. Of course, in practice, π_0 is typically unknown. However, we can get an estimate of π_0 through (2.3).

A.5 Solving the Integral Equation (A.1)

Define $\mathbf{Z} = \mathbf{S} - E(\mathbf{S} \mid D) - \mathbf{u}_0\{n_0 - nI(D = 0)\}/n_1$. An equivalent expression of (A.1) is

$$\mathbf{a}(G) - E[E\{\mathbf{a}(G) \mid X\} \mid G] \;\; = \;\; E(\mathbf{Z} \mid G) - E\{E(\mathbf{Z} \mid X) \mid G\}. \qquad \text{(A.6)}$$

For fixed $\mathbf{u}_0$, all the quantities in $\mathbf{Z}$ are known or have explicit form except $E(\mathbf{S} \mid D)$. With the weighted kernel density $\widehat{\eta}_1, \widehat{\eta}_2$, estimated non-disease rate $\widehat{\pi}_0$ and disease rate $\widehat{\pi}_1$, we can estimate it by $\widehat{E}(\mathbf{S} \mid D = d) = \widehat{\pi}_d^{-1} \int \mathbf{S}(d, g, x)\widehat{\eta}_1(g), \widehat{\eta}_2(x) d\mu(g) d\mu(x)$.

A.5.1 Discrete G with finite number of levels

Assume G is discrete with mass at m_g points $g_1, \cdots, g_{m_g}$. We computed each term in (A.6) under the weighted nonparametric densities $\widehat{\eta}_1, \widehat{\eta}_2$.

$$\begin{aligned}
\widehat{E}\{\mathbf{a}(G) \mid x\} &= \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)\kappa(g_j, x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j, x)\widehat{\eta}_1(g_j)}, \\
\widehat{E}[\widehat{E}\{\mathbf{a}(G) \mid X\} \mid g_k] &= \int \left\{ \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)\kappa(g_j, x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j, x)\widehat{\eta}_1(g_j)} \right\} \frac{\kappa(g_k, x)\widehat{\eta}_2(x)}{\int \kappa(g_k, x)\widehat{\eta}_2(x) d\mu(x)} d\mu(x).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\widehat{E}\{\mathbf{Z}(D, G, X) \mid x\} &= \frac{\sum_{j=1}^{m_g} \sum_{d=0}^{1} n_d/(n\pi_d)\mathbf{Z}(d, g_j, x)H(d, g_j, x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j, x)\widehat{\eta}_1(g_j)}, \\
\widehat{E}[\widehat{E}\{\mathbf{Z}(D, G, X) \mid X\} \mid g_k] &= \int \left\{ \frac{\sum_{j=1}^{m_g} \sum_{d=0}^{1} n_d/(n\pi_d)\mathbf{Z}(d, g_j, x)H(d, g_j, x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j, x)\widehat{\eta}_1(g_j)} \right\} \\
&\quad \times \frac{\kappa(g_k, x)\widehat{\eta}_2(x)}{\int \kappa(g_k, x)\widehat{\eta}_2(x) d\mu(x)} d\mu(x), \qquad \text{(A.7)}
\end{aligned}$$

and

$$\widehat{E}\{\mathbf{Z}(D,G,X) \mid g_k\} \;=\; \sum_{d=0}^{1} \int \mathbf{Z}(d,g_k,x) \frac{n_d/(n\pi_d) H(d,g_k,x)\widehat{\eta}_2(x)}{\int \kappa(g_k,x)\widehat{\eta}_2(x) d\mu(x)} d\mu(x). \quad \text{(A.8)}$$

Consequently, the integral equation (A.6) reduces to the linear equations

$$(I-B)A^{\mathrm{T}} = C^{\mathrm{T}},$$

where A is the $(p+1) \times m_g$ matrix $\{\mathbf{a}(g_1), \cdots, \mathbf{a}(g_{m_g})\}$, corresponding to the solution of the integral equation, I is an $m_g \times m_g$ identity matrix, B is an $m_g \times m_g$ matrix whose (i,j)th element is given by

$$B_{ij} = \int \left\{ \frac{\kappa(g_j,x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j,x)\widehat{\eta}_1(g_j)} \right\} \frac{\kappa(g_i,x)\widehat{\eta}_2(x)}{\int \kappa(g_i,x)\widehat{\eta}_2(x) d\mu(x)} d\mu(x),$$

and C is a $(p+1) \times m_g$ matrix whose kth column is

$$\widehat{E}\{\mathbf{Z}(D,G,X) \mid g_k\} - \widehat{E}[\widehat{E}\{\mathbf{Z}(D,G,X) \mid X\} \mid g_k]$$

defined in (A.7) and (A.8).

After obtaining $\mathbf{a}$, we set

$$\begin{aligned} \mathbf{b}(x) &= \widehat{E}(\mathbf{Z} - \mathbf{a} \mid x) \\ &= \frac{\sum_{j=1}^{m_g} \sum_{d=0}^{1} n_d/(n\pi_d) \mathbf{Z}(d,g_j,x) H(d,g_j,x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j,x)\widehat{\eta}_1(g_j)} \\ &\quad - \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)\kappa(g_j,x)\widehat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} \kappa(g_j,x)\widehat{\eta}_1(g_j)}. \end{aligned}$$

Then we compute $\mathbf{u}_0 = \widehat{E}(\mathbf{a} + \mathbf{b} \mid D = 0)$, where

$$\begin{aligned}
\widehat{E}(\mathbf{a} \mid D = 0) &= \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)\widehat{\eta}_1(g_j) \int H(0, g_j, x)\widehat{\eta}_2(x)d\mu(x)}{\int \sum_{j=1}^{m_g} H(0, g_j, x)\widehat{\eta}_1(g_j)\widehat{\eta}_2(x)d\mu(x)}, \\
\widehat{E}(\mathbf{b} \mid D = 0) &= \int \mathbf{b}(x) \frac{\sum_{j=1}^{m_g} H(0, g_j, x)\widehat{\eta}_1(g_j)\widehat{\eta}_2(x)}{\int \sum_{j=1}^{m_g} H(0, g_j, x)\widehat{\eta}_1(g_j)\widehat{\eta}_2(x)d\mu(x)} d\mu(x).
\end{aligned}$$

A.5.2 Continuous G or Discrete G with Infinite Number of Levels

When G is a continuous variable, we discretize it at a finite number of equally distributed points, say, $g_1 \leq \cdots \leq g_{m_g}$ with $g_{i+1} - g_i \equiv \Delta_g$ for $i = 1, \cdots, m_g - 1$, such that

$$\textstyle\sum_{i=1}^{m_g} f_{G|D}(g_i)\Delta_g \approx 1.$$

Similarly, when G is discrete with infinite number of levels, we simply choose a sufficient number of points from its support to get an overall probability close to 1.

Then the sequential procedures are exactly the same as that described in the case where G is discrete with finite number of levels.

A.6 Proof of Theorem 2

We divided the n observations randomly into two data sets with sample sizes $N_1 = n - n^{1-\delta}$ and $N_2 = n^{1-\delta}$, where $\delta > 0$ is a small positive number. The first data set with size N_1 is used to form the estimating equation while the second data set with size N_2 is used for weighted nonparametric density estimation.

Before proving Theorem 2, we first establish some preliminary results in Lemmas 1-4.

<u>**Lemma**</u> **1.** *As an equation to solve for π,*

$$\begin{aligned}\pi &= \int H(0,g,x)\{\pi f_{G|D=0}(g)+(1-\pi)f_{G|D=1}(g)\}\\ &\quad \times\{\pi f_{X|D=0}(x)+(1-\pi)f_{X|D=1}(x)\}d\mu(g)d\mu(x) \qquad \text{(A.9)}\end{aligned}$$

has at most two roots. Only one root lies between 0 and 1, and it is the true non-disease rate π_0. Here $f_{G|D=d}(g)$ is the density/mass function of G in the control (d=0) and case ($d=1$) subpopulations respectively, and $f_{X|D=d}(x)$ is similarly defined.

Proof. It is obvious that π_0 satisfies (A.9) hence is a solution for the quadratic equation

$$\begin{aligned}\pi &= \int H(0,g,x)\{\pi f_{G|D=0}(g)+(1-\pi)f_{G|D=1}(g)\}\\ &\quad \times\{\pi f_{X|D=0}(x)+(1-\pi)f_{X|D=1}(x)\}d\mu(g)d\mu(x)\\ &= a\pi^2+b\pi(1-\pi)+c(1-\pi)^2,\end{aligned}$$

where

$$\begin{aligned}a &= \int H(0,g,x)f_{G|D=0}(g)f_{X|D=0}(x)d\mu(g)d\mu(x),\\ b &= \int H(0,g,x)f_{G|D=0}(g)f_{X|D=1}(x)d\mu(g)d\mu(x)\\ &\quad +\int H(0,g,x)f_{G|D=1}(g)f_{X|D=0}(x)d\mu(g)d\mu(x),\\ c &= \int H(0,g,x)f_{G|D=1}(g)f_{X|D=1}(x)d\mu(g)d\mu(x).\end{aligned}$$

Equivalently, we have

$$(a-b+c)\pi^2+(b-2c-1)\pi+c=0. \qquad \text{(A.10)}$$

We now show (A.10) has only one root in $(0,1)$. Obviously, if $a-b+c=0$, (A.10) is a linear equation hence it only has one root, which is shown to be π_0. If $a-b+c\neq 0$, as a quadratic function, its delta value is $(b-2c-1)^2-4(a-b+c)c=(b-1)^2+4c(1-a)>0$, as a,b,c are all positive, and $a=\int H(0,g,x)f_{G|D=0}(g)f_{X|D=0}(x)d\mu(g)d\mu(x)<1$. Hence (A.10) has two roots. One of the two root is $\pi_0\in(0,1)$. Now (A.10) evaluated at $\pi=0$ and $\pi=1$ are respectively $c>0$ and $a-1<0$, hence (A.10) has a second root either in $(\infty,0)$ or in $(1,\infty)$. This proves that π_0 is the unique root in $(0,1)$. □

Lemma 2. *Assuming G is discrete, then for fixed g and $d=0,1$. $|\widehat{f}_{G|D=d}(g)-f_{G|D=d}(g)|=O_p(1/\sqrt{N_2})$, where $\widehat{f}_{G|D=d}(g)$ is the empirical estimator of $f_{G|D=d}(g)$.*

Proof. For g on the support of $f_{G|D=d}(g)$, we have

$$\widehat{f}_{G|D=d}(g)=\textstyle\sum_{i=1}^{N_2}I(G_i=g,D_i=d)/\sum_{i=1}^{N_2}I(D_i=d).$$

Define $p_1=\text{pr}(G=g,D=d),p_2=\text{pr}(D=d)$, then we have $E\{N_2^{-1}\sum_{i=1}^{N_2}I(G_i=g,D_i=d)\}=p_1,\ \text{var}\{N_2^{-1}\sum_{i=1}^{N_2}I(G_i=g,D_i=d)\}=p_1(1-p_1)/N_2$. Similarly, $E\{N_2^{-1}\sum_{i=1}^{N_2}I(D_i=d)\}=p_2,\ \text{var}\{N_2^{-1}\sum_{i=1}^{N_2}I(D_i=d)\}=p_2(1-p_2)/N_2$.

According to the central limit theorem and the law of large numbers, we have $N_2^{-1/2}\{\sum_{i=1}^{N_2}I(G_i=g,D_i=d)-p_1\}\xrightarrow{\mathcal{D}}\text{Normal}\{0,p_1(1-p_1)\},N_2^{-1}\sum_{i=1}^{N_2}I(D_i=d)\xrightarrow{P}p_2$, and hence $N_2^{1/2}\{\widehat{f}_{G|D=d}(g)-f_{G|D=d}(g)\}\xrightarrow{\mathcal{D}}\text{Normal}\{0,p_1(1-p_1)/p_2^2\}$ following Slutsky's theorem. Thus, Lemma 2 is shown. □

Lemma 3. *As an equation to solve for π,*

$$\begin{aligned}\pi &= \int H(0,g,x)\{\pi\widehat{f}_{G|D=0}(g)+(1-\pi)\widehat{f}_{G|D=1}(g)\}\\ &\quad\times\{\pi\widehat{f}_{X|D=0}(x)+(1-\pi)\widehat{f}_{X|D=1}(x)\}d\mu(g)d\mu(x)\end{aligned}$$

has at most two roots. Here $\widehat{f}_{G|D=d}(g)$ and $\widehat{f}_{X|D=d}(x)$ are respectively the nonparametric *estimates of $f_{G|D=d}(g)$ and $f_{X|D=d}(x)$. Let $\widehat{\pi}_0$ denote the solution that is closest to π_0, then under the regularity conditions defined in Section 2.4, $\widehat{\pi}_0 = \pi_0 + O_p\{h^2 + (N_2 h)^{-1/2}\}$.*

Proof. We will only prove the result for discrete G and continuous X. Because of the symmetry, the result for continuous G and discrete X automatically holds. The proof for continuous G and X is similar. We rewrite the equation of π as

$$\begin{aligned}
\pi &= \int H(0,g,x)\{\pi \widehat{f}_{G|D=0}(g) + (1-\pi)\widehat{f}_{G|D=1}(g)\} \\
&\quad \times\{\pi \widehat{f}_{X|D=0}(x) + (1-\pi)\widehat{f}_{X|D=1}(x)\}d\mu(g)d\mu(x) \\
&= \int H(0,g,x)\{\pi f_{G|D=0}(g) + (1-\pi) f_{G|D=1}(g)\} \\
&\quad \times\{\pi f_{X|D=0}(x) + (1-\pi) f_{X|D=1}(x)\}d\mu(g)d\mu(x) \\
&\quad + \int H(0,g,x)[\pi\{\widehat{f}_{G|D=0}(g) - f_{G|D=0}(g)\} + (1-\pi)\{\widehat{f}_{G|D=1}(g) - f_{G|D=1}(g)\}] \\
&\quad \times\{\pi f_{X|D=0}(x) + (1-\pi) f_{X|D=1}(x)\}d\mu(g)d\mu(x) \\
&\quad + \int H(0,g,x)\{\pi f_{G|D=0}(g) + (1-\pi) f_{G|D=1}(g)\} \\
&\quad \times[\pi\{\widehat{f}_{X|D=0}(x) - f_{X|D=0}(x)\} + (1-\pi)\{\widehat{f}_{X|D=1}(x) - f_{X|D=1}(x)\}]d\mu(g)d\mu(x) \\
&\quad + \int H(0,g,x)[\pi\{\widehat{f}_{G|D=0}(g) - f_{G|D=0}(g)\} + (1-\pi)\{\widehat{f}_{G|D=1}(g) - f_{G|D=1}(g)\}] \\
&\quad \times[\pi\{\widehat{f}_{X|D=0}(x) - f_{X|D=0}(x)\} + (1-\pi)\{\widehat{f}_{X|D=1}(x) - f_{X|D=1}(x)\}]d\mu(g)d\mu(x).
\end{aligned}$$

Now we examine the last three integrals in the above equation one by one. First,

$$\begin{aligned}
&\Big|\int H(0,g,x)[\pi\{\widehat{f}_{G|D=0}(g)-f_{G|D=0}(g)\}+(1-\pi)\{\widehat{f}_{G|D=1}(g)-f_{G|D=1}(g)\}]\\
&\quad\times\{\pi f_{X|D=0}(x)+(1-\pi)f_{X|D=1}(x)\}d\mu(g)d\mu(x)\Big|\\
&\quad\le\int\Big|\pi\{\widehat{f}_{G|D=0}(g)-f_{G|D=0}(g)\}+(1-\pi)\{\widehat{f}_{G|D=1}(g)-f_{G|D=1}(g)\}\Big|\\
&\qquad\times\{\pi f_{X|D=0}(x)+(1-\pi)f_{X|D=1}(x)\}d\mu(g)d\mu(x)\\
&\quad=\int\Big|\pi\{\widehat{f}_{G|D=0}(g)-f_{G|D=0}(g)\}+(1-\pi)\{\widehat{f}_{G|D=1}(g)-f_{G|D=1}(g)\}\Big|d\mu(g)\\
&\quad=O_p(N_2^{-1/2}),
\end{aligned}$$

where the last equality is a direct result of Lemma 2. Similarly,

$$\begin{aligned}
&\Big|\int H(0,g,x)\{\pi f_{G|D=0}(g)+(1-\pi)f_{G|D=1}(g)\}\\
&\quad\times[\pi\{\widehat{f}_{X|D=0}(x)-f_{X|D=0}(x)\}+(1-\pi)\{\widehat{f}_{X|D=1}(x)-f_{X|D=1}(x)\}]d\mu(g)d\mu(x)\Big|\\
&\quad\le\int\Big|\pi\{\widehat{f}_{X|D=0}(x)-f_{X|D=0}(x)\}+(1-\pi)\{\widehat{f}_{X|D=1}(x)-f_{X|D=1}(x)\}\Big|d\mu(x)\\
&\quad=O_p\{h^2+(N_2h)^{-1/2}\},
\end{aligned}$$

where the last equality is the direct result of the nonparametric density estimation property. Following the same procedure, we have

$$\begin{aligned}
&\Big|\int H(0,g,x)[\pi\{\widehat{f}_{G|D=0}(g)-f_{G|D=0}(g)\}+(1-\pi)\{\widehat{f}_{G|D=1}(g)-f_{G|D=1}(g)\}]\\
&\quad\times[\pi\{\widehat{f}_{X|D=0}(x)-f_{X|D=0}(x)\}+(1-\pi)\{\widehat{f}_{X|D=1}(x)-f_{X|D=1}(x)\}]d\mu(g)d\mu(x)\Big|\\
&\quad\le\int\Big|\pi\{\widehat{f}_{G|D=0}(g)-f_{G|D=0}(g)\}+(1-\pi)\{\widehat{f}_{G|D=1}(g)-f_{G|D=1}(g)\}\Big|d\mu(g)\\
&\qquad\times\int\Big|\pi\{\widehat{f}_{X|D=0}(x)-f_{X|D=0}(x)\}+(1-\pi)\{\widehat{f}_{X|D=1}(x)-f_{X|D=1}(x)\}\Big|d\mu(x)\\
&\quad\le O_p(1/\sqrt{N_2})\{O_p(h^2)+O_p(1/\sqrt{N_2h})\}.
\end{aligned}$$

As a result,

$$\begin{aligned}\pi &= \int H(0,g,x)\{\pi f_{G|D=0}(g) + (1-\pi) f_{G|D=1}(g)\} \\ &\quad \times\{\pi f_{X|D=0}(x) + (1-\pi) f_{X|D=1}(x)\} d\mu(g) d\mu(x) + O_p(h^2) + O_p(1/\sqrt{N_2 h}).\end{aligned}$$

Lemma 1 then immediately leads to the conclusion. □

Remark 6. *Lemma 3 can be further generalized to the case where the discrete covariates can have infinitely many levels and the continuous covariates can have noncompact support, as long as the tails of these distributions are sufficiently thin. The proofs will be more complicated, involving splitting the domain of the covariates into two parts, where the treatment in one part controls the error rate via the nonparametric estimation rate, and the treatment in the other part controls error rate via the tail behavior itself. To avoid extremely technicality, we skip the detailed proofs.*

The above results directly lead to Lemma 4.

Lemma 4. *The weighted nonparametric density estimates $\widehat{\eta}_1(\cdot), \widehat{\eta}_2(\cdot)$ defined in Algorithm 1 Step 4 have at least the usual nonparametric convergence rate. Specifically,*

$$\begin{aligned}\widehat{\eta}_1(\cdot) - \eta_1(\cdot) &= O_p(h^2) + O_p(1/\sqrt{nh}), \\ \widehat{\eta}_2(\cdot) - \eta_2(\cdot) &= O_p(h^2) + O_p(1/\sqrt{nh}).\end{aligned}$$

Now we are ready to prove Theorem 2, the main theoretic result of our work.

Proof of Theorem 2.

The efficient score is

$$\begin{aligned}
&\mathbf{S}_{\text{eff}}\{d, g, x, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} \\
&\quad = \mathbf{S}(d, g, x, \boldsymbol{\theta}) - \mathbf{a}\{g, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} - \mathbf{b}\{x, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} - E_{\boldsymbol{\eta}(\cdot)}\{\mathbf{S}(d, G, X, \boldsymbol{\theta}) \mid d\} \\
&\qquad + E_{\boldsymbol{\eta}(\cdot)}[\mathbf{a}\{G, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} + \mathbf{b}\{X, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} \mid d],
\end{aligned}$$

where

$$\begin{aligned}
&E_{\boldsymbol{\eta}}\{\mathbf{a}(G, \boldsymbol{\theta}, \boldsymbol{\eta}) \mid x\} + \mathbf{b}(x, \boldsymbol{\theta}, \boldsymbol{\eta}) - E\{E(\mathbf{a} + \mathbf{b} \mid D) \mid x\} \\
&\quad = E_{\boldsymbol{\eta}}(\mathbf{S} \mid x) - E_{\boldsymbol{\eta}}\{E(\mathbf{S} \mid D) \mid x\}, \\
&\mathbf{a}(g, \boldsymbol{\theta}, \boldsymbol{\eta}) + E_{\boldsymbol{\eta}}\{\mathbf{b}(X, \boldsymbol{\theta}, \boldsymbol{\eta}) \mid g\} - E_{\boldsymbol{\eta}}\{E_{\boldsymbol{\eta}}(\mathbf{a} + \mathbf{b} \mid D) \mid g\} \\
&\quad = E_{\boldsymbol{\eta}}(\mathbf{S} \mid g) - E_{\boldsymbol{\eta}}\{E_{\boldsymbol{\eta}}(\mathbf{S} \mid D) \mid g\}.
\end{aligned}$$

Here $\boldsymbol{\eta}(\cdot) = \{\eta_1(\cdot), \eta_2(\cdot)\}$ stands for the nuisance parameter, which is free of $\boldsymbol{\theta}$. The notation $\mathbf{a}(g, \boldsymbol{\theta}, \boldsymbol{\eta}), \mathbf{b}(x, \theta, \boldsymbol{\eta})$ and $E_{\boldsymbol{\eta}}(\cdot)$ emphasize that they are calculated under the nuisance parameter $\boldsymbol{\eta}$.

When we adopt the weighted nonparametric estimation for the density of G and X, the estimated efficient score is $\mathbf{S}_{\text{eff}}\{d, g, x, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\}$, where $\widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta}) = \{\widehat{\eta}_1(\cdot, \boldsymbol{\theta}), \widehat{\eta}_2(\cdot, \boldsymbol{\theta})\}^{\mathrm{T}}$. Here $\widehat{\eta}_1(\cdot, \boldsymbol{\theta})$ and $\widehat{\eta}_2(\cdot, \boldsymbol{\theta})$ are defined in Algorithm 1. Note that $\widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})$ is computed using a subset of the data with N_2 observations, while the final estimating equation is computed using the rest of the data with $N_1 = n - N_2$ observations.

Let $\widehat{\boldsymbol{\theta}}$ denote the solution of

$$\mathbf{0} = \textstyle\sum_{i=1}^{N_1} \mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\}.$$

Then

$$
\begin{aligned}
\mathbf{0} &= N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}(\cdot, \widehat{\boldsymbol{\theta}})\} \\
&= N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\} \\
&\quad + N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\dfrac{d}{d\boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta}^*)\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&= N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\} \\
&\quad + \left(E\left[\dfrac{d}{d\boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\}\right] + o_p(1)\right) N_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad \text{(A.11)}
\end{aligned}
$$

Here $\boldsymbol{\theta}^*$ is on the line connecting $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

We first examine the first term in equation (A.11).

$$
\begin{aligned}
& N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\} \\
&\quad = N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} \\
&\qquad + N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\displaystyle\sum_{j=1}^{2}\frac{\partial \mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}\}}{\partial \eta_j}\{\widehat{\eta}_j(\cdot, \boldsymbol{\theta}) - \eta_j\} \\
&\qquad + N_1^{-1/2}(1/2)\textstyle\sum_{i=1}^{N_1}\displaystyle\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{\partial^2 \mathbf{S}_{\text{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}^*)}{\partial \eta_j \eta_k} \\
&\qquad\qquad \times\{\widehat{\eta}_j(\cdot, \boldsymbol{\theta}) - \eta_j\}\{\widehat{\eta}_k(\cdot, \boldsymbol{\theta}) - \eta_k\} \\
&\quad = N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot)\} + o_p(1),
\end{aligned}
$$

where $\boldsymbol{\eta}^*$ is an intermediate value between $\boldsymbol{\eta}(\cdot)$ and $\widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})$, and the derivative with respect to $\boldsymbol{\eta}$ is the Gâteaux derivative. Here in the last equality, the last term is of order $o_p(1)$ because of the convergence rate of $\widehat{\eta}_1, \widehat{\eta}_2$ in Lemma 4 and the regularity condition $N_1 h^8 \to 0$ and $N_1 h^2 \to \infty$. The second term is of order $o_p(1)$ because

$$
N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\dfrac{\partial}{\partial \boldsymbol{\eta}}\mathbf{S}_{\text{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta})
$$

converges to a normal distribution with bounded variance and mean

$$E\left\{\frac{\partial}{\partial\boldsymbol{\eta}}\mathbf{S}_{\text{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta})\right\} = \mathbf{0}$$

due to the central limit theorem. The above expectation vanishes because the efficient score lies in $\Lambda^{\perp}$ and any directional derivative with respect to any component in the nuisance space has mean zero. We then examine the second term in (A.11). We have

$$\begin{aligned}&N_1^{-1}\textstyle\sum_{i=1}^{N_1}\frac{d}{d\boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}(\cdot, \boldsymbol{\theta})\}\\ &\qquad = N_1^{-1}\textstyle\sum_{i=1}^{N_1}\frac{d}{d\boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot, \boldsymbol{\theta})\} + o_p(1),\end{aligned}$$

due to the convergence of $\widehat{\boldsymbol{\eta}}$. Thus, (A.11) yields

$$\begin{aligned}&N_1^{-1/2}\textstyle\sum_{i=1}^{N_1}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot, \boldsymbol{\theta})\}\\ &\qquad = -\left(E\left[\frac{d}{d\boldsymbol{\theta}^{\mathrm{T}}}\mathbf{S}_{\text{eff}}\{D_i, G_i, X_i, \boldsymbol{\theta}, \boldsymbol{\eta}(\cdot, \boldsymbol{\theta})\}\right] + o_p(1)\right) N_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}),\end{aligned}$$

and the results follow. □

APPENDIX B

SKETCH OF TECHNICAL ARGUMENTS FOR SECTION 3

B.1 Dimension Reduction Model for $\epsilon\boldsymbol{\mu}_s$

The dimension reduction assumption (3.9) on $\epsilon\boldsymbol{\mu}_s$ is more complicated than (3.8) or (3.10), and requires careful attention.

In the usual case, we have that

$$\begin{aligned} S^* &= \begin{bmatrix} \partial \text{log}\{H(d,\mathbf{X},Y,\boldsymbol{\alpha}))\}/\partial\boldsymbol{\alpha} \\ \partial \text{log}\{\eta_2^*(\epsilon,\mathbf{X}))\}/\partial\boldsymbol{\beta} \end{bmatrix} \\ &= \begin{bmatrix} \{d - H(1,\mathbf{X},Y,\boldsymbol{\alpha})\}\,(1,\mathbf{X}^{\mathrm{T}},Y)^{\mathrm{T}} \\ -\{\eta_2^*(\epsilon,\mathbf{X})^{-1}\partial\eta_2^*(\epsilon,\mathbf{X})/\partial\epsilon\} \times \{\partial m(\mathbf{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\} \end{bmatrix}, \end{aligned}$$

where $y = m(\mathbf{x},\boldsymbol{\beta}) + \epsilon$. Here η_2^* is the posited conditional density of ϵ given $\mathbf{X}$, not necessarily the true model. Let $w(d,\mathbf{x},y;\boldsymbol{\alpha}) = d - H(1,\mathbf{X},Y,\boldsymbol{\alpha})$, so that

$$\boldsymbol{\mu}_s = E(S^* \mid \epsilon,\mathbf{X}) = \begin{bmatrix} r(\mathbf{x},y;\boldsymbol{\alpha})(1,\mathbf{X}^{\mathrm{T}},Y)^{\mathrm{T}} \\ -\{\eta_2^*(\epsilon,\mathbf{X})^{-1}\partial\eta_2^*(\epsilon,\mathbf{X})/\partial\epsilon\} \times \{\partial m(\mathbf{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\} \end{bmatrix},$$

where

$$\begin{aligned} r(\mathbf{x},y;\boldsymbol{\alpha}) &= E\{w(D,\mathbf{X},Y) \mid \mathbf{X},Y\} \\ &= \textstyle\sum_{d=0}^{1} n_d H(d,\mathbf{X},Y) w(d,\mathbf{X},Y)\kappa(\mathbf{X},Y)/(n\pi_d), \\ &= n^{-1}\left(n_1/\pi_1 - n_0/\pi_0\right) H(0,\mathbf{X},Y)H(1,\mathbf{X},Y)\kappa(\mathbf{X},Y); \\ \kappa(\mathbf{X},Y) &= \left\{\textstyle\sum_{d=0}^{1} n_d H(d,\mathbf{X},Y)/(n\pi_d)\right\}^{-1}. \end{aligned}$$

Hence,

$$E_{\text{true}}\{\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y)\mid\mathbf{X}\} = \begin{bmatrix} E_{\text{true}}\{\epsilon r(\mathbf{X},Y;\boldsymbol{\alpha})\mid\mathbf{X}\}(1,\mathbf{X}^{\mathrm{T}})^{\mathrm{T}} \\ E_{\text{true}}\{\epsilon r(\mathbf{X},Y;\boldsymbol{\alpha})m(\mathbf{X},\boldsymbol{\beta})+\epsilon^2 r(\mathbf{X},Y;\boldsymbol{\alpha})\mid\mathbf{X}\} \\ -E_{\text{true}}\left\{\epsilon\eta_2^*(\epsilon,\mathbf{X})^{-1}\partial\eta_2^*(\epsilon,\mathbf{X})/\partial\epsilon\mid\mathbf{X}\right\}\left\{\partial m(\mathbf{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\right\} \end{bmatrix}.$$

We assume the following models hold.

$$\begin{aligned} E_{\text{true}}\{\epsilon r(\mathbf{X},Y;\boldsymbol{\alpha})\mid\mathbf{X}\} &= \zeta_{21}(\mathbf{Z}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\gamma}_{21}); && \text{(B.1)} \\ E_{\text{true}}\{\epsilon^2 r(\mathbf{X},Y;\boldsymbol{\alpha})\mid\mathbf{X}\} &= \zeta_{22}(\mathbf{Z}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\gamma}_{22}); && \text{(B.2)} \\ E_{\text{true}}\left\{\epsilon\eta_2^*(\epsilon,\mathbf{X})^{-1}\partial\eta_2^*(\epsilon,\mathbf{X})/\partial\epsilon\mid\mathbf{X}\right\} &= \zeta_{23}(\mathbf{Z}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\gamma}_{23}), && \text{(B.3)} \end{aligned}$$

where $\mathbf{Z}=\{\mathbf{X}^{\mathrm{T}},m(\mathbf{X},\boldsymbol{\beta})\}^{\mathrm{T}}$ when m is nonlinear while $\mathbf{Z}=\mathbf{X}$ when m is linear. For identifiability, the lower square blocks of $\boldsymbol{\gamma}_{2j}, j=1,2,3$ are fixed to be identity.

In models (B.1)-(B.3), $\zeta_{21},\zeta_{22},\zeta_{23}$ can be estimated by

$$\begin{aligned} \widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{21}) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}\epsilon_i r(\mathbf{X}_i,Y_i;\boldsymbol{\alpha})K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{21}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{21})}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{21}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{21})}; && \text{(B.4)} \\ \widehat{\zeta}_{22}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{22}) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}\epsilon_i^2 r(\mathbf{X}_i,Y_i;\boldsymbol{\alpha})K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{22}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{22})}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{22}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{22})}; && \text{(B.5)} \\ \widehat{\zeta}_{23}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{23}) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}\epsilon_i\frac{\partial\eta_2^*(\epsilon_i,\mathbf{X}_i)/\partial\epsilon_i}{\eta_2^*(\epsilon_i,\mathbf{X}_i)}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{23}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{23})}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d\sum_{i=1}^{N}I\{D_i=d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{23}-\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_{23})}. && \text{(B.6)} \end{aligned}$$

To get a consistent estimate of $\boldsymbol{\gamma}_{21,-1}$, we solve

$$\begin{aligned} \mathbf{0} \;=\; & \sum_{d=0}^{1}\frac{\widehat{\pi}_d}{n_d}\sum_{j=1}^{N}I(D_j=d) \\ & \times\left\{\epsilon_j(\mathbf{X}_j,Y_j,\boldsymbol{\beta})r(\mathbf{X}_j,Y_j,\boldsymbol{\alpha})-\widehat{\zeta}_{21}(\mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_{21})\right\}\left\{\mathbf{Z}_{\boldsymbol{\beta},j}^*-\widehat{E}_{\text{true}}^{\widehat{\pi}}(\mathbf{Z}_{\boldsymbol{\beta},j}^*\mid\mathbf{Z}_{\boldsymbol{\beta},j}^*\gamma)\right\}. \end{aligned}$$

Similar results work for $\boldsymbol{\gamma}_{22,-1}$ and $\boldsymbol{\gamma}_{23,-1}$. Denote the resulting estimators by $\widehat{\boldsymbol{\gamma}}_{2j,-1}$ and let $\widehat{\gamma}_{2j} = (\widehat{\gamma}^{\mathrm{T}}_{2j,-1}, 1)^{\mathrm{T}}$ for $j = 1, 2, 3$. Then $E_{\text{true}}\{\epsilon \boldsymbol{\mu}_s(\mathbf{X}, Y) \mid \mathbf{X}\}$ can be estimated by

$$\widehat{E}_{\text{true}}\{\epsilon \widehat{\boldsymbol{\mu}}_s(\mathbf{X}, Y) \mid \mathbf{X}\} = \begin{bmatrix} \widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{21})(1, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}} \\ \widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{21})m(\mathbf{X}, \boldsymbol{\beta}) + \widehat{\zeta}_{22}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{22}) \\ -\widehat{\zeta}_{23}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{23})\{\partial m(\mathbf{X}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\} \end{bmatrix}.$$

In all of our simulations, $m(\mathbf{X}, \boldsymbol{\beta}) = \beta_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1$. In addition, the posited model is standard normal, and simplifications result. Thus, $\partial\{\log\eta_2^*(\epsilon, \mathbf{X})\}/\partial\epsilon$ is simply $-\epsilon$. In our simulations, we further take $\boldsymbol{\gamma}_{21} = \boldsymbol{\gamma}_{22} = \boldsymbol{\gamma}_{23} = \boldsymbol{\gamma}_2$ for computational and programming simplicity. As a result, we have that

$$\begin{aligned} S^* &= \left[\{d - H(1, \mathbf{X}, Y, \boldsymbol{\alpha})\}(1, \mathbf{X}^{\mathrm{T}}, Y), \epsilon(1, \mathbf{X}^{\mathrm{T}})\right]^{\mathrm{T}}; \\ \boldsymbol{\mu}_s &= E\{S^* \mid \epsilon, \mathbf{X}\} = \{r(\mathbf{X}, Y; \boldsymbol{\alpha})(1, \mathbf{X}^{\mathrm{T}}, Y), \epsilon(1, \mathbf{X}^{\mathrm{T}})\}^{\mathrm{T}}. \end{aligned}$$

Then

$$E_{\text{true}}\{\epsilon \boldsymbol{\mu}_s(\mathbf{X}, Y) \mid \mathbf{X}\} = \begin{bmatrix} E_{\text{true}}\{\epsilon r(\mathbf{X}, Y; \boldsymbol{\alpha}) \mid \mathbf{X}\}(1, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}} \\ E_{\text{true}}\{\epsilon r(\mathbf{X}, Y; \boldsymbol{\alpha}) \mid \mathbf{X}\}m(\mathbf{X}, \boldsymbol{\beta}) + E_{\text{true}}\{\epsilon^2 r(\mathbf{X}, Y; \boldsymbol{\alpha}) \mid \mathbf{X}\} \\ E_{\text{true}}\{\epsilon^2 \mid \mathbf{X}\}(1, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}} \end{bmatrix}.$$

Under the assumption $\boldsymbol{\gamma}_{21} = \boldsymbol{\gamma}_{22} = \boldsymbol{\gamma}_{23} = \boldsymbol{\gamma}_2$ in (B.4) -(B.6), $\zeta_{21}, \zeta_{22}, \zeta_{23}$ can be estimated by

$$\begin{aligned} \widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i r(\mathbf{X}_i, Y_i; \boldsymbol{\alpha})K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}; \\ \widehat{\zeta}_{22}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i^2 r(\mathbf{X}_i, Y_i; \boldsymbol{\alpha})K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}; \\ \widehat{\zeta}_{23}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2) &= \frac{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i^2 K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}{\sum_{d=0}^{1}\widehat{\pi}_d/n_d \sum_{i=1}^{N} I\{D_i = d\}K_h(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_2 - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\gamma}_2)}, \end{aligned}$$

where again the lower square block of $\boldsymbol{\gamma}_2$ is fixed to be identity. The consistent estimator of $\boldsymbol{\gamma}_{2,-1}$ can be obtained through solving

$$\begin{aligned}\mathbf{0} \;=\; & \sum_{d=0}^{1}\frac{\widehat{\pi}_d}{n_d}\textstyle\sum_{j=1}^{N} I(D_j=d)\Big[\{\epsilon_j(\mathbf{X}_j,Y_j,\boldsymbol{\beta})r(\mathbf{X}_j,Y_j,\boldsymbol{\alpha})-\widehat{\zeta}_{21}(\mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_2)\}\\ & \qquad +\{\epsilon_i^2 r(\mathbf{X}_i,Y_i;\boldsymbol{\alpha})-\widehat{\zeta}_{22}(\mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_2)\}+\{\epsilon_i^2-\widehat{\zeta}_{23}(\mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_2)\}\Big]\\ & \qquad \times\left\{\mathbf{Z}^*_{\boldsymbol{\beta},j}-\widehat{E}^{\widehat{\pi}}_{\mathrm{true}}(\mathbf{Z}^*_{\boldsymbol{\beta},j}\mid \mathbf{Z}^*_{\boldsymbol{\beta},j}\gamma)\right\}.\end{aligned}$$

Denote the resulting estimators $\widehat{\boldsymbol{\gamma}}_{2,-1}$ and let $\widehat{\boldsymbol{\gamma}}_2=(\widehat{\boldsymbol{\gamma}}_{2,-1}^{\mathrm{T}},1)^{\mathrm{T}}$. Then $E_{\mathrm{true}}\{\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y)\mid \mathbf{X}\}$ can be estimated by

$$\widehat{E}_{\mathrm{true}}\{\epsilon\widehat{\boldsymbol{\mu}}_s(\mathbf{X},Y)\mid\mathbf{X}\}=\begin{bmatrix}\widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_2)(1,\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}\\ \widehat{\zeta}_{21}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_2)m(\mathbf{X},\boldsymbol{\beta})+\widehat{\zeta}_{22}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_2)\\ -\widehat{\zeta}_{23}(\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_2)\{\partial m(\mathbf{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}\end{bmatrix}.$$

B.2 Details for the Algorithm in Section 3.4

B.2.1 Algorithm Using Different Indices

1. Posit a model for $\eta_2(\epsilon,\mathbf{x})$ which has mean zero. Under this posited model, calculate S^* from (3.6).

2. Solve $\widehat{\pi}_0(\boldsymbol{\alpha})=\sum_{i=1}^{n}H(0,\mathbf{X}_i,Y_i,\boldsymbol{\alpha})[n_0H(0,\mathbf{X}_i,Y_i,\boldsymbol{\alpha})/\widehat{\pi}_0(\boldsymbol{\alpha})+n_1H(1,\mathbf{X}_i,Y_i,\boldsymbol{\alpha})/\{1-\widehat{\pi}_0(\boldsymbol{\alpha})\}]^{-1}$, and set $\widehat{\pi}_1(\boldsymbol{\alpha})=1-\widehat{\pi}_0(\boldsymbol{\alpha})$.

3. Obtain

$$\begin{aligned}
\widehat{\kappa}_i &= \widehat{\kappa}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) = [\textstyle\sum_d n_d H(d, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})/\{n\widehat{\pi}_d(\boldsymbol{\alpha})\}]^{-1} \\
\widehat{f}_{0i} &= \widehat{f}_{D|X,Y}(0, \mathbf{X}_i, Y_i, \boldsymbol{\alpha}) = n_0 H(0, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})\widehat{\kappa}_i/\{n\widehat{\pi}_0(\boldsymbol{\alpha})\} \\
\widehat{f}_{1i} &= \widehat{f}_{D|X,Y}(1, \mathbf{X}_i, Y_i, \boldsymbol{\alpha}) = n_1 H(1, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})\widehat{\kappa}_i/\{n\widehat{\pi}_1(\boldsymbol{\alpha})\} \\
\widehat{\boldsymbol{\mu}}_{si} &= \widehat{E}(\mathbf{S}_i^* \mid \epsilon_i, \mathbf{X}_i, \boldsymbol{\alpha}) = \textstyle\sum_d n_d H(d, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})\mathbf{S}^*(d, \mathbf{X}_i, Y_i, \boldsymbol{\alpha})\widehat{\kappa}_i/\{n\widehat{\pi}_d(\boldsymbol{\alpha})\} \\
\widehat{b}_0 &= \textstyle\sum_{i=1}^N \widehat{f}_{1i}\widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i} \\
\widehat{b}_1 &= \textstyle\sum_{i=1}^N \widehat{f}_{0i}\widehat{f}_{1i}/\sum_{i=1}^N \widehat{f}_{1i} \\
\widehat{\mathbf{c}}_0 &= \textstyle\sum_{i=1}^N \{\mathbf{S}^*(0, \mathbf{X}_i, Y_i, \boldsymbol{\alpha}) - \widehat{\boldsymbol{\mu}}_{si}\}\, \widehat{f}_{0i}/\sum_{i=1}^N \widehat{f}_{0i} \\
\widehat{\mathbf{c}}_1 &= \textstyle\sum_{i=1}^N \{\mathbf{S}^*(1, \mathbf{X}_i, Y_i, \boldsymbol{\alpha}) - \widehat{\boldsymbol{\mu}}_{si}\}\, \widehat{f}_{1i}/\sum_{i=1}^N \widehat{f}_{1i}.
\end{aligned}$$

4. Estimate $E_{\text{true}}\{\epsilon^2\widehat{\kappa}(\mathbf{X}, Y) \mid \mathbf{X}\}$ using nonparametric regression under the dimension reduction model assumption (3.8).

   (a) Let

$$\begin{aligned}
&\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \boldsymbol{\theta}) \\
&\quad \equiv \frac{\sum_{d=0}^1 \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha})K_h(\mathbf{Z}_{\boldsymbol{\beta},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\boldsymbol{\beta},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)}{\sum_{d=0}^1 \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d)K_h(\mathbf{Z}_{\boldsymbol{\beta},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\boldsymbol{\beta},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)},
\end{aligned}$$

   for $j = 1, \cdots, n$. Here $\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta}) = Y_i - m(\mathbf{X}_i, \boldsymbol{\beta})$. $\mathbf{Z}_{\boldsymbol{\beta},i} = \mathbf{X}$ if $m(\cdot)$ is linear in $\mathbf{X}$; $\mathbf{Z} = \{\mathbf{X}_i^{\mathrm{T}}, m(\mathbf{X}_i, \boldsymbol{\beta})\}^{\mathrm{T}}$, otherwise.

   (b) Estimate $\boldsymbol{\gamma}_{1,-1}$ through solving

$$\begin{aligned}
\mathbf{0} &= \sum_{d=0}^1 \widehat{\pi}_d(\boldsymbol{\alpha})/n_d \textstyle\sum_{j=1}^N I(D_j = d)\{\epsilon_j^2(\mathbf{X}_j, Y_j, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_j, Y_j, \boldsymbol{\alpha}) \\
&\qquad -\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \boldsymbol{\theta})\}\{\mathbf{Z}_{\boldsymbol{\beta},j}^* - \widehat{E}_{\text{true}}^{\widehat{\pi}}(\mathbf{Z}_{\boldsymbol{\beta},j}^* \mid \mathbf{Z}_{\boldsymbol{\beta},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\},
\end{aligned}$$

where $\mathbf{Z}^*_{\boldsymbol{\beta},j}$ is the subvector or submatrix of $\mathbf{Z}_{\boldsymbol{\beta},j}$ without the lower square block and

$$\widehat{E}^{\widehat{\pi}}_{\text{true}}(\mathbf{Z}^*_{\boldsymbol{\beta},j} \mid \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_1) = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d) \mathbf{Z}^*_{\boldsymbol{\beta},i} K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_1)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_1)}.$$

Let the solution be $\widehat{\boldsymbol{\gamma}}_{1,-1}$. Denote $\widehat{\boldsymbol{\gamma}}_1 = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}_{1,-1}, 1)^{\mathrm{T}}$.

(c) Form

$$\begin{aligned}&\widehat{E}_{\text{true}}\left\{\epsilon^2(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \\ &\quad = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i=1}^{N} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}}_1 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}}_1)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i=1}^{N} I(D_i = d) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}}_1 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}}_1)}.\end{aligned}$$

5. Estimate $E_{\text{true}}\{\epsilon\widehat{\boldsymbol{\mu}}_s(\mathbf{X}, Y) \mid \mathbf{X}\}$ using nonparametric regression under the dimension reduction model assumption (3.9). Because $E_{\text{true}}\{\epsilon\widehat{\boldsymbol{\mu}}_s(\mathbf{X}, Y) \mid \mathbf{X}\}$ actually consists of three separate dimension reduction models, its estimation is slightly complex. We give the estimation details in Appendix B.1 and denote the resulting estimator by $\widehat{E}_{\text{true}}\left\{\epsilon(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{\boldsymbol{\mu}}_s(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\}$.

6. Estimate $E_{\text{true}}\{\epsilon\widehat{f}_0(\mathbf{X}, Y) \mid \mathbf{X}\}$ using nonparametric regression under the dimension reduction model assumption (3.10).

(a) Let

$$\begin{aligned}&\widehat{E}^{\widehat{\pi}}_3(\mathbf{X}_j, \boldsymbol{\gamma}_3, \boldsymbol{\theta}) \\ &\quad \equiv \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d)\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_3 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I(D_i = d) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_3 - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3)},\end{aligned}$$

for $j = 1, \cdots, n$.

(b) Estimate $\boldsymbol{\gamma}_{3,-1}$ by solving

$$\begin{aligned}\mathbf{0} \;=\; \sum_{d=0}^{1} & \widehat{\pi}_d(\boldsymbol{\alpha})/n_d \textstyle\sum_{j=1}^{N} I(D_j = d)\{\epsilon_j(\mathbf{X}_j, Y_j, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_j, Y_j, \boldsymbol{\alpha}) \\ & -\widehat{E}_3^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_3, \boldsymbol{\theta})\}\{\mathbf{Z}^*_{\boldsymbol{\beta},j} - \widehat{E}^{\widehat{\pi}}_{\rm true}(\mathbf{Z}^*_{\boldsymbol{\beta},j} \mid \mathbf{Z}^{\rm T}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3)\},\end{aligned}$$

where

$$\widehat{E}^{\widehat{\pi}}_{\rm true}(\mathbf{Z}^*_{\boldsymbol{\beta},j} \mid \mathbf{Z}^{\rm T}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3) = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i\neq j} I(D_i = d)\mathbf{Z}^*_{\boldsymbol{\beta},i} K_h(\mathbf{Z}^{\rm T}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_3 - \mathbf{Z}^{\rm T}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i\neq j} I(D_i = d) K_h(\mathbf{Z}^{\rm T}_{\boldsymbol{\beta},i}\boldsymbol{\gamma}_3 - \mathbf{Z}^{\rm T}_{\boldsymbol{\beta},j}\boldsymbol{\gamma}_3)}.$$

Let the minimizer be $\widehat{\boldsymbol{\gamma}}_{3,-1}$. Denote $\widehat{\boldsymbol{\gamma}}_3 = (\widehat{\boldsymbol{\gamma}}^{\rm T}_{3,-1}, 1)^{\rm T}$.

(c) Form

$$\begin{aligned}&\widehat{E}_{\rm true}\left\{\epsilon(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \\ &\quad = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i=1}^{N} I(D_i = d)\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\rm T}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}}_3 - \mathbf{Z}^{\rm T}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}}_3)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i=1}^{N} I(D_i = d) K_h(\mathbf{Z}^{\rm T}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}}_3 - \mathbf{Z}^{\rm T}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}}_3)}.\end{aligned}$$

7. (a) Form $\widehat{t}_1(\mathbf{X}) = \{\widehat{E}_{\rm true}(\epsilon^2\widehat{\kappa}(\mathbf{X}, Y)|\mathbf{X})\}^{-1}$, $\widehat{\mathbf{t}}_2(\mathbf{X}) = \widehat{E}_{\rm true}(\epsilon\widehat{\boldsymbol{\mu}}_s \mid \mathbf{X}) - (\widehat{\mathbf{c}}_0/\widehat{b}_0) \times \widehat{E}_{\rm true}(\epsilon\widehat{f}_0 \mid \mathbf{X})$ and $\widehat{t}_3(\mathbf{x}) = -\widehat{b}_0^{-1}\widehat{E}_{\rm true}(\epsilon\widehat{f}_0 \mid \mathbf{x})$.

   (b) Form $\widehat{E}\{\epsilon t_1(\mathbf{X})t_3(\mathbf{X})\kappa(\mathbf{X}, Y) \mid D = 0\} = \sum_{i=1}^{n} \epsilon_i\widehat{t}_1(\mathbf{X}_i)\widehat{t}_3(\mathbf{X}_i)\widehat{\kappa}(\mathbf{X}_i, Y_i) \times \widehat{f}_{0i}/\sum_{i=1}^{n} \widehat{f}_{0i}$, $\widehat{E}\{\epsilon t_1(\mathbf{X})\mathbf{t}_2(\mathbf{X})\kappa(\mathbf{X}, Y) \mid D = 0\} = \sum_{i=1}^{n} \epsilon_i\widehat{t}_1(\mathbf{X}_i)\widehat{\mathbf{t}}_2(\mathbf{X}_i) \times \widehat{\kappa}(\mathbf{X}_i, Y_i)\widehat{f}_{0i}/\sum_{i=1}^{n} \widehat{f}_{0i}$ and $\widehat{\mathbf{u}}_0 = \Big(1 - \widehat{E}\left[\epsilon t_1(\mathbf{x})t_3(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\right]\Big)^{-1} \times \widehat{E}\left[\epsilon t_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})\kappa(\mathbf{x}, y) \mid D = 0\right]$.

   (c) Form $\widehat{\mathbf{u}}_1 = -(n_0/n_1)\widehat{\mathbf{u}}_0$, $\widehat{\mathbf{v}}_0 = (\widehat{\pi}_1/\widehat{b}_0)(\widehat{\mathbf{u}}_0 + \widehat{\mathbf{c}}_0)$ and $\widehat{\mathbf{v}}_1 = -(\widehat{\pi}_0/\widehat{b}_0)(\widehat{\mathbf{u}}_0 + \widehat{\mathbf{c}}_0)$.

   (d) Form $\widehat{\mathbf{a}}(\mathbf{x}) = \widehat{t}_1(\mathbf{x})\{\widehat{\mathbf{t}}_2(\mathbf{x}) + \widehat{t}_3(\mathbf{x})\widehat{\mathbf{u}}_0)$.

   (e) Form $\widehat{\mathbf{g}}_i = \widehat{\boldsymbol{\mu}}_{si} - \epsilon_i\widehat{\mathbf{a}}(\mathbf{X}_i)\widehat{\kappa}_i - \widehat{\mathbf{v}}_0\widehat{f}_{0i} - \widehat{\mathbf{v}}_1\widehat{f}_{1i}$.

   (f) Form $\widehat{\mathbf{v}}_{D_i} = (1 - D_i)\widehat{\mathbf{v}}_0 + D_i\mathbf{v}_1$.

(g) Form $\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i) = \mathbf{S}^*_i - \widehat{\mathbf{g}}_i - \widehat{\mathbf{v}}_{D_i}$ and solve the corresponding estimating equation.

B.2.2 Algorithm Using A Common Index

Specifically, we replace the steps 4-6 of Appendix B.2.1 with the following three steps.

1. Define

$$\widehat{E}^{\widehat{\pi}}_1(\mathbf{X}_j, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I\{D_i = d\}\epsilon_i^2(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha})K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\gamma - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\gamma)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I\{D_i = d\}K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\gamma - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\gamma)}; \tag{B.7}$$

$$\widehat{E}^{\widehat{\pi}}_3(\mathbf{X}_j, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I\{D_i = d\}\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_i, Y_i, \boldsymbol{\alpha})K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\gamma - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\gamma)}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i \neq j} I\{D_i = d\}K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\gamma - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\gamma)}. \tag{B.8}$$

Construct $\widehat{E}^{\widehat{\pi}}_2(\mathbf{X}_j, \boldsymbol{\gamma}_2, \boldsymbol{\theta}) = \widehat{E}_{\text{true}}\{\epsilon_j \widehat{\boldsymbol{\mu}}_s(\mathbf{X}_j, Y_j) \mid \mathbf{X}_j\}$ for $j = 1, \cdots, n$, with the method given in Appendix B.1.

2. Estimate $\boldsymbol{\gamma}_{-1}$ by solving

$$\begin{aligned}
\mathbf{0} \;=\; & \sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \textstyle\sum_{j=1}^{N} I(D_j = d) \\
& \times \Big[\epsilon_j^2(\mathbf{X}_j, Y_j, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_j, Y_j, \boldsymbol{\alpha}) - \widehat{E}^{\widehat{\pi}}_1(\mathbf{X}_j, \boldsymbol{\gamma}, \boldsymbol{\theta}) \\
& \qquad + \mathbf{1}^{\mathrm{T}}_{\dim(\boldsymbol{\theta})} \Big\{\epsilon_j(\mathbf{X}_j, Y_j, \boldsymbol{\beta})\widehat{\boldsymbol{\mu}}_s(\mathbf{X}_j, Y_j, \boldsymbol{\alpha}) - \widehat{E}^{\widehat{\pi}}_2(\mathbf{X}_j, \boldsymbol{\gamma}, \boldsymbol{\theta})\Big\} \\
& \qquad + \epsilon_j(\mathbf{X}_j, Y_j, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_j, Y_j, \boldsymbol{\alpha}) - \widehat{E}^{\widehat{\pi}}_3(\mathbf{X}_j, \boldsymbol{\gamma}, \boldsymbol{\theta})\Big] \\
& \times \Big\{\mathbf{Z}^*_{\boldsymbol{\beta},j} - \widehat{E}^{\widehat{\pi}}_{\text{true}}(\mathbf{Z}^*_{\boldsymbol{\beta},j} \mid \mathbf{Z}^*_{\boldsymbol{\beta},j}\gamma)\Big\},
\end{aligned}$$

where

$$\widehat{E}^{\widehat{\pi}}_{\text{true}}(\mathbf{Z}^*_{\boldsymbol{\beta},j} \mid \mathbf{Z}^*_{\boldsymbol{\beta},j}\boldsymbol{\gamma}) = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i\neq j, 1\leq i\leq n} I(D_i = d)\mathbf{Z}^*_{\boldsymbol{\beta},i} K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma})}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d(\boldsymbol{\alpha})}{n_d} \sum_{i\neq j, 1\leq i\leq n} I(D_i = d) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\boldsymbol{\gamma} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},j}\boldsymbol{\gamma})}.$$

Denote the solution by $\widehat{\boldsymbol{\gamma}}_{-1}$ and let $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}_{-1}, 1)^{\mathrm{T}}$.

3. Form

$$\begin{aligned}
&\widehat{E}_{\text{true}}\left\{\epsilon^2(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \\
&\quad = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i^2(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{\kappa}(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\} K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}; \\
&\widehat{E}_{\text{true}}\left\{\epsilon(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{\boldsymbol{\mu}}_s(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \\
&\quad = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{\boldsymbol{\mu}}_s(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\} K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}; \\
&\widehat{E}_{\text{true}}\left\{\epsilon(\mathbf{X}, Y, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \\
&\quad = \frac{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\}\epsilon_i(\mathbf{X}_i, Y_i, \boldsymbol{\beta})\widehat{f}_0(\mathbf{X}_i, Y_i, \boldsymbol{\alpha}) K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}{\sum_{d=0}^{1} \frac{\widehat{\pi}_d}{n_d} \sum_{i=1}^{N} I\{D_i = d\} K_h(\mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta},i}\widehat{\boldsymbol{\gamma}} - \mathbf{Z}^{\mathrm{T}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}})}.
\end{aligned}$$

B.3 Regularity Conditions

Let ℓ be the dimensionality of the kernel regressions in our method after dimension reduction. In our simulations and example, we took $\ell = 1$. The set of regularity conditions required by Theorem 3 is listed below.

C1. The univariate kernel function is a function that integrates to 1 and has support $(-1, 1)$ and order r, i.e., $\int K(u)u^t du = 0$ if $1 \leq t < r$ and $\int K(u)u^r du \neq 0$. The ℓ-dimensional kernel function, still represented with K, is a product of ℓ univariate kernel functions, that is, $K(\mathbf{u}) = \prod_{i=1}^{\ell} K(u_i)$ for a ℓ-dimensional $\mathbf{u}$.

C2. Let $\xi^{\text{true}}_{i,\widetilde{\boldsymbol{\beta}}}$ be the true population density of $\mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_i$ for $i = 1, 2, 3$ and $\widetilde{\boldsymbol{\beta}}$ in a local neighborhood of $\boldsymbol{\beta}$. Assume that $\xi^{\text{true}}_{i,\widetilde{\boldsymbol{\beta}}}$'s are bounded away from 0 and they all have third order bounded and continuous derivatives.

C3. At any fixed $\widetilde{\boldsymbol{\alpha}}$ in a local neighborhood of $\boldsymbol{\alpha}$, $\zeta_1(\cdot,\widetilde{\boldsymbol{\alpha}})$, $\boldsymbol{\zeta}_2(\cdot,\widetilde{\boldsymbol{\alpha}})$ and $\zeta_3(\cdot,\widetilde{\boldsymbol{\alpha}})$ are functions of $\cdot$ with second order bounded and continuous derivatives.

C4. $E_{\text{true}}\{\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})|\mathbf{X}\}$, $E_{\text{true}}\{\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\boldsymbol{\mu}_s(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})^{\otimes 2}|\mathbf{X}\}$ and $E_{\text{true}}\{\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})f_0^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})|\mathbf{X}\}$ are bounded for any $\widetilde{\boldsymbol{\theta}}$ in a local neighborhood of $\boldsymbol{\theta}$.

C5. $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}}) \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}}) \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2\right\}$ and $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon f_0(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}}) \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3\right\}$ have $(r+1)$th order bounded and continuous derivatives for any $\widetilde{\boldsymbol{\theta}}$ in a local neighborhood of $\boldsymbol{\theta}$.

C6. $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1\right\}$, $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon\boldsymbol{\mu}_s(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2\right\}$ and $E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon f_0(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3\right\}$ have $(r+1)$th order bounded and continuous derivatives for any $\widetilde{\boldsymbol{\theta}}$ in a local neighborhood of $\boldsymbol{\theta}$.

C7. $E_{\text{true}}\left[\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1,\mathbf{X}\right]$,
$E_{\text{true}}\left[\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2,\mathbf{X}\right]$,
$E_{\text{true}}\left[\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3,\mathbf{X}\right]$,
$E_{\text{true}}\left[\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_1,\mathbf{X}\right]$,
$E_{\text{true}}\left[\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\boldsymbol{\mu}_s(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\boldsymbol{\mu}_s(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_2,\mathbf{X}\right]$,
and $E_{\text{true}}\left[\epsilon^2 f_0^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}\{\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}-\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}\}^{\text{T}} \mid \mathbf{Z}^{\text{T}}_{\widetilde{\boldsymbol{\beta}}}\boldsymbol{\gamma}_3,\mathbf{X}\right]$ all have bounded entries for any $\widetilde{\boldsymbol{\theta}}$ in a local neighborhood of $\boldsymbol{\theta}$, where $\mathbf{Z}'_{\widetilde{\boldsymbol{\beta}}}$ is an independent and identically distributed copy of $\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}$.

C8. $\pi_d(\widetilde{\boldsymbol{\alpha}})/(n_d/n)$ are bounded for $d = 0, 1$.

C9. $\pi_d(\widetilde{\boldsymbol{\alpha}})/\pi_d(\boldsymbol{\alpha})$ are bounded for $d = 0, 1$.

C10. The bandwidth $h = n^{-\tau}$ where $1/(2\ell) > \tau > 1/(4r)$. This includes the optimal bandwidth $h = O(n^{-1/(2r+\ell)})$ as long as we choose a kernel of order $2r > \ell$.

C11. There exists a positive constant C such that $\lim_{n\to\infty} n_0/n_1 = C < \infty$.

Conditions C1 and C10 are standard requirements on an rth order kernel function and the bandwidth, which ensure the resulting nonparametric regression estimators to be consistent (Ma and Zhu, 2013).

B.4 Proof of Proposition 1

We provide a detailed proof that the first dimension reduction model (3.8) satisfies Proposition 1. Proving that the other two dimension reduction models (3.9) and (3.10) also satisfy Proposition 1 is similar.

In (3.8), $\kappa(\mathbf{x}, y, \boldsymbol{\alpha})$ is a function of the weighted sum of $H(d, \mathbf{x}, y)$ with $d = 0, 1$. As a result,

$$\begin{aligned} \kappa(\mathbf{x}, y, \boldsymbol{\alpha}) &= h\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + m(\mathbf{x}, \boldsymbol{\beta})\alpha_2 + \epsilon\alpha_2\} \\ &= h[\{\mathbf{x}^{\mathrm{T}}, m(\mathbf{x}, \boldsymbol{\beta})\}(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}} + \epsilon\alpha_2], \end{aligned}$$

where $h(\cdot)$ is a differentiable function.

For $\epsilon = Q(\mathbf{X}^{\mathrm{T}}\omega, \epsilon^*)$, where $Q(\cdot)$ is an arbitrary function,

$$\begin{aligned}
&E\{\epsilon^2\kappa(\mathbf{X}, Y, \boldsymbol{\alpha}) \mid \mathbf{X}\} \\
&\quad = E(\epsilon^2 h[\{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \boldsymbol{\beta})\}(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}} + \epsilon\alpha_2] \mid \mathbf{X}) \\
&\quad = E\{Q(\mathbf{X}^{\mathrm{T}}\omega, \epsilon^*)^2 h[\{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \boldsymbol{\beta})\}(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}} + Q(\mathbf{X}^{\mathrm{T}}\omega, \epsilon^*)\alpha_2] \mid \mathbf{X}\} \\
&\quad = \zeta_1(\mathbf{X}^{\mathrm{T}}\omega, \{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \boldsymbol{\beta})\}(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}}) \\
&\quad = \zeta_1(\mathbf{Z}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\gamma}_1),
\end{aligned}$$

where $\zeta_1(\cdot)$ is a smooth function, $\mathbf{Z}_{\boldsymbol{\beta}} = \mathbf{X}$ and $\boldsymbol{\gamma}_1 = (\omega, \boldsymbol{\alpha}_1 + \alpha_2\boldsymbol{\beta})$ is a $p \times 2$ matrix if $m(\cdot)$ is linear in $\mathbf{X}$; otherwise, $\mathbf{Z}_{\boldsymbol{\beta}} = \{\mathbf{X}^{\mathrm{T}}, m(\mathbf{X}, \boldsymbol{\beta})\}^{\mathrm{T}}$ and $\boldsymbol{\gamma}_1 = \{(\omega^{\mathrm{T}}, 0)^{\mathrm{T}}, (\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}}\}$ is a $(p+1) \times 2$ matrix.

B.5 Background and Technical Results

B.5.1 Introduction

Following Ma and Carroll (2016), we divide the n observations randomly into three sets, where the first set contains $N_1 = n - n^{1-\delta} - n^{1-2\delta}$ observations, the second set contains $N_2 = n^{1-\delta}$ observations and the third set contains $N_3 = n^{1-2\delta}$ observations, where δ is a small positive number. For convenience of proof, we require the disease proportion in the third data set to be the same as the whole data set. That is, $N_{30}/N_{31} = n_0/n_1$, where N_{30} and N_{31} are the numbers of controls and cases in the third set of data, respectively. We form and solve the estimating equation (3.5) using data in the first set while calculating all the estimated quantities described in Appendix B.2 steps 1-3 using data in the second set and the other estimated quantities defined in Appendix B.2 steps 4-6 using the data in the third set.

B.5.2 Lemmas

Before proving Theorem 3, we first state several lemmas, which ensure the quantities defined in Appendix B.2 steps 4-6 have the desired orders of bias and mean square error, i.e., the same as that of the usual nonparametric estimators.

From (3.12), we can easily show that

<u>Lemma</u> 5. *For some* $\sigma^2_{\pi_d(\widetilde{\boldsymbol{\alpha}})} < \infty$, $\sqrt{N_2}\{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}}) - \pi_d(\widetilde{\boldsymbol{\alpha}})\} \xrightarrow{d}$ *Normal*$(0, \sigma^2_{\pi_d(\widetilde{\boldsymbol{\alpha}})})$,*as* $n \to \infty$.

We now analyze the property of our estimators defined in Appendix B.2 steps 4-6. For notational brevity, we only focus on the first conditional expectation $E_{\text{true}}\{\epsilon^2\kappa(\mathbf{X}, Y)|\mathbf{X}\}$. The other two conditional expectations have similar properties. We split the analysis into three parts: i) analyze the properties of $\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})$; ii) analyze the properties of $\widehat{\boldsymbol{\gamma}}_1(\widetilde{\boldsymbol{\theta}})$ for $\widetilde{\boldsymbol{\theta}}$ near $\boldsymbol{\theta}$; iii) show that $\widehat{E}_{\text{true}}\{\epsilon^2(\mathbf{X}, Y, \widetilde{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}, Y, \widetilde{\boldsymbol{\alpha}}) \mid \mathbf{X}\}$ has desired bias order and standard deviation order.

For the first part of the analysis, we establish the following lemma.

<u>Lemma</u> 6. *Under the regularity conditions C1-C10,*

$$\begin{aligned}
\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) &= \widehat{E}_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) + O_p(n_2^{-1/2}) \\
&= E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) + O_p(h^r) + O_p\left(N_3^{-1/2}h^{-\ell/2}\right),
\end{aligned}$$

where

$$
\begin{aligned}
&\widehat{E}_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) \\
&\quad = \frac{\sum_{d=0}^{1} \frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)}{\sum_{d=0}^{1} \frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)}; \\
&E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) \\
&\quad = \frac{E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\epsilon_j^2(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\alpha}})|\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}}{E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})} \mid \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}}.
\end{aligned}
$$

Proof. Denote the numerator and denominator of $\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})$ by q_{num} and q_{den} respectively. We can replace $\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}})$ in q_{num} and q_{den} with $\pi_d(\widetilde{\boldsymbol{\alpha}})$ without changing the error order due to the data partition scheme we use. That is,

$$
\begin{aligned}
q_{\text{num}} &= (N_3 - 1)^{-1} \sum_{d=0}^{1} \frac{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}}) \\
&\qquad K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&= (N_3 - 1)^{-1} \sum_{d=0}^{1} \frac{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\beta}}) \\
&\qquad \times\{\widehat{\kappa}(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}}) - \kappa(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}})\}K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&\quad +(N_3 - 1)^{-1} \sum_{d=0}^{1} \frac{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}}) - \pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}}) \\
&\qquad \times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&\quad +(N_3 - 1)^{-1} \sum_{d=0}^{1} \frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i \neq j, 1 \leq i \leq N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widetilde{\boldsymbol{\alpha}}) \\
&\qquad \times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1).
\end{aligned}
$$

With further calculations, this means that

$$
\begin{aligned}
q_{\text{num}} &= O_p(N_2^{-1/2})(N_3-1)^{-1}{\textstyle\sum_{d=0}^{1}}\frac{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i=d)\epsilon_i^2(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\beta}}) \\
&\quad\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&\quad+O_p(N_2^{-1/2})(N_3-1)^{-1}{\textstyle\sum_{d=0}^{1}}\frac{1}{N_{3d}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i=d)\epsilon_i^2(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\beta}}) \\
&\quad\times\kappa(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&\quad+(N_3-1)^{-1}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i=d)\epsilon_i^2(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\beta}}) \\
&\quad\times\kappa(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&= (N_3-1)^{-1}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i=d)\epsilon_i^2(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_i,Y_i,\widetilde{\boldsymbol{\alpha}}) \\
&\quad\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)+O_p(N_2^{-1/2}).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
q_{\text{den}} &= (N_3-1)^{-1}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i=d)K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},i}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \\
&\quad+O_p(N_2^{-1/2}).
\end{aligned}
$$

We now analyze the conditional expectations of q_{num} and q_{den} given $\mathbf{X}_j$ one by one.

First,

$$\begin{aligned}
&E(q_{\text{num}}|\mathbf{X}_j)\\
&\quad= \textstyle\sum_{d=0}^{1}\pi_d(\widetilde{\boldsymbol{\alpha}})E\left\{\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid D=d,\mathbf{X}_j\right\}\\
&\qquad+O_p(N_2^{-1/2})\\
&\quad= E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right\}\\
&\qquad+O_p(N_2^{-1/2})\\
&\quad= E_{\text{true}}\left[E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1,\mathbf{X}_j\right\}\right.\\
&\qquad\left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right]+O_p(N_2^{-1/2})\\
&\quad= E_{\text{true}}\left[E_{\text{true}}\left\{\frac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}\right.\\
&\qquad\left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right]+O_p(N_2^{-1/2})\\
&\quad= E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}})\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}\xi_1^{\text{true}}(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)+O_p(h^r)\\
&\qquad+O_p(N_2^{-1/2}).
\end{aligned}$$

Here we used the regularity conditions C1-C2, C5, C8-C10.

In addition, with the regularity conditions C1-C4 and C8-C10, we have

$$
\begin{aligned}
&\text{var}\left(q_{\text{num}} \mid \mathbf{X}_j\right) \\
&\quad = (N_3-1)^{-1}\text{var}\left\{\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} I(D=d)\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\right. \\
&\qquad\qquad \left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \mid \mathbf{X}_j\right\} + O_p(N_2^{-1}) \\
&\quad = (N_3-1)^{-1}\left(E\left[\left\{\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} I(D=d)\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\right.\right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\right\}^2 \mid \mathbf{X}_j\right] \\
&\qquad\qquad\qquad -E\left\{\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} I(D=d)\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1) \mid \mathbf{X}_j\right\}^2\right) \\
&\qquad\qquad\qquad +O_p(N_2^{-1}) \\
&\quad = (N_3-1)^{-1}\left(E\left[\left\{\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3} I(D=d)\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\right.\right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\times K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1 - \mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\right\}^2 \mid \mathbf{X}_j\right] \\
&\qquad\qquad\qquad \left.-E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}}) \mid \mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}^2 \xi_1^{\text{true}}(\mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\gamma}_1)^2\right) \\
&\qquad\qquad\qquad +O_p\{(N_3-1)^{-1}h^r\} + O_p(N_2^{-1}) \\
&\quad = O_p\left(N_3^{-1}h^{-\ell}\right).
\end{aligned}
$$

The last equality is because

$$\begin{aligned}
&E\left[\left\{\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}I(D=d)\epsilon^2(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\right\}^2\mid\mathbf{X}_j\right]\\
&\quad\le 2E\left[\textstyle\sum_{d=0}^{1}\dfrac{\pi_d^2(\widetilde{\boldsymbol{\alpha}})}{N_{3d}^2/N_3^2}I(D=d)\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h^2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right]\\
&\quad= 2E_{\text{true}}\left[\dfrac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{N_{3D}/N_3}\dfrac{\pi_D(\widetilde{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h^2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right]\\
&\quad\le CE_{\text{true}}\left\{\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})K_h^2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right\}\\
&\quad= CE_{\text{true}}\left[E_{\text{true}}\left\{\epsilon^4(\mathbf{X},Y,\widetilde{\boldsymbol{\beta}})\kappa^2(\mathbf{X},Y,\widetilde{\boldsymbol{\alpha}})\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}K_h^2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right]\\
&\quad\le C'E_{\text{true}}\left\{K_h^2(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1-\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)\mid\mathbf{X}_j\right\}\\
&\quad= O_p(h^{-\ell}),
\end{aligned}$$

where C, C' are constants.

Similarly, we have that

$$\begin{aligned}
E(q_{\text{den}}|\mathbf{X}_j) &= E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}\xi_1^{\text{true}}(\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1)+O_p(h^r)+O_p(N_2^{-1/2});\\
\text{var}(q_{\text{den}}|\mathbf{X}_j) &= O_p\left(N_3^{-1}h^{-\ell}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}}) &= \frac{E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}})|\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}}{E_{\text{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\mid\mathbf{Z}_{\widetilde{\boldsymbol{\beta}},j}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}}\\
&\quad +O_p(h^r)+O_p\left(N_3^{-1/2}h^{-\ell/2}\right).
\end{aligned}$$

Particularly, when $\widetilde{\boldsymbol{\theta}}=\boldsymbol{\theta}$, we have $\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta}) = E_{\text{true}}\{\epsilon_j^2\kappa(\mathbf{X}_j,Y_j)|\mathbf{X}_j\}+O_p(h^r)+O_p\left(N_3^{-1/2}h^{-\ell/2}\right)$. □

Lemma 7. *Under the regularity conditions C1-C10,*

$$\begin{aligned}\widehat{E}^{\widehat{\pi}}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1) &= \widehat{E}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1) + O_p(n_2^{-1/2}) \\ &= E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) + O_p(h^r) + O_p\left(N_3^{-1/2}h^{-\ell/2}\right),\end{aligned}$$

where

$$\begin{aligned}&\widehat{E}_{\mathrm{true}}\left(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1\right) \\ &\quad = \frac{N_3^{-1}\sum_{r=0}^{1}\frac{\pi_r(\widetilde{\boldsymbol{\alpha}})}{N_{3r}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i = r)\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},i}K_h(\mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},i}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)}{N_3^{-1}\sum_{r=0}^{1}\frac{\pi_r(\widetilde{\boldsymbol{\alpha}})}{N_{3r}/N_3}\sum_{i\neq j,1\leq i\leq N_3} I(D_i = r)K_h(\mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},i}\boldsymbol{\gamma}_1 - \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)}; \\ &E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) = \frac{E_{\mathrm{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})}\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} | \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1\right\}}{E_{\mathrm{true}}\left\{\frac{\pi_{D_j}(\widetilde{\boldsymbol{\alpha}})}{\pi_{D_j}(\boldsymbol{\alpha})} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1\right\}}.\end{aligned}$$

We skip the proof of the Lemma 3 here since it is similar to the proof of Lemma 2. Next, we establish the root-N_3 consistency of $\widehat{\boldsymbol{\gamma}}_{j,-1}$ for $j = 1, \cdots, 3$.

Lemma 8. *Under the regularity conditions C1-C10,*

$$\begin{aligned}\sqrt{N_3}\{\widehat{\boldsymbol{\gamma}}_{1,-1}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_{1,-1}\} &\to \mathit{Normal}(0, \Sigma_{\boldsymbol{\gamma}_{1,-1}}), \\ \sqrt{N_3}\{\widehat{\boldsymbol{\gamma}}_{2,-1}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_{2,-1}\} &\to \mathit{Normal}(0, \Sigma_{\boldsymbol{\gamma}_{2,-1}}), \\ \sqrt{N_3}\{\widehat{\boldsymbol{\gamma}}_{3,-1}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_{3,-1}\} &\to \mathit{Normal}(0, \Sigma_{\boldsymbol{\gamma}_{3,-1}}),\end{aligned}$$

when $n \to \infty$. Here $\Sigma_{\boldsymbol{\gamma}_{1,-1}}$, $\Sigma_{\boldsymbol{\gamma}_{2,-1}}$ and $\Sigma_{\boldsymbol{\gamma}_{3,-1}}$ are positive definite matrices.

Proof. Here we only provide the proof of the root-N_3 consistency of $\widehat{\boldsymbol{\gamma}}_{1,-1}$ below. Similar derivations can be used to prove the results regarding $\boldsymbol{\gamma}_{2,-1}$ and $\boldsymbol{\gamma}_{3,-1}$. The estimator $\widehat{\boldsymbol{\gamma}}_{1,-1}$

solves

$$\begin{aligned}
\mathbf{0} &= N_3^{-1/2}\textstyle\sum_{d=0}^{1}\dfrac{\widehat{\pi}_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\displaystyle\sum_{j=1}^{N_3} I(D_j=d)\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}})-\widehat{E}_1^{\widehat{\pi}}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})\right\}\\
&\qquad\times\left\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}-\widehat{E}^{\widehat{\pi}}_{\rm true}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}\mid\mathbf{Z}^{\rm T}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\right\}\\
&= N_3^{-1/2}\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\displaystyle\sum_{j=1}^{N_3} I(D_j=d)\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}})-\widehat{E}_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})\right\}\\
&\qquad\times\left\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}-\widehat{E}_{\rm true}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}\mid\mathbf{Z}^{\rm T}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\right\}\\
&\quad+O_p\{(N_3/N_2)^{1/2}\},
\end{aligned}$$

where we used Lemma 6 and 7. Simple calculation shows that the above equation can be further expanded as

$$
\begin{aligned}
0 \;=\;& N_3^{-1/2}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{j=1}^{N_3} I(D_j = d)\Big\{\epsilon_j^2(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\alpha}}) - E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) \\
&\qquad + E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\}\Big\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} - E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) \\
&\qquad + E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\Big\} + o_p(1) \\
=\;& N_3^{-1/2}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{j=1}^{N_3} I(D_j = d) \\
&\qquad \times\Big\{\epsilon_j^2(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\alpha}}) - E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\} \\
&\qquad \times\Big\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} - E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\} \\
& + N_3^{-1/2}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{j=1}^{N_3} I(D_j = d) \\
&\qquad \times\Big\{\epsilon_j^2(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j, Y_j, \widetilde{\boldsymbol{\alpha}}) - E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\} \\
&\qquad \times\Big\{E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\Big\} \\
& + N_3^{-1/2}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{j=1}^{N_3} I(D_j = d) \qquad\qquad\qquad\qquad\qquad\qquad \text{(B.9)} \\
&\qquad \times\Big\{E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\}\Big\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} - E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\} \\
& + N_3^{-1/2}{\textstyle\sum_{d=0}^{1}}\frac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\sum_{j=1}^{N_3} I(D_j = d) \\
&\qquad \times\Big\{E_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_1(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}})\Big\} \qquad\qquad\qquad\qquad\qquad\qquad \text{(B.10)} \\
&\qquad \times\Big\{E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j, \boldsymbol{\gamma}_1, \widetilde{\boldsymbol{\theta}}) - \widehat{E}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j} \mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\Big\} \\
& + o_p(1).
\end{aligned}
$$

Using Lemmas 6 and 7 and the regularity condition C10, we have that the fourth term in

(B.9)

$$\begin{aligned} & \left\| N_3^{-1/2}\textstyle\sum_{d=0}^{1} \dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\displaystyle\sum_{j=1}^{N_3} I(D_j=d)\left\{E_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})-\widehat{E}_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})\right\}\right. \\ & \qquad \left. \times\left\{E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})-\widehat{E}_{\mathrm{true}}(\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}\mid \mathbf{Z}^{\mathrm{T}}_{\widetilde{\boldsymbol{\beta}},j}\boldsymbol{\gamma}_1)\right\}\right\| \\ = & \left| N_3^{1/2}\left\{O_p(h^r)+O_p\left(N_3^{-1/2}h^{-\ell/2}\right)\right\}^2\right| = o_p(1). \end{aligned}$$

By applying Lemma A1 in Ma and Zhu (2012), we obtain that the second and third terms in (B.9) are of order $O_p\{h^r + N_3^{1/2}h^{2r} + \log^2 N_3/\sqrt{N_3 h^{2\ell}}\} = o_p(1)$. Hence, the estimating equation can be written as

$$\begin{aligned} \mathbf{0} &= N_3^{-1/2}\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\widetilde{\boldsymbol{\alpha}})}{N_{3d}/N_3}\displaystyle\sum_{j=1}^{N_3} I(D_j=d) \\ &\qquad \times\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\beta}})\kappa(\mathbf{X}_j,Y_j,\widetilde{\boldsymbol{\alpha}})-E_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})\right\} \\ &\qquad \times\left\{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}},j}-E_{\mathbf{Z}^*_{\widetilde{\boldsymbol{\beta}}}}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\widetilde{\boldsymbol{\theta}})\right\}+o_p(1). \end{aligned} \qquad \text{(B.11)}$$

We now show that the influence function given in (B.11) has mean 0 at $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

$$\begin{aligned} & E\left[\textstyle\sum_{d=0}^{1}\dfrac{\pi_d(\boldsymbol{\alpha})}{N_{3d}/N_3} I(D_j=d)\right. \\ & \qquad \left.\times\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\boldsymbol{\beta})\kappa(\mathbf{X}_j,Y_j,\boldsymbol{\alpha})-E_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\left\{\mathbf{Z}^*_j-E_{\mathbf{Z}^*}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\right] \\ = & E_{\mathrm{true}}\left[\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\boldsymbol{\beta})\kappa(\mathbf{X}_j,Y_j,\boldsymbol{\alpha})-E_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\left\{\mathbf{Z}^*_j-E_{\mathbf{Z}^*}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\right] \\ = & E_{\mathrm{true}}\left(E_{\mathrm{true}}\left[\left\{\epsilon_j^2(\mathbf{X}_j,Y_j,\boldsymbol{\beta})\kappa(\mathbf{X}_j,Y_j,\boldsymbol{\alpha})-E_1(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\mid \mathbf{X}_j\right]\right. \\ & \qquad \left.\times\left\{\mathbf{Z}^*_j-E_{\mathbf{Z}^*}(\mathbf{X}_j,\boldsymbol{\gamma}_1,\boldsymbol{\theta})\right\}\right) \\ = & \mathbf{0}. \end{aligned}$$

The last equality is because of the single index model assumption (3.8). In practical operation, we will replace $\widetilde{\boldsymbol{\theta}}$ by $\widehat{\boldsymbol{\theta}}$, the solution of the estimating equation defined in (3.13). As long as $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in probability, the above expectation approaches $\mathbf{0}$.

Hence, we have that

$$\sqrt{N_3}\{\widehat{\boldsymbol{\gamma}}_{1,-1}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_{1,-1}\} \to \text{Normal}(0, \Sigma_{\boldsymbol{\gamma}_{1,-1}})$$

when $n \to \infty$, where $\Sigma_{\boldsymbol{\gamma}_{1,-1}}$ is a positive definite matrix. □

We now analyze $\widehat{E}_{\text{true}}\{\epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{X}\}$. We will show that it has bias order $O_p(h^r)$ and standard deviation $O_p\left(N_3^{-1/2}h^{-\ell/2}\right)$ as given in the following lemma.

Lemma 9. *Under the regularity conditions C1-C10,*

$$\begin{aligned}
&\widehat{E}_{\text{true}}\{\epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{X}\} \\
&\quad = \frac{E_{\text{true}}\left\{\frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})}\epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}}{E_{\text{true}}\left\{\frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\boldsymbol{\gamma}_1\right\}} + O_p(h^r) \\
&\qquad + O_p\left(N_3^{-1/2}h^{-l/2}\right) + O_p\left(N_3^{-1}h^{-\ell/2-1}\right).
\end{aligned}$$

Proof. Similar to the proof of Lemma 6, we have that

$$\begin{aligned}
&\widehat{E}_{\text{true}}\{\epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{X}\} \\
= &\frac{\sum_{d=0}^{1}\frac{\widehat{\pi}_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}}\sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\widehat{\kappa}(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}})K_h\{\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})\}}{\sum_{d=0}^{1}\frac{\widehat{\pi}_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}}\sum_{i=1}^{N_3} I(D_i = d)K_h\{\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})\}} \\
= &\frac{\sum_{d=0}^{1}\frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}}\sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}})K_h\{\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})\}}{\sum_{d=0}^{1}\frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}}\sum_{i=1}^{N_3} I(D_i = d)K_h\{\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})\}} \\
&+ O_p(N_2^{-1/2}).
\end{aligned}$$

We first inspect the numerator.

$$
\begin{aligned}
& N_3^{-1} {\textstyle\sum_{d=0}^{1}} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}} \sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}}) K_h \left\{ \mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) \right\} \\
& = N_3^{-1} h^{-(\ell+1)} {\textstyle\sum_{d=0}^{1}} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}}) \\
& \quad \times K' \left\{ \mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \boldsymbol{\gamma}_1^*/h - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1^*/h \right\} (\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i} - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}})^{\mathrm{T}} \{ \widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_1 \} \\
& \quad + N_3^{-1} {\textstyle\sum_{d=0}^{1}} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}}) K_h \left(\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \boldsymbol{\gamma}_1 - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right) \\
& = N_3^{-1} {\textstyle\sum_{d=0}^{1}} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i=1}^{N_3} I(D_i = d)\epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}}) K_h \left(\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \boldsymbol{\gamma}_1 - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right) \\
& \quad + O_p(N_3^{-1/2}) + O_p(N_3^{-1} h^{-\ell/2-1}) \\
& = E_{\mathrm{true}} \left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\} \xi_1^{\mathrm{true}}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1) + O_p(h^r) \\
& \quad + O_p\left(N_3^{-1/2} h^{-l/2} \right) + O_p(N_3^{-1} h^{-\ell/2-1}) \\
& = E_{\mathrm{true}} \left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\} \xi_1^{\mathrm{true}}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1) + O_p(h^r) \\
& \quad + O_p\left(N_3^{-1/2} h^{-l/2} \right) + O_p(N_3^{-1} h^{-\ell/2-1}).
\end{aligned}
$$

Here $\boldsymbol{\gamma}_1^*$ is on the interval connecting $\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})$ and $\boldsymbol{\gamma}_1$. In the second equality above, we used condition C10, the root-N_3 consistency of $\widehat{\boldsymbol{\gamma}}_1(\widehat{\boldsymbol{\theta}})$, the regularity conditions C5-C7 and the fact that

$$
\begin{aligned}
& N_3^{-1} h^{-(\ell+1)} {\textstyle\sum_{d=0}^{1}} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}/N_3} \sum_{i=1}^{N_3} I(D_i = d) \\
& \qquad \times \epsilon_i^2(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}_i, Y_i, \widehat{\boldsymbol{\alpha}}) K'\{ h^{-1}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \boldsymbol{\gamma}_1 - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1)\} (\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i} - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}) \\
& \quad = -\frac{\partial}{\partial \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1} \left[E_{\mathrm{true}} \left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mathbf{Z}_{\widehat{\boldsymbol{\beta}}} \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\} \xi_1^{\mathrm{true}}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1) \right] \\
& \qquad + \frac{\partial}{\partial \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1} \left[E_{\mathrm{true}} \left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}})\kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\} \xi_1^{\mathrm{true}}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1) \right] \mathbf{Z}_{\widehat{\boldsymbol{\beta}}} \\
& \qquad + O_p(h^2) + O_p\{(N_3 h^{\ell+2})^{-1/2}\}.
\end{aligned}
$$

Similarly, for the denominator, we have that

$$\begin{aligned}
&\textstyle\sum_{d=0}^{1} \frac{\pi_d(\widehat{\boldsymbol{\alpha}})}{N_{3d}} \sum_{i=1}^{N_3} I(D_i = d) K_h\{\mathbf{Z}_{\widehat{\boldsymbol{\beta}},i}^{\mathrm{T}} \widehat{\gamma}_1(\widehat{\boldsymbol{\theta}}) - \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \widehat{\gamma}_1(\widehat{\boldsymbol{\theta}})\} \\
&\qquad = E_{\mathrm{true}}\left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\} \xi_1^{\mathrm{true}}(\mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1) + O_p(h^r) + O_p\left(N_3^{-1/2} h^{-l/2}\right) \\
&\qquad\qquad + O_p\left(N_3^{-1} h^{-\ell/2-1}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\widehat{E}_{\mathrm{true}}(\epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}}) \widehat{\kappa}(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{X}) \\
&\quad = \frac{E_{\mathrm{true}}\left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \epsilon^2(\mathbf{X}, Y, \widehat{\boldsymbol{\beta}}) \kappa(\mathbf{X}, Y, \widehat{\boldsymbol{\alpha}}) \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\}}{E_{\mathrm{true}}\left\{ \frac{\pi_D(\widehat{\boldsymbol{\alpha}})}{\pi_D(\boldsymbol{\alpha})} \mid \mathbf{Z}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right\}} + O_p(h^r) + O_p\left(N_3^{-1/2} h^{-l/2}\right) \\
&\qquad + O_p\left(N_3^{-1} h^{-\ell/2-1}\right).
\end{aligned}$$

□

B.6 Proof of Theorem 3

Through the analyses in the lemmas, we proved that all the estimated quantities defined in Appendix B.2.1 have desired bias order and standard deviation orders. Specifically, the difference between the quantities with hat and without hat either have mean zero, standard deviation $O_p(n_2^{-1/2}) = O_p(N_3^{-1/2})$ or have bias $O_p(h^r)$ and standard deviation $O_p\left(N_3^{-1/2} h^{-\ell/2}\right)$ or $O_p\left(N_3^{-1/2} h^{-\ell/2}\right) + O_p\left(N_3^{-1} h^{-(\ell+2)/2}\right)$. Now we are ready to prove our main theorem.

$$
\begin{aligned}
\mathbf{0} &= N_1^{-1/2}\sum_{i=1}^{N_1}\widehat{\mathbf{S}}^*_{\text{eff}}\left(D_i,\mathbf{X}_i,Y_i,\widehat{\boldsymbol{\theta}}\right)\\
&= N_1^{-1/2}\sum_{i=1}^{N_1}\mathbf{S}^*_{\text{eff}}\left[D_i,\mathbf{X}_i,Y_i,\widehat{\boldsymbol{\theta}},\widehat{\pi}_d(\widehat{\boldsymbol{\alpha}}),\widehat{E}\{\widehat{\pi}_d(\widehat{\boldsymbol{\alpha}}),\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})\}\right]\\
&= N_1^{-1/2}\sum_{i=1}^{N_1}\mathbf{S}^*_{\text{eff}}\left[D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta},\widehat{\pi}_d(\boldsymbol{\alpha}),\widehat{E}\{\widehat{\pi}_d(\widehat{\boldsymbol{\alpha}}),\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\}\right]\\
&\quad +N_1^{-1}\sum_{i=1}^{N_1}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{S}^*_{\text{eff}}\left[D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}^*,\widehat{\pi}_d(\boldsymbol{\alpha}^*),\widehat{E}\{\widehat{\pi}_d(\boldsymbol{\alpha}^*),\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}^*)\}\right]\sqrt{N_1}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\
&= N_1^{-1/2}\sum_{i=1}^{N_1}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})\\
&\quad +N_1^{-1/2}\sum_{i=1}^{N_1}\left\{\widehat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})\right\}\\
&\quad +E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})+o_p(1)\right\}\sqrt{N_1}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\
&= N_1^{-1/2}\sum_{i=1}^{N_1}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})+E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta})+o_p(1)\right\}\sqrt{N_1}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\\
&\quad +o_p(1),
\end{aligned}
$$

where $\boldsymbol{\alpha}^*$ is a point on the line connecting $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}$. Simple calculation lead to the proof of Theorem 3. □

APPENDIX C

SKETCH OF TECHNICAL ARGUMENTS FOR SECTION 4

C.1 Identifiability

We prove Proposition 2 through contradition. Assume the problem is not identifiable. Then there must exist $\alpha_c, \boldsymbol{\alpha}_1, \alpha_2, \boldsymbol{\beta}_\tau, \eta_1, \eta_2$ and $\widetilde{\alpha}_c, \widetilde{\boldsymbol{\alpha}}_1, \widetilde{\alpha}_2, \widetilde{\boldsymbol{\beta}}_\tau, \widetilde{\eta}_1, \widetilde{\eta}_2$ such that

$$\begin{aligned} &\pi_d^{-1}\eta_1(\mathbf{x})\eta_2(y-\beta_{\tau,c}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{x})\frac{\exp\{d(\alpha_c+\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}+\alpha_2 y)\}}{1+\exp(\alpha_c+\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}+\alpha_2 y)} \\ &\quad = \widetilde{\pi}_d^{-1}\widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(y-\widetilde{\beta}_{\tau,c}-\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau,\mathbf{x})\frac{\exp\{d(\widetilde{\alpha}_c+\widetilde{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{x}+\widetilde{\alpha}_2 y)\}}{1+\exp(\widetilde{\alpha}_c+\widetilde{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{x}+\widetilde{\alpha}_2 y)}, \end{aligned} \tag{C.1}$$

for all $(\mathbf{x}, y, d)$, where

$$\begin{aligned} \pi_d &= \int \eta_1(\mathbf{x})\eta_2(y-\beta_{\tau,c}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{x})\frac{\exp\{d(\alpha_c+\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}+\alpha_2 y)\}}{1+\exp(\alpha_c+\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}+\alpha_2 y)}d\mu(\mathbf{x})d\mu(y), \\ \widetilde{\pi}_d &= \int \widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(y-\widetilde{\beta}_{\tau,c}-\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau,\mathbf{x})\frac{\exp\{d(\widetilde{\alpha}_c+\widetilde{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{x}+\widetilde{\alpha}_2 y)\}}{1+\exp(\widetilde{\alpha}_c+\widetilde{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{x}+\widetilde{\alpha}_2 y)}d\mu(\mathbf{x})d\mu(y). \end{aligned}$$

Take the ratio of the expression (C.1) at $d = 1$ and $d = 0$ respectively, we obtain that for all $(\mathbf{x}, y)$,

$$\frac{\pi_0}{\pi_1}\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2) = \frac{\widetilde{\pi}_0}{\widetilde{\pi}_1}\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}}_1+y\widetilde{\alpha}_2).$$

This yields that $\alpha_c - \widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1 - \widetilde{\boldsymbol{\alpha}}_1) + y(\alpha_2 - \widetilde{\alpha}_2)$ is a constant. Hence, we have $\boldsymbol{\alpha}_1 = \widetilde{\boldsymbol{\alpha}}_1$ and $\alpha_2 = \widetilde{\alpha}_2$. Furthermore, $\exp(\alpha_c)\pi_0/\pi_1 = \exp(\widetilde{\alpha}_c)\widetilde{\pi}_0/\widetilde{\pi}_1$ and

$$\frac{1}{\pi_0}\frac{\eta_1(\mathbf{x})\eta_2(y-\beta_{\tau,c}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{x})}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)} = \frac{1}{\widetilde{\pi}_0}\frac{\widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(y-\widetilde{\beta}_{\tau,c}-\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau,\mathbf{x})}{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}$$

for all $(\mathbf{x}, y)$. This gives

$$\begin{aligned}
&\widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(y - \widetilde{\beta}_{\tau,c} - \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau, \mathbf{x}) \\
&\quad = \frac{\widetilde{\pi}_0}{\pi_0}\frac{1 + \exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}{1 + \exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}\eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x})
\end{aligned} \tag{C.2}$$

Now consider the case that $\alpha_c = \widetilde{\alpha}_c$, then $\pi_d = \widetilde{\pi}_d$ for $d = 0, 1$ and (C.2) leads to

$$\widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(y - \widetilde{\beta}_{\tau,c} - \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau, \mathbf{x}) = \eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x}).$$

Integrating the above equation with respect to y, we obtain $\eta_1(\mathbf{x}) = \widetilde{\eta}_1(\mathbf{x})$. This subsequently yields $\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x}) = \widetilde{\eta}_2(y - \widetilde{\beta}_{\tau,c} - \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau, \mathbf{x})$. Hence,

$$\tau = \int_{-\infty}^{\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau} \eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x})d\mu(y) = \int_{-\infty}^{\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau} \widetilde{\eta}_2(y - \widetilde{\beta}_{\tau,c} - \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau, \mathbf{x})d\mu(y).$$

Therefore, $\beta_{\tau,c} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau = \widetilde{\beta}_{\tau,c} + \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau$ for all $\mathbf{x}$. As a result, $\beta_{\tau,c} = \widetilde{\beta}_{\tau,c}$, $\boldsymbol{\beta}_\tau = \widetilde{\boldsymbol{\beta}}_\tau$, and $\eta_2 = \widetilde{\eta}_2$, which contradicts our assumptions.

Thus, we must have $\alpha_c \neq \widetilde{\alpha}_c$. Without loss of generity, in the following, we assume $\widetilde{\alpha}_c > \alpha_c$.

Integrating (C.2) with respect to y on $\mathcal{R}$ and $(-\infty, \widetilde{\beta}_{\tau,c}+\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau)$ respectively, we obtain

$$\widetilde{\eta}_1(\mathbf{x}) = \int \frac{\widetilde{\pi}_0}{\pi_0}\frac{1 + \exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}{1 + \exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)} \tag{C.3}$$

$$\times\eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x})d\mu(y); \tag{C.4}$$

$$\tau\widetilde{\eta}_1(\mathbf{x}) = \int_{-\infty}^{\widetilde{\beta}_{\tau,c}+\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau} \frac{\widetilde{\pi}_0}{\pi_0}\frac{1 + \exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}{1 + \exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}$$

$$\times\eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x})d\mu(y). \tag{C.5}$$

Taking the ratio of (C.4) and (C.5), using change of variable $y = \epsilon_\tau + \beta_{\tau,c} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_\tau$, we get

$$\begin{aligned}
\tau &= \frac{\int_{-\infty}^{\widetilde{\beta}_{\tau,c}+\mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_\tau} \frac{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(y-\beta_{\tau,c}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{x})d\mu(y)}{\int \frac{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(y-\beta_{\tau,c}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{x})d\mu(y)} \\
&= \frac{\int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \frac{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}{\int \frac{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}. \qquad \text{(C.6)}
\end{aligned}$$

Denote the numerator $h_1(\mathbf{x})$ and the denominator $h_2(\mathbf{x})$. Then for all $\mathbf{x}$,

$$\begin{aligned}
h_1(\mathbf{x}) &= \int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \frac{1+\exp(\widetilde{\alpha}_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) \\
&= \int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \frac{1-\exp(\widetilde{\alpha}_c-\alpha_c)}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) \\
&\quad +\exp(\widetilde{\alpha}_c-\alpha_c)\int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) \\
&\leq |1-\exp(\widetilde{\alpha}_c-\alpha_c)|\int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) \\
&\quad +\exp(\widetilde{\alpha}_c-\alpha_c)\int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}+\mathbf{x}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)
\end{aligned}$$

Let β_τ^ℓ, $\widetilde{\beta}_\tau^\ell$, and x^ℓ denote the ℓ^{th} element of $\boldsymbol{\beta}_\tau$, $\widetilde{\boldsymbol{\beta}}_\tau$, and $\mathbf{x}$, respectively, for $\ell = 1, \cdots, \dim(\mathbf{x})$. If there is ℓ such that $\widetilde{\beta}_\tau^\ell - \beta_\tau^\ell > 0$, then $\lim_{x^\ell \to -\infty} h_1(\mathbf{x}) = 0$. Meanwhile, as $\widetilde{\alpha}_c > \alpha_c$, we have

$$h_2(\mathbf{x}) \geq \int \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) = 1.$$

Consequently, $\tau = 0$, which contradicts the basic assumption on τ. Similary, if there is ℓ such that $\widetilde{\beta}_\tau^\ell - \beta_\tau^\ell < 0$, letting $x^\ell \to \infty$ will lead to the same conclusion, i.e., $\tau = 0$.

Hence, we obtain $\boldsymbol{\beta}_\tau = \widetilde{\boldsymbol{\beta}}_\tau$ and

$$\tau = \frac{\int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}} \frac{1+\exp\{\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}{1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}{\int \frac{1+\exp\{\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}{1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}. \tag{C.7}$$

Let α_1^ℓ denote the ℓ^{th} element of $\boldsymbol{\alpha}_1$. If there is ℓ such that $\alpha_1^\ell + \beta_\tau^\ell\alpha_2 > 0$, then the numerator $h_1(\mathbf{x}) \to \int_{-\infty}^{\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)$ and the denominator $h_2(\mathbf{x}) \to 1$ as $x^\ell \to \infty$. Consequently, $\widetilde{\beta}_{\tau,c} - \beta_{\tau,c} = 0$ according to the definition of the quantile regression model. If there is ℓ such that $\alpha_1^\ell + \beta_\tau^\ell\alpha_2 < 0$, by letting $x^\ell \to -\infty$, we obtain $\widetilde{\beta}_{\tau,c} - \beta_{\tau,c} = 0$ as well.

Thus, $\widetilde{\beta}_{\tau,c} = \beta_{\tau,c}$ and

$$\tau = \frac{\int_{-\infty}^{0} \frac{1+\exp\{\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}{1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}{\int \frac{1+\exp\{\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}{1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}\eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)} \tag{C.8}$$

for all $\mathbf{x}$. On the other hand, we have

$$\tau = \int_{-\infty}^{0} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau) = \frac{\int_{-\infty}^{0} \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)}{\int \eta_2(\epsilon_\tau,\mathbf{x})d\mu(\epsilon_\tau)},$$

for all $\mathbf{x}$. Let

$$c(\epsilon_\tau,\mathbf{x}) \quad = \quad \frac{1+\exp\{\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}{1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)+\epsilon_\tau\alpha_2\}}.$$

If $\alpha_2 > 0$, then $c(\epsilon_\tau,\mathbf{x})$ is a strictly increasing function of ϵ_τ for all $\mathbf{x}$. For all $\epsilon_\tau < 0$, $c(\epsilon_\tau,\mathbf{x}) < c(0,\mathbf{x})$, and for all $\epsilon_\tau > 0$, $c(\epsilon_\tau,\mathbf{x}) > c(0,\mathbf{x})$, where $c(0,\mathbf{x}) = [1+\exp\{\widetilde{\alpha}_c + \beta_{\tau,c}\alpha_2 + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)\}]/[1+\exp\{\alpha_c+\beta_{\tau,c}\alpha_2+\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_1+\boldsymbol{\beta}_\tau\alpha_2)\}]$ is positive and finite.

As a result,

$$\begin{aligned}
\int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) < c(0, \mathbf{x}) \int_{-\infty}^{0} \eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) = \tau c(0, \mathbf{x}),\\
\int_{0}^{\infty} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) > c(0, \mathbf{x}) \int_{0}^{\infty} \eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) = (1-\tau) c(0, \mathbf{x}).
\end{aligned}$$

Hence, (C.8) leads to

$$\begin{aligned}
\tau &= \frac{\int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}{\int c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}\\
&= \frac{\int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}{\int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) + \int_{0}^{\infty} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}\\
&= \frac{1}{1 + \int_{0}^{\infty} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) / \int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}\\
&< \frac{1}{1 + (1-\tau)/\tau} = \tau,
\end{aligned}$$

which is impossible.

Similary, if $\alpha_2 < 0$, then $c(\epsilon_\tau, \mathbf{x})$ is a strictly decreasing function of ϵ_τ and similar derivation regarding (C.8) then leads to

$$\begin{aligned}
\tau &= \frac{1}{1 + \int_{0}^{\infty} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) / \int_{-\infty}^{0} c(\epsilon_\tau, \mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau)}\\
&> \frac{1}{1 + (1-\tau)/\tau} = \tau,
\end{aligned}$$

which is also impossible.

Hence, the problem is identifiable. □

C.2 Nonidentifiability when $\alpha_2 = 0$

Assume we have a parametric set $\{\alpha_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta_c, \boldsymbol{\beta}_\tau, \eta_1, \eta_2\}$. Choose any $\widetilde{\alpha}_c$ with $\alpha_c \neq \widetilde{\alpha}_c$, let

$$\begin{aligned}
\widetilde{\pi}_0 &= \frac{\exp(\alpha_c)\pi_0}{\exp(\alpha_c)\pi_0 + \exp(\widetilde{\alpha}_c)\pi_1}, \\
\widetilde{\pi}_1 &= 1 - \widetilde{\pi}_0.
\end{aligned}$$

Note that $\widetilde{\pi}_0$ is the solution of $\exp(\alpha_c)\pi_0/\pi_1 = \exp(\widetilde{\alpha}_c)\widetilde{\pi}_0/(1-\widetilde{\pi}_0)$, Obviously, $0 < \widetilde{\pi}_0 < 1$. Let

$$\begin{aligned}
\widetilde{\beta}_{\tau,c} &= \beta_{\tau,c}, \\
\widetilde{\boldsymbol{\beta}}_\tau &= \boldsymbol{\beta}_\tau, \\
\widetilde{\boldsymbol{\alpha}}_1 &= \boldsymbol{\alpha}_1, \\
\widetilde{\eta}_2(\epsilon_\tau, \mathbf{x}) &= \eta_2(\epsilon_\tau, \mathbf{x}), \\
\widetilde{\eta}_1(\mathbf{x}) &= \frac{\widetilde{\pi}_0}{\pi_0}\frac{1+\exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1)}{1+\exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1)}\eta_1(\mathbf{x}).
\end{aligned}$$

We can easily verify that $\widetilde{\eta}_1(\mathbf{x})$ is a valid density function through

$$\begin{aligned}
\int \widetilde{\eta}_1(\mathbf{x})d\mu(\mathbf{x}) &= \int \frac{\widetilde{\pi}_0}{\pi_0}\frac{1+\exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1)}{1+\exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1)}\eta_1(\mathbf{x})d\mu(\mathbf{x}) \\
&= \frac{\widetilde{\pi}_0}{\pi_0}\{\pi_0 + \exp(\widetilde{\alpha}_c - \alpha_c)\pi_1\} \\
&= \widetilde{\pi}_0 + \widetilde{\pi}_1 = 1.
\end{aligned}$$

Besides,

$$\begin{aligned}
&\int \widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(\epsilon_\tau, \mathbf{x})\frac{1}{1+\exp(\widetilde{\alpha}_c + \mathbf{x}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}}_1)} d\mu(\mathbf{x})d\mu(\epsilon_\tau) \\
&= \int \frac{\widetilde{\pi}_0}{\pi_0}\eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x})\frac{1}{1+\exp(\alpha_c + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1)} d\mu(\mathbf{x})d\mu(\epsilon_\tau) \\
&= \frac{\widetilde{\pi}_0}{\pi_0}\pi_0 = \widetilde{\pi}_0.
\end{aligned}$$

It is clear that the parameter set $\{\alpha_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta_c, \boldsymbol{\beta}_\tau, \eta_1, \eta_2\}$ is different from the parameter set $\{\widetilde{\alpha}_c, \widetilde{\boldsymbol{\alpha}}_1, \widetilde{\boldsymbol{\alpha}}_2, \widetilde{\beta}_c, \widetilde{\boldsymbol{\beta}}_\tau, \widetilde{\eta}_1, \widetilde{\eta}_2\}$, while they both satisfy (C.1), hence the problem is not identifiable.

C.3 Nonidentifiability of the Case $\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_\tau \alpha_2 = 0$

We first exclude the special case $\alpha_2 = 0$ for two reasons: (a) the nonidentifiability of the case α_2 has been proved in Section C.2, and (b) α_2 here further implies $\boldsymbol{\alpha}_1 = 0$, and hence the case-control sampling is simply random sampling.

The nonidentifiability proof of the case $\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_\tau \alpha_2 = 0$ is similar to that of the case $\alpha_2 = 0$. Assume we have a parameter set $\{\alpha_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta_c, \boldsymbol{\beta}_\tau, \eta_1, \eta_2\}$. Choose any $\widetilde{\alpha}_c$ with $\alpha_c \neq \widetilde{\alpha}_c$ and set $\widetilde{\pi}_0, \widetilde{\pi}_1$ identically as in the $\alpha_2 = 0$ case. Note that $\exp(\alpha_c)\pi_0/\pi_1 = \exp(\widetilde{\alpha}_c)\widetilde{\pi}_0/\widetilde{\pi}_1$.

Consider the special case when $\eta_2(\epsilon_\tau, \mathbf{x}) = \eta_2(\epsilon_\tau)$, i.e., the distribution of the quantile regression error does not depend on $\mathbf{x}$, and define

$$\begin{aligned}
h_3(\widetilde{\beta}_{\tau,c}) &= \int_{-\infty}^{0} \frac{\widetilde{\pi}_0}{\pi_0}\frac{1+\exp(\widetilde{\alpha}_c + \widetilde{\beta}_{\tau,c}\alpha_2 + \alpha_2\epsilon_\tau)}{1+\exp(\alpha_c + \widetilde{\beta}_{\tau,c}\alpha_2 + \alpha_2\epsilon_\tau)}\eta_2(\epsilon_\tau + \widetilde{\beta}_{\tau,c} - \beta_{\tau,c})d\mu(\epsilon_\tau) \\
&= \int_{-\infty}^{\widetilde{\beta}_{\tau,c} - \beta_{\tau,c}} \frac{\widetilde{\pi}_0}{\pi_0}\frac{1+\exp(\widetilde{\alpha}_c + \beta_{\tau,c}\alpha_2 + \alpha_2\epsilon_\tau)}{1+\exp(\alpha_c + \beta_{\tau,c}\alpha_2 + \alpha_2\epsilon_\tau)}\eta_2(\epsilon_\tau)d\mu(\epsilon_\tau).
\end{aligned}$$

We claim $h_3(\widetilde{\beta}_{\tau,c}) = \tau$ has at least one solution. This is because $h_3(\widetilde{\beta}_{\tau,c}) \to 0$ as $\widetilde{\beta}_{\tau,c} \to$

$-\infty$, and

$$\begin{aligned}
h_3(\widetilde{\beta}_{\tau,c}) &\to \int_{-\infty}^{\infty} \frac{\widetilde{\pi}_0}{\pi_0} \frac{1+\exp(\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)}{1+\exp(\alpha_c+\beta_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)} \eta_2(\epsilon_\tau) d\mu(\epsilon_\tau) \\
&= \int \frac{\widetilde{\pi}_0}{\pi_0} \frac{1+\exp(\widetilde{\alpha}_c+\beta_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)}{1+\exp(\alpha_c+\beta_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)} \eta_2(\epsilon_\tau) d\mu(\epsilon_\tau) \\
&= \frac{\widetilde{\pi}_0}{\pi_0}\{\pi_0+\exp(\widetilde{\alpha}_c-\alpha_c)\pi_1\} \\
&= \widetilde{\pi}_0+\widetilde{\pi}_1 = 1,
\end{aligned}$$

as $\widetilde{\beta}_{\tau,c} \to \infty$.

Set

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_\tau &= \boldsymbol{\beta}_\tau, \\
\widetilde{\boldsymbol{\alpha}}_1 &= \boldsymbol{\alpha}_1, \\
\widetilde{\alpha}_2 &= \alpha_2, \\
\widetilde{\eta}_1(\mathbf{x}) &= \eta_1(\mathbf{x}), \\
\widetilde{\eta}_2(\epsilon_\tau) &= \frac{\widetilde{\pi}_0}{\pi_0} \frac{1+\exp(\widetilde{\alpha}_c+\widetilde{\beta}_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)}{1+\exp(\alpha_c+\widetilde{\beta}_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)} \eta_2(\epsilon_\tau+\widetilde{\beta}_{\tau,c}-\beta_{\tau,c}),
\end{aligned}$$

where $\widetilde{\alpha}_c \neq \alpha_c$ and $\widetilde{\beta}_{\tau,c}$ is the solution of $h_3(\widetilde{\beta}_{\tau,c}) = \tau$. It is easy to verify that $\widetilde{\eta}_2(\epsilon_\tau)$ is a valid density function, $\widetilde{\pi}_0 = \int \widetilde{\eta}_1(\mathbf{x})\widetilde{\eta}_2(\epsilon_\tau)/\{1+\exp(\widetilde{\alpha}_c+\widetilde{\beta}_{\tau,c}\alpha_2+\alpha_2\epsilon_\tau)\}d\mu(\mathbf{x})d\mu(\epsilon_\tau)$, and $\widetilde{\eta}_2 \neq \eta_2$. Furthermore, the two parameter sets $\{\alpha_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \beta_c, \boldsymbol{\beta}_\tau, \eta_1, \eta_2\}$ and $\{\widetilde{\alpha}_c, \widetilde{\boldsymbol{\alpha}}_1, \widetilde{\boldsymbol{\alpha}}_2, \widetilde{\beta}_c, \widetilde{\boldsymbol{\beta}}_\tau, \widetilde{\eta}_1, \widetilde{\eta}_2\}$ satisfy (C.1), hence the problem is not identifiable.

C.4 Nuisance Tangent Space Λ and its Orthogonal Complement $\Lambda^\perp$

The nuisance tangent space Λ with nuisance parameter $\eta = (\eta_1, \eta_2)$ is a subspace of the Hilbert space $\mathcal{H}$. It is defined as the mean squared closure of parametric submodel nuisance tangent spaces, where a parametric submodel nuisance tangent space with a finite-dimensional parameter $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$ is the set of all elements of the form $\mathbf{W}\mathbf{S}_\gamma$. Here $\mathbf{W}$ is an arbitrary $p \times q$ matrix with $p = \dim(\boldsymbol{\theta})$ and $q = \dim(\boldsymbol{\gamma})$, and $\mathbf{S}_\gamma$ is the nuisance score function, i.e., $\mathbf{S}_\gamma = (\mathbf{S}_{\gamma_1}^{\mathrm{T}}, \mathbf{S}_{\gamma_2}^{\mathrm{T}})^{\mathrm{T}}$, with

$$\begin{aligned}
\mathbf{S}_{\gamma_1} &= \eta_1(\mathbf{x};\boldsymbol{\gamma}_1)^{-1}\partial\eta_1(\mathbf{x};\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1 - \pi_d^{-1}E\{\eta_1(\mathbf{x};\boldsymbol{\gamma}_1)^{-1}\partial\eta_1(\mathbf{x};\boldsymbol{\gamma}_1)/\partial\boldsymbol{\gamma}_1 \mid d\}, \\
\mathbf{S}_{\gamma_2} &= \eta_2(\epsilon_\tau, \mathbf{x};\boldsymbol{\gamma}_1)^{-1}\partial\eta_2(\epsilon_\tau, \mathbf{x};\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 \\
&\quad -\pi_d^{-1}E\{\eta_2(\epsilon_\tau, \mathbf{x};\boldsymbol{\gamma}_1)^{-1}\partial\eta_2(\epsilon_\tau, \mathbf{x};\boldsymbol{\gamma}_2)/\partial\boldsymbol{\gamma}_2 \mid d\}.
\end{aligned}$$

$\eta_1(\cdot)$ is an arbitrary density function, i.e., $\eta_1 \geq 0$ and $\int \eta_1(\mathbf{x})d\mu(\mathbf{x}) = 1$, while $\eta_2(\cdot, \mathbf{x})$ is an arbitrary density function with τ^{th} quantile zero, i.e., $\eta_2 \geq 0, \int \eta_2(\epsilon_\tau, \mathbf{x})d\mu(\epsilon_\tau) = 1$, and $\int u_\tau \eta_2(\epsilon_\tau, \mathbf{x})d\epsilon_\tau = 0$. It is easy to show that show the nuisance tangent space Λ with nuisance parameter $\eta = (\eta_1, \eta_2)$ can be written as $\Lambda = \Lambda_1 \oplus \Lambda_2$, where

$$\begin{aligned}
\Lambda_1 &= [\mathbf{g}(\mathbf{x}) - E\{\mathbf{g}(\mathbf{X}) \mid d\} : \mathbf{g}(\mathbf{x}) \in \mathcal{R}^p, E_{\text{true}}\{\mathbf{g}(\mathbf{X})\} = \mathbf{0}], \\
\Lambda_2 &= [\mathbf{g}(\epsilon_\tau, \mathbf{x}) - E\{\mathbf{g}(\epsilon_\tau, \mathbf{X}) \mid d\} : \mathbf{g}(t, \mathbf{x}) \in \mathcal{R}^p, E_{\text{true}}\{\mathbf{g}(\epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\} = \mathbf{0}, \\
&\qquad E_{\text{true}}\{u_\tau \mathbf{g}(\epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\} = \mathbf{0}, a.s.].
\end{aligned}$$

We can further write

$$\begin{aligned}
\Lambda &= [\mathbf{g}(\epsilon_\tau, \mathbf{x}) - E\{\mathbf{g}(\epsilon_\tau, \mathbf{X}) \mid d\} : \mathbf{g} \in \mathcal{R}^p, E_{\text{true}}(\mathbf{g}) = \mathbf{0}, \\
&\qquad E_{\text{true}}\{u_\tau \mathbf{g}(\epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\} = \mathbf{0}, a.s.]\,.
\end{aligned}$$

For any $\mathbf{h} \in \mathcal{R}^p$, if $\mathbf{h} \in \Lambda_1^{\perp}$, then

$$\begin{aligned} 0 &= E(\mathbf{h}^{\mathrm{T}}[\mathbf{g}(\mathbf{X}) - E\{\mathbf{g}(\mathbf{X}) \mid D\}]) \\ &= E(\{\mathbf{h} - E(\mathbf{h} \mid D)\}^{\mathrm{T}}[\mathbf{g}(\mathbf{X}) - E\{\mathbf{g}(\mathbf{X}) \mid D\}]) \\ &= E[\{\mathbf{h} - E(\mathbf{h} \mid D)\}^{\mathrm{T}}\mathbf{g}(\mathbf{X})] \\ &= E[E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{X}\}^{\mathrm{T}}\mathbf{g}(\mathbf{X})]. \end{aligned}$$

Therefore, $E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{X}\}\sum_d \int f_{X,Y,D}(\mathbf{x}, y, d) d\mu(y)/\eta_1(\mathbf{x}) = \mathbf{c}$ a.s., for some constant $\mathbf{c}$. Notice that $E[E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{X}\}] = \mathbf{0}$, we further obtain

$$\mathbf{0} = \int E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{x}\}\sum_d \int f_{X,Y,D}(\mathbf{x}, y, d) d\mu(y) d\mu(\mathbf{x}) = \int \mathbf{c}\eta_1(\mathbf{x}) d\mu(\mathbf{x}) = \mathbf{c}.$$

Hence, $\mathbf{c} = \mathbf{0}$ and $E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{x}\} = \mathbf{0}$ a.s..

For any $\mathbf{h} \in \Lambda_1^{\perp}$, if $\mathbf{h} \in \Lambda_2^{\perp}$, then

$$\begin{aligned} 0 &= E(\mathbf{h}^{\mathrm{T}}[\mathbf{g}(\epsilon_\tau, \mathbf{X}) - E\{\mathbf{g}(\epsilon_\tau, \mathbf{X}) \mid D\}]) \\ &= E(\{\mathbf{h} - E(\mathbf{h} \mid D)\}^{\mathrm{T}}[\mathbf{g}(\epsilon_\tau, \mathbf{X}) - E\{\mathbf{g}(\epsilon_\tau, \mathbf{X}) \mid D\}]) \\ &= E[\{\mathbf{h} - E(\mathbf{h} \mid D)\}^{\mathrm{T}}\mathbf{g}(\epsilon_\tau, \mathbf{X})] \\ &= E[E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{X}\}^{\mathrm{T}}\mathbf{g}(\epsilon_\tau, \mathbf{X})]. \end{aligned}$$

Consequently, $E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{X}\}\sum_d f_{X,Y,D}(\mathbf{X}, Y, d)/\{\eta_1(\mathbf{X})\eta_2(\epsilon, \mathbf{X})\} = u_\tau \mathbf{a}(\mathbf{X}) + \mathbf{c}(\mathbf{X})$ a.s.. Since $\mathbf{h} \in \Lambda_1^{\perp}$, we have $\mathbf{0} = E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \mathbf{X}\} = E[E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid$

$\epsilon_\tau, \mathbf{X}\} \mid \mathbf{X}],$ a.s.. Thus,

$$\begin{aligned}\mathbf{0} &= \int E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{X}\} \frac{\sum_d f_{X,Y,D}(\mathbf{X}, y, d)}{\int \sum_d f_{X,Y,D}(\mathbf{X}, y, d) d\mu(y)} d\mu(y) \\ &= \frac{\int \{u_\tau \mathbf{a}(\mathbf{X}) + \mathbf{c}(\mathbf{X})\} \eta_1(\mathbf{x}) \eta_2(\epsilon_\tau, \mathbf{X}) d\mu(y)}{\int \sum_d f_{X,Y,D}(\mathbf{X}, y, d) d\mu(y)} \\ &= \frac{\mathbf{c}(\mathbf{X}) \eta_1(\mathbf{X})}{\int \sum_d f_{X,Y,D}(\mathbf{X}, y, d) d\mu(y)} \text{ a.s.,}\end{aligned}$$

which implies $\mathbf{c}(\mathbf{X}) = \mathbf{0}$ a.s..

Furthermore, $E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{X}\} \sum_d f_{X,Y,D}(\mathbf{X}, Y, d) / \{\eta_1(\mathbf{X}) \eta_2(\epsilon_\tau, \mathbf{X})\} = u_\tau \mathbf{a}(\mathbf{X})$ a.s., or equivalently, $E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon, \mathbf{X}\} \sum_d (n_d/n) H(d, \mathbf{X}, Y)/\pi_d = u_\tau \mathbf{a}(\mathbf{X})$ a.s..

Hence,

$$\begin{aligned}\Lambda^\perp &= \Lambda_1^\perp \cap \Lambda_2^\perp = \Big[\mathbf{h}(d, \epsilon_\tau, \mathbf{x}) : E(\mathbf{h}) = \mathbf{0}, E\{\mathbf{h} - E(\mathbf{h} \mid D) \mid \epsilon_\tau, \mathbf{x}\} \\ &\quad \times \sum_{d=0}^{1} \frac{n_d H(d, \mathbf{x}, y, \boldsymbol{\alpha})}{n \pi_d} = \mathbf{a}(\mathbf{x}) u_\tau, a.s.\ \forall \mathbf{a}\Big].\end{aligned}$$

C.5 Efficient Score Function $\mathbf{S}_{\text{eff}}$

We now derive the efficient score $\mathbf{S}_{\text{eff}}$ through orthogonally decomposing $\mathbf{S}_{\boldsymbol{\theta}}$ into a function in Λ and a function in $\Lambda^\perp$.

We write $\mathbf{S}_\theta = \mathbf{S} - E(\mathbf{S} \mid D) = \mathbf{g}(\epsilon_\tau, \mathbf{x}) - E(\mathbf{g} \mid D) + \mathbf{S}_{\text{eff}}$, where $E_{\text{true}}(u_\tau \mathbf{g} \mid \mathbf{x}) = \mathbf{0}$. We alternatively write $\mathbf{S}_{\text{eff}} = \mathbf{S} - \mathbf{g}(\epsilon_\tau, \mathbf{x}) - E(\mathbf{S} - \mathbf{g} \mid D)$ and $\mathbf{S}_{\text{eff}}$ satisfies

$$E\{\mathbf{S}_{\text{eff}} - E(\mathbf{S}_{\text{eff}} \mid D) \mid \epsilon_\tau, \mathbf{x}\} \sum_{d=1}^{1} \frac{n_d H(d, \mathbf{x}, y)}{n \pi_d} = \mathbf{a}(\mathbf{x}) u_\tau$$

and $E(\mathbf{S}_{\text{eff}}) = 0$. However, $E(\mathbf{S}_{\text{eff}} \mid d) = 0$ automatically, hence we can ignore the second requirement $E(\mathbf{S}_{\text{eff}}) = 0$. The property $E(\mathbf{S}_{\text{eff}} \mid d) = 0$ also simplies the first requirement to

$$E(\mathbf{S}_{\text{eff}} \mid \epsilon_\tau, \mathbf{x}) \sum_d \frac{n_d H(d, \mathbf{x}, y)}{n\pi_d} = \mathbf{a}(\mathbf{x}) u_\tau.$$

This gives

$$\mathbf{a}(\mathbf{x}) u_\tau \left\{ \sum_d \frac{n_d H(d, \mathbf{x}, y)}{n\pi_d} \right\}^{-1} = E(\mathbf{S} - \mathbf{g} \mid \epsilon_\tau, \mathbf{x}) - E\left\{ E(\mathbf{S} - \mathbf{g} \mid D) \mid \epsilon_\tau, \mathbf{x} \right\}.$$

From our model (4.3), we have

$$f_{D|X,Y}(d, \mathbf{x}, y) = \frac{n_d H(d, \mathbf{x}, y)}{n\pi_d} \left\{ \sum_d \frac{n_d H(d, \mathbf{x}, y)}{n\pi_d} \right\}^{-1}.$$

Here, explicitly,

$$\begin{aligned} \pi_0 &= \text{pr}^{\text{true}}(D = 0) = \int \eta_1(\mathbf{x}) \eta_2(\epsilon_\tau, \mathbf{x}) H(0, \mathbf{x}, y) d\mu(\mathbf{x}) d\mu(y); \\ \pi_1 &= \text{pr}^{\text{true}}(D = 1) = \int \eta_1(\mathbf{x}) \eta_2(\epsilon_\tau, \mathbf{x}) H(1, \mathbf{x}, y) d\mu(\mathbf{x}) d\mu(y). \end{aligned}$$

To simplify notation, in the following calculation we denote

$$\begin{aligned}
b_0 &= E\{f_{D|X,Y}(1,\mathbf{X},Y) \mid D=0\};\\
b_1 &= E\{f_{D|X,Y}(0,\mathbf{X},Y) \mid D=1\};\\
\mathbf{c}_0 &= E(\mathbf{S} \mid D=0) - E\{E(\mathbf{S} \mid \epsilon_\tau,\mathbf{X}) \mid D=0\};\\
\mathbf{c}_1 &= E(\mathbf{S} \mid D=1) - E\{E(\mathbf{S} \mid \epsilon_\tau,\mathbf{X}) \mid D=1\};\\
\kappa(\mathbf{x},y) &= \left[\textstyle\sum_{d=0}^{1}\{n_d H(d,\mathbf{x},y)\}/(n\pi_d)\right]^{-1};\\
\mathbf{u}_0 &= E\left\{u_\tau \mathbf{a}(\mathbf{X})\kappa(\mathbf{X},Y) \mid D=0\right\};\\
\mathbf{u}_1 &= E\left\{u_\tau \mathbf{a}(\mathbf{X})\kappa(\mathbf{X},Y) \mid D=1\right\};\\
\mathbf{v}_0 &= E(\mathbf{S}-\mathbf{g} \mid D=0);\\
\mathbf{v}_1 &= E(\mathbf{S}-\mathbf{g} \mid D=1).
\end{aligned}$$

Note that $\pi_0+\pi_1=1$, $b_0 n_0 = b_1 n_1$, $\mathbf{c}_0 n_0 + \mathbf{c}_1 n_1 = \mathbf{0}$ and $\mathbf{v}_0\pi_0 + \mathbf{v}_1\pi_1 = \mathbf{0}$.

Under a true model, $\pi_0, \pi_1, b_0, b_1, \mathbf{c}_0, \mathbf{c}_1$ are known quantities, while $\mathbf{u}_0, \mathbf{u}_1, \mathbf{v}_0, \mathbf{v}_1$ are not known because $\mathbf{g} = \mathbf{g}(\epsilon_\tau, \mathbf{x})$ and $\mathbf{a} = \mathbf{a}(\mathbf{x})$ are not specified. To further obtain $\mathbf{u}_0, \mathbf{u}_1, \mathbf{v}_0, \mathbf{v}_1$, we rewrite

$$\begin{aligned}
u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x},y) &= E(\mathbf{S}-\mathbf{g} \mid \epsilon_\tau,\mathbf{x}) - \mathbf{v}_0 f_{D|X,Y}(0,\mathbf{x},y) - \mathbf{v}_1 f_{D|X,Y}(1,\mathbf{x},y)\\
&= E(\mathbf{S} \mid \epsilon_\tau,\mathbf{x}) - \mathbf{g} - \mathbf{v}_0 f_{D|X,Y}(0,\mathbf{x},y) - \mathbf{v}_1 f_{D|X,Y}(1,\mathbf{x},y).
\end{aligned}$$

as

$$\mathbf{g}(\epsilon_\tau,\mathbf{x}) = E(\mathbf{S} \mid \epsilon_\tau,\mathbf{x}) - u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x},y) - \mathbf{v}_0 f_{D|X,Y}(0,\mathbf{x},y) - \mathbf{v}_1 f_{D|X,Y}(1,\mathbf{x},y). \quad \text{(C.9)}$$

Since $\mathbf{v}_0 = E(\mathbf{S} - \mathbf{g} \mid D = 0)$, we obtain

$$\begin{aligned} \mathbf{v}_0 &= E(\mathbf{S} \mid D = 0) - E\big\{E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) - u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, Y) \\ &\qquad\qquad -\mathbf{v}_0 f_{D|X,Y}(0, \mathbf{x}, Y) - \mathbf{v}_1 f_{D|X,Y}(1, \mathbf{x}, Y) \mid D = 0\big\} \\ &= \mathbf{c}_0 + \mathbf{u}_0 + \mathbf{v}_0(1 - b_0) + \mathbf{v}_1 b_0. \end{aligned}$$

Thus, we have $b_0\mathbf{v}_0 - b_0\mathbf{v}_1 - \mathbf{u}_0 = \mathbf{c}_0$. Similarly, from $\mathbf{v}_1 = E(\mathbf{S} - \mathbf{g} \mid D = 1)$, we obtain

$$\begin{aligned} \mathbf{v}_1 &= E(\mathbf{S} \mid D = 1) - E\big\{E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) - \epsilon_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, Y) \\ &\qquad\qquad -\mathbf{v}_0 f_{D|X,Y}(0, \mathbf{x}, Y) - \mathbf{v}_1 f_{D|X,Y}(1, \mathbf{x}, Y) \mid D = 1\big\} \\ &= \mathbf{c}_1 + \mathbf{u}_1 + \mathbf{v}_0 b_1 + \mathbf{v}_1(1 - b_1). \end{aligned}$$

Thus, we have $-b_1\mathbf{v}_0 + b_1\mathbf{v}_1 - \mathbf{u}_1 = \mathbf{c}_1$. Since

$$E\left\{u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, Y)\right\} = \mathbf{0},$$

we have

$$\mathbf{u}_0 n_0 + \mathbf{u}_1 n_1 = \mathbf{0}.$$

Combining the above relations, we have obtained $n_0\mathbf{u}_0 + n_1\mathbf{u}_1 = \mathbf{0}$, $\pi_0\mathbf{v}_0 + \pi_1\mathbf{v}_1 = \mathbf{0}$, $b_0\mathbf{v}_0 - b_0\mathbf{v}_1 - \mathbf{u}_0 = \mathbf{c}_0$ and $-b_1\mathbf{v}_0 + b_1\mathbf{v}_1 - \mathbf{u}_1 = \mathbf{c}_1$. The last two equations are equivalent so one is redundant. Using these relations, we can rewrite $\mathbf{u}_1, \mathbf{v}_0, \mathbf{v}_1$ as a function of $\mathbf{u}_0$:

$$\mathbf{u}_1 = -(n_0/n_1)\mathbf{u}_0, \quad \mathbf{v}_0 = (\pi_1/b_0)(\mathbf{u}_0 + \mathbf{c}_0), \quad \mathbf{v}_1 = -(\pi_0/b_0)(\mathbf{u}_0 + \mathbf{c}_0). \tag{C.10}$$

We cannot obtain a more explicit expression for $\mathbf{u}_0$ at this stage, but we can further obtain

$\mathbf{a}(\mathbf{x})$ as a function of $\mathbf{u}_0$. Using (C.9) and since $E_{\text{true}}(u_\tau \mathbf{g} \mid \mathbf{x}) = \mathbf{0}$, we have

$$
\begin{aligned}
&E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\} - E_{\text{true}}\left\{u_\tau^2 \kappa(\mathbf{x}, Y) \mid \mathbf{x}\right\} \mathbf{a}(\mathbf{x}) \\
&-\mathbf{v}_0 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\} - \mathbf{v}_1 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(1, \mathbf{x}, Y) \mid \mathbf{x}\right\} = \mathbf{0}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathbf{a}(\mathbf{x}) &= \left[E_{\text{true}}\left\{u_\tau^2 \kappa(\mathbf{x}, Y) \mid \mathbf{x}\right\}\right]^{-1} \\
&\quad \left[E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\} - \mathbf{v}_0 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\}\right. \\
&\qquad\qquad \left. -\mathbf{v}_1 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(1, \mathbf{x}, Y) \mid \mathbf{x}\right\}\right] \\
&= \left[E_{\text{true}}\left\{u_\tau^2 \kappa(\mathbf{x}, Y) \mid \mathbf{x}\right\}\right]^{-1} \left[E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\}\right. \\
&\qquad -(\pi_1/b_0)(\mathbf{u}_0 + \mathbf{c}_0) E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\} \\
&\qquad\qquad \left. +(\pi_0/b_0)(\mathbf{u}_0 + \mathbf{c}_0) E_{\text{true}}\left\{u_\tau f_{D|X,Y}(1, \mathbf{x}, Y) \mid \mathbf{x}\right\}\right].
\end{aligned}
$$

To further simplify notation, denote

$$
\begin{aligned}
t_1(\mathbf{x}) &= \left[E_{\text{true}}\left\{u_\tau^2 \kappa(\mathbf{x}, Y) \mid \mathbf{x}\right\}\right]^{-1}; \qquad \text{(C.11)} \\
\mathbf{t}_2(\mathbf{x}) &= E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\} - (\pi_1/b_0)\mathbf{c}_0 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\} \\
&\qquad +(\pi_0/b_0)\mathbf{c}_0 E_{\text{true}}\left\{u_\tau f_{D|X,Y}(1, \mathbf{x}, Y) \mid \mathbf{x}\right\} \\
&= E_{\text{true}}\left\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\right\} - (\mathbf{c}_0/b_0) E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\}; \\
t_3(\mathbf{x}) &= -(\pi_1/b_0) E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\} + (\pi_0/b_0) E_{\text{true}}\left\{u_\tau f_{D|X,Y}(1, \mathbf{x}, Y) \mid \mathbf{x}\right\} \\
&= -b_0^{-1} E_{\text{true}}\left\{u_\tau f_{D|X,Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\right\}.
\end{aligned}
$$

Then

$$\mathbf{a}(\mathbf{x}) = t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x}) + t_3(\mathbf{x})\mathbf{u}_0\}, \tag{C.12}$$

hence the definition of $\mathbf{u}_0$ yields

$$\begin{aligned} \mathbf{u}_0 &= E\left(u_\tau \left[t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x}) + t_3(\mathbf{x})\mathbf{u}_0\}\right] \kappa(\mathbf{x}, Y) \mid D = 0\right) \\ &= E\left\{u_\tau t_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})\kappa(\mathbf{x}, Y) \mid D = 0\right\} + E\left\{u_\tau t_1(\mathbf{x}) t_3(\mathbf{x})\kappa(\mathbf{x}, Y) \mid D = 0\right\} \mathbf{u}_0. \end{aligned}$$

This yields

$$\begin{aligned} \mathbf{u}_0 &= \left[1 - E\left\{u_\tau t_1(\mathbf{x}) t_3(\mathbf{x})\kappa(\mathbf{x}, Y) \mid D = 0\right\}\right]^{-1} && \text{(C.13)} \\ &\quad \times E\left\{u_\tau t_1(\mathbf{x})\mathbf{t}_2(\mathbf{x})\kappa(\mathbf{x}, Y) \mid D = 0\right\}. && \text{(C.14)} \end{aligned}$$

Combining the above results, we have obtained the analytic form of $\mathbf{S}_{\text{eff}} = \mathbf{S} - \mathbf{g} - E(\mathbf{S} - \mathbf{g} \mid D = d)$, where g is given in (C.9), $\mathbf{a}(\mathbf{x})$ is given in (C.12), $\mathbf{v}_0, \mathbf{v}_1$ are given in (C.10) $\mathbf{u}_0$ is given in (C.13) and the functions $t_1, \mathbf{t}_2, t_3$ are given in (C.11).

In forming the estimating equation $\sum_{i=1}^N \mathbf{S}_{\text{eff}} = 0$, we will have $\sum_{i=1}^N \{\mathbf{S}(\mathbf{x}_i, Y_i, D_i) - \mathbf{g}(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_\tau, \mathbf{x}_i)\} - n_0 E(\mathbf{S} - \mathbf{g} \mid D = 0) - n_1 E(\mathbf{S} - \mathbf{g} \mid D = 1) = \mathbf{0}$. Using (C.9), we obtain

$$\begin{aligned} E(\mathbf{S} - \mathbf{g} \mid D = 0) &= E(\mathbf{S} \mid D = 0) - E\{E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid D = 0\} \\ &\quad + E\{u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, Y) \mid D = 0\} \\ &\quad + \mathbf{v}_0 E\{f_{D|X,Y}(0, \mathbf{x}, Y) \mid D = 0\} \\ &\quad + \mathbf{v}_1 E\{f_{D|X,Y}(1, \mathbf{x}, Y) \mid D = 0\} \\ &= \mathbf{c}_0 + \mathbf{u}_0 + \mathbf{v}_0(1 - b_0) + \mathbf{v}_1 b_0 \end{aligned}$$

and

$$\begin{aligned} E(\mathbf{S}-\mathbf{g}\mid D=1) &= E(\mathbf{S}\mid D=1)-E\{E(\mathbf{S}\mid \epsilon_\tau,\mathbf{x})\mid D=1\} \\ &\quad +E\{u_\tau\mathbf{a}(\mathbf{x})\kappa(\mathbf{x},Y)\mid D=1\} \\ &\quad +\mathbf{v}_0E\{f_{D|X,Y}(0,\mathbf{x},Y)\mid D=1\} \\ &\quad +\mathbf{v}_1E\{f_{D|X,Y}(1,\mathbf{x},Y)\mid D=1\} \\ &= \mathbf{c}_1+\mathbf{u}_1+\mathbf{v}_0b_1+\mathbf{v}_1(1-b_1), \end{aligned}$$

hence

$$\begin{aligned} & n_0E(\mathbf{S}-\mathbf{g}\mid D=0)+n_1E(\mathbf{S}-\mathbf{g}\mid D=1) \\ = \ & n_0\{\mathbf{c}_0+\mathbf{u}_0+\mathbf{v}_0(1-b_0)+\mathbf{v}_1b_0\}+n_1\{\mathbf{c}_1+\mathbf{u}_1+\mathbf{v}_0b_1+\mathbf{v}_1(1-b_1)\} \\ = \ & (n_0\mathbf{c}_0+n_1\mathbf{c}_1)+(n_0\mathbf{u}_0+n_1\mathbf{u}_1)+(n_0\mathbf{v}_0+n_1\mathbf{v}_1)+(\mathbf{v}_1-\mathbf{v}_0)(n_0\mathbf{b}_0-n_1\mathbf{b}_1) \\ = \ & n_0\mathbf{v}_0+n_1\mathbf{v}_1. \end{aligned}$$

Thus, the estimating equation simplifies to

$$\textstyle\sum_{i=1}^N\{\mathbf{S}(\mathbf{X}_i,Y_i,D_i)-\mathbf{g}(Y_i-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_\tau,\mathbf{X}_i)\}-n_0\mathbf{v}_0-n_1\mathbf{v}_1=\mathbf{0}.$$

C.6 Proof of Theorem 4

Proof: For simplicity of proof, we split the n observations randomly into two sets. The first set contains $n-n^{1-\delta}$ observations and the second set contains $n^{1-\delta}$ observations, where $0<\delta<(1/2-d\tau)$ is a small positive number. We form and solve the estimating equation using data in the first set, while calculating all the hatted quantities described in the algorithm in Section 4.4 using data in the second set. We use this only as a technical device, although in our simulations and empirical example we used all the data.

In the algorithm, the approximations involve either replacing expectation with averaging or standard kernel regression estimation, hence the differences between the quantities with hat and without hat have either mean zero, standard deviation $O(n^{-(1-\delta)/2})$, or mean $O(h^r)$, standard deviation $O\{(n^{1-\delta}h^d)^{-1/2}\}$. In particular, $\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) - \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)$ has bias $O(h^r)$ and standard deviation $O\{(n^{1-\delta}h^d)^{-1/2}\}$. Thus,

$$
\begin{aligned}
\mathbf{0} &= (n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}} \widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \widehat{\boldsymbol{\theta}}) \\
&= (n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}} \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) \\
&\quad +(n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}} \left\{\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) - \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)\right\} \\
&\quad +E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} + o_p(1)\right\}(n-n^{1-\delta})^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&= (n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}} \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) \\
&\quad +E\left\{\frac{\partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}(n-n^{1-\delta})^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\quad +(n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}} \left\{\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) - \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)\right\} + o_p(1).
\end{aligned}
$$

We see that $\widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)$ differs from $\mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0)$ in that all the unknown quantities, except $\mathbf{S}^*$, are estimated. This is equivalent to estimating the unknown functions $\eta_1(\mathbf{x})$, $\eta_2(\epsilon_\tau, \mathbf{x})$ in (4.3) and using the estimate $\widehat{\eta}_1(\mathbf{x})$, $\widehat{\eta}_2(\epsilon_\tau, \mathbf{x})$ in calculating $\mathbf{S}^*_{\text{eff}}$

from the posited $\mathbf{S}^*$. Thus, denoting $\widehat{\eta} = (\widehat{\eta}_1, \widehat{\eta}_2)$, we can approximate

$$
\begin{aligned}
& (n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n-n^{1-\delta}} \Big\{ \widehat{\mathbf{S}}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) - \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0) \Big\} \\
= \;& (n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n-n^{1-\delta}} \{ \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \widehat{\eta}) - \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \eta_0) \} \\
= \;& \{ (n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n-n^{1-\delta}} \partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \eta_0)/\partial\eta \} (\widehat{\eta} - \eta_0) \qquad \text{(C.15)} \\
& + O_p\{ (n - n^{1-\delta})^{1/2} (\widehat{\eta} - \eta_0)^2 \} + o_p(1), \qquad \text{(C.16)}
\end{aligned}
$$

where $\partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \eta_0)/\partial\eta$ is pathwise derivative. However, $\mathbf{S}^*_{\text{eff}}$ is the projection of $\mathbf{S}^*$ to $\Lambda^\perp$ so $\mathbf{S}^*_{\text{eff}} \in \Lambda^\perp$. Thus, for any parametric submodel of η involving parameter γ, we have

$$
\begin{aligned}
& E\{ \partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \gamma)/\partial\gamma^{\mathrm{T}} \} \\
= \;& \int \frac{\partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \gamma)}{\partial\gamma^{\mathrm{T}}} f_{\mathbf{X},Y,D}(\mathbf{x}, y, d) d\mu(\mathbf{x}) \mu(y) d\mu(d) \\
= \;& -\int \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \gamma) \frac{\partial \log\{ f_{\mathbf{X},Y,D}(\mathbf{x}, y, d) \}}{\partial\gamma^{\mathrm{T}}} f_{\mathbf{X},Y,D}(\mathbf{x}, y, d) d\mu(\mathbf{x}) \mu(y) d\mu(d) \\
= \;& -E\{ \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \gamma) \mathbf{S}_\gamma^{\mathrm{T}} \} = \mathbf{0}.
\end{aligned}
$$

The last equality is because by definition $\mathbf{S}_\gamma \in \Lambda$ which is orthogonal to $\Lambda^\perp$ and $\mathbf{S}^*_{\text{eff}} \in \Lambda^\perp$. Here, $f_{\mathbf{X},Y,D}(\mathbf{x}, y, d)$ is defined in (4.3). Because γ is parameter of any arbitrary submodel of η, we actually have obtained

$$
E\{ \partial \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \eta_0)/\partial\eta \} = -E\{ \mathbf{S}^*_{\text{eff}}(D_i, \mathbf{X}_i, Y_i, \boldsymbol{\theta}_0, \eta_0) \mathbf{S}_\eta^{\mathrm{T}} \} = \mathbf{0},
$$

where $\mathbf{S}_\eta$ is the nuisance score function along the arbitrarily chosen specific path of the pathwise derivative. Thus, the first term of (C.15) is of order $o_p(1)$. On the other hand, $O_p\{(n-n^{1-\delta})^{1/2}(\widehat{\eta}-\eta_0)^2\} = O_p\{n^{1/2}h^{2r}+n^{1/2}(n^{1-\delta}h^d)^{-1}\} = O_p(n^{1/2-2r\tau}+n^{-1/2+\delta+d\tau}) = o_p(1)$. We therefore obtain

$$\begin{aligned}\mathbf{0} &= (n-n^{1-\delta})^{-1/2}\sum_{i=1}^{n-n^{1-\delta}}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)\\ &\quad +E\left\{\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\}(n-n^{1-\delta})^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(1).\end{aligned}$$

This yields $(n-n^{1-\delta})^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to \text{Normal}\{\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$, and hence

$$n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to \text{Normal}\{\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

when $n\to\infty$.

□