Title

Essays on Causal Inference and Econometrics

Author

Xie, Haitian

Publication Date

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays on Causal Inference and Econometrics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Haitian Xie

Committee in charge:

Professor Graham Elliott, Co-Chair
Professor Yixiao Sun, Co-Chair
Professor Songzi Du
Professor Dimitris Politis
Professor Wenxin Zhou
Professor Ying Zhu

2023

The Dissertation of Haitian Xie is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents, Jun Xie and Xin Guo.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my advisors, Professor Graham Elliott and Professor Yixiao Sun, for their exceptional mentorship, encouragement, and support throughout my PhD journey in econometrics. Their vision, insights, and expertise have been crucial in leading the direction of my research and providing detailed guidance on its implementation. This dissertation would not have been possible without their unwavering support and belief in me.

I would also like to thank my committee member Professor Ying Zhu for her exceptional support and collaboration on my research projects. I am also grateful to my other committee members, Professor Songzi Du, Professor Dimitris Politis, and Professor Wenxin Zhou, for their constructive and insightful comments that have greatly improved the quality of this work.

I want to extend my heartfelt thanks to my parents and grandparents for their love, support, and encouragement throughout my years of study. Their firm belief in me has been a constant source of motivation.

Chapter 1, in full, is currently being prepared for submission for publication of the material. Xie, Haitian. “Nonlinear and Nonseparable Structural Functions in Fuzzy Regression Discontinuity Designs.” The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, has been submitted for publication. Xie, Haitian. “Uniform Convergence Results for the Local Linear Regression Estimation of the Conditional Distribution.” The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, has been submitted for publication. Xie, Haitian. “Efficient and Robust Estimation of the Generalized LATE Model.” The dissertation author was the primary investigator and author of this material.

VITA

2017	Bachelor of Arts and Bachelor of Science, Wuhan University
2023	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays on Causal Inference and Econometrics

by

Haitian Xie

Doctor of Philosophy in Economics

University of California San Diego, 2023

Professor Graham Elliott, Co-Chair
Professor Yixiao Sun, Co-Chair

This dissertation is a collection of three essays on the econometric analysis of causal inference methods. Chapter 1 examines the identification and estimation of the structural function in fuzzy RD designs with a continuous treatment variable. We show that the nonlinear and nonseparable structural function can be nonparametrically identified at the RD cutoff under shape restrictions, including monotonicity and smoothness conditions. Based on the nonparametric identification equation, we propose a three-step semiparametric estimation procedure and establish the asymptotic normality of the estimator. The semiparametric estimator achieves the same convergence rate as in the case of a binary treatment variable. As an application of the method, we estimate the causal effect of sleep time on health status by using the discontinuity in natural light timing at time zone boundaries.

Chapter 2 examines the local linear regression (LLR) estimate of the conditional distribution function $F(y|x)$. We derive three uniform convergence results: the uniform bias expansion, the uniform convergence rate, and the uniform asymptotic linear representation. The uniformity in the above results is with respect to both x and y and therefore has not previously been addressed in the literature on local polynomial regression. Such uniform convergence results are especially useful when the conditional distribution estimator is the first stage of a semiparametric estimator.

Chapter 3 studies the estimation of causal parameters in the generalized local average treatment effect model, a generalization of the classical LATE model encompassing multi-valued treatment and instrument. We derive the efficient influence function (EIF) and the semiparametric efficiency bound for two types of parameters: local average structural function (LASF) and local average structural function for the treated (LASF-T). The moment condition generated by the EIF satisfies two robustness properties: double robustness and Neyman orthogonality. Based on the robust moment condition, we propose the double/debiased machine learning (DML) estimators for LASF and LASF-T. We also propose null-restricted inference methods that are robust against weak identification issues. As an empirical application, we study the effects across different sources of health insurance by applying the developed methods to the Oregon Health Insurance Experiment.

Chapter 1

Nonlinear and Nonseparable Structural Functions in Fuzzy Regression Discontinuity Designs

1.1 Introduction

The *regression discontinuity* (RD) design is one of the most important approaches to causal inference in non-experimental settings. In an RD design, the researcher is interested in the effect of a treatment T on some outcome Y. The basic idea is that there is an observed running variable R (also called score or index or forcing variable) such that the treatment varies discontinuously when the running variable crosses some cutoff (also called threshold) value $\bar{r}$. By utilizing the variation induced by this discontinuity, the researcher has the power to identify and estimate the causal impact of interest.

Most theoretical studies of the RD design assume that the treatment is a binary intervention. However, in empirical studies of RD design, researchers may be interested in a continuous treatment that takes values inside an interval. Such examples include sleep time [Giuntella and Mazzonna, 2019], air pollution level [Chen et al., 2013, Ebenstein et al., 2017], and medical expense for infants [Almond et al., 2010, Barreca et al., 2011]. Our study provides methods for examining the causal effect of a continuous treatment variable in an RD setting.

With a binary treatment, the sharp design is the case where the treatment can be predicted with certainty by whether or not a running variable R is above or below some cutoff (Figure 1.1(a)). The fuzzy design refers to the case where the treatment probability is a function of the running variable R and changes discretely at the cutoff (Figure 1.1(b)). Regression discontinuity methods exploit this predicted variation in the treatment above and below the cutoff to identify the effect of the treatment.

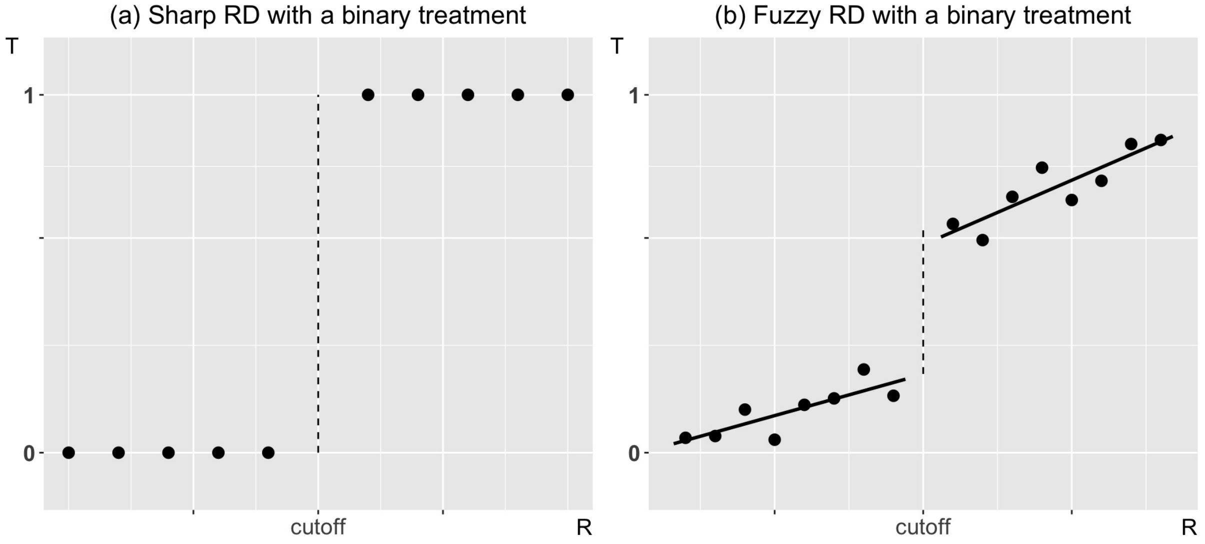

Figure 1.1. Demonstration of RD designs with a binary treatment.

Graph (a) demonstrates the sharp RD design using a raw scatter plot. Graph (b) demonstrates the fuzzy RD design using a binscatter plot, where each dot represents the treatment probability in the respective bin.

With a continuous treatment, where all observations are treated to some extent, the distribution of the treatment is a function of the running variable *R*, and the distribution of treatments shifts discretely at the cutoff. It is this variation, as in the binary treatment case, that allows the identification of the effect of the treatment at the cutoff.

The essence of the idea is shown in Figure 1.2. For any *R*, we have a distribution of the treatment T. Suppose that T is increasing in R, as illustrated in Figure 1.2(a). As R increases, the level of treatment becomes larger, as shown by the increasing quantiles of T. At $R = \bar{r}$, the distribution of treatment levels increases discretely, as shown by the discontinuous jump in the quantiles. This is the discontinuity that will be exploited for estimating the effect of the treatment

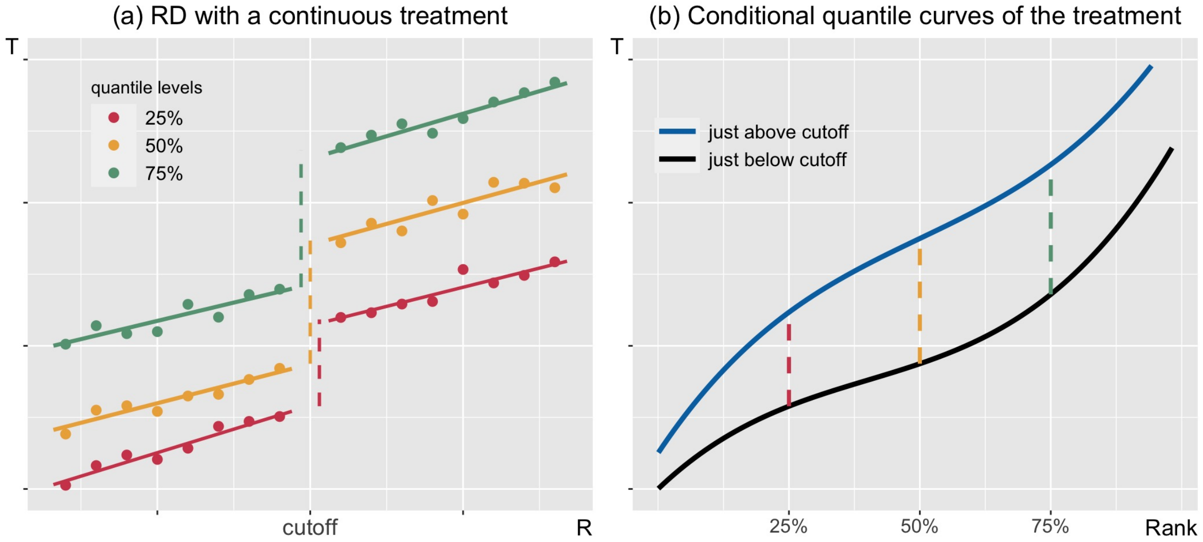

Figure 1.2. Demonstration of RD designs with a continuous treatment.

Graph (a) demonstrates the regression discontinuities of a continuous treatment variable at different quantile levels. The plot is a binscatter plot, where each dot represents the corresponding quantile treatment level in the respective bin. Different quantile regressions bring different discontinuities. Graph (b) plots the conditional quantile curve of the treatment from just below and just above the cutoff. The horizontal axis specifies which quantile level we are looking at. The entire difference between these two curves constitutes the content of RD of a continuous treatment variable. If there is no regression discontinuity, then the two quantile curves would completely overlap. Notice that we use the same color to denote the corresponding jumps between the two plots.

on the outcome variable. Consider now the distribution of treatments just below and just above the cutoff. Figure 1.2(b) shows that for R just below the cutoff, the distribution of treatment is below that of the distribution of treatment immediately above the cutoff. Thus the treatment levels differ, and we can separate levels of treatment when there is a small variation in R around this cutoff point.

To capture the causal effect of the treatment T on the outcome Y, we introduce the structural function

$$Y = g^*(T,R,\varepsilon),$$

where ε contains unobserved causal factors (for easy reference, ε will be called the error term hereafter). The structural function g^* specifies how the treatment T causes the outcome Y

together with the running variable R and error term ε. The identification of g^* is a difficult task, especially without restricting the functional form of g^* or the correlation between ε and (T,R). The goal of our paper is to identify the structural function $g^*(\cdot,\bar{r},\cdot)$ by utilizing the discrete shift in treatment distribution at the cutoff $\bar{r}$. The proposed method fully describes g^* at the cutoff. Summary statistics that are functions of $g^*(\cdot,\bar{r},\cdot)$ can also be constructed.

We provide a nonparametric identification result for the structural function. The identification assumptions are shape restrictions, including monotonicity and smoothness conditions. Under these assumptions, the structural function is still allowed to be nonlinear in the treatment and nonseparable between the treatment and the error term. We propose a three-step semiparametric estimation procedure for the identified structural function $g^*(\cdot,\bar{r},\cdot)$ and derive the asymptotic normal distribution of the estimator.

Consider our empirical study for concreteness: we study the causal impact of sleep time T on health status Y by exploiting the discontinuity in the timing of natural light at time zone boundaries. In this empirical study, the running variable R is the distance to the time zone boundary, and the cutoff $\bar{r}$ is at the time zone boundary. The causal identification is based on the following fact: Individuals living on the late sunset side of the time zone boundary tend to go to bed at a later time, while in the morning, everyone gets up and goes to work at 8 am. Consequently, there is an exogenous variation in sleep time across the time zone boundary. This RD design based on the time zone system is first proposed by [Giuntella and Mazzonna, 2019]. Figure 1.3(a) shows the histogram of sleep time based on the American Time Use Survey (ATUS) and demonstrates that sleep time is indeed a continuous treatment variable. Figure 1.3(b) shows the nonparametric estimates of the conditional quantiles of sleep time. The distribution of the sleep time for individuals living on the early sunset side first-order stochastically dominates the distribution on the late sunset side, which is evidence supporting the above identification strategy.

The empirical study can demonstrate the importance of having a general functional form specification of the structural function. First, there are reasons for one to believe that the

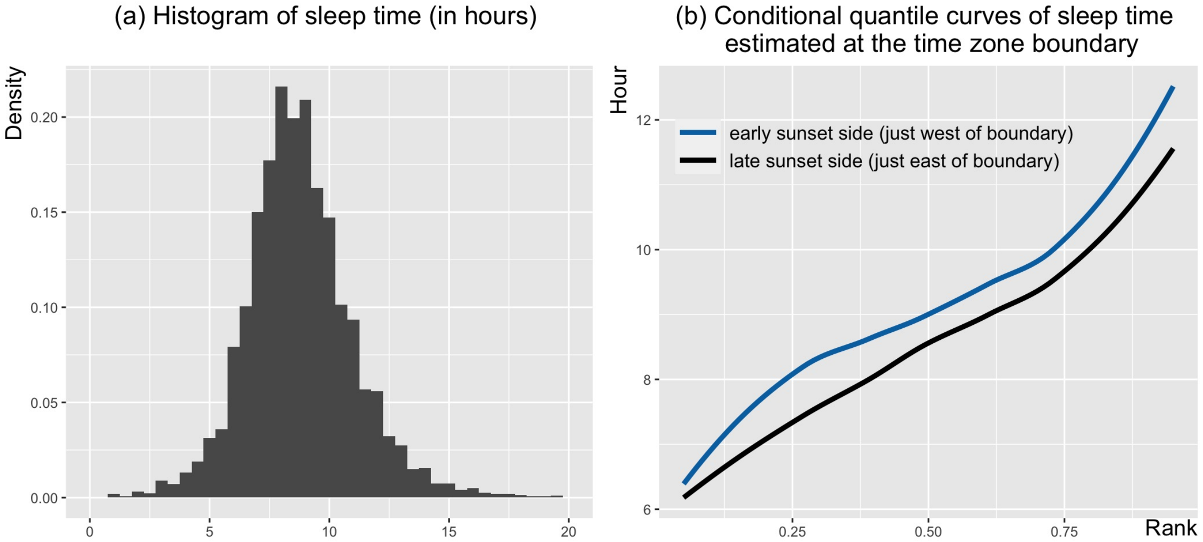

Figure 1.3. Empirical illustration of RD designs with a continuous treatment.

Graph (a) shows the histogram of sleep time. Evidently, this variable is better modeled as continuous rather than discrete. Graph (b) shows the estimated conditional quantile curves of sleep time given that the geographical location is just west and east of the time zone boundary. The nonparametric estimator used here is the local constant quantile regression. The RD is clearly observed as the blue curve first-order stochastically dominates the black curve, a similar situation as demonstrated in Figure 1.2(b).

causal effect from sleep time to health is nonlinear and nonseparable.[1] Second, a nonlinear structural function is required if the researcher wants to find the optimal sleep time for policy recommendations. A fully linear structural function can only find if more or less sleep is better, and would recommend sleeping 24 or 0 hours every day. This is obviously not the correct policy implication, and is at odds with experimental findings. In Section 1.4, we present the novel empirical results for the estimated structural derivative $\frac{\partial}{\partial t} g^*(\cdot, \bar{r}, \cdot)$ based on the proposed estimator and derive the corresponding optimal sleep time.

The rest of the paper is organized as follows. The remaining part of this section discusses the literature. Section 1.2 introduces the RD model with a continuous treatment and presents the nonparametric identification result. Section 1.3 proposes the semiparametric estimation procedure and derives its asymptotic properties. Section 1.4 presents the empirical study and

[1] The nonlinearity can be due to the fact that both undersleeping and oversleeping are harmful to health [Hairston et al., 2010]. The nonseparability can be due to the effect heterogeneity caused by unobserved eating habits, which affect both sleep time and health.

simulation results. The technical proofs for the identification and estimation results are collected in Appendices A.1 and A.2, respectively.

1.1.1 Relation to the literature

The RD method is first introduced by Thistlethwaite and Campbell [1960] into the literature. Hahn et al. [2001] establish the theoretical foundation of RD designs by using the potential outcome framework and demonstrate the identification of the *local average treatment effect* (LATE) for compliers local to the cutoff. Early reviews of the RD design can be found in Imbens and Lemieux [2008] and Lee and Lemieux [2010]. For more recent reviews, see Cattaneo and Escanciano [2017] and Cattaneo and Titiunik [2021].

In the empirical studies, researchers typically use the *two-stage least squares* (TSLS) method to estimate the following *Wald ratio* around the cutoff:

$$\text{Wald ratio} = \frac{\lim_{r\uparrow\bar{r}}\mathbb{E}[Y|R=r] - \lim_{r\downarrow\bar{r}}\mathbb{E}[Y|R=r]}{\lim_{r\uparrow\bar{r}}\mathbb{E}[T|R=r] - \lim_{r\downarrow\bar{r}}\mathbb{E}[T|R=r]}.$$

There are two motivations behind this procedure. First, in the binary treatment case, the Wald ratio would identify the treatment effect.[2] Second, in the continuous treatment case, if the structural function is linear and separable in the treatment, that is, g^* can be decomposed as $g^*(T,R,\varepsilon) = \beta T + \tilde{g}(R,\varepsilon)$, then the Wald ratio would identify the slope coefficient β of the treatment.[3] However, the Wald ratio, as calculated in the literature, cannot identify the structural function in general because the structural function is infinite-dimensional, while the Wald ratio only provides one-dimensional summary information. Under general functional form assumptions on g^*, the Wald ratio is identified as a function of $g^*(\cdot,\bar{r},\cdot)$, which represents a weighted average effect of the treatment T on the outcome Y at the cutoff $R=\bar{r}$. Important information in the original structural function, for example, the optimal level of treatment, is lost

[2] As shown in Hahn et al. [2001], the Wald ratio identifies the average treatment effect in the sharp design and the local average treatment effect (for the compliers) in the fuzzy design.

[3] To see that, we can plug the structural function into the definition of the Wald ratio.

when the structural function is condensed into this scalar weighted average. Hence, there is room for potential improvement in the above empirical studies by using our semiparametric estimator.

There are a few theoretical studies that extend RD design beyond the standard binary treatment. Caetano et al. [2020] study RD designs with a multi-valued discrete treatment variable. Section 3.4.2 of Lee and Lemieux [2010] discusses a fully linear model of RD design with a continuous treatment. The most relevant paper to ours is the recent work by Dong et al. [2021], which studies RD designs specifically with a continuous treatment variable. They propose a way to identify and estimate the *Quantile specific LATE* (Q-LATE) defined as

$$\frac{\lim_{r\uparrow\bar{r}}\mathbb{E}[Y|U=u,R=r]-\lim_{r\downarrow\bar{r}}\mathbb{E}[Y|U=u,R=r]}{\lim_{r\uparrow\bar{r}}\mathbb{E}[T|U=u,R=r]-\lim_{r\downarrow\bar{r}}\mathbb{E}[T|U=u,R=r]}.$$

The Q-LATE parameter is the Wald ratio given a particular quantile level of the treatment. This parameter is a weighted average of the derivative of the structural function [Dong et al., 2021, Section 2.2]. The identification of the Q-LATE parameter is achieved based on assumptions weaker than our paper. In the current study, we make an effort to directly identify the structural function and its derivative at the expense of making stronger assumptions. The structural function itself can be more informative for policy design purposes than a weighted average. In particular, by identifying the structural function, we can recover not only the Q-LATE parameter but also many other policy-relevant quantities. For example, in the empirical study in Section 1.4, we show how the estimated structural function helps determine the optimal sleep time.

Since RD design can be interpreted as a local *instrumental variable* (IV) approach, our paper is naturally connected to the large body of nonparametric IV literature. In particular, the identification result in this paper is related to the literature on instruments with small support: Torgovitsky [2015, 2017], D'Haultfœuille and Février [2015], Masten and Torgovitsky [2016] and Ishihara [2021]. These papers examine a model with a discrete IV and a continuous treatment variable. To directly apply the IV identification result to our setting, we would require the running variable R to be a valid instrument: R is (conditionally) independent with ε and is excluded from the structural function g^*.

The challenge in the RD setting is that the independence and the exclusion restriction for the running variable R are typically violated. In the time zone example, the location one lives is correlated with the unobserved eating habits (contained in ε) and may directly affect the health status Y. The contribution of our paper is to formally establish the identification of the structural function in the RD setting by utilizing only the discontinuity at the cutoff. In particular, the running variable R is allowed to be correlated with ε and included in g^*. The estimation procedure in our RD design is also more challenging than IV estimation because we focus on a local neighborhood of the cutoff. Unlike in the IV setup, where $\sqrt{n}$-consistent estimators exist [Torgovitsky, 2017], the proposed semiparametric estimator in our RD design is $n^{-2/5}$-consistent.

The problem studied by this paper is also related to the broad literature on nonseparable models. Relevant papers include Matzkin [2003], Hoderlein and Mammen [2007, 2009], Sasaki [2015], and Su et al. [2019]. The identification there relies on the exogeneity of the treatment, which is not assumed in RD designs.

1.2 RD Design with a continuous treatment

This section describes the RD model with a continuous treatment, explains the assumptions of the model, and discusses the nonparametric identification of the structural function local to the cutoff.

1.2.1 The model

We study the following causal equation:

$$Y = g^*(T, R, \varepsilon), \tag{1.1}$$

where Y is the outcome of interest, T is the treatment, and R is the running variable. The scalar variable ε represents unobserved causal factors in the outcome equation. We assume all the random variables are absolutely continuous. The function g^* is the unknown true structural function.

The running variable R partly determines the treatment T by the following treatment choice function:

$$T = \begin{cases} m_0(R, U_0), R < \bar{r}, \\ m_1(R, U_1), R \geq \bar{r}, \end{cases} \tag{1.2}$$

where $\bar{r}$ is the cutoff value, and U_0 and U_1 are scalar variables, representing other factors that are not observable to an econometrician. For easy reference, they will be referred to as the error terms hereafter. The important feature of the RD design is that the treatment varies discontinuously when the running variable crosses the cutoff $\bar{r}$. The functions m_0 and m_1 represent respectively the treatment choice mechanism when R is below and above the cutoff.

It is important to point out that the variables U_0, U_1, T and R are allowed to be correlated with the error term ε. If we assume ε to be independent of (T, R), then we can follow Matzkin [2003] or Hoderlein and Mammen [2007] to identify the structural function. If we assume $\varepsilon \perp R$ and R is excluded from m_0, m_1 and g, then we can follow Torgovitsky [2015] to identify the structural function by treating the binary variable $\mathbf{1}\{R \geq \bar{r}\}$ as the instrument.

We make the following assumptions on the model imposed by (1.1) - (1.2). Let $\mathscr{G}$ be the set of candidate structural functions such that the true g^* is contained in $\mathscr{G}$. That is, $\mathscr{G}$ is the infinite-dimensional parameter space where the structural function belongs to. Denote the conditional distribution function by $F_{\cdot|\cdot}(\cdot|\cdot)$, the conditional density function by $f_{\cdot|\cdot}(\cdot|\cdot)$, and the conditional quantile function by $F^{-1}_{\cdot|\cdot}(\cdot|\cdot)$.

Assumption 1.1 (Dual Monotonicity)**.**

(i) Every $g \in \mathscr{G}$ satisfies that for each given $T = t$ and $R = r$, g is strictly increasing in ε.

(ii) *For each given $R = r$, m_0 is strictly increasing in U_0 and m_1 is strictly increasing in U_1.*

Assumption 1.2 (Smoothness)**.**

(i) *The functions m_0, m_1 and every $g \in \mathscr{G}$ are continuous on their respective domains.*

(ii) *The conditional quantile functions $F^{-1}_{U_0|R}(u|r)$ and $F^{-1}_{U_1|R}(u|r)$ are strictly increasing in u and continuous in (u, r).*

(iii) *The conditional distribution functions $F_{\varepsilon|U_0,R}(e|u,r)$ and $F_{\varepsilon|U_1,R}(e|u,r)$ are strictly increasing in e and are continuous in r at $\bar{r}$.*

(iv) *The running variable R is absolutely continuous, and its density is strictly positive around the cutoff $\bar{r}$.*

Assumption 1.3 (Rank Similarity)**.** *$U_0|(\varepsilon, R = \bar{r}^-)$ has the same distribution as $U_1|(\varepsilon, R = \bar{r}^+)$. That is, for every u in the common support of U_0 and U_1 and every e in the support of ε,*

$$\lim_{r\uparrow\bar{r}} f_{U_0|\varepsilon,R}(u|e,r) = \lim_{r\downarrow\bar{r}} f_{U_1|\varepsilon,R}(u|e,r).$$

Assumption 1.1 restricts the shape of the structural function and the treatment choice function. Assumption 1.1(i) has implications on the treatment effect heterogeneity. In particular, it requires the error term ε to be one-dimensional. Such a condition is imposed in Torgovitsky [2015] and D'Haultfœuille and Février [2015] for the IV setting and Matzkin [2003] for the setting with an exogenous treatment. This condition is considered a reasonable sacrifice for deriving a strong result of directly identifying the structural function. Assumption 1.1(ii) imposes monotonicity on the unobserved heterogeneity in the treatment choice. This condition is similar to Assumption 1 in Dong et al. [2021]. In fact, the monotonicity assumption in the treatment choice model is common in the IV literature [e.g., Imbens and Newey, 2009]. From the theoretical perspective, the entirety of Assumption 1.1 suggests that there is a one-to-one mapping between (Y, T) and (ε, U_0, U_1) for a given value of the running variable R.

Assumption 1.2 states that except for the discontinuity introduced in (1.2), everything else is reasonably smooth. In particular, the conditional distribution of ε given (U,R) needs to be smooth with respect to the running variable R at the cutoff $\bar{r}$. This condition indicates that the discontinuity in the outcome Y is purely generated by the discontinuity in the treatment T.

Assumption 1.3 is similar to Assumption 3 in Dong et al. [2021]. It imposes the rank similarity condition on the potential treatments. In the time zone example, this assumption requires that the probability for sleep time to stay at a certain rank remains the same regardless of whether the individual lives just east or west of the time zone boundary. [4]

The treatment choice functions (m_0, m_1) are not identified. Rather than trying to identify them, it is more convenient to consider a normalization to a quantile representation. By using the monotonicity of m_0 and m_1 in Assumption 1.1(ii), we define

$$U = \mathbf{1}\{R < \bar{r}\} F_{U_0|R}(U_0|R) + \mathbf{1}\{R \geq \bar{r}\} F_{U_1|R}(U_1|R) = F_{T|R}(T|R), \tag{1.3}$$

as the conditional rank of T given R.[5] Then the treatment choice model in (1.2) can be written as

$$T = h(R,U) = \begin{cases} h_0(R,U), R < \bar{r}, \\ h_1(R,U), R \geq \bar{r}, \end{cases}$$

where

$$h_0(r,u) = m_0\left(r, F^{-1}_{U_0|R}(u|r)\right) \text{ and } h_1(r,u) = m_1\left(r, F^{-1}_{U_1|R}(u|r)\right).$$

By using $[r_0, r_1]$ to denote the support of R, we can write the domains of h_0 and h_1 respectively as $[r_0, \bar{r}] \times [0,1]$ and $[\bar{r}, r_1] \times [0,1]$.

The following lemma shows that the function h defined above is the conditional quantile

[4] Notice that Assumption 1.3 is different from the rank similarity condition in the IV quantile regression model Chernozhukov and Hansen [2005], where the similarity is imposed on the rank of potential outcomes.

[5] The second equality in Equation (1.3) is proved in Lemma 1.1

function of T given R, and the quantile representation is a valid normalization in the sense that it preserves the monotonicity and smoothness conditions. Consequently, the function h (including both h_0 and h_1) and the conditional rank $U = h^{-1}(R,T)$ are identified from the data, where h^{-1} denotes the inverse of h with respect to the second argument U. In fact, the two functions $h_0(\bar{r},\cdot)$ and $h_1(\bar{r},\cdot)$ are the curves shown in Figure 1.2(b), in which the horizontal axis represents the conditional rank U.

Lemma 1.1 (Quantile Representation). *The following statements hold under Assumptions 1.1 - 1.3:*

(i) $U \perp R$, $U|R \sim \mathit{Unif}[0,1]$, *and* $\mathbb{P}(T \leq h(R,u)|R) = u, u \in [0,1]$.

(ii) For each $R = r$, h_0 *and* h_1 *are strictly increasing in* U.

(iii) The functions h_0 *and* h_1 *are continuous.*

(iv) The conditional distribution function $F_{\varepsilon|U,R}(e|u,r)$ *is strictly increasing in* e *and is continuous in* r *at* $\bar{r}$*, that is,*

$$\lim_{r\uparrow\bar{r}} F_{\varepsilon|U,R}(e|u,r) = \lim_{r\downarrow\bar{r}} F_{\varepsilon|U,R}(e|u,r), \textit{ for every } (e,u). \tag{1.4}$$

By construction, U is independent of R, but U and ε are possibly correlated even after conditioning on R. The following assumption states that the support of the unobserved ε does not vary with U or R. This invariance of the support is not strong since it still allows ε to be correlated with U or R in any way.

Assumption 1.4 (Support Invariance). *Supp*$(\varepsilon|U = u, R = r)$ *does not depend on* u *or* r *in the neighborhood of* $\bar{r}$. *This common support is denoted by* $\mathscr{E}$.

We use the following notation to denote the left and right limits. Let $F_{Y|T,R}$ be the conditional distribution function of Y given T and R. We define

$$F^{-}_{Y|T,R}(y|t,r) = \begin{cases} F_{Y|T,R}(y|t,r), & \text{if } r < \bar{r}, \\ \lim_{r\uparrow\bar{r}} F_{Y|T,R}(y|t,r), & \text{if } r = \bar{r}. \end{cases}$$

$$F^{+}_{Y|T,R}(y|t,r) = \begin{cases} F_{Y|T,R}(y|t,r), & \text{if } r > \bar{r}, \\ \lim_{r\downarrow\bar{r}} F_{Y|T,R}(y|t,r), & \text{if } r = \bar{r}. \end{cases}$$

The conditional density functions $f^{-}_{T|R}$ and $f^{+}_{T|R}$ are analogously defined. These left and right limits exist in view of Assumptions 1.1 and 1.2.

1.2.2 Nonparametric identification

We derive the key identification equation starting from (1.4), which is presented below for reference:

$$\lim_{r\uparrow\bar{r}} F_{\varepsilon|U,R}(e|u,r) = \lim_{r\downarrow\bar{r}} F_{\varepsilon|U,R}(e|u,r).$$

This equation means that the conditional distribution of ε is smooth at the cutoff of the running variable. Then we recall that Assumption 1.1 (the dual monotonicity condition), together with Lemma 1.1, establishes a one-to-one mapping between (ε,U) and (Y,T). Consequently, we can relate the unobserved distribution of ε to the observed distribution of (Y,T) as follows:

$$\lim_{r\uparrow\bar{r}} F_{\varepsilon|U,R}(e|u,r) = F^{-}_{Y|T,R}(g^*(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}),$$
$$\lim_{r\downarrow\bar{r}} F_{\varepsilon|U,R}(e|u,r) = F^{+}_{Y|T,R}(g^*(h_1(\bar{r},u),\bar{r},e)|h_1(\bar{r},u),\bar{r}).$$

Combing the above equations, we obtain that

$$F^{-}_{Y|T,R}(g^*(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}) = F^{+}_{Y|T,R}(g^*(h_1(\bar{r},u),\bar{r},e)|h_1(\bar{r},u),\bar{r}), \text{ for every } (e,u).$$

Given that we can identify $F^{\pm}_{Y|T,R}$ and (h_0, h_1), the above condition imposes a constraint on the structural function $g^*(\cdot, \bar{r}, \cdot)$. This constraint helps us identify $g^*(\cdot, \bar{r}, \cdot)$.

In order to precisely refer to the above condition, we introduce the following definition: A function $g \in \mathscr{G}$ is said to satisfy Condition (1.5) if for every $e \in \mathscr{E}$ and $u \in [0,1]$,

$$F^{-}_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}) = F^{+}_{Y|T,R}(g(h_1(\bar{r},u),\bar{r},e)|h_1(\bar{r},u),\bar{r}). \tag{1.5}$$

The previous analysis shows that the true structural function g^* satisfies Condition (1.5), which is formally presented in the following lemma.

Lemma 1.2 (Local Control Function)**.** *Under Assumptions 1.1 - 1.3, g^* satisfies Condition (1.5) with both sides of the equation equal to $F_{\varepsilon|U,R}(e|u,\bar{r})$.*

For the true structural function g^*, Condition (1.5) can also be written as

$$\lim_{r \uparrow \bar{r}} \mathbb{P}(\varepsilon | T = h_0(r,u), R = r) = \lim_{r \downarrow \bar{r}} \mathbb{P}(\varepsilon | T = h_1(r,u), R = r).$$

The above equation leads to another interpretation of Lemma 1.2: U can serve as a control function local to the cutoff. After fixing the value of U, the variation in the treatment T becomes locally exogenous. This is because given U and R, the treatment T becomes deterministic. The only variation left in T around the cutoff is due to the discontinuity in the treatment choice function. Lemma 1.2 is related to Lemma 1 in Dong et al. [2021], but unlike in that paper, here we explicitly express the distribution of ε in terms of the observed distribution of (Y,T) and the structural function g^*.

From the IV perspective, we can relate Lemma 1.2 to two results in the literature. First, Lemma 1.2 corresponds to Theorem 1 in Imbens and Newey [2009], where they show that the normalized unobserved heterogeneity in the first stage is a valid control variable. Second, Lemma 1.2 is similar to Theorem 1 in Torgovitsky [2015] in the sense that it provides a (necessary) characterization of the identified set of the structural function.[6]

Lemma 1.2 can be used to invalid a candidate structural function g in the parameter space $\mathscr{G}$. Particularly, if

$$F^{-}_{Y|U,R}(g(h_0(\bar{r},u),\bar{r},e)|u,\bar{r}) \neq F^{+}_{Y|U,R}(g(h_1(\bar{r},u),\bar{r},e)|u,\bar{r}), \text{ for some } e \text{ and } u,$$

then g can not be the true structural function. This means that Condition (1.5) is a necessary condition for the identification of g^*. To make Condition (1.5) also a sufficient condition that uniquely specifies the true g^*, we further introduce some regularity conditions below.

Assumption 1.5.

(i) (Fuzzy RD). The support of $T|R$ from just below and above the cutoff are intervals denoted respectively by Supp$(h_0(\bar{r},U)) = [t_0', t_0'']$ and Supp$(h_1(\bar{r},U)) = [t_1', t_1'']$. The two supports are overlapping: $[t_0', t_0''] \cap [t_1', t_1''] \neq \emptyset$.[7]

(ii) (Strong Discontinuity). Local to the cutoff, the functions h_0 and h_1 intersects and only intersects finitely many times. That is, the following set is nonempty and finite:

$$\{h_0(\bar{r},u) : h_0(\bar{r},u) = h_1(\bar{r},u) \in [t_0', t_0''] \cap [t_1', t_1''], u \in [0,1]\}.$$

Assumption 1.5 imposes restrictions on the nature of the discontinuity. Assumption 1.5(i) requires that the RD design is fuzzy in that there are treatment levels that are taken both below and above the cutoff. Assumption 1.5(ii) imposes restrictions on the strength of the discontinuity. It requires that the conditional quantile functions $h_0(\bar{r},\cdot)$ and $h_1(\bar{r},\cdot)$ only intersects finitely many times. The two curves should intersect but not overlap. If the two functions $h_0(\bar{r},\cdot)$ and $h_1(\bar{r},\cdot)$

[6] The identified set can be defined as the subset of $\mathscr{G}$ that contains the functions g that can generate the observed distribution of (Y,T,R). However, it is rather a detour to formally define such a set because in Section 1.3 we directly use Condition (1.5) for estimation.

[7] Infinite intervals are also allowed. For example, Supp$(h_0(\bar{r},U))$ can be $(-\infty, t_0'']$, $[t_0', \infty)$, or $\mathbb{R}$. We use the notation $[t_0', t_0'']$ to represent all these cases.

overlaps on some interval, then the structural function is not identified on that interval because there is no exogenous variation in the treatment inside that interval.[8] In the extreme case where the two curves completely overlap, there is no discontinuity. Figure 1.4 provides two examples and one counterexample of the strong discontinuity condition.

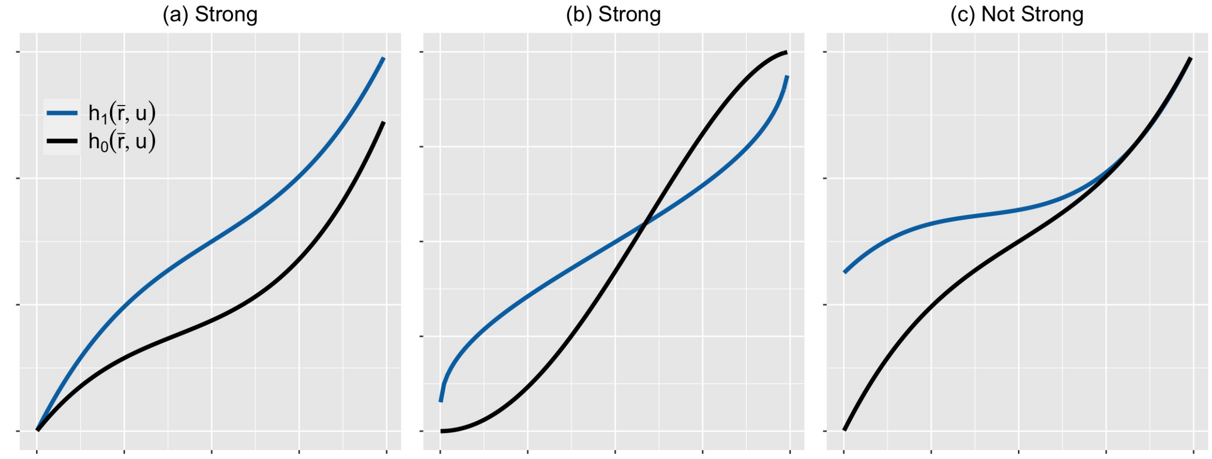

Figure 1.4. Graphical illustration of the strong discontinuity condition.

Under the strong discontinuity condition stated in Assumption 1.5(ii), the two curves $h_0(\bar{r},\cdot)$ and $h_1(\bar{r},\cdot)$ can either follow the stochastic dominance relationship as in (a) or cross each other for multiple times as in (b). Graph (c) is a violation of the strong discontinuity condition, where the two curves completely overlap on an interval.

With the above assumptions, we present the main identification result of the paper. The following theorem shows that Condition (1.5) identifies the true structural function up to a monotone transformation of the error term ε.

Theorem 1.1 (Nonparametric Identification)**.** *Let Assumptions 1.1 - 1.5 hold. If $g \in \mathscr{G}$ satisfies Condition (1.5), then there exists a continuous and strictly increasing function λ^g such that for every $t \in [t_0', t_0''] \cup [t_1', t_1'']$, $e \in \mathscr{E}$, and $u \in [0,1]$, $g^*(t,\bar{r},e) = g(t,\bar{r},\lambda^g(e))$. The specific form of λ^g is given in the proof.*

Remark. *To prove Theorem 1.1, we take any $g \in \mathscr{G}$ that satisfies Condition (1.5). Then we apply the sequencing approach developed in the proof of Theorem 2 in Torgovitsky [2015] to show that g and g^* are related by the transformation λ^g.*

[8] In that case, it is possible to partially identify the structural function.

Theorem 1.1 is the best one can achieve in terms of identifying the nonseparable structural function because the error term is unobserved. Any $g \in \mathscr{G}$ that satisfies Condition (1.5) is equally good as the true g^*. The only difference is that the error term is rescaled by the monotone transformation λ^g. Therefore, such a function g can also be seen as a "version" of g^*.

Inspecting the conditions of Theorem 1.1, we can see that no independence assumption is needed. However, in some theoretical studies of RD designs, a local independence assumption is imposed on the running variable around the cutoff. For example, Assumption A3(i) in Hahn et al. [2001] requires (ε, U) to be jointly independent of R conditioning on R near $\bar{r}$. In the binary treatment case, Dong [2018a] shows that this local independence condition is not needed to achieve identification.

Based on the observation made in Lemma 1.2, we can recover the conditional distribution of ε given U and $R = \bar{r}$. For any $g \in \mathscr{G}$, if g is the true structural function, then the corresponding conditional distribution of ε is

$$F^g_{\varepsilon|U,R}(e|u,\bar{r}) = F^-_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}) = F^+_{Y|T,R}(g(h_1(\bar{r},u),\bar{r},e)|h_1(\bar{r},u),\bar{r}). \quad (1.6)$$

In fact, the above conditional distribution $F^g_{\varepsilon|U,R}$ is a transformed version of the true conditional distribution $F_{\varepsilon|U,R}$, where the transformation is the λ^g defined in Theorem 1.1. This means that the conditional distribution of $\varepsilon|U, R = \bar{r}$ is identified up to the same monotone transformation as the structural function. To eliminate such inconvenience caused by the error term, we can integrate out ε and obtain a unique *conditional average structural function* (CASF):

$$\beta^*(t) = \mathbb{E}[g^*(t,r,\varepsilon)|R = \bar{r}],$$

where the expectation is taken with respect to the true conditional distribution of ε given $R = \bar{r}$. The following corollary summarizes the above analysis.

Corollary 1.1 (CASF). *Let Assumptions 1.1 - 1.5 hold. For any $g^\circ \in \mathscr{G}$ that satisfies Condition*

(1.5), let λ^g be the transformation defined in Theorem 1.1. The following two statements hold true.[9]

(i) For every $e \in \mathscr{E}$ and $u \in [0,1]$, $F^{g^\circ}_{\varepsilon|U,R}(\lambda^{g^\circ}(e)|u,\bar{r}) = F_{\varepsilon|U,R}(e|u,\bar{r})$.

(ii) For any $t \in [t_0', t_0''] \cup [t_1', t_1'']$, the CASF $\beta^*(t,\bar{r})$ is uniquely identified as

$$\beta^*(t) = \int g^\circ(t,\bar{r},e)dF^{g^\circ}_{\varepsilon|U,R}(e|u,\bar{r})du.$$

The CASF $\beta^*(t)$ gives the average outcome the policy-maker can achieve when the treatment level for individuals with characteristic $R = \bar{r}$ is set to t. It is worth noting the difference between the CASF and the *local average structural function* (LASF) commonly seen in the LATE literature. The LASF represents the average outcome for the so-called compliers, an unobservable subpopulation. Therefore, the policy-maker cannot assign treatment to the compliers even when the LASF is identified. On the other hand, the identified CASF can directly guide the treatment assignment to the subpopulation with $R = \bar{r}$. The derivative of the CASF is not the causal effect specific to any subpopulation. Following the spirit of, for example, Heckman and Vytlacil [2001], we may call CASF a policy-relevant parameter.

1.3 Semiparametric estimation

In this section, we consider parametrizations of the structural function that maintain nonlinearity and nonseparability. We propose a semiparametric estimation procedure and derive its large-sample properties. The estimator is semiparametric because the structural function is parametrically specified, while the treatment choice model is left nonparametrically specified.

We do not consider a fully nonparametric estimator since such a procedure can be too data-demanding for practical use, which is especially true for the RD design since the estimation

[9] Since the conditional distribution of ε given $U, R = \bar{r}$ is identified. We can also identify other structural parameters, including the conditional quantile structural function.

is in the local neighborhood of the cutoff.[10]

1.3.1 Construction of the estimator

Consider the following parametrization of $\mathscr{G}$ local to the cutoff.

Assumption 1.6 (Semiparametric Specification)**.** *There is a one-to-one mapping from the class $\{(t,e) \mapsto g(t,\bar{r},e) : g \in \mathscr{G}\}$ of functions to a finite-dimensional parameter space $\Gamma \subset \mathbb{R}^{d_\Gamma}$. We write such parametrization as $\{g_\gamma(\cdot,\bar{r},\cdot) : \gamma \in \Gamma\}$. Assume this parametric model is correctly specified, that is, there exists $\gamma^* \in \Gamma$ such that $g_{\gamma^*}(\cdot,\bar{r},\cdot) = g^*(\cdot,\bar{r},\cdot)$.*

Assumption 1.7 (Normalization of Γ)**.** *For any $\gamma,\gamma' \in \Gamma$, if there exists a transformation λ such that*

$$g_\gamma(\cdot,\bar{r},\cdot) = g_{\gamma'}(\cdot,\bar{r},\lambda(\cdot)),$$

then $\gamma = \gamma'$ and λ is the identity transformation.

Assumption 1.7 is a normalization condition that fixes the scale of the error term ε. An example is provided below to illustrate the parametrization of the structural function. One way to achieve such normalization is to have some treatment value $\tilde{t}$ such that $g_\gamma(\tilde{t},\bar{r},e) = e,$ for all $\gamma \in \Gamma$.

Example 1.1. *Let $\tilde{t} \in [t_0',t_0''] \cap [t_1',t_1'']$. We can specify the structural function by*

$$g_\gamma(T,\bar{r},\varepsilon) = \gamma_1(T-\tilde{t}) + \gamma_2(T-\tilde{t})^2 + \gamma_3(T-\tilde{t})\varepsilon + \varepsilon.$$

The parameter $\gamma = (\gamma_1,\gamma_2,\gamma_3)$ is three-dimensional. The function $g_\gamma(T,\bar{r},\varepsilon)$ is strictly increasing in ε for all γ satisfying $\mathbb{P}(\gamma_3(T-\tilde{t})+1>0)=1$. The parametrization satisfies Assumption 1.7

[10] From the theoretical perspective, it can be challenging to construct a fully nonparametric estimator. If we follow the sieve approach, for example, we would need to consider a basis of functions that are strictly increasing in one of the arguments to accommodate the monotonicity of the structural function, which is a non-trivial task.

because, by construction, $g_\gamma(\tilde{t},\bar{r},e) = e$ for all values of γ and e.[11] *The model is quadratic in T and nonseparable between T and ε. The effect of T on Y is allowed to be nonlinear and contain unobserved heterogeneity. The distribution of ε is not parametrized and thus can be very general.*

The true parameter γ^* in the normalized semiparametric model can be identified as follows. Hereafter, we use $h^* = (h_0^*, h_1^*)$ to signify the true conditional quantile functions and $h = (h_0, h_1)$ a generic pair of conditional quantile functions. Let $w(e,u)$ be a weighting function defined on $\mathbb{R} \times [0,1]$. Define the criterion function as

$$\left\| D_{\gamma,h} \right\|_w = \left(\int_0^1 \int_{\mathbb{R}} |D_{\gamma,h}(e,u)|^2 w(e,u) dedu \right)^{1/2}, \tag{1.7}$$

where $D_{\gamma,h}(e,u)$ is defined to be

$$\int_0^u \left(F^-_{Y|T,R}(g_\gamma(h_0(\bar{r},v),\bar{r},e)|h_0(\bar{r},v),\bar{r}) - F^+_{Y|T,R}(g_\gamma(h_1(\bar{r},v),\bar{r},e)|h_1(\bar{r},v),\bar{r}) \right) dv. \tag{1.8}$$

This criterion function is based on Equation (1.5), which by Lemma 1.2 is a necessary characterization of the identified set. We take an integral form of Condition (1.5) because it gives a faster convergence rate of the resulting estimator.

Assumption 1.8 (Weighting function w)**.** *The function w is nonnegative, integrates to one, and is bounded on $\mathbb{R} \times [0,1]$. The support of w contains $\mathscr{E} \times [0,1]$. The criterion function $\left\| D_{\gamma,h} \right\|_w$ defined in (1.7) is finite.*

In practice, we may use a weighting function w that is supported on the entire domain $\mathbb{R} \times [0,1]$ since $\mathscr{E}$ is unknown. For the numerical results in Section 1.4, we use the normal density function as the weighting w. The following corollary provides the semiparametric identification result, which is based on the nonparametric identification result in Section 1.2. It shows that the criterion function, when evaluated at the true nuisance parameter value h^*, is uniquely minimized by the true γ^*.

[11] This normalization strategy is presented in Equation (2.5) of Matzkin [2003].

Corollary 1.2 (Semiparametric Identification). *Let Assumptions 1.1 - 1.8 hold. For any* $\gamma \in \Gamma$ *such that* $\gamma \neq \gamma^*$, *we have* $\left\|D_{\gamma,h^*}\right\|_w > \left\|D_{\gamma^*,h^*}\right\|_w = 0$. *In other words,* γ^* *is the unique minimizer of* $\left\|D_{\gamma^*,h^*}\right\|_w$.

Assume there is an *independent and identically distributed* (iid) sample $(Y_i, T_i, R_i)_{i=1}^n$ available. We propose an estimation procedure based on the above semiparametric identification result. The idea is that we first estimate the nonparametric components (h_0, h_1) and $(F^-_{Y|T,R}, F^+_{Y|T,R})$ that appear in the criterion function. Then we construct an empirical version of the criterion function and take its minimizer to be the estimator.

The estimation procedure of γ is more specifically divided into three steps. The first step is to estimate the conditional quantile functions h_0 and h_1. The second step uses *local linear regression* (LLR) to estimate the conditional distributions $F^-_{Y|T,R}$ and $F^+_{Y|T,R}$. It is standard to use local polynomials in the estimation of RD designs [Porter, 2003, Sun, 2005]. The challenge here is that, unlike in the classical RD methods where local polynomial is used to estimate the conditional expectation function of Y given R, here we estimate the conditional distribution of Y.[12] The third step constructs an estimate of the criterion function by replacing the nonparametric nuisance parameters in (1.8) by their estimated counterparts and then finds the estimate of γ^* by minimizing the estimated criterion function. We describe the detail of the estimation procedure as follows. Denote $\mathscr{Y}$ as the range of the outcome Y.

- **STEP 1.** Estimate the conditional quantile of T given R. Choose estimators $\hat{h}_0(\bar{r}, \cdot)$ and $\hat{h}_1(\bar{r}, \cdot)$ of the corresponding conditional quantile processes, $h_0(\bar{r}, \cdot)$ and $h_1(\bar{r}, \cdot)$. Specific constructions are provided in Section 1.3.3.

- **STEP 2.** Estimate the conditional distribution of Y given T, R. Choose two bandwidth sequences $b_1 = b_{1n}$ and $b_2 = b_{2n}$ and three kernel functions k_Y, k_T, and k_R. Define $K_Y(y) = \int_{-\infty}^{y} k_Y(\tilde{y}) d\tilde{y}$. For each $y \in \mathscr{Y}$ and $t \in [t_0', t_0'']$, solve the following minimization

[12] Local linear estimation of the conditional distribution function can be found in Hansen [2004] and Chapter 2 of this dissertation.

problem:

$$\min_{a^-,a_T^-,a_R^-} \sum_{i:R_i<\bar{r}} \left(K_Y\left(\frac{y-Y_i}{b_2}\right) - a^- - a_T^-(T_i-t) - a_R^-(R_i-\bar{r}) \right)^2 \times k_T\left(\frac{T_i-t}{b_1}\right) k_R\left(\frac{R_i-\bar{r}}{b_1}\right).$$

The minimizer $\hat{a}^-$ is the estimate $\hat{F}^-_{Y|T,R}(y|t,\bar{r})$. For each $y \in \mathscr{Y}$ and $t \in [t_1', t_1'']$, solve the following minimization problem:

$$\min_{a^+,a_T^+,a_R^+} \sum_{i:R_i\geq\bar{r}} \left(K_Y\left(\frac{y-Y_i}{b_2}\right) - a^+ - a_T^+(T_i-t) - a_R^+(R_i-\bar{r}) \right)^2 \times k_T\left(\frac{T_i-t}{b_1}\right) k_R\left(\frac{R_i-\bar{r}}{b_1}\right).$$

The minimizer $\hat{a}^+$ is the estimate $\hat{F}^+_{Y|T,R}(y|t,\bar{r})$.

- **STEP 3.** Construct the empirical version of the criterion function:

$$\left\|\hat{D}_{\gamma,\hat{h}}\right\|_w = \left(\int_0^1 \int_{\mathbb{R}} |\hat{D}_{\gamma,\hat{h}}(e,u)|^2 w(e,u) dedu \right)^{1/2},$$

where $\hat{D}_{\gamma,\hat{h}}(e,u)$ is defined to be

$$\int_0^u \left(\hat{F}^-_{Y|T,R}(g_\gamma(\hat{h}_0(\bar{r},v),\bar{r},e)|\hat{h}_0(\bar{r},v),\bar{r}) - \hat{F}^+_{Y|T,R}(g_\gamma(\hat{h}_1(\bar{r},v),\bar{r},e)|\hat{h}_1(\bar{r},v),\bar{r}) \right) dv.$$

The estimator $\hat{\gamma}$ is any parameter value in Γ that satisfies

$$\left\|\hat{D}_{\hat{\gamma},h}\right\|_w \leq \inf_{\gamma\in\Gamma} \left\|\hat{D}_{\gamma,h}\right\|_w + O_p(\alpha_n), \tag{1.9}$$

where $\alpha_n \to 0$ is a deterministic sequence specified by Equation (A.3) in Appendix A.2.

1.3.2 Asymptotic normality

In order to derive the asymptotic distribution of the semiparametric estimator, we consider the regularity assumptions listed below.

Assumption 1.9 (Distributions of Y, T, and R)**.**

(i) The support of T does not vary with R except when crossing the cutoff $\bar{r}$, i.e., Supp$(T|R = r) = [t'_0, t''_0]$ for $r < \bar{r}$ and Supp$(T|R = r) = [t'_1, t''_1]$ for $r > \bar{r}$. The density functions $f^-_{T,R}$ and $f^+_{T,R}$ are bound away from zero.

(ii) The density functions $f^-_{T,R}$ and $f^+_{T,R}$ are twice continuously differentiable, and $\frac{\partial^2}{\partial t^2} f^-_{T,R}(t, \bar{r})$ and $\frac{\partial^2}{\partial t^2} f^+_{T,R}(t, \bar{r})$ are Lipschitz continuous with respect to t.

(iii) The support of Y, $\mathscr{Y}$, is compact. The conditional distribution functions $F^-_{Y|T,R}$ and $F^+_{Y|T,R}$ are three-times continuously differentiable over $\mathscr{Y} \times [t'_0, t''_0] \times [r_0, \bar{r}]$ and $\mathscr{Y} \times [t'_1, t''_1] \times [\bar{r}, r_1]$, respectively.

Assumption 1.10 (Complexity of the Parametric Model)**.** *The parametrization $\{g_\gamma(\cdot, \bar{r}, \cdot) : \gamma \in \Gamma\}$ satisfies the following conditions.*

(i) The parameter space Γ is compact.

(ii) The class of functions $\{T \mapsto g_\gamma(T + v, \bar{r}, e) : \gamma \in \Gamma, v \in (-1, 1), e \in \mathscr{E}\}$ is finite-dimensional.

(iii) The function $g_\gamma(t, \bar{r}, e)$ is twice continuously differentiable over $\gamma \in \Gamma$, $t \in [t'_0, t''_0] \cup [t'_1, t''_1]$, and $e \in \mathscr{E}$.

(iv) The gradient $\nabla_\gamma D_{\gamma^*, h^*}(e, u)$ is a vector of linearly independent functions of (e, u).

Assumption 1.11 (Kernels)**.**

(i) The kernel functions k_T and k_R are (1) supported on $[-1, 1]$, (2) strictly greater than zero in the interior of the support, (3) of bounded variation, (4) continuously differentiable on $\mathbb{R}$.

(ii) *The kernel function k_Y is (1) nonnegative and (2) integrable on $\mathbb{R}$ with $\int k_Y(y)dy = 1$ and satisfies (3) $\int yk_Y(y)dy = 0$.*

Assumption 1.12 (Bandwidth)**.** *The bandwidth b_1 and b_2 satisfy the following conditions:*

(i) $b_1 \asymp b_2$.[13]

(ii) $(n \log n)b_1^6 = o(1)$.

(iii) $nb_1^{\frac{13}{3}+\varepsilon} \to \infty$, *for some sufficiently small $\varepsilon > 0$.*

Assumption 1.13 (First-step Conditional Quantile Estimators)**.** *The estimators $\hat{h}_0$ and $\hat{h}_1$ satisfy the following conditions.*

(i) *Monotonicity and smoothness: for every n sufficiently large, there exist $C > 0$ and deterministic and finite partitions $\mathscr{P}_0^n$ and $\mathscr{P}_1^n$ on $(0,1)$ such that*

$$\mathbb{P}\left(\hat{h}_0(\bar{r},\cdot) \notin \mathscr{H}_0(\mathscr{P}_0^n)\right), \mathbb{P}\left(\hat{h}_1(\bar{r},\cdot) \notin \mathscr{H}_1(\mathscr{P}_1^n)\right) = O\left(\sqrt{b_1}\right),$$

where

$$\begin{aligned}\mathscr{H}_0(\mathscr{P}_0^n) = \{ & \text{function } h \text{ from } [0,1] \text{ into } [t_0', t_0''] : \text{on each element of } \mathscr{P}_0^n, h \text{ is strictly} \\ & \text{increasing, its inverse } h^{-1} \text{ is three-times continuously differentiable,} \\ & \text{and } (h^{-1})^{(3)} \text{ is Lipschitz continuous } \},\end{aligned}$$

and $\mathscr{H}_1(\mathscr{P}_1^n)$ is defined analogously by replacing $\mathscr{P}_0^n$ with $\mathscr{P}_1^n$.

[13] The notation $b_1 \asymp b_2$ means that there exists $C > 1$ such that $b_1/b_2 \in [1/C, C]$.

(ii) Uniform Bahadur representation:

$$\begin{aligned}\hat{h}_0(\bar{r},u)-h_0^*(\bar{r},u) &= b_1^2\nu_0(u)+O_p(b_1^3)\\ &\quad+\frac{1}{nb_1}\sum_{i=1}^{n}q_0(T_i,R_i;u)k_{Q,0}\left(\frac{R_i-\bar{r}}{b_1}\right)\mathbf{1}\{R_i<\bar{r}\}+o_p\left(1/\sqrt{nb_1}\right),\\ \hat{h}_1(\bar{r},u)-h_1^*(\bar{r},u) &= b_1^2\nu_1(u)+O_p(b_1^3)\\ &\quad+\frac{1}{nb_1}\sum_{i=1}^{n}q_1(T_i,R_i;u)k_{Q,1}\left(\frac{R_i-\bar{r}}{b_1}\right)\mathbf{1}\{R_i\geq\bar{r}\}+o_p\left(1/\sqrt{nb_1}\right),\end{aligned}$$

uniformly over $u\in(0,1)$. The functions ν_0 and ν_1 are bounded. The functions q_0 and q_1 are (1) bounded, (2) centered, that is, $\mathbb{E}[q_0(T,R;u)|T,R]=\mathbb{E}[q_1(T,R;u)|T,R]=0$, and (3) does not vary with n. The functions $k_{Q,0}$ and $k_{Q,1}$ are bounded.

(iii) Uniform convergence rate:

$$\begin{aligned}\|\hat{h}-h^*\|_\infty &= \sup_{u\in(0,1)}|\hat{h}_0(\bar{r},u)-h_0^*(\bar{r},u)|\vee|\hat{h}_1(\bar{r},u)-h_1^*(\bar{r},u)|\\ &= O_p\left(\sqrt{\log n/(nb_1)}+b_1^2\right).\end{aligned}$$

A brief discussion of the assumptions is in order. Assumption 1.9 imposes smoothness restrictions on the joint distribution of (Y,T,R). In the previous section, the identification result only requires continuity of the relevant functions. For estimation, we need higher-order smoothness regarding the distribution functions. Assumption 1.10 imposes restrictions on the parametric model of the structural function. Part (ii) restricts the complexity of the model. Part (iii) imposes high-order smoothness on the structural function. Part (iv) is similar to Assumption D4 in Torgovitsky [2017] and requires that $\nabla_\gamma D_{\gamma^*,h^*}$ to carry information about each component of the parameter.

Assumption 1.11 imposes restrictions on the kernel functions k_T, k_R, and k_Y. The differentiability is needed to prove a stochastic equicontinuity condition. Assumption 1.12

restricts that b_1 and b_2 are of the same asymptotic order, which is slightly faster than $n^{-1/6}$ and slightly slower than $n^{-3/13}$. This assumption is not restrictive and allows for the asymptotic mean squared error (AMSE) optimal bandwidth as well as undersmoothing.

Assumption 1.13 imposes high-level restrictions on the first-stage nonparametric conditional quantile estimators. Part (i) assumes that the quantile estimators are piece-wise monotonic and smooth with a high probability. Part (ii) and (iii) give the uniform Bahadur representation and the uniform convergence rate, which are fairly standard in the quantile estimation literature. In Section 1.3.3, we discuss a specific nonparametric quantile estimator that satisfies Assumption 1.13.

Theorem 1.2 (Asymptotic Distribution of the Semiparametric Estimator)**.** *Let Assumptions 1.1 - 1.13 hold. Then* $\|\hat{\gamma}-\gamma^*\|_2 = O_p(b_1^2 + 1/\sqrt{nb_1})$ *and*

$$\left(\sqrt{nb_1}(\Sigma_- + \Sigma_+)^{-1/2}\right)(\Delta(\hat{\gamma}-\gamma^*) - b_1^2(B_- - B_+)) \xrightarrow{d} N(0, \boldsymbol{I}_{d_\Gamma}),$$

where I_{d_Γ} is the d_Γ-dimensional identity matrix. The exact forms of Δ, B_-, B_+, Σ_- and Σ_+ are given in Equations (A.11), (A.12), (A.13), (A.14) and (A.15) in Appendix A.2, respectively.

Remark. *The convergence rate of $\hat{\gamma}$ is $b_1^2 + 1/\sqrt{nb_1}$, which is equal to $n^{-2/5}$ when $b_1 \asymp n^{-1/5}$. This rate is the same as the one obtained in the classical RD design with a binary treatment variable [Hahn et al., 2001]. Having a continuous treatment does not slow down the convergence rate of the estimator in this case. This is due to the integral smoothing in the definition of the criterion function $D_{\gamma,h}$ in (1.8).*

Remark. *The proof of Theorem 1.2 follows the general steps of proving asymptotic normality of semiparametric estimators as in, for example, Torgovitsky [2017] and Chen et al. [2003]. The main difficulty is that the usual stochastic equicontinuity condition is not sharp because the criterion function is nonparametrically estimated. More specifically, the consistency of $\hat{\gamma}$ by itself is not enough to eliminate the estimation errors asymptotically. To overcome this issue,* *we first use the empirical process theory to derive a uniform convergence rate for the estimated criterion function, which gives an initial bound on the convergence rate of $\hat{\gamma}$. A sharper stochastic equicontinuity result is then derived based on this initial bound together with more applications of the empirical process theory. This sharper stochastic equicontinuity result helps demonstrate that the usual linearization of the criterion function is valid. See the proof in Appendix A.2 for details.*

Once the asymptotic normal distribution of the estimator $\hat{\gamma}$ is established, we can conduct inference for γ^*. Theorem 1.2 together with the undersmoothing condition that $nb_1^5 = o(1)$ gives that

$$\sqrt{nb_1}(\hat{\gamma} - \gamma^*) \xrightarrow{d} N(0, \Delta^{-1}(\Sigma_- + \Sigma_+)\Delta^{-1}).$$

A linear null hypothesis regarding γ can be written as $H\gamma = \eta$, where $\eta \in \mathbb{R}^{d_\eta}$ and H is a $d_\eta \times d_\gamma$ full-rank matrix. Consider the test statistic

$$nb_1(\hat{\gamma} - \gamma^*)' \left(H\hat{\Delta}^{-1}(\hat{\Sigma}_- + \hat{\Sigma}_+)\hat{\Delta}^{-1}H'\right)^{-1}(\hat{\gamma} - \gamma^*),$$

where $\hat{\Delta}$, $\hat{\Sigma}_-$, and $\hat{\Sigma}_+$ are consistent estimators of Δ, Σ_-, and Σ_+, respectively. By Slutsky's theorem, the above test statistic converges in distribution to the χ^2 distribution with d_η degrees of freedom. In Appendix A.2, we discuss how to construct consistent estimators for Δ, Σ_-, and Σ_+.

1.3.3 First-step nonparametric quantile estimators

This section discusses how to construct nonparametric conditional quantile estimators that satisfy Assumption 1.13. Consider the following two-step estimation procedure introduced by Qu and Yoon [2015]. Define $\rho_u(t) = t(u - \mathbf{1}\{t < 0\})$.

- **STEP 1.** Choose a bandwidth sequence $b_3 = b_{3n} = o(1)$ and a kernel function k_{FS}.

Partition the unit interval $(0,1)$ into a grid of equally spaced points $\{u_1, \cdots, u_{J_n}\}$, where $J_n/(nb_3)^{1/4} \to \infty$. Solve the following optimization problem:

$$\min_{\{\mathrm{h}_j, \mathrm{h}'_j\}_{j=1}^{J_n}} \sum_{j=1}^{J_n} \sum_{i=1}^{n} \rho_{u_j} \left(T_i - \mathrm{h}_j - \mathrm{h}'_j (R_i - \bar{r}) \right) k_{FS} \left(\frac{R_i - \bar{r}}{b_3} \right) \mathbf{1}\{R_i < \bar{r}\}. \tag{1.10}$$

Denote the minimizers by $(\hat{h}_0(\bar{r}, u_1), \cdots, \hat{h}_0(\bar{r}, u_{J_n}))$.

- **STEP 2.** Let $u_0 = 0$ and $u_{J_{n+1}} = 1$. Let $\hat{h}_0(\bar{r}, u_0) = \min_{i:R_i<\bar{r}} T_i$ and $\hat{h}_0(\bar{r}, u_{J_n+1}) = \max_{i:R_i<\bar{r}} T_i$. Linearly interpolate between the estimates to obtain an estimate for the entire quantile process. That is, for any $u \in (u_j, u_{j+1})$, define

$$\hat{h}_0(\bar{r}, u) = \frac{u_{j+1} - u}{u_{j+1} - u_j} \hat{h}_0(\bar{r}, u_j) + \frac{u - u_j}{u_{j+1} - u_j} \hat{h}_0(\bar{r}, u_{j+1}).$$

The estimator $\hat{h}_1(\bar{r}, \cdot)$ can be analogously defined by using the data with $R_i \geq \bar{r}$.[14] We can verify Assumption 1.13 for the estimator constructed above. Denote

$$\Omega_{Q,0} = \int (1,x)(1,x)' \mathbf{1}\{x < 0\} k_{FS}(x) dx,$$
$$\Omega_{Q,1} = \int (1,x)(1,x)' \mathbf{1}\{x \geq 0\} k_{FS}(x) dx.$$

Proposition 1.1. *Let Assumptions 1.2(iv) and 1.9(i)-(ii) hold. Assume that the third-order derivatives* $\frac{\partial^3}{\partial R^3} h_0^*(r,u)$ *and* $\frac{\partial^3}{\partial R^3} h_1^*(r,u)$ *are Lipschitz continuous respectively on* $[r_0, \bar{r}] \times [0,1]$ *and* $[\bar{r}, r_1] \times [0,1]$. *Assume that the bandwidth* $b_3 = cb_1$ *for some constant* $c > 0$. *Assume that the kernel* k_{FS} *is nonnegative, of bounded variation, compactly supported, having finite first-order derivatives and satisfying*

$$\int k_{FS}(x) dx = 1, \int x k_{FS}(x) dx = 0, \int x^2 k_{FS}(x) dx < \infty.$$

[14] There are three estimators of conditional quantile process in Qu and Yoon [2015]. The estimator explained here is their second one, denoted by $\hat{\alpha}^*$ in that paper. Their third estimator imposes a monotonicity constrain to the minimization problem (1.10).

Then the estimators $\hat{h}_0(\bar{r},\cdot)$ and $\hat{h}_1(\bar{r},\cdot)$ described above satisfy Assumption 1.13. The specific forms of ν_0 and ν_1 are

$$\nu_0(u) = \frac{c^2}{2}\frac{\partial^2}{\partial r^2}h_0^*(\bar{r},u)\iota'\Omega_{Q,0}^{-1}\int x^2(1,x)'k_{FS}(x)\mathbf{1}\{x<0\}dx,$$
$$\nu_1(u) = \frac{c^2}{2}\frac{\partial^2}{\partial r^2}h_1^*(\bar{r},u)\iota'\Omega_{Q,1}^{-1}\int x^2(1,x)'k_{FS}(x)\mathbf{1}\{x\geq 0\}dx.$$

The specific forms of $k_{Q,0}$ and $k_{Q,1}$ are

$$k_{Q,j}(x) = \iota'\Omega_{Q,j}^{-1}(1,x/c)'k_{FS}(x/c)/c, j=0,1.$$

The functions q_0 and q_1 are

$$q_0(T,R;u) = (u-\mathbf{1}\{T\leq h_0^*(\bar{r},u)\})/(f_R(\bar{r})f_{T|R}^-(h_0^*(\bar{r},u)|\bar{r})),$$
$$q_1(T,R;u) = (u-\mathbf{1}\{T\leq h_1^*(\bar{r},u)\})/(f_R(\bar{r})f_{T|R}^+(h_1^*(\bar{r},u)|\bar{r})),$$

which take the form of an influence function for quantiles.

Other quantile estimation methods are also available. For example, one can consider the generic framework proposed by Chernozhukov et al. [2010] for rearrangement. In particular, they show that the rearrangement of a preliminary estimated quantile process delivers a monotonic estimator that preserves the asymptotic properties. This result gives a different way to generate estimators that satisfy Assumption 1.13. We can start with an estimator with desired asymptotic properties that give rise to Assumption 1.13(ii) and (iii), and then apply the rearrangement procedure. The resulting estimator would be monotonic on the entire domain, and partitioning is unnecessary.

1.4 Numerical results

This section presents the empirical study and the simulation results. The empirical study shows that the semiparametric estimator is considerably better than the simple TSLS estimator in discovering quantitative information regarding the structural function. The simulation studies show that the semiparametric procedure can accurately estimate the parameters with moderate sample size.[15]

1.4.1 Empirical study

In this empirical study, we examine the causal effect of sleep time on health status by exploiting the discontinuity in the timing of natural light at time zone boundaries. The unit of observation is the individual in the American Time Use Survey (ATUS). The outcome Y is the individual's health status measured by the body-mass index (BMI).[16] The treatment T is the sleep time. The running variable R is the longitudinal distance to the nearest time zone boundary, with cutoff $\bar{r} = 0$ denoting the time zone border. As explained in the introduction, the identification is based on the exogenous variation in the sleep time around the time zone boundary. This exogenous variation is due to the difference in the timing of natural light on each side of the time zone boundary.

Many studies in the medical literature examine the effect of sleep time on overweight issues. See Beccuti and Pannain [2011] and the references therein. These studies typically use survey or laboratory data. This problem is first studied by using the RD design in Giuntella and Mazzonna [2019].[17] The relevant outcome variable they use is a binary indicator of the obesity (or overweight) status indicating whether the BMI is above some threshold. They use the TSLS procedure to estimate the Wald ratio across the time zone boundary. We consider two

[15] Replication files for the empirical and simulation studies are available from the author upon request.

[16] BMI is a person's weight in kilograms divided by the square of height in meters. The Centers for Disease Control and Prevention define overweight as BMI ¿ 25 and obesity as BMI ¿ 30.

[17] Giuntella and Mazzonna [2019] study many health and economics-related issues. Here we only mention the relevant ones.

improvements based on their work. First, we directly use BMI as the outcome variable, providing a more quantitative measure of the health status. Second, we use the proposed semiparametric estimator to estimate a nonlinear structural function. Previous medical studies have provided evidence of the nonlinearity of the structural function. For example, Hairston et al. [2010] show that both undersleeping and oversleeping lead to an increase in BMI while sleeping around 8 hours leads to a more healthy BMI level.

The data for this empirical study is collected from IPUMS CPS [Flood et al., 2020] and IPUMS ATUS [Hofferth et al., 2020] during the periods 2006 - 2008 and 2014 - 2016. By linking these datasets, we can locate the county where the individual lives and then use the county's centroid as the location of the individual. We focus on counties near the time zone boundary between the Eastern and Central time zone. The counties are divided into two regions based on their latitude. We estimate the model separately for each region.

The estimated marginal effects of sleep on BMI from the semiparametric estimator and the TSLS estimator are shown in Figure 1.5. Several interesting findings are observed based on the semiparametric estimates. First, the marginal effects are increasing and increase from negative to positive. This lends some support to the previous argument that the structural function is nonlinear and neither sleeping too little nor too much is preferable. Second, we can determine the optimal (in terms of BMI) sleep time by finding the zero of the marginal effect curve. In both cases, the optimal sleep time is estimated to be between 7 and 8 hours, which also aligns with the findings in previous medical studies. Third, the results from the two regions are similar, meaning that the variation across different latitudes is small.

From Figure 1.5, we can also see that the TSLS estimates are not capable of demonstrating the above results. First, the TSLS procedure only provides a constant estimate of the marginal effect across all levels of sleep time. This means an extra hour of sleep would lead to the same effect on health regardless of the person's current sleep time, which is inappropriate in this setting. Moreover, we cannot estimate the optimal sleep time based on the linear structural function. Second, the magnitude of the TSLS estimates is small. This is because the TSLS

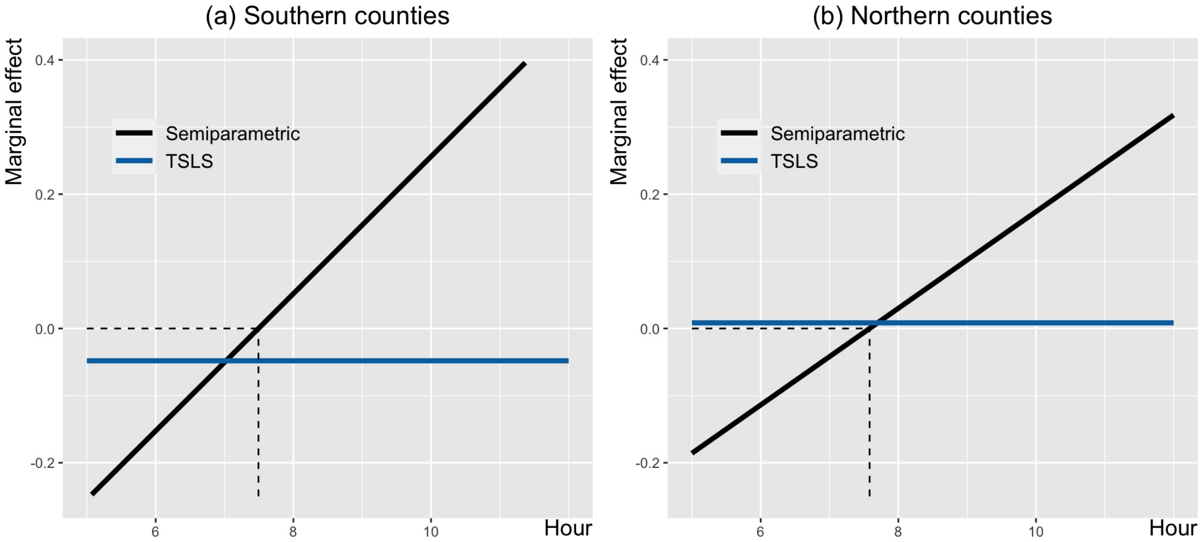

Figure 1.5. Estimated marginal effects of sleep time on BMI (kg/m^2).

The plots show the estimated marginal effects based on the semiparametric and the TSLS estimator. Graph (a) is computed based on counties with latitude ¡ 37. Graph (a) is computed based on counties with latitude ¿ 37. We can see that the marginal effects are increasing, indicating a nonlinear (U-shaped) structural function. The optimal sleep time computed as the zero of the marginal effect curve is between 7 and 8 hours. However, the marginal effect estimated from the TSLS procedure is constant across different sleep times. These estimates are small in magnitude and less informative for the researcher.

provides a weighted average of the marginal effects across the entire range of sleep time. By averaging the negative and positive effects, the TSLS delivers an estimate attenuated toward zero, which is not informative for the researcher.[18]

1.4.2 Simulations

We use simulation studies to investigate the performance of the proposed semiparametric method and compare it with the performance of the TSLS estimator. The data generating process (DGP) for these simulations was chosen to roughly approximate the ATUS data used in the empirical study. Let marginal distributions of U and ε be given by $F_U = Unif(0,1)$ and $F_\varepsilon = Beta(2,2)$, respectively. Two marginal distributions for R are considered, $Unif(0,1)$ and $N(0,1)$. The joint distribution of (R,ε,U) be characterized by the Gaussian copula with

[18] Giuntella and Mazzonna [2019] find a more significant effect of sleep time on obesity. There are two possible reasons: they consider the binary indicator of obesity, and they include more control variables in the regression.

correlation structure $corr(R,U)=0$, $corr(\varepsilon,R)=\rho_R$, and $corr(\varepsilon,U)=\rho_U$. The treatment choice model is given by $h_0^*(r,u)=r+2\sin(\pi u/2)$ and $h_1^*(r,u)=r+2u^3$. The structural function is given by

$$g_\gamma(T,R,\varepsilon)=\gamma_1(T-0.5)+\gamma_2(T^2-0.5^2)+\gamma_3(T-0.5)\varepsilon+\varepsilon+R.$$

The true γ^* is taken to be $(1,1,1)$.

Further implementation details are described below. Construct a kernel function k that is an even function given by

$$k(x)=\begin{cases}2x^3-3x^2+1, & \text{if } x\in[0,1],\\ 0, & \text{if } x>1.\end{cases}$$

We can verify that k is continuously differentiable on the real line and compactly supported on $[0,1]$. Within the interior of its support, k is strictly positive. We use this function k to be the kernels k_T, k_R, k_Y, and k_{FS} in the estimation. The bandwidth is chosen to be $b_1=b_2=b_3=2n^{-1/5}$. The weighting function $w(e,u)$ is chosen to be constant in u and equal to the standard normal density function with respect to e.

Table 1.1 contains the simulation results of the performance of the semiparametric estimator for different choices of the marginal distribution of R, the correlation parameters (ρ_U,ρ_R) and the sample size. In each case, the number of replications is set at 500. We can see that the estimator performs well with a moderate sample size ($n=1000$). The marginal distribution of R does not have a large impact on the performance. When the sample size is small, larger values of ρ_R or ρ_U can lead to poorer performance of the estimator. The plausible reason is that larger values of the correlation parameters would lead to more severe endogeneity issues in finite samples. When the sample size becomes large, the performance of the estimator does not vary significantly with the choices of (ρ_R,ρ_U).

Table 1.1. Performance of the semiparametric estimator.

Dist. R	ρ_U	ρ_R	Param	$n=500$			$n=1000$			$n=1500$		
				bias	sd	mse	bias	sd	mse	bias	sd	mse
$U(0,1)$	.3	.3	γ_1	-.257	.206	.108	-.233	.134	.072	-.219	.119	.062
			γ_2	.038	.153	.025	.062	.098	.013	.068	.081	.011
			γ_3	.172	.139	.049	.158	.091	.033	.154	.073	.029
		.5	γ_1	-.260	.464	.283	-.231	.134	.071	-.215	.116	.060
			γ_2	.027	.514	.265	.053	.097	.012	.060	.080	.010
			γ_3	.178	.168	.060	.159	.097	.035	.155	.076	.030
	.5	.3	γ_1	-.256	.199	.105	-.225	.129	.067	-.210	.111	.056
			γ_2	.039	.157	.026	.061	.098	.013	.068	.078	.011
			γ_3	.189	.140	.055	.171	.086	.037	.163	.070	.031
		.5	γ_1	-.248	.456	.269	-.217	.125	.063	-.202	.104	.052
			γ_2	.015	.512	.262	.044	.095	.011	.052	.075	.008
			γ_3	.222	.419	.225	.178	.100	.042	.171	.076	.035
$N(0,1)$	.3	.3	γ_1	-.224	.479	.280	-.228	.156	.076	-.216	.136	.065
			γ_2	.006	.527	.278	.059	.114	.016	.067	.093	.013
			γ_3	.152	.170	.052	.152	.105	.034	.148	.085	.029
		.5	γ_1	-.254	.362	.196	-.224	.153	.074	-.212	.124	.060
			γ_2	.030	.268	.073	.048	.112	.015	.058	.091	.012
			γ_3	.167	.178	.060	.152	.113	.036	.149	.086	.030
	.5	.3	γ_1	-.243	.233	.113	-.221	.149	.071	-.209	.127	.060
			γ_2	.027	.185	.035	.059	.113	.016	.069	.090	.013
			γ_3	.176	.172	.061	.163	.102	.037	.157	.082	.031
		.5	γ_1	-.230	.500	.303	-.210	.144	.065	-.198	.118	.053
			γ_2	-.012	.619	.383	.037	.113	.014	.049	.086	.010
			γ_3	.193	.248	.099	.172	.116	.043	.167	.088	.036

The structural function follows the three-parameter specification in Example 1.1. The number of replications is 500. The marginal distribution of R is chosen to be the uniform distribution on $[0,1]$ or the standard normal distribution. The two correlation parameters $\rho_R = \text{corr}(\varepsilon, R)$ and $\rho_U = \text{corr}(\varepsilon, U)$ are chosen from $\{0.3, 0.5\}$. The results demonstrate the following points. First, the semiparametric estimator performs well with a moderate sample size of 1000. Second, the performance of the estimator is not affected by The marginal distribution of R. Third, when the sample size is as large as 1000, the performance of the estimator does not vary significantly with the choices of (ρ_R, ρ_U).

It is also of interest to compare the semiparametric estimator with the TSLS estimator. Directly comparing the two estimators can be difficult since they are of different dimensions and converge to different limits. Instead, we can compare their performance in estimating the marginal effect. For the structural function $g_\gamma(t,\bar{r},e) = \gamma_1 t + \gamma_2 t^2 + \gamma_3 te + e$, the marginal effect of the treatment on the outcome is $\frac{\partial}{\partial t} g_\gamma(t,\bar{r},e) = \gamma_1 + 2\gamma_2 t + \gamma_3 e$, which takes on different values for different treatment and outcome levels. For a given treatment level t, we can use the semiparametric estimator to obtain an estimate $\hat{\gamma}_1 + 2\hat{\gamma}_2 t + \hat{\gamma}_3 e$ of the marginal effect. However, a TSLS procedure would deliver a scalar estimate that is a mixture of marginal effects across different treatment and outcome levels. Figure 1.6 shows that with a nonlinear specification, the semiparametric estimator outperforms the TSLS estimator. We compare the two estimates of the marginal effect at four quantile levels of the treatment: 20%, 40%, 60%, and 80%. The simulated distribution of the semiparametric estimator is correctly centered, while the TSLS estimator incurs a large bias. Figure 1.7 shows similar findings with a fully nonlinear and nonseparable specification, where we compare the two estimates at four levels of e: 0.2,0.4,0.6,0.8.

Next, we compare the semiparametric estimator with the TSLS estimator when the structural function g is linear. This is achieved by imposing $\gamma_2 = \gamma_3 = 0$. The remaining slope coefficient γ_1 is equal to the marginal effect. In this case, the TSLS estimator is consistent for the coefficient γ_1. However, the identification of the TSLS estimator is based solely on the difference between the two means. If the two distributions corresponding to h_0 and h_1 have the same mean, then the TSLS procedure suffers from weak identification issues. In contrast, the identification of the semiparametric estimator $\hat{\gamma}$ is based on the entire difference between h_0 and h_1. As a result, the semiparametric estimator continues to work under the linear specification even if the estimand of the TSLS estimator is weakly identified. For the simulation, we let h_0 be the quantile function of $Beta(0.1,0.1)$, and h_1 be the quantile function of $Beta(10,10)$. These two distributions are significantly different, but they have the same mean, which is equal to 0.5. Figure 1.8 shows that, in this case, the semiparametric estimator outperforms the TSLS estimator in estimating the marginal effect.

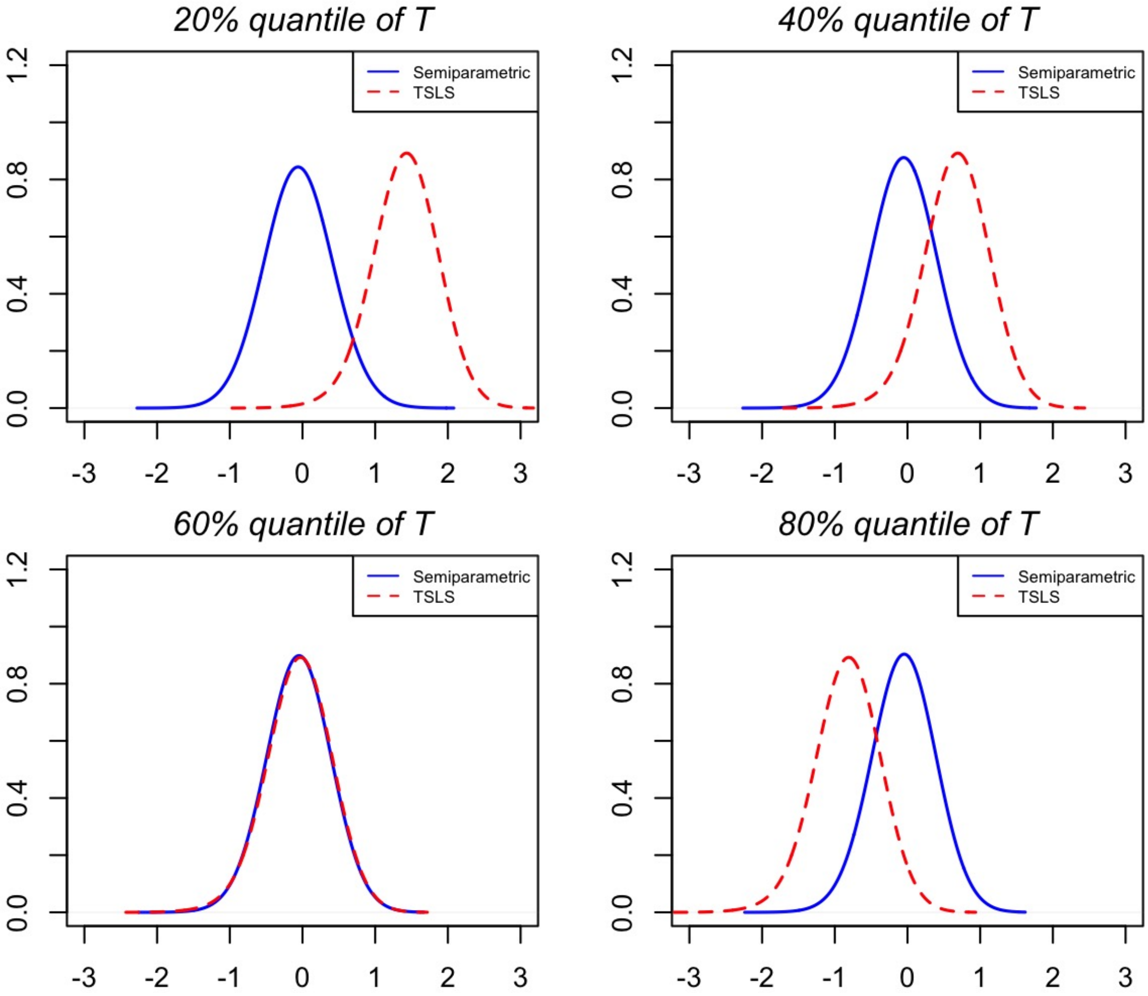

Figure 1.6. Marginal effects comparison with a nonlinear structural function.

Kernel density estimates of estimated marginal effect by the semiparametric and TSLS estimators (minus the true marginal effect) based on 500 replications. The sample size is 1000. The true structural function is specified to be $g(t,\bar{r},e) = t/2 + t^2 + e$, where the marginal effect is $1/2 + 2t$. The graphs show the estimation results of four quantile levels of the treatment: 20%, 40%, 60%, and 80%. The distribution of the semiparametric estimator is correctly centered while the TSLS estimator incurs a large bias. The TSLS estimator gives an approximately unbiased estimate of the marginal effect only around the 60% quantile level.

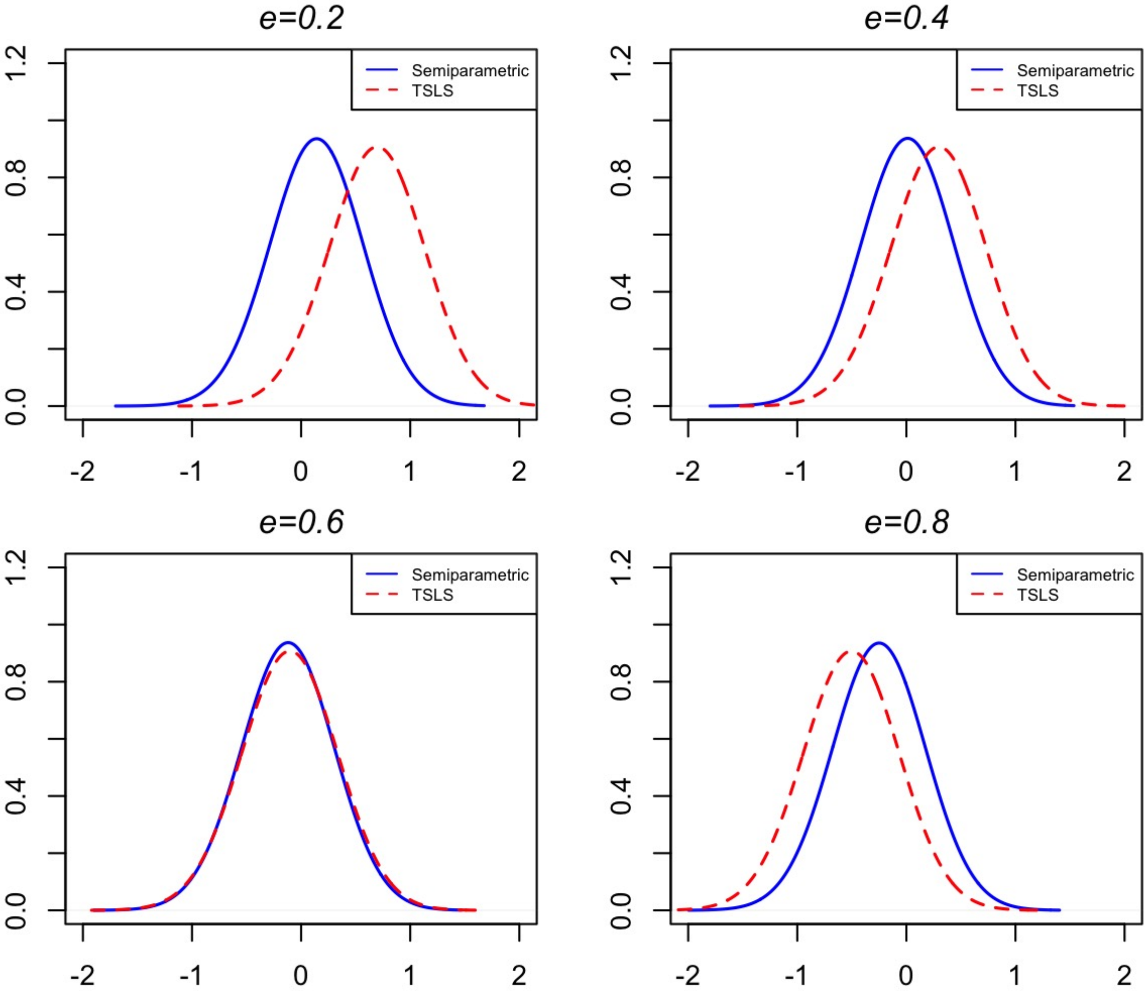

Figure 1.7. Marginal effects comparison with a nonseparable structural function.

Kernel density estimates of estimated marginal effect by the semiparametric and TSLS estimators (minus the true marginal effect) based on 500 replications. The sample size is 1000. The true structural function is specified to be $g(t,\bar{r},e) = t/2 + t^2 + 2te + e$, where the marginal effect is $1/2 + 2t + 2e$. The treatment level is specified to be the median. The graphs show the estimation results of four levels of e: 0.2, 0.4, 0.6, 0.8. The distribution of the semiparametric estimator is correctly centered while the TSLS estimator incurs a large bias. The TSLS estimator gives an unbiased estimate of the marginal effect only around $e = 0.6$.

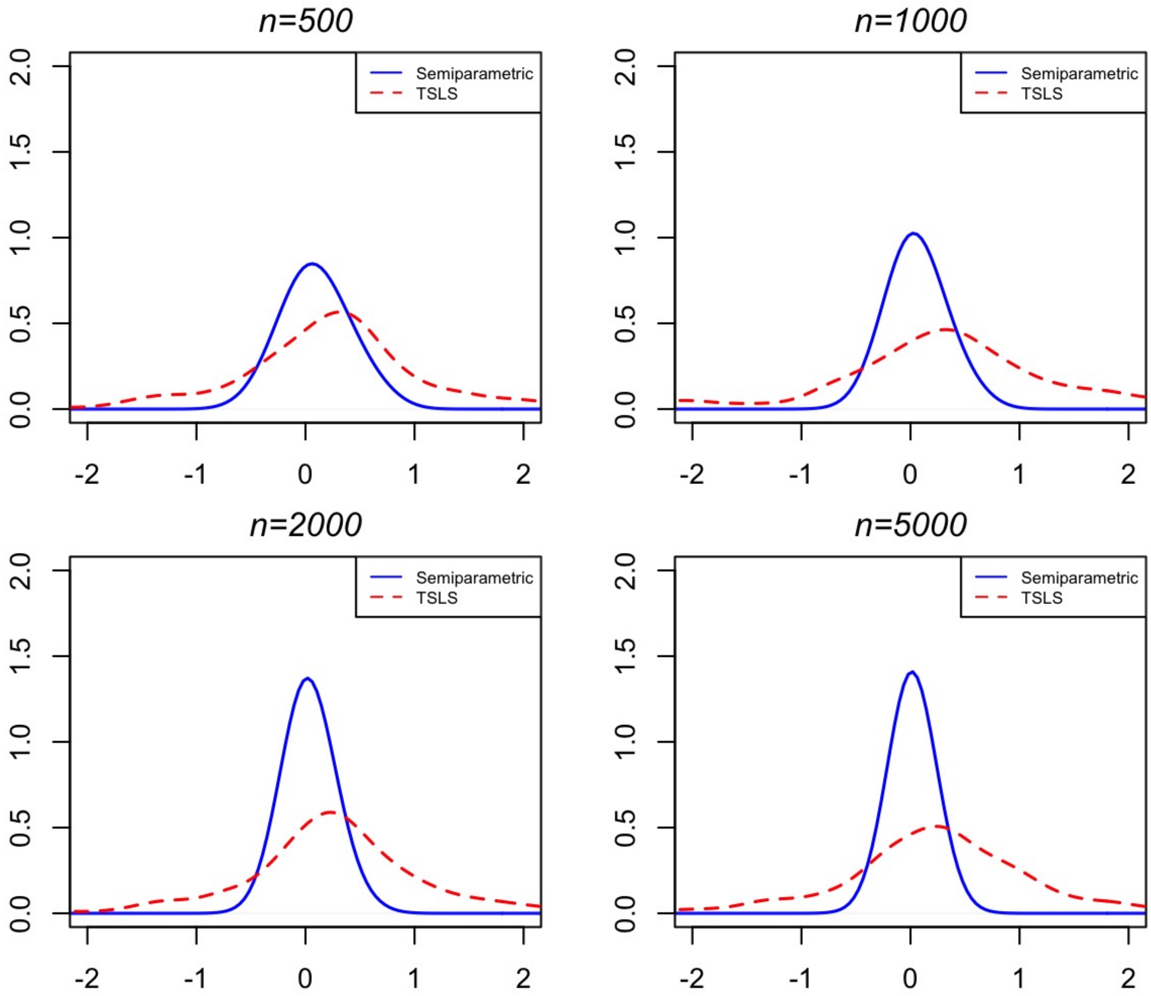

Figure 1.8. Semiparametric and TSLS estimators when the structural function is linear.

Kernel density estimates of the semiparametric and TSLS estimators (minus the true $\gamma^* = 1$) based on 500 replications. In the DGP, h_0 is equal to the quantile function of $Beta(0.1, 0.1)$, and h_1 is equal to the quantile function of $Beta(10, 10)$. These two distributions are significantly different, but they have the same mean. In this case, the semiparametric estimator outperforms the TSLS estimator because the latter is weakly identified.

1.5 Conclusion

In this study, we have examined the identification and estimation of the structural function in an RD design with a continuous treatment variable. We have established the nonparametric identification result and proposed a semiparametric estimator for the possibly nonlinear and nonseparable structural function. The estimator is proven to be consistent and asymptotically normal. The empirical study and simulation results demonstrate the advantage of the semiparametric estimator compared to the TSLS estimator.

There are two promising ways to extend the results in this paper in the future. First, we can consider extrapolating the identification result away from the cutoff. The extrapolation can be done by identifying the derivative of the structural function with respect to the running variable at the cutoff as in Dong and Lewbel [2015]. Second, we can apply the methodology developed in this paper to the regression kink design model studied by Card et al. [2015], Dong [2018b], where the treatment choice function exhibits a kink instead of a discontinuity at the cutoff.

1.6 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material. Xie, Haitian. “Nonlinear and Nonseparable Structural Functions in Fuzzy Regression Discontinuity Designs.” The dissertation author was the primary investigator and author of this material.

Chapter 2

Uniform Convergence Results for the Local Linear Regression Estimation of the Conditional Distribution

2.1 Introduction

This paper studies the nonparametric estimation of the conditional distribution function. The analysis concerns a random variable $Y \in \mathbb{R}$ and a random vector of covariates $X \in \mathbb{R}^d$. The conditional distribution function of Y given $X = x$ is denoted by $F(\cdot|x)$, that is,

$$F(y|x) = \mathbb{P}(Y \leq y|X = x), y \in \mathbb{R}.$$

When the conditional distribution function $F(\cdot|\cdot)$ is assumed to be smooth, it is natural to consider using the *local linear regression* (LLR) method to estimate F.

The main subject of this study is the uniform convergence of the LLR estimator with respect to both y and x. In particular, we derive the uniform bias expansion, characterize the uniform convergence rate, and present the uniform asymptotic linear representation of the estimator. As explained in, for example, Hansen [2008] and Kong et al. [2010], these uniform results are often useful for semiparametric estimation based on nonparametrically estimated components.

The estimation of the conditional distribution is an important area of research. Hansen

[2004] studies the asymptotic properties of both the Nadaraya-Watson (local constant) estimator and the LLR estimator, and obtains point-wise convergence results. It is well-known that the LLR estimator has the better boundary properties of the two, but unlike the Nadaraya-Watson estimator, the LLR estimator is not guaranteed to be a proper distribution function.[1] Recently, Das and Politis [2020] propose a way to correct the LLR estimator. The conditional distribution estimation is also useful for estimating conditional quantiles. For example, Yu [1997] and Yu and Jones [1998] first estimate the conditional distribution function and then invert it to obtain the conditional quantile function.

The local polynomial estimators have been studied extensively, but the uniform convergence results for the estimation of F are new to the literature. In the general setup of local polynomial estimators, there is only one regressand, namely, Y. However, in the conditional distribution estimation, there is a class of regressands, namely, $\mathbf{1}\{Y \leq y\}, y \in \mathbb{R}$. For example, Masry [1996] establishes the uniform convergence rate for general local polynomial estimators, but the uniformity is with respect to the values of the regressors. Therefore, their results can only be applied to an estimate of $F(y|\cdot)$ for a fixed $y \in \mathbb{R}$. For the same reason, the results in Kong et al. [2010] cannot be used to provide a uniform asymptotic linear representation for $y \in \mathbb{R}$. Our paper aims to solve these issues and prove that under suitable conditions, the desired results are uniform with respect to both y and x. We make use of the recent discovery by Fan and Guerre [2016] on the support of the covariates, ensuring that the uniform results are valid over the entire support.

The second contribution of the paper is the presentation of a novel way of proving the uniform convergence rate via empirical process theory. This theory was developed by Giné and Guillou [2001] and Giné and Guillou [2002] and supports the uniform almost sure convergence of the kernel density estimator. In this paper, we simplify their method and make it more accessible to users who are only concerned with the notion of uniform convergence in probability.

[1] To solve this problem Hall et al. [1999] propose a weighted Nadaraya-Watson estimator that has the same asymptotic distribution as the LLR estimator, but these weights require extensive computation.

The remaining parts of the paper are organized as follows. Section 2.2 introduces the statistical model and the assumptions. Section 2.3 establishes the uniform bias expansion result. Section 2.4 introduces empirical process theory and uses it to prove the uniform convergence rate. Section 2.5 presents the uniform asymptotic linear representation and provides a simple example to illustrate the result. The proofs are contained in the Supplementary Material.

2.2 Model and Assumptions

Let $\{(Y_i, X_i), 1 \leq i \leq n\}$ be a random sample of (Y, X). The estimation procedure is described as follows. Let w and k be two kernel functions and $K(v) = \int_{-\infty}^{v} k(u)du$. Let $h_1 = h_{1n} = o(1)$ and $h_2 = h_{2n} = o(1)$ be two scalar sequences of bandwidths. Let $\boldsymbol{r}(u) = (1, u^\top)^\top, u \in \mathbb{R}^d$ and $\boldsymbol{e}_0 = (1, 0, \cdots, 0)$ be the first $(d+1)$-dimensional unit vector. The proposed estimator is $\hat{F}(y|x) = \boldsymbol{e}_0^\top \hat{\boldsymbol{\beta}}(y, x, h_1, h_2)$, where

$$\begin{aligned}\hat{\boldsymbol{\beta}}(y,x,h_1,h_2) &= \left(\hat{\boldsymbol{\beta}}_0(y,x,h_1,h_2), \hat{\boldsymbol{\beta}}_1(y,x,h_1,h_2), \cdots, \hat{\boldsymbol{\beta}}_d(y,x,h_1,h_2)\right)^\top \\ &= \underset{\boldsymbol{\beta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(K\left(\frac{y - Y_i}{h_2}\right) - \boldsymbol{r}(X_i - x)^\top \boldsymbol{\beta} \right)^2 w\left(\frac{X_i - x}{h_1}\right). \qquad (2.1)\end{aligned}$$

Let H_1 be the $(d+1) \times (d+1)$ diagonal matrix with diagonal elements: $(1, h_1, \cdots, h_1)$. The first-order condition of the above minimization problem gives

$$H_1 \hat{\boldsymbol{\beta}}(y,x,h_1,h_2) = \hat{\Xi}(x,h_1)^{-1} \hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2), \qquad (2.2)$$

where

$$\begin{aligned}\hat{\Xi}(x,h_1) &= \frac{1}{nh_1^d} \sum_{i=1}^{n} \boldsymbol{r}\left(\frac{X_i - x}{h_1}\right) \boldsymbol{r}\left(\frac{X_i - x}{h_1}\right)^\top w\left(\frac{X_i - x}{h_1}\right), \\ \hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2) &= \frac{1}{nh_1^d} \sum_{i=1}^{n} \boldsymbol{r}\left(\frac{X_i - x}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) w\left(\frac{X_i - x}{h_1}\right).\end{aligned}$$

In the construction of the estimator, we do not use the indicator $\mathbf{1}\{Y_i \leq y\}$. Instead, we use the smoothed version $K\left((y-Y_i)/h_2\right)$, which requires the selection of another bandwidth h_2 and additional smoothness assumptions on the conditional distribution function. However, there are several advantages to using the smoothed version. First, the estimator constructed from the indicators is not smooth in y. When we believe that the true distribution function is smooth, it is customary to use the smoothed estimator. Second, from the asymptotic perspective, the indicator $\mathbf{1}\{Y_i \leq y\}$ can be considered to be the limiting case of $K\left((y-Y_i)/h_2\right)$ for $h_2 = 0$. As Hansen [2004] shows, the asymptotic mean squared error is strictly decreasing for $h_2 = 0$; hence, there are efficiency gains from smoothing. Third, as the simulation results in Yu [1997] and Yu and Jones [1998] demonstrate, the estimates are not very sensitive to the value of h_2. Lastly, as we show in Section 2.5, the smoothed estimator exhibits a stochastic equicontinuity condition in y. This condition is particularly useful when the conditional distribution estimation is an intermediate step in a semiparametric estimation procedure. For example, Chen et al. [2003] provide results on using the stochastic equicontinuity condition to derive the asymptotic distribution of two-step semiparametric estimators.

The following assumptions are maintained throughout the paper.

Assumption X (Distribution of X)**.** *The support of X, denoted by $\mathscr{X}$, is convex and compact. The marginal density f_X is bounded away from zero on $\mathscr{X}$. The restriction of f_X to $\mathscr{X}$ is twice continuously differentiable. There exist $\lambda_0, \lambda_1 \in (0,1]$ such that for any $x \in \mathscr{X}$ and all $\varepsilon \in (0, \lambda_0]$, there is $x^\top \in \mathscr{X}$ satisfying $B(x^\top, \lambda_1\varepsilon) \subset B(x,\varepsilon) \cap \mathscr{X}$, where $B(x,\varepsilon)$ denotes the ball centered at x with radius ε.*

Assumption Y (Conditional distribution of $Y|X$)**.** *The conditional distribution function $F(y|x)$ restricted to $\mathbb{R} \times \mathscr{X}$ is twice continuously differentiable in y and x. Moreover, this second-order derivative of F restricted to $\mathbb{R} \times \mathscr{X}$ is uniformly continuous.*

Assumption K (Kernel functions)**.**

(i) The kernel function w is a product kernel, that is, $w(u) = w_1(u_1)w_2(u_2)\cdots w_d(u_d)$. Each w_ℓ (1) is a symmetric density function with compact support $[-1,1]$; (2) has its second moment normalized to one, that is, $\int u_\ell^2 w_\ell(u_\ell)du_\ell = 1$; (3) is positive in the interior of the support $(-1,1)$; and (4) is of bounded variation.

(ii) The kernel function k (1) is a symmetric density function with a compact support and (2) has its second moment normalized to one, that is, $\int v^2 k(v)dv = 1$.

A brief discussion of the above assumptions is in order. Assumption **X** is introduced by Fan and Guerre [2016] as a regularity condition on the support $\mathscr{X}$. It ensures that there are sufficient observations around every estimation location, including the boundary points. Assumption **Y** imposes smoothness conditions on the conditional distribution function F. Under this assumption, the Hessian matrix of F is uniformly continuous on the compact support $supp(Y,X)$. Assumption **K** contains standard conditions on the kernel functions k and w. The bounded variation condition is imposed for the application of empirical process theory.

2.3 Uniform Bias Expansion

We denote the true value of the conditional distribution function and its derivative with respect to x as

$$\boldsymbol{\beta}^*(y,x) = \left(\boldsymbol{\beta}_0^*(y,x), \boldsymbol{\beta}_1^*(y,x), \cdots, \boldsymbol{\beta}_d^*(y,x)\right)^\top = \left(F(y|x), \nabla_x F(y|x)^\top\right)^\top,$$

where $\nabla_x F(y|x) = \left(\frac{\partial}{\partial x_1}F(y|x), \cdots, \frac{\partial}{\partial x_d}F(y|x)\right)^\top$ is the gradient of $F(y|x)$ with respect to x. A convenient way to analyze the estimator $\hat{\boldsymbol{\beta}}(y,x,h_1,h_2)$ is to consider it as an estimator of the pseudo-true value defined by

$$\begin{aligned}\bar{\boldsymbol{\beta}}(y,x,h_1,h_2) &= \Big(\bar{\boldsymbol{\beta}}_0(y,x,h_1,h_2), \bar{\boldsymbol{\beta}}_1(y,x,h_1,h_2), \cdots, \bar{\boldsymbol{\beta}}_d(y,x,h_1,h_2)\Big)^\top \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{d+1}} \mathbb{E}\left[\left(K\left(\frac{y-Y}{h_2}\right) - \boldsymbol{r}(X-x)^\top \boldsymbol{\beta}\right)^2 w\left(\frac{X-x}{h_1}\right)\right]. \qquad (2.3)\end{aligned}$$

This pseudo-true value $\bar{\boldsymbol{\beta}}$ is deterministic and converges to the true value $\boldsymbol{\beta}^*$ as $n \to \infty$.[2] We can break the asymptotic analysis of $\hat{\boldsymbol{\beta}}(y,x,h_1,h_2) - \boldsymbol{\beta}^*(y,x)$ into two parts:

$$\hat{\boldsymbol{\beta}}(y,x,h_1,h_2) - \boldsymbol{\beta}^*(y,x) = \underbrace{\hat{\boldsymbol{\beta}}(y,x,h_1,h_2) - \bar{\boldsymbol{\beta}}(y,x,h_1,h_2)}_{\text{stochastic term}} + \underbrace{\bar{\boldsymbol{\beta}}(y,x,h_1,h_2) - \boldsymbol{\beta}^*(y,x)}_{\text{bias term}}.$$

In this section, we study the bias term, which is the difference between the pseudo-true value and the true value. The first-order condition of (2.3) gives an explicit expression of the pseudo-true value: $H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2) = \Xi(x,h_1)^{-1}\boldsymbol{\upsilon}(y,x,h_1,h_2)$, where

$$\begin{aligned}\Xi(x,h_1) &= \frac{1}{h_1^d}\mathbb{E}\left[\boldsymbol{r}\left(\frac{X-x}{h_1}\right)\boldsymbol{r}\left(\frac{X-x}{h_1}\right)^\top w\left(\frac{X-x}{h_1}\right)\right], \\ \boldsymbol{\upsilon}(y,x,h_1,h_2) &= \frac{1}{h_1^d}\mathbb{E}\left[\boldsymbol{r}\left(\frac{X-x}{h_1}\right)K\left(\frac{y-Y}{h_2}\right)w\left(\frac{X-x}{h_1}\right)\right].\end{aligned}$$

Define $\Omega(x,h_1) = \int \boldsymbol{r}(u)\boldsymbol{r}(u)^\top w(u)\mathbf{1}\{x+h_1u \in \mathscr{X}\}du$. The following lemma shows that the matrices $\Xi(x,h_1)$ and $\Omega(x,h_1)$ are always bounded and invertible.

Lemma 2.1. *Under Assumptions **X** and **K**, there exists $C > 0$ such that the eigenvalues of $\Xi(x,h_1)$ and $\Omega(x,h_1)$ are in $[1/C,C]$ for all $x \in \mathscr{X}$ and $h_1 \geq 0$ small enough.*

[2] The terminology “pseudo-true” is adopted from Fan and Guerre [2016].

Theorem 2.1. *Let Assumptions **X**, **Y**, and **K** hold. Then*

$$
\begin{aligned}
&H_1\left(\bar{\boldsymbol{\beta}}(y,x)-\boldsymbol{\beta}^*(y,x)\right)\\
=&\frac{h_1^2}{2}\Omega(x,h_1)^{-1}\sum_{\ell,\ell'=1}^{d}\frac{\partial^2}{\partial x_\ell \partial x_{\ell'}}F(y|x)\int \boldsymbol{r}(u)u_\ell u_{\ell'}w(u)\mathbf{1}\{x+h_1u\in\mathscr{X}\}du\\
+&\frac{h_2^2}{2}\Omega(x,h_1)^{-1}\frac{\partial^2}{\partial y^2}F(y|x)\int \boldsymbol{r}(u)w(u)\mathbf{1}\{x+h_1u\in\mathscr{X}\}du+o(h_1^2+h_2^2),
\end{aligned}
\tag{2.4}
$$

uniformly over $y\in\mathbb{R}$ and $x\in\mathscr{X}$. In particular, we have

$$
\bar{\boldsymbol{\beta}}_0(y,x)-\boldsymbol{\beta}_0^*(y,x)=\frac{h_1^2}{2}\sum_{\ell=1}^{d}\frac{\partial^2}{\partial x_\ell^2}F(y|x)+\frac{h_2^2}{2}\frac{\partial^2}{\partial y^2}F(y|x)+o(h_1^2+h_2^2),
\tag{2.5}
$$

uniformly over $y\in\mathbb{R}$ and $x\in\mathring{\mathscr{X}}_{h_1}$, where $\mathring{\mathscr{X}}_{h_1}=\{x\in\mathscr{X}:x\pm h_1=(x_1\pm h_1,\cdots,x_d\pm h_1)\in\mathscr{X}\}$ denotes the set of interior points with respect to the bandwidth h_1.

The novelty of Theorem 2.1 is that it provides a uniform bias expansion for the LLR estimator over the entire region $(y,x)\in\mathbb{R}\times\mathscr{X}$. For the boundary points $x\notin\mathring{\mathscr{X}}_{h_1}$, the bias is $O(h_1^2+h_2^2)$. For the interior points $x\in\mathring{\mathscr{X}}_{h_1}$, the bias expression (2.5) is the same as in Hansen [2004] and Chapter 6 of Li and Racine [2007], which contains the curvature of $F(y|x)$.

2.4 Uniform Convergence Rate

In this section, we derive the uniform convergence rate of the stochastic term

$$
\hat{\boldsymbol{\beta}}(y,x,h_1,h_2)-\bar{\boldsymbol{\beta}}(y,x,h_1,h_2).
$$

We make use of empirical process theory, which is a powerful tool for studying the uniform convergence of random sequences. Some auxiliary concepts and results are introduced below.

Let $\mathscr{G}$ be a class of uniformly bounded measurable functions defined on some subset of $\mathbb{R}^d$, that is, there exists $M>0$ such that $|g|\leq M$ for all $g\in\mathscr{G}$. We say $\mathscr{G}$ is *Euclidean* with coefficients (A,v), where $A,v>0$, if for every probability measure P and every $\varepsilon \in (0,1]$, $N(\mathscr{G},P,\varepsilon) \leq A/\varepsilon^v$, where $N(\mathscr{G},P,\varepsilon)$ is the ε-covering of the metric space $(\mathscr{G},L_2(P))$, that is, $N(\mathscr{G},P,\varepsilon)$ is defined as the minimal number of open $\|\cdot\|_{L_2(P)}$-balls of radius ε and centers in $\mathscr{G}$ required to cover $\mathscr{G}$. By definition, if $\mathscr{G}$ is Euclidean with coefficients (A,v), then any subset of $\mathscr{G}$ is also Euclidean with coefficients (A,v).

The above definition of Euclidean classes is introduced by Nolan and Pollard [1987]. The same concept is also studied in Giné and Guillou [1999], but they refer to what we call "Euclidean" as "VC." There is a slight difference that Nolan and Pollard [1987] use the L_1-norm, while Giné and Guillou [1999] use the L_2-norm. We ignored the envelope in their definition because we only work with uniformly bounded $\mathscr{G}$. The following lemma is useful for deriving the uniform convergence results.

Lemma 2.2. *Let $\xi_1,\cdots,\xi_n$ be an iid sample of a random vector ξ in $\mathbb{R}^d$. Let $\mathscr{G}_n$ be a sequence of classes of measurable real-valued functions defined on $\mathbb{R}^d$. Assume that there is a uniformly bounded Euclidean class $\mathscr{G}$ with coefficients A and v such that $\mathscr{G}_n \subset \mathscr{G}$ for all n. Let σ_n^2 be a positive sequence such that $\sigma_n^2 \geq \sup_{g\in\mathscr{G}_n} \mathbb{E}[g(\xi)^2]$. Then*

$$\Delta_n = \sup_{g\in\mathscr{G}_n} \left| \sum_{i=1}^n (g(\xi_i) - \mathbb{E}g(\xi_i)) \right| = O_p\left(\sqrt{n\sigma_n^2 |\log \sigma_n|} + |\log \sigma_n| \right).$$

In particular, if $n\sigma_n^2/|\log \sigma_n| \to \infty$, then $\Delta_n = O_p\left(\sqrt{n\sigma_n^2|\log\sigma_n|}\right)$.

The above lemma is based on the results developed by Giné and Guillou [2001] and Giné and Guillou [2002]. These two papers focus on proving the almost sure convergence of kernel density estimators based on empirical process theory. We simplify their method and make it available to users who are only interested in convergence in probability. Based on Lemma 2.2, deriving the uniform convergence rate of kernel-based nonparametric estimators boils down to two parts: proving the relevant function classes are Euclidean and computing a uniform bound for the variance.

Controlling the stochastic term does not require the smoothness of the conditional distribution function F. We only need the assumptions regarding the support $\mathscr{X}$ and kernel functions. The following theorem establishes the uniform convergence rate for the stochastic term of the LLR estimator. Then, combining this result with Theorem 2.1, we obtain the uniform convergence rate of the LLR estimator as a corollary.

Theorem 2.2. *Let Assumptions **X** and **K** hold. If the bandwidth satisfies $nh_1^d/|\log h_1| \to \infty$, then*

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}} \left|H_1\left(\hat{\boldsymbol{\beta}}(y,x,h_1,h_2)-\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)\right| = O_p\left(\sqrt{\frac{|\log h_1|}{nh_1^d}}\right). \tag{2.6}$$

Corollary 2.1. *Let Assumptions **X**, **Y**, and **K** hold. Then*

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}} \left|\hat{F}(y|x)-F(y|x)\right| = O_p\left(h_1^2+h_2^2+\sqrt{\frac{|\log h_1|}{nh_1^d}}\right).$$

We want to compare the above uniform convergence result with the one in Masry [1996]. In Masry [1996], the covariates X are supported on the entire space $\mathbb{R}^d$ while the convergence result is only uniform for x in a compact subset of $\mathbb{R}^d$. In our case, the support $\mathscr{X}$ is compact, and the convergence result is uniform over the entire support $(y,x)\in\mathbb{R}\times\mathscr{X}$.

Corollary 2.1 shows that the uniformity over $y\in\mathbb{R}$ does not have an impact on the convergence rate. This is similar to the fact that we can uniformly estimate the unconditional distribution function under the $n^{-1/2}$-rate. The conditional distribution estimation is a nonparametric problem concerning only the covariates.

So far we have been studying the smoothed estimator. It is also of interest to study the unsmoothed version defined as $\check{F}(y|x)=\boldsymbol{e}_0^\top\check{\boldsymbol{\beta}}(y,x,h_1)$, where

$$\check{\boldsymbol{\beta}}(y,x,h_1)=\underset{\boldsymbol{\beta}\in\mathbb{R}^{d+1}}{\operatorname{argmin}}\sum_{i=1}^n\left(\mathbf{1}\{Y_i\le y\}-\boldsymbol{r}(X_i-x)^\top\boldsymbol{\beta}\right)^2 w\left(\frac{X_i-x}{h_1}\right).$$

The above minimization problem is constructed by replacing the term $K((y-Y_i)/h_2)$ by the term $\mathbf{1}\{Y_i \leq y\}$. For this estimator, we only need to chose one bandwidth h_1. Since $\mathbb{E}[\mathbf{1}\{Y_i \leq y\}|X = x] = F(y|x)$, the bias term of this estimator is only $O(h_1^2)$. The bias associated with the smoothing in y no longer exists. The uniform convergence of $\check{F}$ can be established without the differentiability of F with respect to y. The trade-off is that the estimator $\check{F}(y|x)$ itself is not smooth in y even if F is. The stochastic term can be analyzed as before. Then we obtain the following uniform convergence rate for the unsmoothed estimator. Notice that we replace Assumption **Y** by a weaker condition which only requires the smoothness of F with respect to x.

Theorem 2.3. *Let Assumptions **X** and **K** hold. Assume that $F(y|x)$ restricted to $\mathbb{R} \times \mathscr{X}$ is twice continuously differentiable in x, and the second-order derivative $\nabla_x^\top \nabla_x F$ is uniformly continuous on $\mathbb{R} \times \mathscr{X}$. If the bandwidth satisfies $nh_1^d/|\log h_1| \to \infty$, then*

$$\sup_{y\in\mathbb{R}, x\in\mathscr{X}} \left|\check{F}(y|x) - F(y|x)\right| = O_p\left(h_1^2 + \sqrt{\frac{|\log h_1|}{nh_1^d}}\right).$$

2.5 Uniform Asymptotic Linear Representation

This section derives the uniform asymptotic linear representation of the smoothed LLR estimator. These results are particularly useful in deriving the asymptotic distribution for complicated estimators.

Theorem 2.4. *Let Assumptions **X**, **Y**, and **K** hold. If the bandwidth satisfies that $nh_1^d/|\log h_1| \to \infty$, $nh_1^{d+4}/|\log h_1|$ bounded, and $h_2 = O(h_1)$, then*

$$H_1\left(\hat{\boldsymbol{\beta}}(y,x,h_1,h_2) - \bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right) = \Xi(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2) + O_p\left(\frac{|\log h_1|}{nh_1^d}\right),$$

uniformly over $y \in \mathbb{R}$ and $x \in \mathscr{X}$, where

$$\boldsymbol{s}(Y_i, X_i; y, x, h_1, h_2) = \boldsymbol{r}\left(\frac{X_i - x}{h_1}\right)\left(K\left(\frac{y - Y_i}{h_2}\right) - \tilde{F}(y|X_i)\right) w\left(\frac{X_i - x}{h_1}\right),$$

$$\tilde{F}(y|x) = \mathbb{E}[K((y - Y)/h_2) \mid X = x].$$

The asymptotic order of the remainder term, $O_p\left(|\log h_1|/nh_1^d\right)$, is the same as in Equation (13) of Kong et al. [2010]. Thus, once more, the uniformity over $y \in \mathbb{R}$ does not have an impact on the convergence rate. Combining the results in Theorem 2.1 and 2.4 and applying the central limit theorem for triangular arrays, we can show that the LLR estimator is asymptotic normal with some asymptotic bias.

We can use this asymptotic linear representation, together with the smoothness of K, to derive the following stochastic equicontinuity condition.

Corollary 2.2. *Let the assumptions of Theorem 2.4 hold. Let $\delta_n = o(1)$. Then the following stochastic equicontinuity condition hold.*

$$\sup_{|y_1 - y_2| \leq \delta_n, x \in \mathscr{X}} \left|\hat{F}(y_1|x) - F(y_1|x) - (\hat{F}(y_2|x) - F(y_2|x))\right| = O_p\left(\sqrt{\frac{\log h_1}{nh_1^d}\frac{\delta_n}{h_2}} + \frac{|\log h_1|}{nh_1^d}\right).$$

We study another simple example to demonstrate how the uniform asymptotic linear representation can be used. Suppose that $d = 1$, $\mathscr{X} = [\underline{x}, \bar{x}]$, and Y is supported on $[\underline{y}, \bar{y}]$. We want to estimate the integrated conditional distribution $\theta = \int_{\underline{y}}^{\bar{y}} \int_{\underline{x}}^{\bar{x}} F(y|x) dx dy$ with the estimator $\hat{\theta} = \int_{\underline{y}}^{\bar{y}} \int_{\underline{x}}^{\bar{x}} \hat{F}(y|x) dx dy$. The theorem below gives the asymptotic distribution of the estimator $\hat{\theta}$.

Corollary 2.3. *Let Assumptions **X**, **Y**, and **K** hold. If the bandwidth satisfies that $\sqrt{n}h_1/|\log h_1| \to \infty$, $\sqrt{n}h_1^2 \to 0$ bounded, and $h_2 = O(h_1)$, then $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$, where*

$$V = \int \left(\int \left(\mathbf{1}\{s \leq y\} - F(y|t)\right) dy\right)^2 f(s,t) dt ds,$$

and $f(y,x)$ *denotes the joint density of* (Y,X).

2.6 Acknowledgements

Chapter 2, in full, has been submitted for publication. Xie, Haitian. “Uniform Convergence Results for the Local Linear Regression Estimation of the Conditional Distribution.” The dissertation author was the primary investigator and author of this material.

Chapter 3

Efficient and Robust Estimation of the Generalized LATE Model

3.1 Introduction

Since the seminal works of Imbens and Angrist [1994] and Angrist et al. [1996], the *local average treatment effect* (LATE) model has become popular for causal inference in economics. Instead of imposing homogeneity of the treatment effects as in the classical instrumental variable (IV) regression model, the LATE framework allows the treatment effect to vary across individuals. Under the monotonicity condition, the average treatment effect can be identified for a subgroup of individuals whose treatment choice complies with the change in instrument levels.

The current form of the LATE model only accepts binary treatment variables. This restriction is inconvenient in many economic settings where the treatment is multi-leveled in nature. For example, parents select different preschool programs for their kids, schools assign students to different classroom sizes, families relocate to various neighborhoods in housing experiments, and people choose different sources of health insurance. To apply the LATE model to these settings, researchers often need to redefine the treatment so that there are only two treatment levels. However, merging the treatment levels can complicate the task of program evaluation and dampen the causal interpretation of the estimates. As pointed out by Kline and Walters [2016], if the original treatment levels are substitutes, then there is ambiguity regarding which causal parameters are of interest. After merging the treatment levels, the heterogeneity in the treatment effect across different treatment levels is lost.

This paper addresses the above issues by generalizing the LATE framework to incorporate the potential multiplicity in treatment levels directly. We call the new framework the *generalized LATE* (GLATE) model. The main assumption of the GLATE model is the unordered monotonicity assumption proposed by Heckman and Pinto [2018a], which is a generalization of the monotonicity assumption in the binary LATE model.[1]

We generalize the identification results in Heckman and Pinto [2018a] to explicitly account for the presence of conditioning covariates, which is often important in practical settings. Recently, Blandhol et al. [2022] point out that linear TSLS, the common way to control for covariates in empirical studies, does not bear the LATE interpretation. The only specifications that have LATE interpretations are the ones that control for covariates nonparametrically. Therefore, it is essential from the causal analysis perspective to incorporate the covariates into the GLATE framework in a nonparametric way.

The causal parameters identifiable in the GLATE model include *local average structural function* (LASF) and *local average structural function for the treated* (LASF-T). LASF is the mean *potential outcome* for specific subpopulations. These subpopulations are defined by their treatment choice behaviors and are generalizations of the concepts *always takers*, *compliers*, and *never takers* in the binary LATE model. The parameter LASF-T further restricts the subpopulation to exclude individuals who do not take up the treatment.

The paper is concerned with the econometric aspects of the GLATE model. The analysis begins by deriving *efficient influence function* (EIF) and *semiparametric efficiency bound* (SPEB) for the identified parameters. The calculation is based on the method outlined in Chapter 3 of Bickel et al. [1993] and Newey [1990]. We then verify that the *conditional expectation projection* (CEP) estimator [e.g., Chen et al., 2008], constructed directly from the identification result, achieves the SPEB and hence is semiparametric efficient. Using these results, we may

[1] To distinguish with the GLATE model, we sometimes use the terminology “binary LATE model” to refer to the LATE model studied by Imbens and Angrist [1994] and Abadie [2003].

efficiently estimate other important parameters of interest by the plug-in method since a standard delta-method argument preserves semiparametric efficiency.

The EIF not only facilitates the efficiency calculation but can also serve as the moment condition for estimation. This is because the EIF is mean zero by construction and is equal to the original identification result plus an adjustment term due to the presence of infinite-dimensional parameters. We show that the moment condition constructed from the EIF satisfies two related robustness properties: double robustness and Neyman orthogonality. Double robustness guarantees that the moment condition is correctly specified in a parametric setting even when some nuisance parameters are not.

The Neyman orthogonality condition means that the moment condition is insensitive to the nuisance parameters. This condition is particularly useful when the conditioning covariates are of high dimension. To further utilize this condition, we study the *double/debiased machine learning* (DML) estimator [Chernozhukov et al., 2018] in the GLATE setting. Under certain conditions regarding the convergence rate of the first-step nonparametric estimators, the DML estimator is asymptotically normally uniformly over a large class of *data generating processes* (DGPs).

The weak identification issue is a practical concern of the GLATE model. This is because both the treatment and instrument are multi-valued, and hence the subpopulation on which LASF and LASF-T are defined can be small in size. To deal with this issue, we propose null-restricted test statistics in one-sided and two-sided testing problems. This procedure is the generalization of the well-known Anderson-Rubin (AR) test. We show that the proposed tests are consistent and uniformly control size across a large class of DGPs, in which the size of the subpopulation mentioned above can be arbitrarily close to zero.

The paper is organized as follows. The remaining part of this section discusses the literature. Section 3.2 introduces the GLATE model and the nonparametric identification results. Section 3.3 calculates the EIF and SPEB. Section 3.4 discusses the robustness properties of the moment condition generated by the EIF. Section 3.5 proposes inference procedures under weak identification issues. Section 3.6 presents the empirical application. Section 3.7 concludes. The proofs for theoretical results in the main text are collected in Appendix A in the Supplementary Material.

3.1.1 Literature Review

The GLATE model provides a way to conduct causal inference under endogeneity when the treatment is multi-valued and unordered. As mentioned above, the identification result (conditional on the covariates) is first established in Heckman and Pinto [2018a] by using the unordered monotonicity condition. Lee and Salanié [2018] proposes another method of identification in a similar model of multi-valued treatment. Their method is concerned with continuous instruments, while the GLATE is framed in terms of discrete-valued instruments. When the treatment levels are ordered, Angrist and Imbens [1995] derives the identification and estimation results for the causal parameter, which is a weighted average of LATEs across different treatment levels.

The literature on semiparametric efficiency in program evaluation starts with the seminal work of Hahn [1998], which studies the benchmark case of estimating the *average treatment effect* (ATE) under *unconfoundedness*. For multi-level treatment, Cattaneo [2010] studies the efficient estimation of causal parameters implicitly defined through over-identified non-smooth moment conditions. In the case where unconfoundedness fails and instruments are present, Frölich [2007] calculates the SPEB for the LATE parameter, and Hong and Nekipelov [2010a] extend to the estimation of parameters implicitly defined by moment restrictions. In a more general framework encompassing missing data, Chen et al. [2008] study semiparametric efficiency bounds and efficient estimation of parameters defined through overidentifying moment restrictions. However, there is currently no theoretical research on semiparametric efficient estimation in models that encompasses endogeneity and unordered multiple treatment levels.

Several ways are available for calculating the EIF for semiparametric estimators, as illustrated by Newey [1990] and Ichimura and Newey [2022]. Semiparametric efficiency calculations can be used to construct robust (Neyman orthogonal) moment conditions. This method is illustrated in Newey [1994] and Chernozhukov et al. [2016]. Based on the Neyman orthogonality condition, Chernozhukov et al. [2018] introduces the DML method that suits high dimensional settings. This is because Donsker properties and stochastic equicontinuity conditions are no longer required in deriving the asymptotic distribution of the semiparametric estimator.

For testing the GLATE model, Sun [2021] proposes a bootstrap test which is the generalization and improvement of the test studied by Kitagawa [2015] in the binary LATE model.

The GLATE model has received attention in the recent empirical literature due to its ability to model multi-valued treatment. Kline and Walters [2016] evaluate the cost-effectiveness of Head Start, classifying Head Start and other preschool programs as different treatment levels against the control group of no preschool. Galindo [2020] assesses the impact of different childcare choice in Colombia on children's development. Pinto [2021] studies the neighborhood effects and voucher effects in housing allocations using data from the Moving to Opportunity experiment. Our theoretical analysis of the GLATE model presents important tools for estimation and inference that can be applied to those empirical settings.

3.2 Identification in the GLATE Model

This section describes the *generalized local average treatment effect* (GLATE) model, discusses identification of the *local average structural function* (LASF) and other parameters, and introduces the notation.

3.2.1 The model

We assume a finite collection of instrument values $\mathscr{Z} = \{z_1, \cdots, z_{N_Z}\}$ and a finite collection of treatment values $\mathscr{T} = \{t_1, \cdots, t_{N_T}\}$, where N_Z and N_T are respectively the total number of instrument and treatment levels. The sets $\mathscr{T}$ and $\mathscr{Z}$ are categorical and unordered. The instrumental variable Z denotes which of the N_Z instrument levels is realized. The random variables $T_{z_1}, \cdots, T_{z_{N_Z}}$, each taking values in $\mathscr{T}$, denote the collection of potential treatments under each instrument status. Thus, the observed treatment level is the random variable $T = T_Z = \sum_{z \in \mathscr{Z}} \mathbf{1}\{Z = z\} T_z$. For each given treatment level $t \in \mathscr{T}$, there is a potential outcome $Y_t \in \mathscr{Y} \subset \mathbb{R}$. The observed outcome is denoted by $Y = Y_T = \sum_{t \in \mathscr{T}} \mathbf{1}\{T = t\} Y_t$. The random vector $X \in \mathscr{X} \subset \mathbb{R}^{dx}$ contains the set of covariates. The observed data is a random sample $(Y_i, T_i, Z_i, X_i), 1 \leq i \leq n$.

The description above establishes a random sampling model where the researcher only observes one potential outcome, the one associated with the observed treatment. This implies that the sample of Y, observed from an individual with treatment $T = t$, comes from the conditional distribution of Y_t given $T = t$ rather than from the marginal distribution of Y_t. In general, this fact leads to identifications issues and presents challenges for causal inference. To overcome these problems, we impose further structures on the model.

Assumption 3.1 (Conditional Independence)**.** $(\{Y_t : t \in \mathscr{T}\}, \{T_z : z \in \mathscr{Z}\}) \perp Z \mid X$.

Assumption 3.2 (Unordered Monotonicity)**.** *For any $t \in \mathscr{T}, z, z' \in \mathscr{Z}$, either*

$$\mathbb{P}(\mathbf{1}\{T_z = t\} \geq \mathbf{1}\{T_{z'} = t\} \mid X) = 1$$

or

$$\mathbb{P}(\mathbf{1}\{T_z = t\} \leq \mathbf{1}\{T_{z'} = t\} \mid X) = 1.$$

Assumption 3.1 and 3.2 provide the multi-valued analog of Assumption 2.1 in Abadie [2003]. Assumption 3.1 restricts that the instrument Z is independent with the potential treatments and outcomes once we condition on X. Assumption 3.2 is the conditional version of the unordered monotonicity condition proposed by Heckman and Pinto [2018a]. It means that when we focus on a particular treatment level t and a pair (z, z') of instrument values, the binary environment should satisfy the usual monotonicity constraint in the LATE model. Specifically, the unordered monotonicity condition requires that a shift in the instrument moves all agents uniformly toward or against each possible treatment value.[2]

We define the type S of an individual as the vector of the potential treatments, that is,

$$S = (T_{z_1} \cdots, T_{z_{N_Z}})'.$$

By construction, S is not observed. Assumption 3.2, the unordered monotonicity condition, is essentially a restriction on $\mathscr{S} \equiv supp(S)$, the support of S. Denote the elements in $\mathscr{S}$ by $s_1, \cdots, s_{N_S}$, where N_S is the cardinality of $\mathscr{S}$. A convenient way to characterize $\mathscr{S}$ is by using the $N_Z \times N_S$ matrix $R \equiv (s_1, \cdots, s_{N_S})$. The matrix R is referred to as the response matrix since it describes how each type of individuals' treatment choice responds to the instrument.

The role of S is to assist the identification of the counterfactual outcomes by dividing the population into a finite number of groups, where identification can be achieved within specific groups. Those groups are defined as follows. For $k = 0, \cdots, N_Z$, let $\Sigma_{t,k}$ be the set of types in which the treatment level t appears exactly k times. That is,

$$\Sigma_{t,k} \equiv \{s \in \mathscr{S} : \textstyle\sum_{i=1}^{N_Z} \mathbf{1}\{s[i] = t\} = k\},$$

where $s[i]$ denotes the ith element of the vector s. In particular, the collection $\Sigma_{t,k}, k = 0, \cdots, N_Z$ forms a partition of $\mathscr{S}$.

For individuals with type S in the same type set $\Sigma_{t,k}$, their treatment response in terms of $T = t$ is in a way homogeneous. Thus, it is easier intuitively to identify the marginal distribution of the potential outcome Y_t within each $\Sigma_{t,k}$. More specifically, we define the *local average structural functions* (LASF) and the *local average structural functions for the treated* (LASF-T)

[2] As pointed out by Vytlacil [2002], the LATE monotonicity condition is a restriction across individuals on the relationship between different hypothetical treatment choices defined in terms of an instrument.

as follows.

$$\text{LASF: } \beta_{t,k} \equiv \mathbb{E}[Y_t \mid S \in \Sigma_{t,k}],$$
$$\text{LASF-T: } \gamma_{t,k} \equiv \mathbb{E}[Y_t \mid S \in \Sigma_{t,k}, T = t].$$

Before presenting the identification results for the above two classes of parameters, we illustrate the GLATE model in the following two examples.

Example 3.1 (Binary LATE model)**.** *In the binary LATE model of Imbens and Angrist [1994], there are two treatment levels $\mathscr{T} = \{0,1\}$ and two instrument levels $\mathscr{Z} = \{0,1\}$. There are three types: $\mathscr{S} = \{s_1 = (0,0)', s_2 = (0,1)', s_3 = (1,1)'\}$, which are referred to in the literature as never-takers, compliers, and always-takers, respectively. The type set $\Sigma_{1,0} = \{s_1\}$ contains the never-takers, $\Sigma_{1,1} = \{s_2\}$ the compliers, and $\Sigma_{1,2} = \{s_3\}$ the always-takers. The response matrix is the following binary matrix*

$$R = (s_1, s_2, s_3) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

The local average treatment effect is the treatment effect for the compliers, which can be written as the difference between two LASFs:

$$\mathbb{E}[Y_1 - Y_0 \mid S = \textit{compliers}] = \mathbb{E}[Y_1 - Y_0 \mid T_1 > T_0] = \beta_{1,1} - \beta_{0,1}.$$

Example 3.2 (Three treatment levels and two instrument levels)**.** *The simplest GLATE model (excluding the binary case in Example 3.1) has three treatment levels $\mathscr{T} = \{t_1, t_2, t_3\}$ and two instrument levels $\mathscr{Z} = \{z_1, z_2\}$. There are five types specified as the columns in the following* *response matrix*

$$R = (s_1, s_2, s_3, s_4, s_5) = \begin{pmatrix} t_1 & t_2 & t_3 & t_1 & t_2 \\ t_1 & t_2 & t_3 & t_3 & t_3 \end{pmatrix}.$$

In this example, a shift from z_1 to z_2 moves all agents uniformly toward the treatment level t_3. The type set $\Sigma_{t_1,2} = \{s_1\}$ contains the type that always choose the treatment t_1 and thus can be referred to as t_1-always taker. The same applies to $\Sigma_{t_2,2} = \{s_2\}$ and $\Sigma_{t_3,2} = \{s_3\}$. The type set $\Sigma_{t1,1} = \{s_4\}$ switches from t_1 to t_3 and hence can be considered as t_1-swticher (or t_1-compliter). Similarly, we can refer to $\Sigma_{t_2,1} = \{s_5\}$ as t_2-switcher and $\Sigma_{t_2,1} = \{s_5\}$ as t_3-switcher. This model is used in Kline and Walters [2016] to study the causal effect of the Head Start preschool program. The instrument indicates whether the household receives a Head Start offer, and the treatment levels are $t_1 =$ Head Start, $t_2 =$ other preschool programs, and $t_3 =$ no preschool. The unordered monotonicity condition means that anyone who changes behavior as a result of the Head Start offer does so to attend Head Start.

3.2.2 Identification Results

We introduce some matrix notations related to the type S. For each treatment level $t \in \mathscr{T}$, let B_t be a binary matrix of the same dimension as the response matrix R with each element of B_t signifying whether the corresponding element in the response matrix is t. That is, $B_t[i,j]$, the (i,j)th element of B_t, is whether T_{z_i} equals t for the subpopulation $S = s_j$. Define $b_{t,k} \equiv \left(\mathbf{1}\{s_1 \in \Sigma_{t,k}\}, \cdots, \mathbf{1}\{s_{N_S} \in \Sigma_{t,k}\}\right) B_t^+$, where B_t^+ is the Moore-Penrose inverse of B_t.

For convenience, we also need some notations regarding conditional expectations. Let

$$\pi(X) \equiv (\pi_{z_1}(X), \cdots, \pi_{z_{N_Z}}(X))' \text{ with } \pi_z(X) \equiv P\left(Z = z \mid X\right)$$

be the vector of functions that describes the conditional distribution of the instrument Z. For each treatment level $t \in \mathscr{T}$, let

$$P_t(X) \equiv (P_{t,z_1}(X), \cdots, P_{t,z_{N_Z}}(X))' \text{ with } P_{t,z}(X) \equiv P(T = t \mid Z = z, X)$$

be the vector that describes the conditional treatment probabilities given each level of the instrument. Denote

$$Q_t(X) \equiv (Q_{t,z_1}(X), \cdots, Q_{t,z_{N_Z}}(X))' \text{ with } Q_{t,z}(X) \equiv \mathbb{E}[Y\mathbf{1}\{T = t\} \mid Z = z, X]$$

as the vector that contains the conditional outcomes for each treatment level t. Notice that the functions π, P_t, and Q_t are all identified.

Theorem 3.1 (Identification of LASF)**.** *Let Assumptions 3.1 - 3.2 hold. Let $t \in \mathscr{T}$ and $k \in \{1, \cdots, N_Z\}$.*

(i) The type set probability is identified by

$$p_{t,k} \equiv \mathbb{P}(S \in \Sigma_{t,k}) = b_{t,k}\mathbb{E}\left[P_t(X)\right].$$

(ii) If $p_{t,k} > 0$, the LASF is identified by:

$$\beta_{t,k} = b_{t,k}\mathbb{E}\left[Q_t(X)\right]/p_{t,k}.$$

Theorem 3.1 identifies $p_{t,k}$, the size of the subpopulation $\Sigma_{t,k}$, and the local structural function for that subpopulation. The only exception when the identification fails is when the type set $\Sigma_{t,0}$, in which case the individual never chooses the treatment t. This identification result is a modification of Theorem T-6 in Heckman and Pinto [2018a] that explicitly accounts for the presence of covariates X. Bayes rule is applied to convert the conditional result into the unconditional one. The following theorem presents the identification result for the LASF-T.

Let $\mathscr{Z}_{t,k} \subset \mathscr{Z}$ be the set of instrument values that induces the treatment level t in the type set $\Sigma_{t,k}$. That is, $\mathscr{Z}_{t,k} \equiv \left\{ z_i \in \mathscr{Z} : s[i] = t, \text{ for all } s \in \Sigma_{t,k} \right\}$, where $s[i]$ denotes the ith element of the vector s. Then define $\pi_{t,k} \equiv \sum_{z \in \mathscr{Z}_{t,k}} \pi_z$ as the total probability of those instrument values.

Theorem 3.2 (Identification of LASF-T)**.** *Let Assumptions 3.1 - 3.2 hold. Let $t \in \mathscr{T}$ and $k \in \{1, \cdots, N_Z\}$. Then $\mathscr{Z}_{t,k}$ is nonempty.*

(i) The treatment probability within the type set is identified by

$$q_{t,k} \equiv P\left(T = t, S \in \Sigma_{t,k}\right) = b_{t,k} \mathbb{E}\left[P_t(X) \pi_{t,k}(X)\right].$$

(ii) If $q_{t,k} > 0$, then the LASF-T is identified by

$$\gamma_{t,k} = b_{t,k} \mathbb{E}\left[Q_t(X) \pi_{t,k}(X)\right] / q_{t,k}. \tag{3.1}$$

The identification results are illustrated using the two examples.

Example 3.3 (continues = eg:binary)**.** *Since the treatment is binary, the matrix B_1 is equal to the response matrix R. The matrix B_1 and its generalized inverse B_1^+ are respectively*

$$B_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \textit{ and } (B_1^+)' = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The matrix B_0 and its generalized inverse B_0^+ are respectively

$$B_0 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \textit{ and } (B_0^+)' = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}.$$

The vectors $b_{1,1}$ and $b_{0,1}$ are respectively

$$b_{1,1} = (-1,1), \text{ and } b_{0,1} = (1,-1).$$

Theorem 3.1 implies that

$$\beta_{1,1} = \frac{\mathbb{E}[Q_{1,1}(X)] - \mathbb{E}[Q_{1,0}(X)]}{\mathbb{E}[P_{1,1}(X)] - \mathbb{E}[P_{1,0}(X)]}, \text{ and } \beta_{0,1} = \frac{\mathbb{E}[Q_{0,0}(X)] - \mathbb{E}[Q_{0,1}(X)]}{\mathbb{E}[P_{0,0}(X)] - \mathbb{E}[P_{0,1}(X)]}.$$

The two denominators in the above expressions are both equal to the type probability of compliers. Then the usual identification of the LATE parameter [e.g., Frölich, 2007] follows:

$$\mathbb{E}[Y_1 - Y_0 \mid T_1 > T_0] = \frac{\int(\mathbb{E}[Y \mid Z = 1, X = x] - \mathbb{E}[Y \mid Z = 0, X = x])f_X(x)dx}{\int(\mathbb{E}[T \mid Z = 1, X = x] - \mathbb{E}[T \mid Z = 0, X = x])f_X(x)dx},$$

where f_X denotes the marginal density function of X.

Example 3.4 (continues = eg:3t2z)**.** *Recall that $\Sigma_{t_1,1} = \{s_4\}$ contains the t_1-switcher. By Theorem 3.1, the LASF for the treatment level t and the subpopulation $S = s_4$ is identified by*[3]

$$p_{t_1,1} = \mathbb{E}[P_{t_1,z_1}(X)] - \mathbb{E}[P_{t_1,z_2}(X)],$$
$$\beta_{t_1,1} = \frac{\mathbb{E}[Q_{t_1,z_1}(X)] - \mathbb{E}[Q_{t_1,z_2}(X)]}{\mathbb{E}[P_{t_1,z_1}(X)] - \mathbb{E}[P_{t_1,z_2}(X)]}.$$

Notice that $\mathscr{Z}_{t_1,1} = \{z_1\}$. Then by Theorem 3.2 we have

$$q_{t_1,1} = \mathbb{E}[(P_{t_1,z_1}(X) - P_{t_1,z_2}(X))\pi_{z_1}(X)],$$
$$\gamma_{t_1,1} = \frac{\mathbb{E}[(Q_{t_1,z_1}(X) - Q_{t_1,z_2}(X))\pi_{z_1}(X)]}{\mathbb{E}[(P_{t_1,z_1}(X) - P_{t_1,z_2}(X))\pi_{z_1}(X)]}.$$

[3] The calculation of $b_{t,k}$ is omitted for brevity, but it can be done in the same way as Example 3.1.

3.3 Semiparametric Efficiency

In this section, we calculate the *semiparametric efficiency bound* (SPEB) and propose estimators that achieve such bounds. We focus on the parameters LASF and LASF-T. In Appendix B in the Supplementary Material, we study general parameters implicitly defined through moment restrictions.

3.3.1 LASF and LASF-T

For the rest of the paper, we assume that $Y_t, t \in \mathscr{T}$ have finite second moments. This is necessary since we are studying efficiency. Let ι denote the column vector of ones and $\zeta(Z,X,\pi)$ the diagonal matrix with the diagonal elements being $\mathbf{1}\{Z=z\}/\pi_z(X), z \in \mathscr{Z}$. The following theorem gives the *efficient influence function* (EIF) and the SPEB for the parameters identified in the preceding section.

Theorem 3.3 (SPEB for LASF and LASF-T)**.** *Let Assumptions 3.1 - 3.2 hold. Let $t \in \mathscr{T}$ and $k \in \{1,\cdots,N_Z\}$. Assume that $p_{t,k}, q_{t,k} > 0$.*

(i) The semiparametric efficiency bound for $\beta_{t,k}$ is given by the variance of the efficient influence function

$$\begin{aligned}&\psi^{\beta_{t,k}}(Y,T,Z,X,\beta_{t,k},p_{t,k},Q_t,P_t,\pi)\\ =&\frac{1}{p_{t,k}}b_{t,k}\Big(\zeta(Z,X,\pi)\left(\iota(Y\mathbf{1}\{T=t\})-Q_t(X)\right)+Q_t(X)\Big)\\ -&\frac{\beta_{t,k}}{p_{t,k}}b_{t,k}\Big(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)+P_t(X)\Big).\end{aligned} \tag{3.2}$$

(ii) The semiparametric efficiency bound for $\gamma_{t,k}$ is given by the variance of the efficient

influence function

$$
\begin{aligned}
&\psi^{\gamma_{t,k}}(Y,T,Z,X,\gamma_{t,k},q_{t,k},Q_t,P_t,\pi)\\
=&\frac{1}{q_{t,k}}b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota(Y\mathbf{1}\{T=t\})-Q_t(X)\right)\pi_{t,k}(X)+Q_t(X)\mathbf{1}\{Z\in\mathscr{Z}_{t,k}\}\right)\\
-&\frac{\gamma_{t,k}}{q_{t,k}}b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)\pi_{t,k}(X)+P_t(X)\mathbf{1}\{Z\in\mathscr{Z}_{t,k}\}\right).
\end{aligned}
$$

(iii) The semiparametric efficiency bound for $p_{t,k}$ is given by the variance of the efficient influence function

$$
\psi^{p_{t,k}}(T,Z,X,p_{t,k},P_t,\pi)=b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)+P_t(X)\right)-p_{t,k}.
$$

(iv) The semiparametric efficiency bound for $q_{t,k}$ is given by the variance of the efficient influence function

$$
\begin{aligned}
&\psi^{q_{t,k}}(T,Z,X,q_{t,k},P_t,\pi)\\
=&b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)\pi_{t,k}(X)+P_t(X)\mathbf{1}\{Z\in\mathscr{Z}_{t,k}\}\right)-q_{t,k}.
\end{aligned}
$$

The EIF in Theorem 3.3 can be interpreted as the moment condition from the identification results modified by an adjustment term due to the presence of unknown infinite-dimensional parameters. Take $\psi^{\beta_{t,k}}$ as an example, the terms

$$
b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota(Y\mathbf{1}\{T=t\})-Q_t(X)\right)\right)/p_{t,k}
$$

and

$$
\beta_{t,k}b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)\right)/p_{t,k}
$$

are respectively the adjustment terms due to the presence of Q_t and P_t.

From the expression of $\psi^{\beta_{t,k}}$, we can see that the SPEB would be large when $p_{t,k}$ is small. This is because $p_{t,k}$ measures the size of the subpopulation $S \in \Sigma_{t,k}$ on which the LASF is estimated. When $p_{t,k}$ is small, we run into the weak identification issue. In Section 3.5, we study inference procedures that are robust against weak identification issues.

One benefit of the EIFs is that we can easily calculate the covariance matrix of different estimators. Consider an example where we are interested in two LASFs β_1 and β_2, whose EIF is given by ψ_1 and ψ_2, respectively. If the two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are both semiparametric efficient, then their covariance matrix equals $\mathbb{E}[\psi_1 \psi_2']$.

Example 3.5 (continues = eg:binary)**.** *In the binary LATE model, the first two parts of Theorem 3.3 reduce to Theorem 2 of Hong and Nekipelov [2010a]. If we assume unconfoundedness by having $T = Z$, then the result further reduces to Theorem 1 of Hahn [1998].*

The derived SPEB helps determine whether an estimation procedure is efficient. In this section, we focus on the *conditional expectation projection* (CEP) estimator.[4] Define

$$h_{Y,t,z}(X) = \mathbb{E}\left[\mathbf{1}\{Z = z\} Y \mathbf{1}\{T = t\} \mid X\right] \text{ and } h_{t,z}(X) = \mathbb{E}\left[\mathbf{1}\{Z = z\} \mathbf{1}\{T = t\} \mid X\right].$$

The CEP procedure first estimates π_z, $h_{Y,t,z}$, and $h_{t,z}$ by using nonparametric estimators $\hat{\pi}_z$, $\hat{h}_{Y,t,z}$, and $\hat{h}_{t,z}$ respectively. These estimators can be constructed based on series or local polynomial estimation. Then $Q_{t,z}$ and $P_{t,z}$ are estimated using $\hat{Q}_{t,z} = \hat{h}_{Y,t,z}/\hat{\pi}_z$ and $\hat{P}_{t,z} = \hat{h}_{t,z}/\hat{\pi}_z$. The vectors of estimators $\hat{Q}_t$ and $\hat{P}_t$, $\hat{\pi}$ are stacked in an obvious way. Let $\hat{\pi}_{t,k} = \sum_{z \in \mathscr{Z}_{t,k}} \hat{\pi}_z$. The CEP estimators for the structural parameters are defined by

$$\begin{aligned} \hat{p}_{t,k} &= \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{P}_t(X_i), & \hat{q}_{t,k} &= \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{P}_t(X_i) \hat{\pi}_{t,k}(X_i), \\ \hat{\beta}_{t,k} &= \frac{1}{\hat{p}_{t,k}} \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{Q}_t(X_i), & \hat{\gamma}_{t,k} &= \frac{1}{\hat{q}_{t,k}} \frac{1}{n} \sum_{i=1}^{n} b_{t,k} \hat{Q}_t(X_i) \hat{\pi}_{t,k}(X_i). \end{aligned}$$

[4] The terminology "conditional expectation projection" is adopted from the papers Chen et al. [2008] and Hong and Nekipelov [2010a], whereas Hahn [1998] refers to these estimators as "nonparametric imputation based estimators."

The next proposition shows that the CEP estimators are semiparametrically efficient. The result is similar in style to Hahn's (1998) Proposition 4 that the low-level regularity conditions are omitted. Instead, the proposition assumes the high-level condition that the CEP estimators are asymptotically linear, which means they are asymptotically equivalent to sample averages. More formally, an estimator $\hat{\beta}$ of β is asymptotically linear if it admits an influence function. That is, there exists an iid sequence ψ_i with zero mean and finite variance such that

$$\sqrt{n}(\hat{\beta}-\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i+o_p(1).$$

Since each element of the conditional expectations $h_{Y,t,z}$, $h_{t,z}$, and π_z can be considered as coming from a binary LATE model, the regularity conditions in Hong and Nekipelov [2010b] should work with little modification.

Proposition 3.1. *Suppose the CEP estimators are asymptotically linear, then they achieve the semiparametric efficiency bound.*

The reason that this type of estimator is efficient is well explained in Ackerberg et al. [2014]. The estimation problem here falls into their general semiparametric model, where the finite-dimensional parameter of interest is defined by unconditional moment restrictions. They show that the semiparametric two-step optimally weighted GMM estimators, the CEP estimators in this case, achieve the efficiency bound since the parameters of interest are exactly identified. Discussions related to this phenomenon can also be found in Chen and Santos [2018].

We next examine the efficient estimation of other policy-relevant parameters that can be derived from the parameters $\left(\beta_{t,k},\gamma_{t,k},p_{t,k},q_{t,k}\right)$. As an example, consider the type set $\Sigma_t\equiv\cup_{k=1}^{N_Z-1}\Sigma_{t,k}$, which is referred to as t-switchers. This subpopulation contains individuals who switch between t and other treatments when given different levels of instruments. It is a generalization of the concept of compliers in the binary LATE framework.[5] The LASF for the

[5] Recall that switchers are also illustrated in Example 3.2.

subpopulation Σ_t is given by

$$\beta_t \equiv \mathbb{E}\left[Y_t \mid S \in \Sigma_t\right] = \frac{\sum_{k=1}^{N_Z-1} \beta_{t,k} p_{t,k}}{\sum_{k=1}^{N_Z-1} p_{t,k}}.$$

Similarly, one can also define

$$\gamma_t = \mathbb{E}\left[Y_t \mid T = t, S \in \Sigma_t\right] = \frac{\sum_{k=1}^{N_Z-1} \gamma_{t,k} p_{t,k}}{\sum_{k=1}^{N_Z-1} p_{t,k}}, \tag{3.3}$$

which represents the LASF-T for the subpopulation of t-treated t-switchers.

For some subpopulations, a treatment effect can be identified. This point is already illustrated with Example 3.2 in the discussion of the identification of the usual LATE parameter. We further illustrate this point with Example 3.2.

Example 3.6 (continues = eg:3t2z)**.** *The quantity*

$$\beta_{t_3,1} - \frac{\beta_{t_1,1} p_{t_1,1} + \beta_{t_2,1} p_{t_2,1}}{p_{t_1,1} + p_{t_2,1}}$$

represents the local average treatment effect of t_3 against other treatments within the subpopulation of t_3-switchers. Analogously, the parameter

$$\gamma_{t_3,1} - \frac{\gamma_{t_3,t_1,1} q_{t_3,t_1,1} + \gamma_{t_3,t_2,1} q_{t_3,t_2,1}}{q_{t_3,t_1,1} + q_{t_3,t_2,1}}$$

is the local average treatment effect of t_3 against other treatments within the subpopulation of t_3-treated t_3-switchers.

To summarize the above examples using a general expression, let $\phi = \phi(\underline{p}, \underline{q}, \underline{\beta}, \underline{\gamma})$ be a finite-dimensional parameter, where $\phi(\cdot)$ is a known continuously differentiable function, and $\underline{p}$ is the vector containing all identifiable $p_{t,k}$'s, that is, $\underline{p} \equiv \{p_{t,k} : t \in \mathscr{T}, 1 \leq k \leq N_Z\}$. Let $\underline{q}, \underline{\beta}$, and $\underline{\gamma}$ be defined analogously. A natural estimator can be defined through the CEP estimates, $\phi(\underline{\hat{p}}, \underline{\hat{q}}, \underline{\hat{\beta}}, \underline{\hat{\gamma}})$. The delta method can help calculate the efficiency bound of ϕ and show the efficiency of $\phi(\underline{\hat{p}}, \underline{\hat{q}}, \underline{\hat{\beta}}, \underline{\hat{\gamma}})$. In fact, by Theorem 25.47 of van der Vaart [1998], we immediately have the following corollary, which shows that plug-in estimators are efficient.

Corollary 3.1. *The semiparametric efficiency bound of ϕ is given by the variance of efficient influence function*

$$\psi^{\phi} = \sum_{p \in \underline{p}} \frac{\partial \phi}{\partial p} \psi^{p} + \sum_{q \in \underline{q}} \frac{\partial \phi}{\partial q} \psi^{q} + \sum_{\beta \in \underline{\beta}} \frac{\partial \phi}{\partial \beta} \psi^{\beta} + \sum_{\gamma \in \underline{\gamma}} \frac{\partial \phi}{\partial \gamma} \psi^{\gamma} \tag{3.4}$$

where the partial derivatives are evaluated at the true parameter value. Moreover, the plug-in estimator $\phi(\underline{\hat{p}}, \underline{\hat{q}}, \underline{\hat{\beta}}, \underline{\hat{\gamma}})$, based on the CEP estimators $\underline{\hat{p}}, \underline{\hat{q}}, \underline{\hat{\beta}}, \underline{\hat{\gamma}}$, achieves the efficiency bound.

3.4 Robustness

In the previous section, the EIF is used as a tool for computing the SPEB. In this section, we directly use the EIF as the moment condition for estimation. These moment conditions are appealing because they satisfy double robustness and local robustness — the two topics of this section.

A word on notation: in the rest of the paper, we use a superscript o to signify the true value whenever necessary. For example, when both π^o and π appear, the former means the true probability while the latter denotes a generic function.

3.4.1 Double Robustness

We focus on the LASF $\beta_{t,k}$. The same analysis can be applied to the other parameters. To avoid notational burden in the main text, we drop the subscript (t,k) in $\beta_{t,k}$, $p_{t,k}$, and $b_{t,k}$, and the subscript t in P_t and Q_t.[6] It is straightforward to verify that the EIF ψ^{β} has zero mean. However, we do not want to use ψ^{β} itself as the estimating equation since it contains $1/p$ as a factor. To

[6] The full subscripts are kept in the Appendices.

deal with this problem, we simply multiply ψ^{β} by p and define

$$\begin{aligned}\psi(Y,T,Z,X,\beta,Q,P,\pi) &= p\psi^{\beta}(Y,T,Z,X,\beta,p,Q,P,\pi)\\ &= b\Big(\zeta(Z,X,\pi)\left(\iota(Y\mathbf{1}\{T=t\})-Q(X)\right)+Q(X)\Big)\\ &\quad -\beta b\Big(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P(X)\right)+P(X)\Big).\end{aligned}$$

The corresponding moment condition is

$$\mathbb{E}\left[\psi(Y,T,Z,X,\beta^{o},Q^{o},P^{o},\pi^{o})\right]=0. \tag{3.5}$$

This moment condition is doubly robust, as demonstrated in the following proposition.

Proposition 3.2 (Double Robustness). *Let (Q,P,π) be an arbitrary vector of functions and (Q^{o},P^{o},π^{o}) the true vector of conditional expectations. Then*

$$\mathbb{E}\left[\psi(Y,T,Z,X,\beta^{o},Q^{o},P^{o},\pi)\right]=0$$

and

$$\mathbb{E}\left[\psi(Y,T,Z,X,\beta^{o},Q,P,\pi^{o})\right]=0.$$

The above proposition divides the nonparametric nuisance parameters into two groups, π and (Q,P). The doubly robust moment condition is valid if either of these two groups of nuisance parameters is true. On the other hand, if the researcher uses parametric models for these nuisance parameters, then the structural parameter β can be recovered provided that at least one of the working nuisance models is correctly specified. Therefore, the doubly robust moment condition is “less demanding” on the researcher’s ability to devise a correctly specified model for the nuisance parameters. The double robustness result in Proposition 3.2 can be seen as the GLATE extension of the existing double robustness results in the binary LATE literature [e.g., Tan, 2006, Okui et al., 2012].

3.4.2 Neyman Orthogonality

The second robustness property is Neyman orthogonality. Moment conditions with this property have reduced sensitivity with respect to the nuisance parameters. Formally, Neyman orthogonality means that the moment condition has zero Gateaux derivative with respect to the nuisance parameters. The result is presented in the following proposition.

Proposition 3.3 (Neyman Orthogonality)**.** *Let (Q,P,π) be an arbitrary set of functions. For $r \in [0,1)$, define $Q^r = Q^o + r(Q - Q^o)$, $P^r = P^o + r(P - P^o)$, and $\pi^r = \pi^o + r(\pi - \pi^o)$. Suppose that $\sup_{r\in[0,1]} \left|\frac{\partial}{\partial r}\psi(Y,T,Z,X,\beta,Q^r,P^r,\pi^r)\right|$ is integrable, then*

$$\frac{\partial}{\partial r}\mathbb{E}\left[\psi(Y,T,Z,X,\beta,Q^r,P^r,\pi^r)\right]\Big|_{r=0} = 0,$$

where β does not need to be the true parameter value.

In many econometrics models, double robustness and Neyman orthogonality come in pairs. Discussions about their general relationships can be found in Chernozhukov et al. [2016]. In practice, double robustness is often used for parametric estimation, as previously explained, whereas Neyman orthogonality is used in estimation with the presence of possibly high-dimensional nuisance parameters.

Next, we apply the double/debiased machine learning (DML) method developed by Chernozhukov et al. [2018] to the moment condition (3.5). This estimation method works even when the nuisance parameter space is complex enough that the traditional assumptions, e.g., Donsker properties, are no longer valid.[7] The implementation details are explained below.

The nuisance parameters Q, P, and π are estimated using a cross-fitting method: Take

[7] In two-step semiparametric estimations, Donsker properties are usually required so that a suitable stochastic equicontinuity condition is satisfied. See, for example, Assumption 2.5 in Chen et al. [2003].

an L-fold random partition of the data such that the size of each fold is n/L. For $l = 1,\cdots,L$, let I_l denote the set of observation indices in the lth fold and $I_l^c = \bigcup_{l'\neq l} I_{l'}$ the set of observation indices not in the lth fold. Define $\check{Q}^l$, $\check{P}^l$, and $\check{\pi}^l$ to be the estimates constructed by using data from I_l^c. The DML estimator of β is constructed following the moment condition (3.5):[8]

$$\check{\beta} = \frac{\sum_{l=1}^{L}\sum_{i\in I_l} b\Big(\zeta(Z_i,X_i,\check{\pi}^l)\Big(\iota(Y_i\mathbf{1}\{T_i=t\}) - \check{Q}^l(X_i)\Big) + \check{Q}^l(X_i)\Big)}{\sum_{l=1}^{L}\sum_{i\in I_l} b\Big(\zeta(Z_i,X_i,\check{\pi}^l)\Big(\iota\mathbf{1}\{T_i=t\} - \check{P}^l(X_i)\Big) + \check{P}^l(X_i)\Big)}. \tag{3.6}$$

To conduct inference, we also need an estimate for the asymptotic variance of $\check{\beta}$, which we denote by σ^2. The asymptotic variance equals to the expectation of the squared efficient influence function: $\sigma^2 = \mathbb{E}\left[\psi^\beta\right]^2 = \mathbb{E}[\psi^2]/p^2$. We first estimate p by using the cross-fitting method, which is essentially given by the denominator of (3.6):

$$\check{p} = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l} b\Big(\zeta(Z_i,X_i,\check{\pi}^l)\Big(\iota\mathbf{1}\{T_i=t\} - \check{P}^l(X_i)\Big) + \check{P}^l(X_i)\Big). \tag{3.7}$$

Then the asymptotic variance can be estimated by

$$\begin{aligned}\check{\sigma}^2 &= \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left(\psi^\beta\left(Y_i,T_i,Z_i,X_i,\check{\beta},\check{p},\check{Q}^l,\check{P}^l,\check{\pi}^l\right)\right)^2 \\ &= \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left(\psi\Big(Y_i,T_i,Z_i,X_i,\check{\beta},\check{Q}^l,\check{P}^l,\check{\pi}^l\Big)/\check{p}\right)^2.\end{aligned}$$

We want to establish the convergence results for the DML estimator uniformly over a class of data generating processes (DGPs) defined as follows. For any two constants $c_1 > c_0 > 0$, let $\mathscr{P}(c_1,c_0)$ be the set of joint distributions of (Y,T,Z,X) such that

(i) $p \in [c_0,1]$,

(ii) $\mathbb{E}[\psi^2], \pi_z^o(X) \geq c_0, z \in \mathscr{Z}$, and $|Y\mathbf{1}\{T=t\}|, |Y\mathbf{1}\{T=t\} - Q_t^o(X)| \leq c_1$.

[8] This is the DML2 estimator defined in Chernozhukov et al. [2018]. Another estimator, the DML1 estimator, is proposed in the same paper. We do not study the DML1 estimator since it is asymptotically equivalent to DML2, and the authors generally recommend DML2.

The first condition excludes the case where β is weakly identified (when p can be arbitrarily close to zero). Inference under weak identification is studied in the next section. The following theorem establishes the asymptotic properties of the DML estimation procedure. In particular, the estimator achieves the SPEB.

Theorem 3.4. *Let Assumptions 3.1 and 3.2 hold. Assume the following conditions on the nuisance parameter estimators* $(\check{Q}^l, \check{P}^l, \check{\pi}^l)$*:*

(i) For $z \in \mathscr{Z}$*,* $|\check{Q}^l|$ *is bounded,* $\check{P}^l_z$ *and* $\check{\pi}^l_z \in [0,1]$*, and* $\check{\pi}^l_z$ *is bounded away from zero.*

(ii) $\max_{z \in \mathscr{Z}} \left(\|\hat{Q} - Q^o\|_2 \vee \|\hat{P} - P^o\|_2 \vee \|\hat{\pi} - \pi^o\|_2 \right) = o_p\left(n^{-1/4}\right)$.

Then the estimator $\check{\beta}$ *obeys that*

$$\sigma^{-1}\sqrt{n}\left(\check{\beta} - \beta\right) \Rightarrow N(0,1),$$

uniformly over the DGPs in $\mathscr{P}(c_0, c_1)$*. Moreover, the above convergence result continues to hold when* σ *is replaced by the estimator* $\check{\sigma}$*.*

The proof verifies the conditions of Theorem 3.1 in Chernozhukov et al. [2018]. The essential restriction is on the uniform convergence rate for the estimators of the nuisance parameters. In low-dimensional settings, one can consider the local polynomial regression for estimation of the conditional expectations. Under suitable conditions [Hansen, 2008, Masry, 1996], the uniform convergence rate of the local polynomial estimators is $(\log n/n)^{2/(d_X+4)}$, which is $o(n^{-1/4})$ if $d_X \leq 3$. In high-dimensional settings, as pointed out by Chernozhukov et al. [2018], the rate $o(n^{-1/4})$ is often available for common machine learning methods under structured assumptions on the nuisance parameters.[9] This means that the asymptotic normality of the DML estimator continues to hold.

[9] This includes the LASSO method under sparsity of the nuisance space. See, for example, Bühlmann and Van De Geer [2011], Belloni and Chernozhukov [2011], and Belloni and Chernozhukov [2013]. However, Chernozhukov et al. [2018] also indicate that to prove that machine learning methods achieve the $o(n^{-1/4})$ rate, one will eventually have to use related entropy conditions.

Theorem 3.4 can be directly used to conduct inference on β. Confidence regions can be constructed by inverting the usual t-tests. These confidence regions are uniformly valid since the convergence results in the above theorem hold uniformly over $\mathscr{P}$. In the next section, we explain why uniform validity is crucial when dealing with weak identification issues.

3.5 Weak Identification

The convergence result established in Theorem 3.4 is uniform over the set of DGPs with type probability p bounded away from zero. However, the identification of β would be weak in the case where p can be arbitrarily close to zero. This leads to distortion of the uniform size of the test and poor asymptotic approximation in finite-sample settings. This section studies this weak identification issue and proposes an inference procedure that is robust against such a problem.

We begin with a heuristic illustration of the weak identification problem. To ease notation, define $\upsilon = \beta p$ and

$$\check{\upsilon} = \check{\beta}\check{p} = \sum_{l=1}^{L}\sum_{i\in I_l} b\Big(\zeta(Z_i, X_i, \check{\pi}^l)\Big(\iota(Y_i\mathbf{1}\{T_i = t\}) - \check{Q}^l(X_i)\Big) + \check{Q}^l(X_i)\Big).$$

After a simple calculation, we can write

$$\check{\beta} - \beta = \frac{\sqrt{n}(\check{\upsilon} - \upsilon) - \beta\sqrt{n}(\check{p} - p)}{\sqrt{n}(\check{p} - p) + \sqrt{n}p}.$$

In the above expression, we can interpret the estimation errors $\sqrt{n}(\check{\upsilon} - \upsilon)$ and $\sqrt{n}(\check{p} - p)$ as the noises, while the signal is the term $\sqrt{n}p$. Under the usual asymptotics where $p > 0$ is fixed, the noise terms are bounded in probability, whereas the signal term $\sqrt{n}p \to \infty$. Hence, the signal dominates the noise, and the estimator $\check{\beta}$ is consistent. However, under asymptotics with a drifting sequence $p = p_n \to 0$ and $\sqrt{n}p$ converging to a finite constant, the signal and the noise are of the same magnitude, which results in the inconsistency of $\check{\beta}$. This problem is the weak identification issue. In the weak IV literature, a common measure of identification strength is the so-called *concentration parameter*. In our case, the concentration parameter is given by $\sqrt{n}p$ where $\sqrt{n}p \to \infty$ corresponds to strong identification, and identification is weak when the limit of $\sqrt{n}p$ is finite.

While weak identification is a finite-sample issue, it is formalized using the asymptotic framework. However, the illustration above using asymptotics under drifting sequences is not meant to model DGPs that vary with the sample size n. Instead, it is a tool used to detect the lack of uniform convergence. In fact, controlling the uniform size of the test is the key to solving weak identification problems.[10] Formally, the uniform size of a test is the large sample limit of the supremum of the rejection probability under the null hypothesis, where the supremum is taken over the nuisance parameter space. When testing a null hypothesis on β in the GLATE model, the supremum mentioned above is taken over all values of $p > 0$. That is, a desirable test should have rejection probability under the null converge to the nominal size uniformly over $p \in (0,1]$. From the previous discussion, we can see that the uniform size can not be controlled using the usual t-statistic $\sqrt{n}(\check{\beta} - \beta)/\check{\sigma}$. This failure of uniform convergence, however, does not conflict with Theorem 3.4, where the uniform convergence of $\check{\beta}$ is established only after restricting p to be bounded away from zero.

Inference procedures that are robust against weak identification can be obtained by directly imposing the null hypothesis in the construction of the test statistic. One such example is the well-known Anderson-Rubin (AR) statistic in the weak IV literature. Its idea can be generalized to the GLATE model. We first consider testing the two-sided hypothesis $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. To control the uniform size of the test, we need the test statistic to converge uniformly on the parameter space where (1) $\beta = \beta_0$, and (2) p is allowed to be arbitrarily close to zero. A null-restricted t-statistic can be obtained as follows. Notice that when $p > 0$, $\beta = \beta_0$

[10] See, for example, Imbens and Manski [2004], Mikusheva [2007], and Andrews et al. [2020].

is equivalent to

$$0 = \upsilon - \beta_0 p = \mathbb{E}\left[\psi(Y,T,Z,X,\beta_0,Q,P,\pi)\right]. \tag{3.8}$$

Its estimate can be written as

$$\check{\upsilon} - \beta_0 \check{p} = (\check{\upsilon} - \upsilon) - \beta(\check{p} - p) + (\beta - \beta_0)p. \tag{3.9}$$

Under the null hypothesis $\beta = \beta_0$, the above estimate does not depend on the concentration parameter $\sqrt{n}p$ and consists only of the noise terms $\check{\upsilon} - \upsilon$ and $\check{p} - p$, whose uniform convergence can be established directly.

For implementation, this test statistic can be obtained as a straightforward application of the DML procedure described in the previous section to the moment condition (3.8). As a consequence of Proposition 3.3, the above moment condition satisfies the Neyman orthogonality condition regardless of the true value of β. More specifically, the null-restricted t-statistic is defined to be

$$\check{\rho} = \sqrt{n}(\check{\upsilon} - \beta_0 \check{p})/\check{\sigma}_\psi,$$

where

$$\check{\sigma}_\psi^2 = \frac{1}{n}\sum_{l=1}^{L}\sum_{i \in I_l} \psi(Y_i,T_i,Z_i,X_i,\beta_0,\check{Q}^l,\check{P}^l,\check{\pi}^l)^2.$$

The corresponding test of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ rejects for large values of $|\check{\rho}|$.

The same methodology can be applied to testing one-sided hypothesis $H_0 : \beta \leq \beta_0$ versus $H_1 : \beta > \beta_0$. Under the null hypothesis, $(\beta - \beta_0)p$ is non-positive, suggesting that the test should reject for large values of $\check{\rho}$. Notice that this relies on knowing the sign of p due to the GLATE model structure. This restriction on the sign of p is similar to knowing the first-stage sign in the linear IV model, which is studied by Andrews and Armstrong [2017] in the context of unbiased estimation.

We now define the set of DGPs that allows p to be arbitrarily close to zero. For any two constants $c_1 > c_0 > 0$, let $\mathscr{P}^{\mathrm{WI}}(c_0, c_1)$ be the set of joint distributions of (Y, T, Z, X) such that

(i) $p \in (0, 1]$,

(ii) $\mathbb{E}[\psi^2], \pi_z^o(X) \geq c_0, z \in \mathscr{Z}$, and $\left|Y\mathbf{1}\{T = t\}\right|, |Y\mathbf{1}\{T = t\} - Q_t^o(X)| \leq c_1$.

For any $\beta' \in \mathbb{R}$, let $\mathscr{P}_{\beta'}^{\mathrm{WI}}(c_0, c_1)$ be the subset of $\mathscr{P}^{\mathrm{WI}}(c_0, c_1)$ in which the true value of the parameter β is β'. In particular, $\mathscr{P}_{\beta_0}^{\mathrm{WI}}(c_0, c_1)$ denotes the subset where the null hypothesis is true. The superscript "WI" denotes weak identification. The difference between $\mathscr{P}(c_0, c_1)$ and $\mathscr{P}^{\mathrm{WI}}(c_0, c_1)$ is that $\mathscr{P}^{\mathrm{WI}}(c_0, c_1)$ allows the type probability p to be arbitrarily small, whereas the type probabilities in $\mathscr{P}(c_0, c_1)$ are uniformly bounded away from zero. Denote $\mathscr{N}_\nu$ as the νth quantile of the standard normal distribution. The following theorem establishes that the above testing procedures have uniformly correct sizes and are consistent.

Theorem 3.5. *Suppose the conditions on the nuisance parameter estimates in Theorem 3.4 hold. Let $\alpha \in (0, 1)$ be the nominal size of the tests.*

(i) The test that rejects $H_0 : \beta = \beta_0$ in favor of $H_0 : \beta \neq \beta_0$ when $|\check{\rho}| > \mathscr{N}_{1-\frac{\alpha}{2}}$ has (asymptotically) uniformly correct size and is consistent. That is,

$$\sup\left\{\mathbb{P}_P\left(|\check{\rho}| > \mathscr{N}_{1-\frac{\alpha}{2}}\right) : P \in \mathscr{P}_{\beta_0}^{WI}(c_0, c_1)\right\} \to \alpha$$

and

$$\mathbb{P}_P\left(|\check{\rho}| > \mathscr{N}_{1-\frac{\alpha}{2}}\right) \to 1, P \in \mathscr{P}_{\beta}^{WI}(c_0, c_1), \beta \neq \beta_0.$$

(ii) The test that rejects $H_0 : \beta \leq \beta_0$ in favor of $H_0 : \beta > \beta_0$ when $\check{\rho} > \mathscr{N}_{1-\alpha}$ has (asymptoti-

cally) uniformly correct size and is consistent. That is,

$$\sup\Big\{\mathbb{P}_P\Big(\check{\rho} > \mathscr{N}_{1-\alpha}\Big) : P \in \mathscr{P}_\beta^{WI}(c_0,c_1), \beta \leq \beta_0\Big\} \to \alpha$$

and

$$\mathbb{P}_P\Big(\check{\rho} > \mathscr{N}_{1-\alpha}\Big) \to 1, P \in \mathscr{P}_\beta^{WI}(c_0,c_1), \beta > \beta_0.$$

3.6 Empirical Application

In this section, we apply the theoretical results to data from the Oregon Health Insurance Experiment [Finkelstein et al., 2012] and examine the effects on the health of different sources of health insurance. The experiment is conducted by the state of Oregon between March and September 2008. A series of lottery draws were administered to award the participants the option of enrolling in the Oregon Health Plan Standard, which is a Medicaid expansion program available for Oregon adult residents that have limited income. Follow-up surveys were sent out in several waves to record, among many variables, the participants' insurance plan and health status. Finkelstein et al. [2012] obtain the effects of insurance coverage by using a LATE model. We apply the GLATE model can study the effect heterogeneity across different sources of insurance.

According to the data, many lottery winners did not choose to participate in the Medicaid program. Instead, they went with other insurance plans or chose not to have any health insurance. Based on this observation, we can set up the GLATE model. The instrument Z is the binary lottery that determines whether an individual is selected. The covariates X include the number of household members and survey waves. Given X, Z is randomly assigned [Finkelstein et al., 2012, p1071].[11] The treatment T is the insurance plan, which contains three categories: Medicaid (*m*), non-Medicaid insurance plans (*nm*), and no health insurance (*no*). The second category includes

[11] Though the covariates are discrete, the methods developed in this paper are still different from linear regressions in Finkelstein et al. [2012].

Medicare, private plans, employer plans, and other plans. The counterfactual health plan choices under different lottery results are the variables T_0 and T_1. The unordered monotonicity condition requires that any participant who changes insurance plan due to winning the lottery does so to enroll in the Medicaid program.

The above setup is the same as Example 3.2. We follow the terminologies in Kline and Walters [2016] and define the following six type sets by their counterfactual insurance plan choices:

1. *no*-never takers: $S \in \Sigma_{no,2} = \{s_1\}$, $T_0 = T_1 = no$;
2. *nm*-never takers: $S \in \Sigma_{nm,2} = \{s_2\}$, $T_0 = T_1 = nm$;
3. always takers: $S \in \Sigma_{m,2} = \{s_3\}$, $T_0 = T_1 = m$;
4. *no*-compliers: $S \in \Sigma_{no,1} = \{s_4\}$, $T_0 = no$, $T_1 = m$;
5. *nm*-compliers: $S \in \Sigma_{nm,1} = \{s_5\}$, $T_0 = nm$, $T_1 = m$;
6. compliers: $S \in \Sigma_{m,1} = \{s_4, s_5\}$, $T_0 \neq m$, $T_1 = m$.

The two groups of never takers choose not to join Medicaid regardless of the offer. Always takers manage to enroll in Medicaid even without an offer. The *no*- and *nm*- compliers switch to Medicaid from no insurance plan and other plans, respectively, upon winning the lottery. Combining these two groups gives the larger set of compliers.

Table 3.1 shows the estimated probabilities of the six types.[12] We can see that half of the population are *no*-never takers, who are never covered by any insurance plan. The compliers make up around one-fifth of the population. There are effectively no *nm*-compliers, meaning that the experiment does not crowd out other insurance plan choices. These findings are consistent with Finkelstein et al. [2012].

[12] We use the data from the 12-month survey. After taking care of the missing values, we are left with 23290 observations. For cross-fitting, we choose $L = 10$.

Table 3.1. Estimated probability of different types.

Type	Probability	Estimate (se)
no-never takers	$p_{no,2}$	.492 (.046)
nm-never takers	$p_{nm,2}$	.208 (.018)
always takers	$p_{m,2}$	.116 (.018)
no-compliers	$p_{no,1}$	.197 (.059)
nm-compliers	$p_{nm,1}$	.010 (.024)
compliers	$p_{m,1}$	.208 (.060)

The outcome of interest Y is health status, which is (inversely) measured by the number of days (out of past 30) when poor health impaired regular activities.[13] The potential outcomes are denoted by Y_{no}, Y_{nm}, and Y_m. By Theorem 3.1, we can identify the distribution of Y_{no} for *no*-never takers and *no*-compliers, the distribution of Y_{nm} for *nm*-never takers and *nm*-compliers, and the distribution of Y_{nm} for always takers and compliers. Table 3.2 reports the estimated LASFs.[14] We can clearly see a pattern of self-selection into the treatment. For example, when there is no insurance coverage, the potential health status of *no*-compliers is worse than *no*-never takers and therefore choose to enroll in Medicaid.

Table 3.2. Estimated LASFs.

Type	Treatment	LASF	Estimate (se)
no-never takers	*no*	$\beta_{no,2}$	6.78 (1.19)
nm-never takers	*nm*	$\beta_{nm,2}$	7.74 (1.05)
always takers	*m*	$\beta_{m,2}$	9.96 (1.75)
no-compliers	*no*	$\beta_{no,1}$	11.50 (2.92)
compliers	*m*	$\beta_{m,1}$	0.48 (3.42)

3.7 Concluding Remarks

In this paper, we considered the estimation of the causal parameters, LASF and LASF-T, in the GLATE model by using the EIF. The proposed DML estimator satisfies the SPEB and can

[13] Other types of outcomes are also studied by Finkelstein et al. [2012], including health care utilization and financial strain. Here we only focus on health status for simplicity.

[14] The LASF $\beta_{nm,1}$ is excluded because there are few *nm*-compliers as reported in Table 3.1.

be applied in situations, such as high-dimensional settings, where Donsker properties fail. For inference, we proposed generalized AR tests robust against weak identification issues. Currently, empirical researchers use the TSLS and control the covariates linearly in models with multi-valued treatments and instruments. This linear specification does not have LATE interpretation, as pointed out by Blandhol et al. [2022]. Therefore, we advocate using the semiparametric methods studied by this paper in those cases.

3.8 Acknowledgements

Chapter 3, in full, has been submitted for publication. Xie, Haitian. “Efficient and Robust Estimation of the Generalized LATE Model.” The dissertation author was the primary investigator and author of this material.

Appendix A

Appendix for Chapter 1

A.1 Proof of identification results

In this section we prove the identification results including Lemma 1.1, Lemma 1.2, Theorem 1.1, Corollary 1.1, and Corollary 1.2.

Proof of Lemma 1.1. We first prove the second equality in Equation (1.3). The conditional distribution function of T given R is

$$\begin{aligned} F_{T|R}(t|r) &= \mathbb{P}(T \le t|R=r)\mathbf{1}\{r<\bar{r}\} + \mathbb{P}(T \le t|R=r)\mathbf{1}\{r \ge \bar{r}\} \\ &= \mathbb{P}(U_0 \le m_0^{-1}(r,t)|R=r)\mathbf{1}\{r<\bar{r}\} + \mathbb{P}(U_1 \le m_1^{-1}(r,t)|R=r)\mathbf{1}\{r \ge \bar{r}\} \\ &= F_{U_0|R}(m_0^{-1}(r,t)|r)\mathbf{1}\{r<\bar{r}\} + F_{U_1|R}(m_1^{-1}(r,t)|r)\mathbf{1}\{r \ge \bar{r}\}, \end{aligned}$$

where the last line follows from the monotonicity of m_0 and m_1. Therefore, we have

$$\begin{aligned} F_{T|R}(T|R) &= F_{U_0|R}(m_0^{-1}(R,T)|R)\mathbf{1}\{R<\bar{r}\} + F_{U_1|R}(m_1^{-1}(R,T)|R)\mathbf{1}\{R \ge \bar{r}\} \\ &= F_{U_0|R}(U_0|R)\mathbf{1}\{R<\bar{r}\} + F_{U_1|R}(U_1|R)\mathbf{1}\{R \ge \bar{r}\} = U. \end{aligned}$$

For (i) of Lemma 1.1, take any $u \in [0,1]$ and $r < \bar{r}$, we have

$$\mathbb{P}(U \le u|R=r) = \mathbb{P}(F_{U_0|R}(U_0|r) \le u|R=r) = u.$$

Similarly, we can show that $\mathbb{P}(U \leq u|R = r) = u$ when $r \geq \bar{r}$. Therefore, $U|R$ follows the uniform distribution. Then the second argument follows from the monotonicity of h with respect to the second argument. The statements (ii) and (iii) are straightforward from the definition of h_0 and h_1 and the monotonicity and continuity assumptions. For (iv), notice that for $r < \bar{r}$,

$$\begin{aligned}
F_{\varepsilon|U,R}(e|u,r) &= \mathbb{P}(\varepsilon \leq e|U = u, R = r) \\
&= \mathbb{P}(\varepsilon \leq e|F_{U_0|R}(U_0|r) = u, R = r) \\
&= \mathbb{P}(\varepsilon \leq e|U_0 = F^{-1}_{U_0|R}(u|r), R = r) \\
&= F_{\varepsilon|U_0,R}\left(e|F^{-1}_{U_0|R}(u|r), r\right),
\end{aligned}$$

where the third equality follows from the strict monotonicity of $F^{-1}_{U_0|R}(u|r)$ in u imposed in Assumption 1.2(ii). Similarly, we can show that for $r \geq \bar{r}$,

$$F_{\varepsilon|U,R}(e|u,r) = \mathbb{P}(\varepsilon \leq e|U = u, R = r) = F_{\varepsilon|U_1,R}\left(e|F^{-1}_{U_1|R}(u|r), r\right).$$

Combining the two equations together, we have

$$F_{\varepsilon|U,R}(e|u,r) = \begin{cases} F_{\varepsilon|U_0,R}\left(e|F^{-1}_{U_0|R}(u|r), r\right), r < \bar{r}, \\ F_{\varepsilon|U_1,R}\left(e|F^{-1}_{U_1|R}(u|r), r\right), r \geq \bar{r}. \end{cases}$$

By Assumption 1.2(iii), we know that $F_{\varepsilon|U,R}(e|u,r)$ is strictly increasing in the first argument e. By Bayes rule, the rank similarity condition in Assumption 1.3 implies that $U_0|R = \bar{r}^-$ has the same distribution as $U_1|R = \bar{r}^+$, and $\varepsilon|U_0, R = \bar{r}^-$ has the same distribution as $\varepsilon|U_1, R = \bar{r}^+$. Then we have

$$\lim_{r \uparrow \bar{r}} F_{\varepsilon|U,R}(e|u,r) = F_{\varepsilon|U_0,R}\left(e|F^{-1}_{U_0|R}(u|\bar{r}), \bar{r}\right) = F_{\varepsilon|U_1,R}\left(e|F^{-1}_{U_1|R}(u|\bar{r}), \bar{r}\right) = \lim_{r \downarrow \bar{r}} F_{\varepsilon|U,R}(e|u,r).$$

□

Lemma A.1. *Let $f : (x,y) \mapsto z$ be a real-valued bivariate function defined on a compact set in $\mathbb{R} \times \mathbb{R}^d$. Assume that f is continuous on its entire domain and strictly increasing in the first argument x. Let f^{-1} denote the inverse of f with respect to the first argument. Then f^{-1} is continuous on its domain and strictly increasing in the first argument.*

Therefore, under Assumptions 1.1 and 1.2, the inverse of g, h_0 and h_1 (with respect to the last argument), which are respectively $g^{-1}, (h_0)^-$ and $(h_1)^-$, are all continuous and strictly increasing with respect to the last argument.

Proof of Lemma A.1. Fix any (z_0, y_0), we want to show that $f^{-1}(z_0, \cdot)$ is continuous at (z_0, y_0). If not, then there exists $\delta > 0$ and a sequence $\{(z_k, y_k)\}$ such that $\|(z_k, y_k) - (z_0, y_0)\| < 1/k$ but

$$|f^{-1}(z_0, y_k) - f^{-1}(z_0, y_0)| > \delta.$$

Denote $x_k = f^{-1}(z_k, y_k)$ and $x_0 = f^{-1}(z_0, y_0)$. Because the sequence $\{x_k\}$ lies in a compact set, it has a convergent subsequence. Without loss of generality, we assume $\{x_k\}$ itself is converging. Then $\lim x_k \neq x_0$. However, by the continuity of f,

$$f(\lim x_k, y_0) = f(\lim x_k, \lim y_k) = \lim f(x_k, y_k) = z_0 = f(x_0, y_0).$$

This leads to a contradiction since $f(\cdot, y_0)$ is strictly increasing.

To show that f^{-1} is strictly increasing with respect to the first argument, take any y and $z_1 > z_0$. If $f^{-1}(z_1, y) \leq f^{-1}(z_0, y)$, then $z_1 = f(f^{-1}(z_1, y), y) \leq f(f^{-1}(z_0, y), y) = z_0$, which leads to a contradiction. □

Proof of Lemma 1.2. By the definition of $F^-_{Y|T,R}$ and the monotonicity and continuity of g^* and

h_0,

$$\begin{aligned}\text{LHS of (1.5) with } g = g^* &= \lim_{r\uparrow\bar{r}} F_{Y|T,R}(g^*(h_0(r,u),r,e)|h_0(r,u),r)\\ &= \lim_{r\uparrow\bar{r}} \mathbb{P}(Y \leq g^*(h_0(r,u),r,e)|T = h_0(r,u), R = r)\\ &= \lim_{r\uparrow\bar{r}} \mathbb{P}(g^*(h_0(r,u),r,\varepsilon) \leq g^*(h_0(r,u),r,e)|T = h_0(r,u), R = r)\\ &= \lim_{r\uparrow\bar{r}} \mathbb{P}(\varepsilon \leq e|U = u, R = r)\\ &= F_{\varepsilon|U,R}(e|u,r)\end{aligned}$$

where the last line follows from the continuity of $F_{\varepsilon|U,R}(e|u,r)$ with respect to the last argument r (Lemma 1.1). Similarly, we can show that the RHS of (1.5) is equal to $F_{\varepsilon|U,R}(e|u,\bar{r})$. Then the result follows. □

Proof of Theorem 1.1. Denote

$$\mathscr{T}^{\times} = \{h_0(\bar{r},u) : h_0(\bar{r},u) = h_1(\bar{r},u) \in [t_0', t_0''] \cap [t_1', t_1''], u \in [0,1]\}.$$

By Assumption 1.5(ii), $\mathscr{T}^{\times}$ is nonempty and finite. Then $[t_0', t_0''] \cap [t_1', t_1'']$ is a closed interval with nonempty interior.[1] Let

$$\inf([t_0', t_0''] \cap [t_1', t_1'']) = t_1 \leq t_2 \leq \cdots \leq t_L = \sup([t_0', t_0''] \cap [t_1', t_1''])$$

denote the unique elements of $\mathscr{T}^{\times} \cup \{\inf([t_0', t_0''] \cap [t_1', t_1'']), \sup([t_0', t_0''] \cap [t_1', t_1''])\}$.

Here is the strategy of the proof. For each $g \in \mathscr{G}$ that satisfies Equation (1.5), define

$$\tilde{\lambda}^g(t,e) = g^{-1}(t,\bar{r},g^*(t,\bar{r},e)). \tag{A.1}$$

[1] Notice that if $[t_0', t_0''] \cap [t_1', t_1'']$ is a singleton, then $\mathscr{T}^{\times}$ is empty.

The goal is to show that $\tilde{\lambda}^g$ is constant as a function of t for every $e \in \mathscr{E}$. We proceed in five steps. Step 1 derives some useful properties of the function $\tilde{\lambda}^g$, including an important identity, Equation (A.2). Step 2 shows that $\tilde{\lambda}^g$ is constant in t on the interval (t_1, t_2). Step 3 shows that $\tilde{\lambda}^g$ is constant in t on the entire region $[t_1, t_L] = [t'_0, t''_0] \cap [t'_1, t''_1]$. Step 4 further expands this constancy to $[t'_0, t''_0] \cup [t'_1, t''_1]$. Step 5 concludes.

Step 1. Since g^* and g^{-1} are continuous and are strictly increasing in the last argument, $\tilde{\lambda}^g$ is also continuous and strictly increasing in the last argument. Also, $F^-_{Y|T,R}(\cdot|t,\bar{r})$ is strictly increasing since

$$F^-_{Y|T,R}(y|t,\bar{r}) = F_{\varepsilon|U,R}((g^*)^{-1}(t,\bar{r},y)|h_0^{-1}(\bar{r},t),\bar{r})$$

is strictly increasing in y (Assumption 1.2(ii)).

Notice that $g(t,\bar{r},\tilde{\lambda}^g(t,e)) = g^*(t,\bar{r},e)$. Then for any $e \in \mathscr{E}$ and $u \in [0,1]$,

$$\begin{aligned}
&F^-_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},\tilde{\lambda}^g(h_0(\bar{r},u),e))|h_0(\bar{r},u),\bar{r}) \\
=&F^-_{Y|T,R}(g^*(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}) \\
=&F^+_{Y|T,R}(g^*(h_1(\bar{r},u),\bar{r},e)|h_1(\bar{r},u),\bar{r}) \\
=&F^+_{Y|T,R}(g(h_1(\bar{r},u),\bar{r},\tilde{\lambda}^g(h_1(\bar{r},u),e))|h_1(\bar{r},u),\bar{r}),
\end{aligned}$$

where the second inequality follows from Lemma 1.2. The above equality implies

$$\tilde{\lambda}^g(h_0(\bar{r},u),e) = \tilde{\lambda}^g(h_1(\bar{r},u),e). \tag{A.2}$$

To see that, suppose there exists e and u such that $\tilde{\lambda}^g(h_0(\bar{r},u),e) \neq \tilde{\lambda}^g(h_1(\bar{r},u),e)$. Since g

satisfies Condition (1.5),

$$\begin{aligned}
&F^{-}_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},\tilde{\lambda}^g(h_1(\bar{r},u),e))|h_0(\bar{r},u),\bar{r})\\
=&F^{+}_{Y|T,R}(g(h_1(\bar{r},u),\bar{r},\tilde{\lambda}^g(h_1(\bar{r},u),e))|h_1(\bar{r},u),\bar{r})\\
=&F^{-}_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},\tilde{\lambda}^g(h_0(\bar{r},u),e))|h_0(\bar{r},u),\bar{r}),
\end{aligned}$$

which violates the fact that $F^{-}_{Y|T,R}(g(h_0(\bar{r},u),\bar{r},\cdot)|h_0(\bar{r},u),\bar{r})$ is strictly increasing.

Step 2. Consider the interval (t_1,t_2). By construction, $\{t_1,t_2\}\cap\mathscr{T}^{\times}\neq\emptyset$. Notice that over (t_1,t_2), $h_0^{-1}(\bar{r},\cdot)$ and $h_1^{-1}(\bar{r},\cdot)$ do not intersect. Then by continuity, one of them is always strictly greater than the other. The goal is to show that $\tilde{\lambda}^g(t,e)$ is constant as a function of t over (t_1,t_2). There are four cases to consider, depending on whether $t_1\in\mathscr{T}^{\times}$ or $t_2\in\mathscr{T}^{\times}$ and whether $h_0^{-1}(\bar{r},\cdot)$ is strictly greater or smaller than $h_1^{-1}(\bar{r},\cdot)$ over (t_1,t_2).

We first focus on the case of $t_1\in\mathscr{T}^{\times}$ and $h_0^{-1}(\bar{r},\cdot)<h_1^{-1}(\bar{r},\cdot)$ over (t_1,t_2). The other cases are essentially the same. Define a mapping $\pi(t)=h_1(\bar{r},h_0^{-1}(\bar{r},t))$. Such a mapping π maps the interval (t_1,t_2) back to itself. To see that, we first notice that $\pi(t)$ is less than t for any $t\in(t_1,t_2)$ since

$$\pi(t)=h_1(\bar{r},h_0^{-1}(\bar{r},t))\leq h_1(\bar{r},h_1^{-1}(\bar{r},t))=t<t_2.$$

Suppose $\pi(t)\leq t_1$, then

$$\begin{aligned}
h_0^{-1}(\bar{r},t_1)&=h_1^{-1}(\bar{r},t_1)\\
&\leq h_1^{-1}(\bar{r},\pi(t))\\
&=h_0^{-1}(\bar{r},t),
\end{aligned}$$

where the first line follows from $t_1\in\mathscr{T}^{\times}$, the second line follows from the monotonicity of h_1^{-1} and $\pi(t)\leq t_1$, and the last line follows from the definition of π. This contradicts the strict monotonicity of $h_0^{-1}(\bar{r},\cdot)$ since $t_1 < t$.

Now pick any $\tilde{t}_0 \in (t_1,t_2)$, the recursive sequence $\tilde{t}_{k+1} = \pi(\tilde{t}_k)$ is well-defined. This sequence is non-increasing and bounded below by t_1. Therefore, $\lim_{k\to\infty}\tilde{t}_k$ exists and lies in the interval $[t_1,t_2)$. By the continuity of h_0^{-1} and h_1^{-1},

$$\begin{aligned} h_0^{-1}(\bar{r},\lim\tilde{t}_k) &= \lim h_0^{-1}(\bar{r},\tilde{t}_k) \\ &= \lim h_1^{-1}(\bar{r},\pi(\tilde{t}_k)) \\ &= \lim h_1^{-1}(\bar{r},\tilde{t}_{k+1}) \\ &= h_1^{-1}(\bar{r},\lim\tilde{t}_{k+1}), \end{aligned}$$

where the second line follows from the definition of π and the third line follows from the construction of the sequence $\{\tilde{t}_k\}$. Then it must be true that $\lim\tilde{t}_k = t_1$ since we are studying the case where $h_0^{-1}(\bar{r},\cdot) < h_1^{-1}(\bar{r},\cdot)$ over (t_1,t_2).

Equation (A.2) implies that $\tilde{\lambda}^g(\cdot,e)$ is invariant with respect to the transformation π:

$$\tilde{\lambda}^g(\pi(t),\bar{r},e) = \tilde{\lambda}^g(t,e),$$

for every $t \in [t_0', t_0'']$ and $e \in \mathscr{E}$. Then $\tilde{\lambda}^g$ is invariant along the sequence $\{\tilde{t}_k\}$. By the continuity of $\tilde{\lambda}^g$,

$$\tilde{\lambda}^g(\tilde{t}_0,e) = \lim_{k\to\infty}\tilde{\lambda}^g(\tilde{t}_0,e) = \lim_{k\to\infty}\tilde{\lambda}^g(\tilde{t}_k,e) = \tilde{\lambda}^g(\lim_{k\to\infty}\tilde{t}_k,e) = \tilde{\lambda}^g(t_1,e) = \lambda^g(e),$$

where the first equality holds since $\tilde{\lambda}^g(\tilde{t}_0,\bar{r},e)$ is constant with respect to k, the second equality holds since $\tilde{\lambda}^g$ is invariant along the sequence $\{\tilde{t}_k\}$, the third equality follows from the continuity of $\tilde{\lambda}^g$, and the last equality is the definition of the function λ on $\mathscr{E}$.

Since the initial point $\tilde{t}_0$ is chosen arbitrarily from the interval (t_1,t_2), the above analysis shows that $\tilde{\lambda}^g(t,e) = \lambda^g(e)$ for $t \in (t_1,t_2)$ (hence for $t \in [t_1,t_2]$, by continuity) and $e \in \mathscr{E}$.

Recall that this analysis is conducted for the case where $t_1 \in \mathscr{T}^{\times}$ and $h_0^{-1}(\bar{r},\cdot)$ is strictly smaller than $h_1^{-1}(\bar{r},\cdot)$ over (t_1,t_2). The other three cases reach the same conclusion that $\tilde{\lambda}^g(t,e)$ is equal to $\lambda^g(e)$ over $[t_1,t_2]\times\mathscr{E}$ through symmetric arguments. More specifically, we can switch h_1 and h_0 in defining π so that the sequence $\{\tilde{t}_k\}$ tends to a point in $\mathscr{T}^{\times}$.

Step 3. Repeat step 2 on each interval $(t_l,t_{l+1}), l=2,\cdots,L$. It follows that $\tilde{\lambda}^g(t,e)=\lambda^g(e)$ over $([t_0',t_0'']\cap[t_1',t_1''])\times[0,1]$.

Step 4. Pick any $t'\in \mathrm{int}([t_0',t_0'']\cup[t_1',t_1''])\setminus[t_1,t_L]$ (if this set is nonempty). There are four cases to consider, depending on whether $t'\in[t_0',t_0'']$ or $t'\in[t_1',t_1'']$ and whether $t_0''<t_1''$ or $t_0''>t_1''$. Without loss of generality, assume that $t'\in[t_0',t_0'']$ and $t_0''<t_1''$. The other three cases can be dealt with symmetric arguments. In this case, $t_1'<t_0''$ because $[t_0',t_0'']\cap[t_1',t_1'']$ is a non-degenerate interval. Denote $t''\in[t_1',t_1'']$ such that $h_0^{-1}(\bar{r},t)=h_1^{-1}(\bar{r},t'')$. It must be the case that $t''\in[t_1',t_0'']$. If that is not the case, then $t''>t_0''$. Then for any $t\in[t_1',t_0'']$,

$$h_0^{-1}(\bar{r},t)>h_0^{-1}(\bar{r},t')=h_1^{-1}(\bar{r},t'')>h_1^{-1}(\bar{r},t), t\in[t_0',t_0'']\cap[t_1',t_1''],$$

by the strict monotonicity of $h_0^{-1}(\bar{r},\cdot)$ and $h_1^{-1}(\bar{r},\cdot)$. However, this contradicts the assumption that $\mathscr{T}^{\times}$ is nonempty. Then by (A.2),

$$\tilde{\lambda}^g(t',e)=\tilde{\lambda}^g(t'',e)=\lambda^g(e).$$

Step 5. By the definition of $\tilde{\lambda}^g$ in (A.1), we now have

$$g^*(t,\bar{r},e)=g(t,\bar{r},\lambda^g(e)),\ \text{for } t\in[t_0',t_0'']\cup[t_1',t_1''], e\in\mathscr{E}.$$

By the properties of $\tilde{\lambda}^g$, we know λ^g is continuous and strictly increasing. The above statement holds for any $g\in\mathscr{G}$ that satisfies Equation (1.5).

□

Proof of Corollary 1.1. Based on the definition of $F^{g^\circ}_{\varepsilon|U,R}$ in (1.6), Theorem 1.1 and Lemma 1.2, we have

$$\begin{aligned} F^{g^\circ}_{\varepsilon|U,R}(\lambda^g(e)|u,\bar{r}) &= F^{-}_{Y|T,R}(g^\circ(h_0(\bar{r},u),\bar{r},\lambda^{g^\circ}(e))|h_0(\bar{r},u),\bar{r}) \\ &= F^{-}_{Y|T,R}(g^\circ(h_0(\bar{r},u),\bar{r},e)|h_0(\bar{r},u),\bar{r}) \\ &= F_{\varepsilon|U,R}(e|u,\bar{r}). \end{aligned}$$

The second claim follows from a change of variable.

□

Proof of Corollary 1.2. By construction, $\left\|D_{\gamma,h^*}\right\|_w \geq 0$ for any $\gamma \in \Gamma$. By Lemma 1.2, we have $\left\|D_{\gamma^*,h^*}\right\|_w = 0$. We want to show that γ^* is the unique zero. Since $w > 0$, we have

$$\begin{aligned} \left\|D_{\gamma,h^*}\right\|_w = 0 &\implies D_{\gamma,h^*}(e,u) = 0, \text{ for all } e \in \mathscr{E}, u \in [0,1] \\ &\implies \text{Condition (1.5) is satisfied by } g_\gamma(\cdot,\bar{r},\cdot), \end{aligned}$$

where the second line follows by taking the partial derivative with respect to u on both sides. By Theorem 1.1, this implies that $g_\gamma(\cdot,\bar{r},\cdot) = g_{\gamma'}(\cdot,\bar{r},\lambda(\cdot))$. By Assumption 1.7, it must be that $\gamma = \gamma^*$. Therefore, γ^* is the unique minimizer of $\left\|D_{\gamma,h^*}\right\|_w$. □

A.2 Proof of estimation results

This section proceeds as follows. Section A.2.1 provides the proofs of Theorem 1.2 and Proposition 1.1. Section A.2.2 introduces the empirical process theory and presents the lemmas on the uniform convergence results used in Section A.2.1. Section A.2.3 discusses the consistent estimation of the asymptotic covariance matrix.

A.2.1 Proofs of Theorem 1.2 and Proposition 1.1

Proof of Theorem 1.2. In this proof, the functions $h_1(r,u)$, $h_0(r,u)$, and $g(t,r,e)$ are only evaluated at $r = \bar{r}$. For simplicity, we omit this argument $\bar{r}$ throughout. The proof proceeds with seven steps:

- Step 1 contains preliminary results on the LLR estimator of the condition distribution $Y|T,R$.
- Step 2 derives the consistency of $\hat{\gamma}$.
- Step 3 derives an initial estimate of the convergence rate of $\hat{\gamma}$.
- Step 4 proves a stochastic equicontinuity condition on the criterion function.
- Step 5 presents a linear approximation of the criterion function.
- Step 6 shows the asymptotic normality of the minimizer of the linearized criterion function.
- Step 7 derives the asymptotic normal distribution of $\hat{\gamma}$.

Step 1. (Preliminary results on LLR.) Let $X_i(t) = (1,(T_i - t)/b_1,(R_i - \bar{r})/b_1)'$ denote the vector containing the regressors in the LLR. For $\boldsymbol{x} = (1,x_1,x_2)$, let

$$\begin{aligned}
k_0(\boldsymbol{x}) &= k_T(x_1)k_R(x_2)\mathbf{1}\{x_2 < 0\}, \\
k_1(\boldsymbol{x}) &= k_T(x_1)k_R(x_2)\mathbf{1}\{x_2 \geq 0\}.
\end{aligned}$$

The kernel weights in the LLR can be written as

$$\begin{aligned}
k_0(X_i(t)) &= k_T\left((T_i - t)/b_1\right) k_R\left((R_i - \bar{r})/b_1\right) \mathbf{1}\{R_i < \bar{r}\}, \\
k_1(X_i(t)) &= k_T\left((T_i - t)/b_1\right) k_R\left((R_i - \bar{r})/b_1\right) \mathbf{1}\{R_i \geq \bar{r}\}.
\end{aligned}$$

From Chapter 2 of this dissertation, we have the following uniform asymptotic linear representation for the LLR estimator of the conditional distribution functions:

$$
\begin{aligned}
\hat{F}^{-}_{Y|T,R}(y|t,\bar{r}) - F^{-}_{Y|T,R}(y|t,\bar{r}) = {} & b_1^2\mu_0(y,t) + \iota'\Xi_0(t)^{-1}\frac{1}{nb_1^2}\sum_{i=1}^{n} s_0(Y_i,T_i,R_i,y,t) \\
& + O_p\left(b_1^3 + \frac{|\log b_1|}{nb_1^2}\right),
\end{aligned}
$$

uniformly over $y \in \mathbb{R}, t \in [t_0', t_0'']$, and

$$
\begin{aligned}
\hat{F}^{+}_{Y|T,R}(y|t,\bar{r}) - F^{+}_{Y|T,R}(y|t,\bar{r}) = {} & b_1^2\mu_1(y,t) + \iota'\Xi_1(t)^{-1}\frac{1}{nb_1^2}\sum_{i=1}^{n} s_1(Y_i,T_i,R_i,y,t) \\
& + O_p\left(b_1^3 + \frac{|\log b_1|}{nb_1^2}\right),
\end{aligned}
$$

uniformly over $y \in \mathbb{R}, t \in [t_1', t_1'']$. In the above expressions, $\iota = (1,0,\cdots,0)$. The functions $\mu_0(y,t)$ and $\mu_1(y,t)$ are defined by

$$
\begin{aligned}
\mu_0(y,t) = {} & \frac{b_1^2}{2}\iota'\Omega_0(t)^{-1}\frac{\partial^2}{\partial t^2}F^{-}_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_1^2k_0(\boldsymbol{x})\mathbf{1}\{t+b_1x_1 \in [t_0', t_0'']\}dx_1dx_2 \\
& + \frac{b_1^2}{2}\iota'\Omega_0(t)^{-1}\frac{\partial^2}{\partial r^2}F^{-}_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_2^2k_0(\boldsymbol{x})\mathbf{1}\{t+b_1x_1 \in [t_0', t_0'']\}dx_1dx_2 \\
& + b_1^2\iota'\Omega_0(t)^{-1}\frac{\partial^2}{\partial t\partial r}F^{-}_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_1x_2k_0(\boldsymbol{x})\mathbf{1}\{t+b_1x_1 \in [t_0', t_0'']\}dx_1dx_2 \\
& + \frac{b_2^2}{2}\Omega_0(t)^{-1}\frac{\partial^2}{\partial y^2}F^{-}_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}k_0(\boldsymbol{x})\mathbf{1}\{t+b_1x_1 \in [t_0', t_0'']\}dx_1dx_2,
\end{aligned}
$$

and

$$
\begin{aligned}
\mu_1(y,t) = &\frac{b_1^2}{2}\iota'\Omega_1(t)^{-1}\frac{\partial^2}{\partial t^2}F^+_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_1^2k_1(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_1',t_1'']\}dx_1dx_2\\
&+\frac{b_1^2}{2}\iota'\Omega_1(t)^{-1}\frac{\partial^2}{\partial r^2}F^+_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_2^2k_1(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_1',t_1'']\}dx_1dx_2\\
&+b_1^2\iota'\Omega_1(t)^{-1}\frac{\partial^2}{\partial t\partial r}F^+_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}x_1x_2k_1(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_1',t_1'']\}dx_1dx_2\\
&+\frac{b_2^2}{2}\Omega_1(t)^{-1}\frac{\partial^2}{\partial y^2}F^+_{Y|T,R}(y|t,\bar{r})\int \boldsymbol{x}k_1(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_1',t_1'']\}dx_1dx_2.
\end{aligned}
$$

The matrices $\Omega_0(t)$, $\Omega_1(t)$, $\Xi_0(t)$, and $\Xi_1(t)$ are defined by

$$
\begin{aligned}
\Omega_0(t) &= \int \boldsymbol{x}\boldsymbol{x}'k_0(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_0',t_0'']\}dx_1dx_2,\\
\Omega_1(t) &= \int \boldsymbol{x}\boldsymbol{x}'k_1(\boldsymbol{x})\mathbf{1}\{t+b_1x_1\in[t_1',t_1'']\}dx_1dx_2,
\end{aligned}
$$

and

$$
\begin{aligned}
\Xi_0(t) &= \int \boldsymbol{x}\boldsymbol{x}'k_0(\boldsymbol{x})f^-_{T,R}(t+b_1x_1,\bar{r}+b_1x_2)dx_1dx_2,\\
\Xi_1(t) &= \int \boldsymbol{x}\boldsymbol{x}'k_1(\boldsymbol{x})f^-_{T,R}(t+b_1x_1,\bar{r}+b_1x_2)dx_1dx_2.
\end{aligned}
$$

The terms s_0 and s_1 are defined by

$$
\begin{aligned}
s_0(Y_i,T_i,R_i,y,t) &= X_i(t)\tilde{K}_Y(Y_i,T_i,R_i;y,t)k_0(X_i(t)),\\
s_1(Y_i,T_i,R_i,y,t) &= X_i(t)\tilde{K}_Y(Y_i,T_i,R_i;y,t)k_1(X_i(t)),\\
\tilde{K}_Y(Y_i,T_i,R_i,y) &= K_Y\left((y-Y_i)/b_1\right)-\mathbb{E}\left[K_Y\left((y-Y_i)/b_1\right)|T_i,R_i\right].
\end{aligned}
$$

Under Assumption 1.9(i) and (ii) and Assumption 1.11(i), we can apply Lemma 2.1 in Chapter 2 of this dissertation, which is a modification of Lemma 11 in Fan and Guerre [2016], and obtain that the eigenvalues of $\Omega_0(t)$, $\Omega_1(t)$, $\Xi_0(t)$ and $\Xi_1(t)$ are bounded and bounded away from zero for all values of t and b_1. Consequently, the norm of these matrices and there inverses are bounded.

Notice that in Chapter 2 of this dissertation, the remainder term from the bias expansion is $o(b_1^2)$ while in the above asymptotic linear representation, the corresponding term is $O(b_1^3)$. This is because we assume that $F^-_{Y|T,R}$ and $F^+_{Y|T,R}$ are three-times continuously differentiable (Assumption 1.9). Under this assumption, we can go through the same steps as in the proof of Theorem 2.1 in Chapter 2 of this dissertation and show that the remainder from the bias expansion is $O(b_1^3)$. The details are omitted for brevity. The bias terms $\mu_0(y,t)$ and $\mu_1(y,t)$ are continuously differentiable under Assumption 1.9. The indicator functions inside the integral, for example, $\mathbf{1}\{t+b_1x_1 \in [t_0', t_0'']\}$, can be eliminated by explicitly indicating the lower and upper limits of the corresponding integral. The derivative can then be taken by using the Leibniz rule.

Step 2. (Consistency of $\hat{\gamma}$.) Because w is positive and integrates to 1, we have $\|\hat{D}_{\gamma,\hat{h}} - D_{\gamma,\hat{h}}\|_w \leq \|\hat{D}_{\gamma,\hat{h}} - D_{\gamma,\hat{h}}\|_\infty$. For any $\gamma \in \Gamma$, we can apply Fubini's theorem to the uniform asymptotic linear representation and obtain that

$$\hat{D}_{\gamma,\hat{h}}(e,u) - D_{\gamma,\hat{h}}(e,u) = \mathrm{I} + \mathrm{II} + O_p\left(b_1^2 + \log n/(nb_1^2)\right)$$

uniformly over $\gamma \in \Gamma, e \in \mathscr{E}$, and $u \in (0,1)$, where

$$\mathrm{I} = \frac{1}{nb_1^2}\sum_{i=1}^n \int_0^u \iota' \Xi_0(\hat{h}_0(v))^{-1} s_0(Y_i, T_i, R_i; g_\gamma(\hat{h}_0(v), e), \hat{h}_0(v))dv,$$
$$\mathrm{II} = \frac{1}{nb_1^2}\sum_{i=1}^n \int_0^u \iota' \Xi_1(\hat{h}_1(v))^{-1} s_1(Y_i, T_i, R_i; g_\gamma(\hat{h}_1(v), e), \hat{h}_1(v))dv.$$

By symmetry, we only need to study the term I. Denote $\mathbf{1}_{\hat{h}_0} = \mathbf{1}\{\hat{h}_0 \in \mathscr{H}_0(\mathscr{P}_0^n)\}$ as the indicator of whether $\hat{h}_0 \in \mathscr{H}_0(\mathscr{P}_0^n)$. Then $\mathrm{I} \leq \mathrm{I.1} + \mathrm{I.2}$, where

$$\mathrm{I.1} = \sup_{\gamma\in\Gamma, h_0\in\mathscr{H}_0(\mathscr{P}_0^n), e\in\mathscr{E}, u\in(0,1)} \left| \frac{1}{nb_1^2}\sum_{i=1}^n \int_0^u \iota'\Xi_0(h_0(v))^{-1} s_0(Y_i,T_i,R_i;g_\gamma(h_0(v),e),h_0(v))dv \right|,$$

$$\mathrm{I.2} = (1-\mathbf{1}_{\hat{h}_0}) \sup_{y\in\mathscr{Y}, t\in[t_0',t_0'']} \left| \frac{1}{nb_1^2}\sum_{i=1}^n \iota'\Xi_0(t)^{-1} s_0(Y_i,T_i,R_i;y,t) \right|$$

Define

$$\tilde{\alpha}_n = n^{-1/2} b_1^{-7/12-\bar{\varepsilon}/5}, \text{ and } \alpha_n = \left(b_1^2 + \sqrt{\log n/(nb_1^4)}\right)(b_1^2+\tilde{\alpha}_n), \tag{A.3}$$

where $\bar{\varepsilon}$ is defined in Assumption 1.12. In Lemma A.6, we show that $\mathrm{I.1} = O_p(\tilde{\alpha}_n)$. Combining Assumption 1.13(i) and Lemma A.8, we have $\mathrm{I.2} = O_p(\tilde{\alpha}_n)$. Therefore,

$$\sup_{\gamma\in\Gamma} \left\| \hat{D}_{\gamma,\hat{h}} - D_{\gamma,\hat{h}} \right\|_w O_p(b_1^2+\tilde{\alpha}_n). \tag{A.4}$$

By the smoothness of $F^{-}_{Y,T,R}$ and g_γ, the following term is $O(\|\hat{h}-h^*\|_\infty)$:

$$\sup_{\gamma\in\Gamma} \left| \int_0^u \left(F^{-}_{Y|T,R}(g_\gamma(\hat{h}_0(v),e)|h_0(v),\bar{r}) - F^{-}_{Y|T,R}(g_\gamma(h_0^*(v),e)|h_0(v),\bar{r}) \right) dv \right|$$

Therefore, we obtain that uniformly over $\gamma \in \Gamma$,

$$\left\| D_{\gamma,\hat{h}} - D_{\gamma,h^*} \right\|_w = O(\|\hat{h}-h^*)\|_\infty = O_p\left(b_1^2 + \sqrt{\log n/(nb_1)}\right).$$

By the triangle inequality, we have

$$\left\| D_{\hat{\gamma},h^*} \right\|_w \leq \left\| D_{\hat{\gamma},h^*} - D_{\hat{\gamma},\hat{h}} \right\|_w + \left\| \hat{D}_{\hat{\gamma},\hat{h}} - D_{\hat{\gamma},\hat{h}} \right\|_w + \left\| \hat{D}_{\hat{\gamma},\hat{h}} \right\|_w.$$

By the definition of $\hat{\gamma}$ in (1.9), we have

$$\left\|\hat{D}_{\hat{\gamma},\hat{h}}\right\|_w \leq \left\|\hat{D}_{\gamma^*,\hat{h}}\right\|_w + o_p\left(\alpha_n\right) \leq \left\|\hat{D}_{\gamma^*,\hat{h}} - D_{\gamma^*,\hat{h}}\right\|_w + \left\|D_{\gamma^*,\hat{h}} - D_{\gamma^*,h^*}\right\|_w + o_p\left(\alpha_n\right).$$

Combining the above two inequalities together, we obtain that

$$\left\|D_{\hat{\gamma},h^*}\right\|_w \leq 2\sup_{\gamma\in\Gamma}\left\|D_{\gamma,\hat{h}} - D_{\gamma,h^*}\right\|_w + 2\sup_{\gamma\in\Gamma}\left\|\hat{D}_{\gamma,\hat{h}} - D_{\gamma,\hat{h}}\right\|_w + o_p\left(\alpha_n\right) = O_p\left(b_1^2 + \tilde{\alpha}_n\right). \tag{A.5}$$

In particular, the above quantity is $o_p(1)$. Because Γ is compact, and $\left\|D_{\cdot,h^*}\right\|_w$ is continuous and has a unique minimizer γ^* (Corollary 1.2), for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\|\gamma - \gamma^*\|_2 > \varepsilon \implies \left\|D_{\gamma,h^*}\right\|_w > \delta$. Therefore, $\mathbb{P}(\|\hat{\gamma} - \gamma^*\|_2 > \varepsilon) \leq \mathbb{P}(\left\|D_{\gamma,h^*}\right\|_w > \delta) = o(1)$. This proves that $\hat{\gamma}$ is a consistent estimator.

Step 3. (Convergence rate of $\hat{\gamma}$.) Since $\hat{\gamma}$ is consistent, we can Taylor expand $D_{\hat{\gamma},h^*}$ around γ^*. Together with the reverse triangle inequality and the fact that $D_{\gamma^*,h^*} = 0$, the expansion gives that

$$\begin{aligned}
\left\|D_{\hat{\gamma},h^*}\right\|_w &= \left\|\nabla_\gamma D_{\gamma^*,h^*}(\hat{\gamma} - \gamma^*) + (\hat{\gamma} - \gamma^*)'\nabla_\gamma^2 D_{\tilde{\gamma},h^*}(\hat{\gamma} - \gamma^*)\right\|_w \\
&\geq \left\|\nabla_\gamma D_{\gamma^*,h^*}(\hat{\gamma} - \gamma^*)\right\|_w - \|\hat{\gamma} - \gamma^*\|_2^2 \int_0^1 \int_{\mathscr{E}} \left\|\nabla_\gamma^2 D_{\tilde{\gamma},h^*}(u,e)\right\|_2 w(e,u)\,de\,du \\
&\geq \left\|\nabla_\gamma D_{\gamma^*,h^*}(\hat{\gamma} - \gamma^*)\right\|_w + O\left(\|\hat{\gamma} - \gamma^*\|_2^2\right),
\end{aligned}$$

where $\tilde{\gamma}$ is some point on the line segment connecting $\hat{\gamma}$ and γ^* and the last line follows from Assumption 1.10(iii) that $\left\|\nabla_\gamma^2 D_{\tilde{\gamma},h^*}(u,e)\right\|_2$ is bounded. We claim that there exists a universal constant $C > 0$ such that

$$\left\|\nabla_\gamma D_{\gamma^*,h^*}\zeta\right\|_w \geq C\|\zeta\|, \text{ for all } \zeta \in \mathbb{R}^{d_\Gamma}. \tag{A.6}$$

If this claim is true, then by Equation (A.5), we obtain a bound on the convergence rate of $\hat{\gamma}$:

$$\|\hat{\gamma}-\gamma^*\|_2 = O_p\left(b_1^2+\tilde{\alpha}_n\right).$$

The remaining part of this step is devoted to the proof of (A.6). Suppose that claim is false, then for each integer $k \geq 1$, there exists $\zeta_k \in \mathbb{R}^{d_\Gamma}$ such that $\left\|\nabla_\gamma D_{\gamma^*,h^*}\zeta_k\right\|_w < 1/k\|\zeta_k\|$. Without loss of generality, we can assume $\|\zeta_k\| = 1$ (or simply redefine the sequence as $\zeta_k/\|\zeta_k\|$). By the Bolzano–Weierstrass theorem, the sequence $\{\zeta_k\}$ has a convergent subsequence. Without loss of generality, we assume $\{\zeta_k\}$ itself is convergent with the limit denoted by ζ_∞. Then it must be the case that $\left\|\nabla_\gamma D_{\gamma^*,h^*}\zeta_\infty\right\|_w = 0$. Since $\nabla_\gamma D_{\gamma^*,h^*}$ is a continuous function, the previous equation implies that $\nabla_\gamma D_{\gamma^*,h^*}\zeta_\infty = 0$. This violates Assumption 1.10(iv) that $\nabla_\gamma D_{\gamma^*,h^*}$ is a vector of linearly independent functions.

Step 4. (Stochastic equicontinuity of the criterion function.) Let $\gamma_n \xrightarrow{p} \gamma^*$ be such that $\|\gamma_n-\gamma^*\| = O_p(b_1^2+\tilde{\alpha}_n)$. We want to find the asymptotic order of the term $\|\hat{D}_{\gamma_n,\hat{h}} - D_{\gamma_n,\hat{h}} - \hat{D}_{\gamma^*,h^*}\|_w$, which is bounded by

$$\sup_{e\in\mathscr{E},u\in(0,1)}\left|\hat{D}_{\gamma_n,\hat{h}}(e,u)-D_{\gamma_n,\hat{h}}(e,u)-(\hat{D}_{\gamma^*,h^*}(e,u)-D_{\gamma^*,h^*}(e,u))\right| \leq \mathrm{I}+\mathrm{II},$$

where

$$\begin{aligned}\mathrm{I} = \sup_{e\in\mathscr{E},u\in(0,1)}&\left|\hat{F}^{-}_{Y|T,R}(g_{\gamma_n}(\hat{h}_0(v),e)|\hat{h}_0(v),\bar{r}) - F^{-}_{Y|T,R}(g_{\gamma_n}(\hat{h}_0(v),e)|\hat{h}_0(v),\bar{r})\right.\\ &\left.-\left(\hat{F}^{-}_{Y|T,R}(g_{\gamma^*}(h_0^*(v),e)|h_0^*(v),\bar{r}) - F^{-}_{Y|T,R}(g_{\gamma^*}(h_0^*(v),e)|h_0^*(v),\bar{r})\right)\right|,\end{aligned}$$

and

$$\begin{aligned}\mathrm{II} = \sup_{e\in\mathscr{E},u\in(0,1)} \Big| & \hat{F}^{+}_{Y|T,R}(g_{\gamma_n}(\hat{h}_1(v),e)|\hat{h}_1(v),\bar{r}) - F^{-}_{Y|T,R}(g_{\gamma_n}(\hat{h}_1(v),e)|\hat{h}_1(v),\bar{r}) \\ & - \left(\hat{F}^{+}_{Y|T,R}(g_{\gamma^*}(h_1^*(v),e)|h_1^*(v),\bar{r}) - F^{+}_{Y|T,R}(g_{\gamma^*}(h_1^*(v),e)|h_1^*(v),\bar{r})\right)\Big|.\end{aligned}$$

By symmetry, we only need to study the term I. The uniform asymptotic linear representation of the LLR estimators gives a bias-variance decomposition that

$$\mathrm{I} \leq \mathrm{I}.1 + \mathrm{I}.2 + O_p\left(b_1^3 + |\log b_1|/nb_1^2\right),$$

where

$$\begin{aligned}\mathrm{I}.1 &= b_1^2\left(\mu_0(g_{\gamma_n}(\hat{h}_0(v),e),\hat{h}_0(v)) - \mu_0(g_{\gamma^*}(h_0^*(v),e),h_0^*(v))\right), \\ \mathrm{I}.2 &= \sup_{e\in\mathscr{E},u\in[0,1]}\Big|\frac{1}{nb_1^2}\sum_{i=1}^{n}\iota'\Xi_0(\hat{h}_0(v))^{-1}s_0(Y_i,T_i,R_i;g_{\gamma_n}(\hat{h}_0(v),e),\hat{h}_0(v)) \\ &\qquad - \iota'\Xi_0(h_0^*(v))^{-1}s_0(Y_i,T_i,R_i;g_{\gamma^*}(h_0^*(v),e),h_0^*(v))\Big|.\end{aligned}$$

By the smoothness of μ_0 (Step 1) and g_γ (Assumption 1.10), we can bound the term I.1 by

$$\mathrm{I}.1 \leq Cb_1^2(\left\|\gamma_n - \gamma^*\right\|_2 + \|\hat{h} - h^*\|_\infty) = O_p(b_1^4 + b_1^2\tilde{\alpha}_n).$$

For the term I.2, consider the decomposition that $\mathrm{I}.2 \leq \mathrm{I}.2.1 + \mathrm{I}.2.2$, where

$$\begin{aligned}\mathrm{I}.2.1 &= \sup_{e\in\mathscr{E},u\in[0,1]}\Big|\iota'\Xi_0(\hat{h}_0(v))^{-1}\frac{1}{nb_1^2}\sum_{i=1}^{n}s_0(Y_i,T_i,R_i;g_{\gamma_n}(\hat{h}_0(v),e),\hat{h}_0(v)) \\ &\qquad - s_0(Y_i,T_i,R_i;g_{\gamma^*}(h_0^*(v),e),h_0^*(v))\Big|, \\ \mathrm{I}.2.2 &= \sup_{e\in\mathscr{E},u\in[0,1]}\Big|\iota'\left(\Xi_0(\hat{h}_0(v))^{-1} - \Xi_0(h_0^*(v))^{-1}\right)\frac{1}{nb_1^2}\sum_{i=1}^{n}s_0(Y_i,T_i,R_i;g_{\gamma^*}(h_0^*(v),e),h_0^*(v))\Big|.\end{aligned}$$

As mentioned in Step 1, we know that $\left\|\Xi(t)^{-1}\right\|_2$ is bounded for $t \in [t_0', t_0'']$ by Lemma 2.1 in Chapter 2 of this dissertation. Applying the mean value theorem, we obtain that

$$\text{I.2.1} \le C(\text{I.2.1.1} + \text{I.2.1.2} + \text{I.2.1.3})$$

where

$$\begin{aligned}
\text{I.2.1.1} &= \sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial y} s_0(Y_i, T_i, R_i; y, t) \right\|_2 \left\|\nabla_\gamma g_{\gamma^*}\right\|_\infty \left\|\gamma_n - \gamma^*\right\|_2, \\
\text{I.2.1.2} &= \sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial y} s_0(Y_i, T_i, R_i; y, t) \right\|_2 \left\| \frac{\partial}{\partial T} g_{\gamma^*} \right\|_\infty \|\hat{h} - h^*\|_\infty, \\
\text{I.2.1.3} &= \sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial t} s_0(Y_i, T_i, R_i; y, t) \right\|_2 \|\hat{h} - h^*\|_\infty.
\end{aligned}$$

In Lemma A.7, we show that the following two terms are of order $O_p\left(\sqrt{\log n/(nb_1^4)}\right)$:

$$\sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial y} s(Y_i, T_i, R_i; y, t) \right|, \quad \sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial t} s(Y_i, T_i, R_i; y, t) \right|.$$

Because $\left\|\nabla_\gamma g_{\gamma^*}\right\|_\infty$ and $\left\|\partial g_{\gamma^*}/\partial T\right\|_\infty$ are finite, we know that

$$\text{I.2.1} = O_p\left(\sqrt{\log n/(nb_1^4)}\right) \times \left(\left\|\gamma_n - \gamma^*\right\|_2 + \|\hat{h} - h^*\|_\infty\right) = O_p\left(\sqrt{\log n/(nb_1^4)}\tilde{\alpha}_n\right).$$

Applying the mean value theorem to I.2.2, we obtain that

$$\text{I.2.2} \le \sup_{t\in[t_0', t_0'']} \left| \iota' \frac{\partial}{\partial t} \Xi_0(t)^{-1} \right| \sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n s_0(Y_i, T_i, R_i; y, t) \right\|_2 \|\hat{h} - h^*\|_\infty.$$

In Lemma A.8, we show that

$$\sup_{y\in\mathbb{R},t\in[t_0',t_0'']}\left\|\sum_{i=1}^{n} s_0(Y_i,T_i,R_i;y,t)/(nb_1^2)\right\|_2 = O_p\left(\sqrt{\log n/(nb_1^2)}\right).$$

Therefore, I.2.1 asymptotically dominates I.2.2. Hence, the term I is of the following order:

$$\mathrm{I} = O_p\left(\left(b_1^2+\sqrt{\log n/(nb_1^4)}\right)\tilde{\alpha}_n\right) = O_p(\alpha_n).$$

Based on the same argument, the above asymptotic order also applies to the term II. Thus, we have the following stochastic equicontinuity result:

$$\|\hat{D}_{\gamma_n,\hat{h}} - D_{\gamma_n,\hat{h}} - \hat{D}_{\gamma^*,h^*}\|_w = O_p(\alpha_n). \tag{A.7}$$

Step 5. (Linearization of the criterion function.) Let $\partial_h^{[\hat{h}-h^*]}D_{\gamma,h^*}(e,u)$ be the Fréchet derivative of $D_{\gamma,h}(e,u)$ with respect to h at h^*, in the direction of $h-h^*$. That is,

$$\partial_h^{[\hat{h}-h^*]}D_{\gamma,h^*}(e,u) = \int_0^u (\phi_\gamma^-(e,v)-\phi_\gamma^+(e,v))(\hat{h}_0(v)-h_0^*(v))dv,$$

where

$$\phi_\gamma^-(e,v) = \frac{\partial}{\partial Y}F_{Y|T,R}^-(g_\gamma(h_0^*(v),e)|h_0^*(v),\bar{r})\frac{\partial}{\partial T}g_\gamma(h_0^*(v),e) + \frac{\partial}{\partial T}F_{Y|T,R}^-(g_\gamma(h_0^*(v),e)|h_0^*(v),\bar{r}),$$
$$\phi_\gamma^+(e,v) = \frac{\partial}{\partial Y}F_{Y|T,R}^+(g_\gamma(h_1^*(v),e)|h_1^*(v),\bar{r})\frac{\partial}{\partial T}g_\gamma(h_1^*(v),e) + \frac{\partial}{\partial T}F_{Y|T,R}^+(g_\gamma(h_1^*(v),e)|h_1^*(v),\bar{r}).$$

It is straightforward to see that $\left\|\partial_h^{[\hat{h}-h^*]}D_{\gamma,h^*}\right\|_\infty = O(\|\hat{h}-h^*\|_\infty)$. Following the same steps as in Lemma 4 of Torgovitsky [2017], we can show that $\|D_{\gamma,\hat{h}} - D_{\gamma,h^*} - \partial_h^{[\hat{h}-h^*]}D_{\gamma,h^*}\|_w =$

$O(\|\hat{h}-h^*\|_\infty^2)$, uniformly over $\gamma \in \Gamma$, and

$$\left\|\partial_h^{[\hat{h}-h^*]} D_{\gamma_n,h^*} - \partial_h^{[\hat{h}-h^*]} D_{\gamma^*,h^*}\right\|_w = O(\left\|\gamma_n - \gamma^*\right\|_2 \|\hat{h}-h^*\|_\infty),$$

for any sequence $\gamma_n \xrightarrow{p} \gamma^*$. Define

$$\hat{L}_\gamma(e,u) = \hat{D}_{\gamma^*,h^*}(e,u) + \nabla_\gamma D_{\gamma^*,h^*}(e,u)(\gamma-\gamma^*) + \partial_h^{[\hat{h}-h^*]} D_{\gamma^*,h^*}(e,u),$$

as a linear approximation of $\hat{D}_{\gamma,\hat{h}}(e,u)$ for γ near γ^*. For any sequence $\|\gamma_n - \gamma^*\|_2 = O_p(b_1^2 + \tilde{\alpha}_n)$, we have

$$\left\|\hat{L}_{\gamma_n}\right\|_w \le \left\|\hat{D}_{\gamma^*,h^*} - D_{\gamma^*,h^*}\right\|_w + \left\|\nabla_\gamma D_{\gamma^*,h^*}\right\|_w \left\|\gamma_n - \gamma^*\right\|_2 + O(\|\hat{h}-h^*\|_\infty) = O_p(b_1^2 + \tilde{\alpha}_n), \quad (A.8)$$

where the asymptotic order of the first term on the RHS is derived in (A.4). We want to bound the approximation error from the linearization of the criterion function. By adding and subtracting terms, we obtain that

$$\begin{aligned} &\left\|\hat{L}_{\gamma_n} - \hat{D}_{\gamma_n,\hat{h}}\right\|_w \\ \le &\left\|\hat{D}_{\gamma^*,h^*} - \hat{D}_{\gamma_n,\hat{h}} - (D_{\gamma^*,h^*} - D_{\gamma_n,\hat{h}})\right\| + \left\|D_{\gamma^*,h^*} + \nabla_\gamma D_{\gamma^*,h^*}(\gamma-\gamma^*) - D_{\gamma_n,h^*}\right\|_w \\ &+ \left\|D_{\gamma_n,h^*} + \partial_h^{[\hat{h}-h^*]} D_{\gamma_n,h^*} - D_{\gamma_n,\hat{h}}\right\|_w + \left\|\partial_h^{[\hat{h}-h^*]} D_{\gamma^*,h^*} - \partial_h^{[\hat{h}-h^*]} D_{\gamma_n,h^*}\right\|_w. \end{aligned} \quad (A.9)$$

The four terms on the RHS of the above inequality can be analyzed as the following. The order of the first term on the RHS of (A.9) is given by (A.7) in the previous step. The second term is $O(\|\gamma_n - \gamma^*\|_2)$ by the smoothness of D_{γ,h^*}. The third term is bounded by

$$\sup_{\gamma\in\Gamma} \left\|D_{\gamma,\hat{h}} - D_{\gamma,h^*} - \partial_h^{[\hat{h}-h^*]} D_{\gamma,h^*}\right\|_w = O(\|\hat{h}-h^*\|_\infty^2).$$

The fourth term is $O(\left\|\gamma_n - \gamma^*\right\|_2 \|\hat{h}-h^*\|_\infty)$. Therefore, the leading term on the RHS of (A.9) is the first term, and hence the approximation error from the linearization of the criterion function is of the following order:

$$\left\|\hat{L}_{\gamma_n} - \hat{D}_{\gamma_n,\hat{h}}\right\|_w = O_p(\alpha_n). \tag{A.10}$$

Step 6. (Minimizer of the linearized criterion function.) Define $\tilde{\gamma}$ as the minimizer of $\left\|\hat{L}_\gamma\right\|_w$. The first-order condition gives that

$$\Delta(\tilde{\gamma} - \gamma^*) = \int \nabla_\gamma D_{\gamma^*,h^*}(e,u)\left(\hat{D}_{\gamma^*,h^*}(e,u) + \partial_h^{[\hat{h}-h^*]} D_{\gamma^*,h^*}(e,u)\right) w(e,u)dedu,$$

where

$$\Delta = \int \nabla_\gamma D_{\gamma^*,h^*}(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)' w(e,u)dedu. \tag{A.11}$$

By the uniform asymptotic linear representation of the LLR estimators and $\hat{h}$, we can write

$$\begin{aligned}
&\int \nabla_\gamma D_{\gamma^*,h^*}(e,u)\left(\hat{D}_{\gamma^*,h^*}(e,u) + \partial_h^{[\hat{h}-h^*]} D_{\gamma^*,h^*}(e,u)\right) w(e,u)dedu \\
=&\frac{1}{nb_1}\sum_{i=1}^n (\zeta_-^{DF}(Y_i,T_i,R_i) + \zeta_-^{Q}(Y_i,T_i,R_i)) - \frac{1}{nb_1}\sum_{i=1}^n (\zeta_+^{DF}(Y_i,T_i,R_i) + \zeta_+^{Q}(Y_i,T_i,R_i)) \\
&+ b_1^2(B_- - B_+) + O_p(b_1^3) + o_p(1/\sqrt{nb_1}).
\end{aligned}$$

The terms B_- and B_+ are deterministic bias terms defined by

$$B_- = \int w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\left(\int_0^u \mu_0(g_{\gamma^*}(h_0^*(v),e),h_0^*(v)) + \phi_{\gamma^*}^-(e,v)\nu_0(v)\right) dedu, \tag{A.12}$$

$$B_+ = \int w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\left(\int_0^u \mu_1(g_{\gamma^*}(h_1^*(v),e),h_1^*(v)) + \phi_{\gamma^*}^+(e,v)\nu_1(v)\right) dedu. \tag{A.13}$$

The functions ζ_-^{DF} and ζ_+^{DF} represent stochastic terms from the LLR estimation of the conditional distribution $F_{Y|T,R}$. They are defined by

$$\begin{aligned}
&\zeta_-^{DF}(Y,T,R)\\
=&\frac{1}{b_1}\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_0^u \iota'\Xi_0(h_0^*(v))^{-1}s_0(Y,T,R,g_{\gamma^*}(h_0^*(v),e),h_0^*(v))dvdedu\\
=&\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_{(T-h_0^*(u))/b_1}^{(T-t_0')/b_1} \iota'\Xi_0(T+b_1v)^{-1}(1,v,(R-\bar{r})/b_1)'\\
&\times \tilde{K}_Y(Y,T,R;g_\gamma(T+b_1v,e))k_T(v)k_R^-((R-\bar{r})/b_1)((h_0^*)^{-1})'(T+b_1v)dv,
\end{aligned}$$

and

$$\begin{aligned}
&\zeta_+^{DF}(Y,T,R)\\
=&\frac{1}{b_1}\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_0^u \iota'\Xi_1(h_1^*(v))^{-1}s_1(Y_i,T_i,R_i,g_{\gamma^*}(h_1^*(v),e),h_1^*(v))dvdedu\\
=&\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_{(T-h_1(u))/b_1}^{(T-t_1')/b_1} \iota'\Xi_1(T+b_1v)^{-1}(1,v,(r-\bar{r})/b_1)'\\
&\times \tilde{K}_Y(Y,T,R;g_{\gamma^*}(T+b_1v,e))k_T(v)k_R^+((R-\bar{r})/b_1)((h_0^*)^{-1})'(T+b_1v)dv.
\end{aligned}$$

In the above notations, $k_R^-(x)=k_R(x)\mathbb{1}\{x<0\}$ and $k_R^+(x)=k_R(x)\mathbb{1}\{x\geq 0\}$. Similarly define $k_{Q,0}^-(x)=k_{Q,0}(x)\mathbb{1}\{x<0\}$ and $k_{Q,1}^+(x)=k_{Q,1}(x)\mathbb{1}\{x\geq 0\}$. The functions ζ_-^Q and ζ_+^Q represent stochastic terms from the nonparametric estimation of the conditional quantile function h^*. They are defined by

$$\begin{aligned}
\zeta_-^Q(T,R)&=\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_0^u \phi_{\gamma^*}^-(e,v)q_0(T,R;v)k_{Q,0}^-\big((R-\bar{r})/b_1\big)\,dvdedu,\\
\zeta_+^Q(T,R)&=\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_0^u \phi_{\gamma^*}^+(e,v)q_1(T,R;v)k_{Q,1}^+\big((R-\bar{r})/b_1\big)\,dvdedu.
\end{aligned}$$

By Fubini's theorem, we have

$$\mathbb{E}[\zeta_\pm^{DF}(Y,T,R)|T,R]=\mathbb{E}[\zeta_\pm^Q(T,R))|T,R]=0.$$

Notice that $((h_0^*)^{-1})'(\cdot) = f_{T|R}^-(\cdot|\bar{r})$ and $((h_1^*)^{-1})'(\cdot) = f_{T|R}^+(\cdot|\bar{r})$. The variance matrix can be computed as follows, where to save space, we use the notation of squaring a vector to mean the tensor product of that vector with itself.

$$\begin{aligned}
&\mathbb{E}\left[(\zeta_-^{DF}(Y,T,R)+\zeta_-^Q(T,R))\otimes(\zeta_-^{DF}(Y,T,R)+\zeta_-^Q(T,R))\right]\\
&=\int_{\mathscr{Y}\times[t_0',t_0'']\times[r_0,\bar{r}]}\left(\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_{(t-h_0(u))/b_1}^{(t-t_0')/b_1}\iota'\Xi_0(t+b_1v)^{-1}(1,v,(r-\bar{r})/b_1)'\right.\\
&\quad\times\tilde{K}_Y(y,t,r;g_{\gamma^*}(t+b_1v,e))k_T(v)k_R^-((r-\bar{r})/b_1)f_{T|R}^-(t+b_1v|\bar{r})dv,\\
&\quad\left.+\int_0^u\phi_{\gamma*}^-(e,v)q_0(t,r;v)k_{Q,0}^-\big((r-\bar{r})/b_1\big)\,dvdedu\right)^2 f_{Y,T,R}^-(y,t,r)dydtdr,
\end{aligned}$$

where $f_{Y,T,R}^\pm$ is defined analogously as $f_{T|R}^\pm$ and $F_{Y|T,R}^\pm$. Applying the change of variables $\tilde{r}=(r-\bar{r})/b_1$, we obtain that the above matrix is equal to b_1 times the matrix

$$\begin{aligned}
&\int_{\mathscr{Y}\times[t_0',t_0'']\times[-1,0]}\left(\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\int_{(t-h_0(u))/b_1}^{(t-t_0')/b_1}\iota'\Xi_0(t+b_1v)^{-1}(1,v,\tilde{r})'\right.\\
&\quad\times\tilde{K}_Y(y,t,\bar{r}+b_1\tilde{r};g_{\gamma^*}(t+b_1v,e))k_T(v)k_R^-(\tilde{r})f_{T|R}^-(t+b_1v|\bar{r})dv\\
&\quad\left.+\int_0^u\phi_{\gamma*}^-(e,v)q_0(t,\bar{r}+b_1\tilde{r};v)k_{Q,0}^-(\tilde{r})\,dvdedu\right)^2 f_{Y,T,R}^-(y,t,\bar{r}+b_1\tilde{r})dydtd\tilde{r}.
\end{aligned}$$

For any $t\in[t_0'+b_1,t_0''-b_1]$, we have $\Xi_0(t)=f_{T,R}^-(t,\bar{r})\bar{\Omega}_0$ with $\bar{\Omega}_0=\int \boldsymbol{x}\boldsymbol{x}'k_0(\boldsymbol{x})dx_1dx_2$. By letting $n\to\infty$ (so that $b_1\to 0$) and using the continuity of the relevant functions and the dominated convergence theorem, we know that the above matrix is asymptotically equivalent to

$$\begin{aligned}
\Sigma_- = \int\Bigg(&\int_{\mathscr{E}}\int_0^1 w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\Bigg(\iota'\bar{\Omega}_0^{-1}(1,0,\tilde{r})'\Big(\mathbf{1}\{y\le g_{\gamma^*}(t,e)\}\\
&-F_{Y|T,R}^-(g_{\gamma^*}(t,e)|t,\bar{r})\Big)k_R^-(\tilde{r})/f_R(\bar{r})\\
&+\int_0^u\phi_{\gamma*}^-(e,v)q_0(t,\bar{r};v)k_{Q,0}^-(\tilde{r})\,dv\Bigg)dedu\Bigg)^2 f_{Y,T,R}^-(y,t,\bar{r})dydtd\tilde{r}. \qquad \text{(A.14)}
\end{aligned}$$

In particular, we have used the following convergence result in the above expression:

$$
\begin{aligned}
K_Y((y-y')/b_1) &\to \mathbf{1}\{y' \leq y\}, \\
\tilde{K}_Y(y,t,\bar{r}+b_1\tilde{r};g_{\gamma^*}(t+b_1v,e)) &\to \mathbf{1}\{y \leq g_{\gamma^*}(t,e)\} - F^{-}_{Y|T,R}(g_{\gamma^*}(t,e)|t,\bar{r}).
\end{aligned}
$$

The above derivation shows that

$$
\mathbb{E}\Big[(\zeta_{-}^{DF}(Y,T,R)+\zeta_{-}^{Q}(T,R))\otimes(\zeta_{-}^{DF}(Y,T,R)+\zeta_{-}^{Q}(T,R))\Big] \sim b_1\Sigma_{-}.
$$

Similarly, we can show that

$$
\mathbb{E}\Big[(\zeta_{+}^{DF}(Y,T,R)+\zeta_{+}^{Q}(T,R))\otimes(\zeta_{+}^{DF}(Y,T,R)+\zeta_{+}^{Q}(T,R))\Big] \sim b_1\Sigma_{+},
$$

where

$$
\begin{aligned}
\Sigma_{+} = \int \Bigg(\int_{\mathscr{E}} \int_0^1 & w(e,u)\nabla_\gamma D_{\gamma^*,h^*}(e,u)\Bigg(\iota'\bar{\Omega}_1^{-1}(1,0,\tilde{r})'\Big(\mathbf{1}\{y \leq g_{\gamma^*}(t,e)\} \\
& - F^{+}_{Y|T,R}(g_{\gamma^*}(t,e)|t,\bar{r})\Big)k_R^{+}(\tilde{r})/f_R(\bar{r}) \\
& + \int_0^u \phi^{+}_{\gamma*}(e,v)q_1(t,\bar{r};v)k^{+}_{Q,1}(\tilde{r})\,dv\Bigg)dedu\Bigg)^2 f^{+}_{Y,T,R}(y,t,\bar{r})dydtd\tilde{r}, \qquad \text{(A.15)}
\end{aligned}
$$

and $\bar{\Omega}_1 = \int \boldsymbol{x}\boldsymbol{x}'k_1(\boldsymbol{x})dx_1dx_2$. The terms $\zeta_{-}^{DF}(Y,T,R)$ and $\zeta_{-}^{Q}(T,R)$ contain the factor $\mathbf{1}\{R<0\}$ while the terms $\zeta_{+}^{DF}(Y,T,R)$ and $\zeta_{+}^{Q}(T,R)$ contain the factor $\mathbf{1}\{R \geq 0\}$. Hence, we can compute the variance matrix of their sum as

$$
var\Big((\zeta_{-}^{DF}(Y,T,R)+\zeta_{-}^{Q}(T,R))-(\zeta_{+}^{DF}(Y,T,R)+\zeta_{+}^{Q}(T,R))\Big) = \Sigma_{-}+\Sigma_{+}.
$$

Since Σ_{-} and Σ_{+} do not vary with n, Chebyshev's inequality implies that the following term is

$O_p\left(1/\sqrt{nb_1}\right)$:

$$\frac{1}{nb_1}\sum_{i=1}^{n}\left(\zeta_-^{DF}(Y_i,T_i,R_i)+\zeta_-^{Q}(T_i,R_i)\right)-\frac{1}{nb_1}\sum_{i=1}^{n}\left(\zeta_+^{DF}(Y_i,T_i,R_i)+\zeta_+^{Q}(T_i,R_i)\right).$$

Moreover, $\mathscr{E}$ is compact and the relevant functions in the expressions of $\zeta_\pm^{DF}$ and $\zeta_\pm^{Q}$ are bounded (Assumptions 1.8, 1.9, 1.10, and 1.13). We can apply the Lyapnov's central limit theorem (for example, Theorem 5.11 in White [2001]) to obtain that

$$\begin{aligned}\left(\sqrt{nb_1}(\Sigma_-+\Sigma_+)^{-1/2}\right)&\left(\frac{1}{nb_1}\sum_{i=1}^{n}\left(\zeta_-^{DF}(Y_i,T_i,R_i)+\zeta_-^{Q}(T_i,R_i)\right)\right.\\&\left.-\frac{1}{nb_1}\sum_{i=1}^{n}\left(\zeta_+^{DF}(Y_i,T_i,R_i)+\zeta_+^{Q}(T_i,R_i)\right)\right)\xrightarrow{d}N(0,\boldsymbol{I}_{d_\Gamma}).\end{aligned}$$

Therefore, we obtain, for $\tilde{\gamma}$, the convergence rate: $\|\tilde{\gamma}-\gamma^*\|_2=O_p\left(b_1^2+1/\sqrt{nb_1}\right)$, and asymptotic normality:

$$\left(\sqrt{nb_1}(\Sigma_-+\Sigma_+)^{-1/2}\right)(\Delta(\tilde{\gamma}-\gamma^*)-b_1^2(B_--B_+))\xrightarrow{d}N(0,\boldsymbol{I}_{d_\Gamma}),\tag{A.16}$$

under the condition that $nb_1^7\to 0$ (Assumption 1.12).

Step 7. (Asymptotic normality of $\hat{\gamma}$.) By Equation (A.10), we can apply the triangle inequality repeatedly and obtain that

$$\left\|\hat{L}_{\hat{\gamma}}\right\|_w\le\left\|\hat{Q}_{\hat{\gamma},\hat{h}}\right\|_w+O_p(\alpha_n)\le\left\|\hat{Q}_{\tilde{\gamma},\hat{h}}\right\|_w+O_p(\alpha_n)\le\left\|\hat{L}_{\tilde{\gamma}}\right\|_w+O_p(\alpha_n),$$

where the second inequality uses the definition of $\hat{\gamma}$ in (1.9). Squaring the above inequality and using (A.8) to bounded $\|\hat{L}_{\tilde{\gamma}}\|_w$, we obtain that

$$\begin{aligned}\left\|\hat{L}_{\hat{\gamma}}\right\|_w^2 &\leq \left\|\hat{L}_{\tilde{\gamma}}\right\|_w^2 + O_p(\alpha_n^2) + O_p(\alpha_n \tilde{\alpha}_n) \leq \left\|\hat{L}_{\tilde{\gamma}}\right\|_w^2 + O_p(\alpha_n \tilde{\alpha}_n) \\ &= \left\|\hat{L}_{\tilde{\gamma}}\right\|_w^2 + O_p\left(\left(b_1^2 + \sqrt{\log n/(nb_1^4)}\right)\left(b_1^4 + n^{-1} b_1^{-7/6-2\bar{\varepsilon}/5}\right)\right).\end{aligned}$$

Thus, we have

$$\left\|\hat{L}_{\hat{\gamma}}\right\|_w^2 - \left\|\hat{L}_{\tilde{\gamma}}\right\|_w^2 = O_p\left(b_1^6 + n^{-1} b_1^{5/6-2\bar{\varepsilon}/5} + b_1^2\sqrt{\log n/n} + \sqrt{\log n} n^{-3/2} b_1^{-3\frac{1}{6}-2\bar{\varepsilon}/5}\right).$$

We want to show that the four terms inside the O_p-notation in the above equation is $o(1/(nb_1))$. Both the terms b_1^6 and $b_1^2\sqrt{\log n/n}$ are $o(1/(nb_1))$ under Assumption 1.12(ii). The term $n^{-1} b_1^{5/6-2\bar{\varepsilon}/5}$ is $o(1/(nb_1))$ since $b_1 = o(1)$. For the fourth term, we have

$$\sqrt{\log n} n^{-3/2} b_1^{-3\frac{1}{6}-2\bar{\varepsilon}/5} = o(1/(nb_1)) \iff nb_1^{4\frac{1}{3}+\bar{\varepsilon}} b_1^{-\bar{\varepsilon}/5} / \log n \to \infty,$$

where the statement on the RHS is true by Assumption 1.12(iii). The above derivations show that

$$\left\|\hat{L}_{\hat{\gamma}}\right\|_w^2 - \left\|\hat{L}_{\tilde{\gamma}}\right\|_w^2 = o_p(b_1^2 + 1/(nb_1)).$$

By adding and subtracting $(\tilde{\gamma} - \gamma^*)\nabla_\gamma D_{\gamma^*,h^*}$, we obtain that

$$\left\|\hat{L}_{\hat{\gamma}}\right\|_w^2 = \left\|\tilde{L}_{\hat{\gamma}}\right\|_w^2 + \left\|(\hat{\gamma} - \tilde{\gamma})\nabla_\gamma D_{\gamma^*,h^*}\right\|_w^2 + 2(\hat{\gamma} - \tilde{\gamma}) \int \tilde{L}_{\tilde{\gamma}}(e,u) \nabla_\gamma D_{\gamma^*,h^*}(e,u) w(e,u) dedu.$$

The last term (the innner product term) above is zero because $\tilde{L}_{\tilde{\gamma}}$ is orthogonal to $\nabla_\gamma D_{\gamma^*,h^*}$ from the projection perspective. This can also be verified by using the definition of $\tilde{\gamma}$. Hence, we have $\left\|(\hat{\gamma} - \tilde{\gamma})\nabla_\gamma D_{\gamma^*,h^*}\right\|_w^2 = o_p(1/(nb_1))$. By the same argument as in Step 3, we can show that $\|\hat{\gamma} - \tilde{\gamma}\|_2 = o_p(1/\sqrt{nb_1})$. Therefore, by (A.16) and Slutsky's theorem, we obtain the desired asymptotic distribution of $\hat{\gamma}$:

$$\begin{aligned}
&\left(\sqrt{nb_1}(\Sigma_- + \Sigma_+)^{-1/2}\right)(\Delta(\hat{\gamma}-\gamma^*) - b_1^2(B_- - B_+)) \\
=&\left(\sqrt{nb_1}(\Sigma_- + \Sigma_+)^{-1/2}\right)(\Delta(\tilde{\gamma}-\gamma^*) - b_1^2(B_- - B_+)) + o_p(1) \xrightarrow{d} N(0, \boldsymbol{I}_{d_\Gamma}).
\end{aligned}$$

□

Proof of Proposition 1.1. We only prove the results for $\hat{h}_0(\bar{r},\cdot)$ since the results for $\hat{h}_1(\bar{r},\cdot)$ can be proved analogously. For part (i) of Assumption 1.13, we can set the partition $\mathscr{P}_0^n$ to be the class of intervals $\{[u_j, u_{j+1}] : j = 0, \cdots, J_n\}$. The estimator $\hat{h}_0(\bar{r},\cdot)$ is a linear function within each interval and hence is contained in the class $\mathscr{H}_0^n(\mathscr{P}_0^n)$.

For part (ii), notice that, under Assumption 1.9(i), the estimator $\hat{h}_0(\bar{r},u_0)$ and $\hat{h}_0(\bar{r},u_{J_n+1})$ converge to t_0' and t_0'', respectively, at the $1/n$ rate. Therefore, we can replace $\hat{h}_0(\bar{r},u_0)$ by t_0' and $\hat{h}_0(\bar{r},u_{J_n+1})$ by t_0'' without affecting the asymptotics. Let $\tilde{h}_0(\bar{r},u)$ denote the solution of (1.10) at given u. The uniform asymptotic linear representation for $\hat{h}_0(\bar{r},u), u \in (0,1)$ follows from Lemma 3 in the Appendix of Dong et al. [2021], which is a slight modification of Theorem 1.2 of Qu and Yoon [2015]. Then we can use Step 2 in the proof of Theorem 2 in Qu and Yoon [2015] to show that the error induced by linear interpolation is asymptotically negligible.

The uniform convergence rate in Part (iii) of Assumption 1.13 can be shown by using the uniform asymptotic linear representation. Since v_0 is bounded, the bias term is $O(b_1^2)$. In Lemma A.9, we show that the stochastic term is satisfies

$$\sup_{u\in[0,1]} \left| \frac{1}{nb_1} \sum_{i=1}^n q_0(T_i, R_i; u) k_{Q,0}\left(\frac{R_i - \bar{r}}{b_1}\right) \mathbf{1}\{R_i < \bar{r}\} \right| = O_p\left(\sqrt{\log n / nb_1}\right).$$

This proves the desired result.

□

A.2.2 Uniform convergence rates and the empirical process theory

Below are some basic concepts and results from the empirical process theory which are used to prove several uniform convergence results.

Let $\mathscr{F}$ be a class of uniformly bounded measurable matrix-valued functions, that is, there exists $M > 0$ such that, for all $f \in \mathscr{F}$, $\|f\|_2 \leq M$. Let $N(\mathscr{F}, P, \varepsilon)$ be the ε-covering number of the metric space $(\mathscr{F}, L_2(P))$, that is, $N(\mathscr{F}, P, \varepsilon)$ is defined as the minimal number of open $\|\cdot\|_{L_2(P)}$-balls of radius ε and centers in $\mathscr{F}$ required to cover $\mathscr{F}$.

We say that a uniformly bounded function class $\mathscr{F}$ is *Euclidean* if there exists $A_1, A_2 > 0$ (that only depend on the uniform bound) such that for every probability measure P and every $\varepsilon \in (0,1]$, $N(\mathscr{F}, P, \varepsilon) \leq A_1/\varepsilon^{A_2}$. We say that a function class $\mathscr{F}$ is *log-Euclidean* with coefficient $\rho \in (0,1)$ if there exists $A > 0$ (that only depends on the uniform bound) such that for every probability measure P and every $\varepsilon \in (0,1]$, $\log N(\mathscr{F}, P, \varepsilon) \leq A/\varepsilon^{2\rho}$.

The above definition of Euclidean classes is introduced by Nolan and Pollard [1987]. The same concept is also studied by Giné and Guillou [1999], but they refer to what we call "Euclidean" as "VC." There is a slight difference that Nolan and Pollard [1987] use the L_1-norm while Giné and Guillou [1999] use the L_2-norm. We ignored the envelope in their definition since we only work with uniformly bounded $\mathscr{F}$. The following two lemmas demonstrates how to generate function classes that are Euclidean and log-Euclidean.

Lemma A.2. *Let $\mathscr{F}_1$ and $\mathscr{F}_2$ be uniformly bounded and Euclidean classes of functions. The following classes of functions are also uniformly bounded and Euclidean.*

(i) $\mathscr{F}_1 \oplus \mathscr{F}_2 = \{f_1 + f_2 : f_1 \in \mathscr{F}_1, f_2 \in \mathscr{F}_2\}$.

(ii) $\mathscr{F}_1\mathscr{F}_2 = \{f_1 \cdot f_2 : f_1 \in \mathscr{F}_1, f_2 \in \mathscr{F}_2\}$.

(iii) $\{\mathbb{E}[f_1(\cdot)|X] : f_1 \in \mathscr{F}_1\}$.

(iv) $\Big\{k\big((\cdot - x)/b\big) : x \in \mathbb{R}, b > 0\Big\}$*, where* $k : \mathbb{R} \to \mathbb{R}$ *is a function of bounded variation.*

Proof of Lemma A.2. See Appendix B in Chapter 2 of this dissertation. □

Lemma A.3. *Let $\mathscr{F}_1$ be a uniformly bounded and Euclidean class of functions and $\mathscr{F}_2$ be a uniformly bounded and log-Euclidean class of functions with coefficient ρ. Then $\mathscr{F}_1\mathscr{F}_2$ is uniformly bounded and log-Euclidean with coefficient $\rho+\varepsilon$ for any $\varepsilon>0$.*

The following two lemmas give the asymptotic order of the supremum of empirical processes generated by Euclidean and log-Euclidean classes, respectively.

Proof of Lemma A.3. This follows from the definition of Euclidean and log-Euclidean classes. □

Lemma A.4. *Let $X_1,\cdots,X_n$ be an iid sample of a random vector X in $\mathbb{R}^d$. Let $\mathscr{G}_n$ be a sequence of classes of measurable real-valued functions defined on $\mathbb{R}^d$. Assume that there is a fixed uniformly bounded Euclidean class $\mathscr{F}$ such that $\mathscr{F}_n \subset \mathscr{F}$ for all n. Let $\sigma_n^2 \geq \sup_{f\in\mathscr{F}_n}\mathbb{E}[f(X)^2]$. Then*

$$\sup_{f\in\mathscr{F}_n}\left|\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X_i))\right| = O_p\left(\sqrt{n\sigma_n^2|\log\sigma_n|}+|\log\sigma_n|\right).$$

In particular, if $n\sigma_n^2/|\log\sigma_n|\to\infty$, then the above rate simplifies to $O_p\left(\sqrt{n\sigma_n^2|\log\sigma_n|}\right)$.

Proof of Lemma A.4. This is Lemma 2.1 in Chapter 2 of this dissertation. □

Lemma A.5. *Let $X_1,\cdots,X_n$ be an iid sample of a random vector X in $\mathbb{R}^d$. Let $\mathscr{F}_n$ be a sequence of classes of measurable real-valued functions defined on $\mathbb{R}^d$. Assume that there is a fixed uniformly bounded log-Euclidean class $\mathscr{F}$ with coefficient ρ such that $\mathscr{F}_n \subset \mathscr{F}$ for all n. Let $\sigma_n^2 = \sup_{f\in\mathscr{F}_n}\mathbb{E}[f(X)^2]$. Then*

$$\sup_{f\in\mathscr{F}_n}\left|\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X_i))\right| = O_p\left(\sqrt{n}\sigma_n^{1-\rho}+n^{\rho/(1+\rho)}\right).$$

Proof of Lemma A.5. Let $M > 0$ be the uniform bound of $\mathscr{F}$. Since $\mathscr{F}$ is log-Euclidean with coefficient ρ, there exists $A > 0$ such that $\log N(\mathscr{F}, P_n, \varepsilon) \leq A/\varepsilon^{\rho}$ for every $\varepsilon \in (0,1]$, where P_n is the empirical measure. Since each $\mathscr{F}_n$ is contained in $\mathscr{F}$, the above result also holds when $\mathscr{F}$ is replaced by $\mathscr{F}_n$. Denote $Rad_i, 1 \leq i \leq n$, as a sequence of iid Rademacher variables. By Equation (3.19) in Koltchinskii [2011] (which is a result of Theorem 3.12 in the same book), there exists a universal constant $C > 0$ such that

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} Rad_i f(X_i) \right| &\leq C A^{\rho} M^{\rho} \sqrt{n} \sigma_n^{1-\rho} \vee C A^{2\rho/(\rho+1)} M n^{\rho/(1+\rho)} \\ &= O_p\left(\sqrt{n} \sigma_n^{1-\rho} + n^{\rho/(1+\rho)} \right).\end{aligned}$$

Then the desired result follows from the usual symmetrization argument (for example, Theorem 2.1 in Koltchinskii [2011]) and Chebyshev's inequality. □

The following three lemmas give uniform convergence results that are used in the proof of Theorem 1.2.

Lemma A.6. *Under the assumptions of Theorem 1.2, the following term is $O_p(\tilde{\alpha}_n)$:*

$$\sup_{e \in \mathscr{E}, u \in [0,1], \gamma \in \Gamma, h_0 \in \mathscr{H}_0(\mathscr{P}_0^n)} \left| \frac{1}{n b_1^2} \sum_{i=1}^{n} \int_0^u \iota' \Xi_0(h_0(v))^{-1} s_0(Y_i, T_i, R_i; g_{\gamma}(h_0(v), \bar{r}, e), h_0(v)) dv \right|.$$

Proof of Lemma A.6. Since $\mathscr{P}_0^n$ is a finite partition, we can without loss of generality assume that $\mathscr{P}_0^n$ only contains the whole interval $[t_0', t_0'']$ so that there is effectively no partition. To simply notation, we omit the term $\mathscr{P}_0^n$. By the change of variables $\tilde{v} = (T_i - h_0(v))/b_1$ and Fubini's theorem, we have

$$\begin{aligned}
&\Big|\frac{1}{nb_1^2}\sum_{i=1}^n\int_0^u \iota'\Xi_0(h_0(v))^{-1}s_0(Y_i,T_i,R_i;g_\gamma(h_0(v),\bar{r},e),h_0(v))dv\Big|\\
&\leq \int\Big|\frac{1}{nb_1}\sum_{i=1}^n \iota'\Xi_0(T_i+b_1\tilde{v})^{-1}s_0(Y_i,T_i,R_i;g_\gamma(T_i+b_1\tilde{v},\bar{r},e),T_i+b_1\tilde{v})\\
&\qquad\times(h_0^{-1})'(T_i+b_1\tilde{v})\mathbf{1}\{(T_i-h_0(u))/b_1<\tilde{v}<(T_i-t_0')/b_1\}d\tilde{v}\Big|\\
&\leq \sup_{\tilde{v}\in(-1,1)}\Big|\frac{1}{nb_1}\sum_{i=1}^n \iota'\Xi_0(T_i+b_1\tilde{v})^{-1}s_0(Y_i,T_i,R_i;g_\gamma(T_i+b_1\tilde{v},\bar{r},e),T_i+b_1\tilde{v})\\
&\qquad\times(h_0^{-1})'(T_i+b_1\tilde{v})\mathbf{1}\{(T_i-h_0(u))/b_1<\tilde{v}<(T_i-t_0')/b_1\}\Big|,
\end{aligned}$$

where, in the last inequality, the supremum is taken over $\tilde{v}\in(-1,1)$ because of the support of k_T. Define the following function of (Y,T,R) indexed by (v,u,e,γ,h_0):

$$\begin{aligned}
\psi_n(Y,T,R;v,u,e,\gamma,h_0)&=\iota'\Xi_0(T+b_1v)^{-1}s_0(Y,T,R;g_\gamma(T+b_1v,\bar{r},e),T+b_1v)\\
&\qquad\times(h_0^{-1})'(T+b_1v)\mathbf{1}\{(T-h_0(u))/b_1<v<(T-t_0')/b_1\}.
\end{aligned}$$

Let $\Psi_n=\{\psi_n(\cdot,\cdot,\cdot;v,u,e,\gamma,h):v\in(-1,1),u\in(0,1),e\in\mathscr{E},\gamma\in\Gamma,h_0\in\mathscr{H}_0\}$. Our goal is to use empirical process theory to derive the asymptotic order of

$$\sup_{\psi_n\in\Psi_n}|\sum_{i=1}^n\psi_n(Y,T,R;v,u,e,\gamma,h_0)|.$$

Consider a larger class Ψ as the product $\Psi = \Psi_{\Xi}\Psi_Y\Psi_{TR}\Psi_{\mathscr{H}_0}$, where

$$\begin{aligned}
\Psi_{\Xi_0} &= \{T \mapsto \iota'\Xi_0(T+v)^{-1} : v \in (-1,1)\}, \\
\Psi_Y &= \{(Y,T,R) \mapsto \tilde{K}_Y(Y,T,R; g_\gamma(T+v,\bar{r},e), T+v) : v \in (-1,1), \gamma \in \Gamma, e \in \mathscr{E}\}, \\
\Psi_{TR} &= \{(Y,T,R) \mapsto (1, v, (R-\bar{r})/b)' k_T(v) k_R^-((R-\bar{r})/b) \\
&\qquad \times \mathbf{1}\{T - h_0(u) < bv < T - t_0'\} : b, u \in (0,1), v \in (-1,1)\}, \\
\Psi_{\mathscr{H}_0} &= \{T \mapsto (h_0^{-1})'(T+v) : h_0 \in \mathscr{H}_0, v \in (0,1)\}.
\end{aligned}$$

Notice that in the above definition of Ψ_{Ξ_0}, Ψ_Y, and $\Psi_{\mathscr{H}_0}$, omitting the parameter b does not change the class under consideration. The class Ψ does not vary with n and $\Psi_n \subset \Psi, n \geq 1$. In the following paragraphs, we show that the classes Ψ_{Ξ_0}, Ψ_Y, and Ψ_{TR} are Euclidean while the class $\Psi_{\mathscr{H}_0}$ is log-Euclidean.

For Ψ_{Ξ_0}, we know that $\|\Xi_0\|$ and $\|\Xi_0^{-1}\|$ are uniformly bounded by Lemma 2.1 in Chapter 2 of this dissertation. By the smoothness of $f_{T,R}^-$ in Assumption 1.9, the class $\{T \mapsto \iota'\Xi_0(T+v) : v \in (-1,1)\}$ is Lipschitz in the parameter $v \in (-1,1)$ and hence, by Theorem 2.7.11 in van der Vaart and Wellner [1996b], has covering numbers bounded by that of one-dimensional intervals. This implies that $\{T \mapsto \iota'\Xi_0(T+v) : v \in (-1,1)\}$ is uniformly bounded and Euclidean. Then by Theorem 3 in Andrews [1994], we know that Ψ_{Ξ_0} is uniformly bounded and Euclidean.

The class Ψ_Y can be written as $\Psi_Y = \Psi_{Y1} + \Psi_{Y2}$, where

$$\begin{aligned}
\Psi_{Y1} &= \{(Y,T) \mapsto K_Y((g_\gamma(T+v,\bar{r},e) - Y)/b) : b, v \in (0,1), \gamma \in \Gamma, e \in \mathscr{E}\}, \\
\Psi_{Y2} &= \{(Y,T,R) \mapsto -\mathbb{E}[K_Y((g_\gamma(T+v,\bar{r},e) - Y)/b) | T, R] : b, v \in (0,1), \gamma \in \Gamma, e \in \mathscr{E}\}.
\end{aligned}$$

In view of Lemma A.2(i) and (iii), we only need to show that Ψ_Y is Euclidean. The class Ψ_Y is uniformly bounded by 1. The function K_Y is increasing since k_Y is positive. The subgraph class of Ψ_Y can be written as

$$\begin{aligned}&\{\{(y,t,s) : K_Y((g_\gamma(t+v,\bar{r},e)-y)/b) \leq s\} : b,v \in (0,1), \gamma \in \Gamma, e \in \mathscr{E}\}\\ =&\{\{(y,t,s) : g_\gamma(t+v,\bar{r},e)-y-bK^{-1}(s) \leq 0\} : b,v \in (0,1), \gamma \in \Gamma, e \in \mathscr{E}\}.\end{aligned}$$

By Assumption 1.10, the following function class is finite-dimensional:

$$\{(t,y,s) \mapsto g_\gamma(t+v,\bar{r},e)-y-bK^{-1}(s) : b,v \in (0,1), \gamma \in \Gamma, e \in \mathscr{E}\}.$$

By Lemma 18(ii) in Nolan and Pollard [1987], the subgraph class of Ψ_Y is a polynomial class, which implies (by Theorem 2.6.7 in van der Vaart and Wellner [1996b]) that Ψ_Y is Euclidean.

For the class Ψ_{TR}, notice that the function k_T, k_R, and the indicator function are all of bounded variation. The kernel functions k_T and k_R are supported on $[-1,1]$. Therefore, the term $(R-\bar{r})/b$ is bounded between $[-1,1]$. By Lemma A.2(ii) and (iv), we know that Ψ_{TR} is uniformly bounded and Euclidean.

Lastly, by Assumption 1.13(i), the class $\Psi_{\mathscr{H}_0}$ is contained in the class of twice continuously diferentiable functions whose second-order derivatives are Lipschitize continuous. By the well-known bounds on the entropy of Lipschitz classes (see, for example, Example 5.11 in Chapter 5 of Wainwright [2019]), we know the class $\Psi_{\mathscr{H}_0}$ is log-Euclidean with coefficient $1/2 \times 1/(2+1) = 1/6$. Then by Lemma A.3, we know that Ψ is log-Euclidean with coefficient $1/6+\varepsilon$ for any small $\varepsilon > 0$.

Next, we want to derive a uniform variance bound for the class Ψ and appeal to Lemma A.5. By the uniform boundedness of the classes studied above and applying the usual change of variables, we obtain that

$$\mathbb{E}[\psi_n(Y,T,R;v,u,e,\gamma,h_0)^2] \leq C\mathbb{E}[k_R((R-\bar{r})/b_1)^2] = Cb_1 \int k_R(\tilde{r}) f_R(\bar{r}+b_1\tilde{r}) d\tilde{r} = O(b_1).$$

Lemma A.5 then gives that

$$\sup_{\psi\in\Psi_n}\Big|\sum_{i=1}^{n}\psi(Y,T,R;v,u,e,\gamma,h_0)\Big| = O_p\left(n^{1/2}b_1^{(1-1/6)/2-\varepsilon}+n^{1/7}\right),$$

for any small $\varepsilon>0$. Notice that, in the rate specified above, the term $n^{1/7}$ is dominated in view of Assumption 1.12. Then the desired convergence rate follows from dividing by nb_1 on both sides.

□

Lemma A.7. *Under the assumptions of Theorem 1.2, we have*

$$\sup_{y\in\mathbb{R},t\in[t_0',t_0'']}\left\|\frac{1}{nb_1^2}\sum_{i=1}^{n}\frac{\partial}{\partial y}s_0(Y_i,T_i,R_i;y,t)\right\|_2 = O_p\left(\sqrt{\log n/(nb_1^4)}\right),$$
$$\sup_{y\in\mathbb{R},t\in[t_0',t_0'']}\left\|\frac{1}{nb_1^2}\sum_{i=1}^{n}\frac{\partial}{\partial t}s_0(Y_i,T_i,R_i;y,t)\right\|_2 = O_p\left(\sqrt{\log n/(nb_1^4)}\right).$$

Proof of Lemma A.7. The partial derivative of s_0 with respect to y is a vector of length three whose generic element can be denoted by $\dot{s}_0(Y_i,T_i,R_i;y,t,b_1,b_2)/b_2$, where

$$\dot{s}_0(Y,T,R;y,t)=\left(\frac{T-t}{b_1}\right)^{\ell_1}\left(\frac{R-\bar{r}}{b_1}\right)^{\ell_2}\left(k_Y\left(\frac{y-Y}{b_2}\right)-\mathbb{E}\left[k_Y\left((y-Y)/b_2\right)|T,R\right]\right)k_0(X(t))$$

with $(\ell_1,\ell_2)=(0,0),(1,0),(0,1)$. We use the empirical process theory to derive the uniform convergence rate of the sample average of $\dot{s}_0$. Recall that the kernel functions k_Y,k_T, and k_R are of bounded variation. Then by Lemma A.2, we know that the following function class is uniformly bounded and Euclidean:

$$\begin{aligned}&\{(Y,T,R)\mapsto\dot{s}_0(Y,T,R;y,t,b,b'):y\in\mathbb{R},t\in[t_0',t_0''],b,b'>0\}\\=&\{(Y,T,R)\mapsto((T-t)/b)^{\ell_1}\left((R-\bar{r})/b\right)^{\ell_2}\left(k_Y\left((y-Y)/b'\right)-\mathbb{E}\left[k_Y\left((y-Y)/b'\right)|T,R\right]\right)\\&\quad\times k_T((T-t)/b)k_R^-((R-\bar{r})/b):y\in\mathbb{R},t\in[t_0',t_0''],b,b'>0\}.\end{aligned}$$

By the law of iterated expectations and differentiation under the integral, we know that $\dot{s}_0$ is centered. By using the fact that k_T and k_R are supported on $[-1,1]$ and k_Y is bounded and applying the standard change of variables, we can bound the variance of $\dot{s}_0$ by

$$2\|k_Y\|_\infty^2 \mathbb{E}\left[k_T((T-t)/b_1)^2 k_R^-((R-\bar{r})/b_1)^2\right] = b_1^2 2\|k_Y\|_\infty^2 \int k_T(x_1)^2 k_R^-(x_2)^2 = O(b_1^2),$$

uniformly over $y \in \mathbb{R}$ and $t \in [t_0', t_0'']$. Then by Lemma 2.2, we know that the uniform convergence rate of the sample average of $\dot{s}_0$ is $O_p\left(\sqrt{nb_1^2 \log n}\right)$. Therefore,

$$\sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n \frac{\partial}{\partial y} s_0(Y_i, T_i, R_i; y, t) \right\|_2 = \frac{1}{nb_1^2 b_2} O_p\left(\sqrt{nb_1^2 \log n}\right) = O_p\left(\sqrt{\log n/(nb_1^4)}\right)$$

under the condition that $b_1/b_2 \in [1/C, C]$ (Assumption 1.12). This proves the first claim of the lemma. For the second claim, the same argument applies. We just want to point out that k_T is differentiable on the entire real line by Assumption 1.11 even though its support is $[-1,1]$.

□

Lemma A.8. *Under the assumptions of Theorem 1.2, we have*

$$\sup_{y\in\mathscr{Y}, t\in[t_0', t_0'']} \left| \frac{1}{nb_1^2} \sum_{i=1}^n \iota' \Xi_0(t)^{-1} s_0(Y_i, T_i, R_i; y, t) \right| = O_p\left(\sqrt{\log n/(nb_1^2)}\right),$$

$$\sup_{y\in\mathbb{R}, t\in[t_0', t_0'']} \left\| \frac{1}{nb_1^2} \sum_{i=1}^n s_0(Y_i, T_i, R_i; y, t) \right\|_2 = O_p\left(\sqrt{\log n/(nb_1^2)}\right).$$

Proof of Lemma A.8. Following the same steps as in the proofs of the previous two lemmas, we can show that the relevant function classes are uniformly bounded and Euclidean. By the usual change of variables, we can show that the uniform variance bound is $O(b_1^2)$ before taking into account the factor $1/(nb_1)$ in the two terms. Then the desired results follow from Lemma 2.2. The details are omitted for brevity. □

Lemma A.9.

$$\sup_{u\in[0,1]}\left|\frac{1}{nb_1}\sum_{i=1}^{n}q_0(T_i,R_i;u)k_{Q,0}\left(\frac{R_i-\bar{r}}{b_1}\right)\mathbf{1}\{R_i<\bar{r}\}\right|=O_p\left(\sqrt{\log n/nb_1}\right).$$

Proof of Lemma A.9. Without loss of generality, let $c=1$. Define

$$\begin{aligned}\psi_n(T,R;u)&=q_0(T,R;u)k_{Q,0}\left(\frac{R-\bar{r}}{b_1}\right)\mathbf{1}\{R<\bar{r}\}\\&=\frac{u-\mathbf{1}\{T\le h_0^*(\bar{r},u)\}}{f_R(\bar{r})f_{T|R}^{-}(h_0^*(\bar{r},u)|\bar{r})}\iota'\Omega_{Q,0}^{-1}(1,(R-\bar{r})/b_1)'K_{FS}((R-\bar{r})/b_1)\end{aligned}$$

and $\Psi_n=\{(T,R)\mapsto\psi_n(T,R;u):u\in[0,1]\}$. By the law of iterated expectations, ψ_n is centered. Let $M=\sup_{u\in[0,1]}|f_R(\bar{r})f_{T|R}^{-}(h_0^*(\bar{r},u)|\bar{r})|$. Define a product class $\Psi=\Psi_T\Psi_R$ where

$$\begin{aligned}\Psi_T&=\{(T,R)\mapsto C(u-\mathbf{1}\{T\le t\}):u\in[0,1],t\in[t_0',t_0''],|C|\le M\},\\\Psi_R&=\{(T,R)\mapsto\iota'\Omega_{Q,0}^{-1}(1,(R-\bar{r})/b)'K_{FS}((R-\bar{r})/b):b>0\}.\end{aligned}$$

The class Ψ does not vary with n, and $\Psi_n\subset\Psi,n\ge1$. The class Ψ_T is uniformly bounded and Euclidean since the set of indicator functions $\mathbf{1}\{T\le t\},t\in[t_0',t_0'']$ is Euclidean. The class Ψ_R is uniformly bounded and Euclidean since K_{FS} is of bounded variation and compactly supported. By the usual change of variables, we can show that the uniform variance bound for Ψ_n is $O(b_1)$. Then the desired convergence rate follows from Lemma 2.2.

□

A.2.3 Covariance Matrix Estimation

In this section, we discuss the estimation of the asymptotic variance matrix of $\hat{\gamma}$, which involves the estimation of Δ, Σ_-, and Σ_+. For concreteness, we consider the first-step nonparametric conditional quantile estimation procedure described in Section 1.3.3 and Proposition 1.1. In the expressions of Δ, Σ_-, and Σ_+, the functions that require estimation include $\nabla_\gamma D_{\gamma^*,h^*}$, $\phi_{\gamma^*}^{\pm}$,

$f^{\pm}_{Y,T,R}$, $f^{\pm}_{T|R}$, and f_R. By definition,

$$\begin{aligned}\nabla_\gamma D_{\gamma^*}(e,u) = \int_0^u \Big[&\frac{\partial}{\partial Y} F^{-}_{Y|T,R}(g_{\gamma^*}(h^*_0(\bar{r},v),\bar{r},e)|h^*_0(\bar{r},v),\bar{r}) \nabla_\gamma g_{\gamma^*}(h^*_0(\bar{r},v),\bar{r},e) \\ &- \frac{\partial}{\partial Y} F^{+}_{Y|T,R}(g_{\gamma^*}(h^*_1(\bar{r},v),\bar{r},e)|h^*_1(\bar{r},v),\bar{r}) \nabla_\gamma g_{\gamma^*}(h^*_1(\bar{r},v),\bar{r},e) \Big] dv.\end{aligned}$$

In the above quantity, we only need to estimate $\frac{\partial}{\partial Y} F^{-}_{Y|T,R} = f^{\pm}_{Y|T,R}$ since we already have estimators for γ^* and h^*. By observing the definition of $\phi^{\pm}_{\gamma^*}$, we know that the additional term that requires estimation is $\frac{\partial}{\partial T} F^{\pm}_{Y|T,R}$. To summarize, we want to estimate $f^{\pm}_{Y,T,R}$ and $\frac{\partial}{\partial T} F^{\pm}_{Y|T,R}$. Once $f^{\pm}_{Y,T,R}$ is obtained, we can operate to get the marginal and conditional density functions.

For estimation of $f^{\pm}_{Y,T,R}$, we can employ the method developed by Cattaneo et al. [2020]. They use the second-order local polynomial regression to estimate the joint density. Due to the nature of local polynomial regressions, the estimator is boundary adaptive and particularly suitable for RD designs. To estimate the partial derivative $\frac{\partial}{\partial T} F^{\pm}_{Y|T,R}$, we can employ a second-order local polynomial regression. The procedure is similar to STEP 2 in the construction of $\hat{\gamma}$. We add two quadratic terms into the minimization problem:

$$\begin{aligned}\sum_{i:R_i<\bar{r}} &\left(K_Y\left(\frac{y-Y_i}{b_2}\right) - a^{-} - a^{-}_T(T_i - t) - a^{-}_{T,2}(T_i-t)^2 - a^{-}_R(R_i-\bar{r}) - a^{-}_{R,2}(R_i-\bar{r})^2 \right)^2 \\ &\times k_T\left(\frac{T_i - t}{b_1}\right) k_R\left(\frac{R_i-\bar{r}}{b_1}\right).\end{aligned}$$

The minimizer $\hat{a}^{-}_T$ is the estimate of $\frac{\partial}{\partial T} F^{-}_{Y|T,R}(y|t,\bar{r})$. The estimate of $\frac{\partial}{\partial T} F^{+}_{Y|T,R}(y|t,\bar{r})$ can be analogously constructed.

We assume that the resulting estimators $\hat{f}^{\pm}_{Y,T,R}$ and $\frac{\partial}{\partial T} \hat{F}^{\pm}_{Y|T,R}$ are uniformly consistent.

That is,

$$\sup_{y,t,r} |\hat{f}^{\pm}(y,t,r) - f^{\pm}(y,t,r)| = o_p(1),$$
$$\sup_{y,t} \left| \frac{\partial}{\partial T} \hat{F}^{\pm}_{Y|T,R}(y|t,\bar{r}) - \frac{\partial}{\partial T} F^{\pm}_{Y|T,R}(y|t,\bar{r}) \right| = o_p(1).$$

Such uniform convergence results can be proved along the lines of, for examples, Fan and Guerre [2016] and Chapter 2 of this dissertation. The details are omitted here. We can construct the following distributional estimates:

$$\hat{f}^{\pm}_{T,R}(t,\bar{r}) = \int \hat{f}^{\pm}_{Y,T,R}(y,t,\bar{r})dy,$$
$$\hat{f}^{\pm}_{Y|T,R}(y,t|\bar{r}) = \hat{f}^{\pm}_{Y,T,R}(y,t,\bar{r}) / \hat{f}^{\pm}_{T,R}(t,\bar{r}).$$

Under the assumption that $f^{\pm}_{T,R}$ is bounded away from zero, the estimator $\hat{f}^{\pm}_{Y|T,R}(y,t|\bar{r})$ is uniformly consistent. Let

$$\hat{\Delta} = \int w(e,u) \bigg(\int_0^u \Big[\hat{f}^{-}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_0(\bar{r},v),\bar{r},e)|\hat{h}_0(\bar{r},v),\bar{r}) \nabla_{\gamma} g_{\hat{\gamma}}(\hat{h}_0(\bar{r},v),\bar{r},e)$$
$$- \hat{f}^{+}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_1(\bar{r},v),\bar{r},e)|\hat{h}_1(\bar{r},v),\bar{r}) \nabla_{\gamma} g_{\hat{\gamma}}(\hat{h}_1(\bar{r},v),\bar{r},e) \Big] dv \bigg)^2 dedu.$$

Under the uniform consistency of $\hat{f}^{\pm}_{Y|T,R}$ and $\hat{h}$ and the consistency of $\hat{\gamma}$, we can show that $\hat{\Delta}$ is a consistent estimator of Δ. For the estimation of Σ_{-}, define

$$\hat{f}_R(\bar{r}) = \int \hat{f}^{\pm}_{T,R}(t,\bar{r})dt,$$
$$\hat{f}^{\pm}_{T|R}(t|\bar{r}) = \hat{f}^{\pm}_{T,R}(t,\bar{r}) / \hat{f}_R(\bar{r}),$$
$$\hat{\phi}^{-}_{\hat{\gamma}}(e,v) = \hat{f}^{-}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_0(v),e)|\hat{h}_0(v),\bar{r}) \frac{\partial}{\partial T} g_{\hat{\gamma}}(\hat{h}_0(v),e) + \frac{\partial}{\partial T} \hat{F}^{-}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_0(v),e)|\hat{h}_0(v),\bar{r}).$$

The above estimators are also uniformly consistent. In particular,

$$\sup_{e,v}\left|\hat{\phi}_{\hat{\gamma}}^{-}(e,v)-\phi_{\gamma^*}^{-}(e,v)\right|=o_p(1).$$

Let

$$\begin{aligned}\hat{\Sigma}_- = \int\Bigg(\int_{\mathscr{E}}\int_0^1 w(e,u)\Bigg(\int_0^u \Big[&\hat{f}^{-}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_0(\bar{r},v),\bar{r},e)|\hat{h}_0(\bar{r},v),\bar{r})\nabla_{\gamma}g_{\hat{\gamma}}(\hat{h}_0(\bar{r},v),\bar{r},e)\\ &-\hat{f}^{+}_{Y|T,R}(g_{\hat{\gamma}}(\hat{h}_1(\bar{r},v),\bar{r},e)|\hat{h}_1(\bar{r},v),\bar{r})\nabla_{\gamma}g_{\hat{\gamma}}(\hat{h}_1(\bar{r},v),\bar{r},e)\Big]dv\Bigg)\\ &\Bigg(\iota'\bar{\Omega}_0^{-1}(1,0,\tilde{r})'\Big(\mathbf{1}\{y\leq g_{\hat{\gamma}}(t,e)\}-\hat{F}^{-}_{Y|T,R}(g_{\hat{\gamma}}(t,e)|t,\bar{r})\Big)k_R^{-}(\tilde{r})/\hat{f}_R(\bar{r})\\ &+\frac{k^{-}_{Q,0}(\tilde{r})}{c\hat{f}^{-}_{T,R}(t,\bar{r})}\int_0^u\hat{\phi}^{-}_{\hat{\gamma}}(e,v)(v-\mathbf{1}\{t\leq\hat{h}_0(\bar{r},v)\})dv\Bigg)dedu\Bigg)^2\hat{f}^{-}_{Y,T,R}(y,t,\bar{r})dydtd\tilde{r}.\end{aligned}$$

Under the uniform consistency of $\hat{f}_R$, $\hat{f}^{\pm}_{T,R}$, $\hat{f}^{\pm}_{Y|T,R}$, $\hat{F}_{Y|T,R}$, $\hat{f}^{\pm}_{Y,T,R}$, $\hat{\phi}^{-}_{\hat{\gamma}}$, and $\hat{h}$, and the consistency of $\hat{\gamma}$, $\hat{\Sigma}_-$ is a consistent estimator of Σ_-. The estimation of Σ_+ can be performed analogously.

Appendix B

Appendix for Chapter 2

The proofs for the theorems in the main text are collected in Appendix B.1. Appendix B.2 contains some preliminary results in empirical process theory.

B.1 Proofs

Proof of Lemma 1. This lemma is almost the same as Lemma 11 in Fan and Guerre [2016]. The only difference is that in this paper we allow the kernel to diminish at the boundary of the support but the proof of Fan and Guerre [2016] nonetheless goes through. In fact, following their steps, we can show that the eigenvalues of $\Xi(x,h_1)$ and $\Omega(x,h_1)$ are larger than

$$\inf_{x\in B(0,1)} \min_{b^\top b=1} b^\top \left(\int \boldsymbol{r}(u)\boldsymbol{r}(u)^\top w(u)\mathbf{1}\{u \in B(x,\lambda_1)\}du \right) b,$$

which is strictly positive since $w > 0$ on $[-1,1]^d$. □

Proof of Theorem 1. By the standard change of variables and the law of iterated expectations, we can write

$$\begin{aligned}\Xi(x,h_1) &= \int \boldsymbol{r}(u)\boldsymbol{r}(u)^\top w(u) f_X(x+h_1u)du,\\ \boldsymbol{\upsilon}(y,x,h_1,h_2) &= \int \boldsymbol{r}(u)w(u)\tilde{F}(y|x+h_1u) f_X(x+h_1u)du,\end{aligned}$$

where $\tilde{F}(y|x) = \mathbb{E}[K((y-Y)/h_2) \mid X = x]$. Because f_X is continuously differentiable on $\mathscr{X}$, we have $\Xi(x,h_1) = f_X(x)\Omega(x,h_1) + o(1)$, uniformly over $x \in \mathscr{X}$. Applying change of variables and integration by parts to $\tilde{F}(y|x+h_1u)$, we have

$$\begin{aligned}\tilde{F}(y|x+h_1u) &= \int K((y-y')/h_2)f(y'|x+h_1u)dy' \\ &= \int K(v)f(y-h_2v|x+h_1u)h_2dv \\ &= \int k(v)F(y-h_2v|x+h_1u)dv.\end{aligned}$$

By Assumption **Y**, $F(y|x)$ restricted to $\mathbb{R} \times \mathscr{X}$ is twice uniformly continuously differentiable. Then for any $y \in \mathbb{R}$, the following expansion holds:

$$\begin{aligned}F(y-h_2v|x+h_1u) = F(y|x+h_1u) &- \frac{\partial}{\partial y}F(y|x+h_1u)h_2v + \frac{1}{2}\frac{\partial^2}{\partial y^2}F(y|x+h_1u)h_2^2v^2 \\ &+ \frac{1}{2}h_2^2\left(\frac{\partial^2}{\partial y^2}F(\tilde{y}|x+h_1u)v^2 - \frac{\partial^2}{\partial y^2}F(y|x+h_1u)v^2\right),\end{aligned}$$

where $\tilde{y}$ is between y and $y-h_2v$. Therefore,

$$\tilde{F}(y|x+h_1u) = F(y|x+h_1u) + \frac{h_2^2}{2}\frac{\partial^2}{\partial y^2}F(y|x+h_1u)\int v^2k(v)dv + o(h_2^2),$$

uniformly over $y \in \mathbb{R}$ and $x+h_1u \in \mathscr{X}$. The remainder term is uniformly $o(h_2^2)$ because $\frac{\partial^2}{\partial y^2}F$ is a continuous function on the compact set *supp*(Y,X). Next, by the smoothness of $F(y|x)$ with respect to x, we have

$$\begin{aligned}F(y|x+h_1u) &= F(y|x) + h_1u^\top\nabla_xF(y|x) + \frac{h_1^2}{2}u^\top[\nabla_x^\top\nabla_xF(y|\tilde{x})]u \\ &= \boldsymbol{r}(u)^\top H_1\boldsymbol{\beta}^*(y,x) + \frac{h_1^2}{2}u^\top[\nabla_x^\top\nabla_xF(y|x)]u \\ &+ \frac{h_1^2}{2}u^\top\left([\nabla_x^\top\nabla_xF(y|\tilde{x})]u - [\nabla_x^\top\nabla_xF(y|x)]u\right)u \\ &= \boldsymbol{r}(u)^\top H_1\boldsymbol{\beta}^*(y,x) + \frac{h_1^2}{2}u^\top[\nabla_x^\top\nabla_xF(y|x)]u + o(h_1^2), \qquad \text{(B.1)}\end{aligned}$$

uniformly over $y \in \mathbb{R}$ and $x, x+h_1u \in \mathscr{X}$. The remainder term is uniformly $o(h_1^2)$ because $\nabla_x^\top \nabla_x F$ is assumed to be uniformly continuous on $\mathbb{R} \times \mathscr{X}$. Similarly, we have

$$f_X(x+h_1u) = f_X(x) + o(1), \tag{B.2}$$

$$\frac{\partial^2}{\partial y^2} F(y|x+h_1u) = \frac{\partial^2}{\partial y^2} F(y|x) + o(1), \tag{B.3}$$

uniformly over $y \in \mathbb{R}$ and $x, x+h_1u \in \mathscr{X}$. Therefore,

$$\begin{aligned}
\boldsymbol{\upsilon}(y,x,h_1,h_2) = {} & \Xi(x,h_1)H_1\boldsymbol{\beta}^*(y,x) \\
& + \frac{h_1^2}{2} f_X(x) \int \boldsymbol{r}(u)w(u)u^\top [\nabla_x^\top \nabla_x F(y|x)]u\mathbf{1}\{x+h_1u \in \mathscr{X}\}du \\
& + \frac{h_2^2}{2}\frac{\partial^2}{\partial y^2} F(y|x) f_X(x) \int v^2 k(v)dv \int \boldsymbol{r}(u)w(u)\mathbf{1}\{x+h_1u \in \mathscr{X}\}du + o(h_1^2+h_2^2),
\end{aligned}$$

uniformly over $y \in \mathbb{R}$ and $x \in \mathscr{X}$. Therefore,

$$\begin{aligned}
& H_1(\bar{\boldsymbol{\beta}}(y,x,h_1,h_2) - \boldsymbol{\beta}^*(y,x)) \\
= & \frac{h_1^2}{2}\Omega(x,h_1)^{-1} \sum_{\ell,\ell'=1}^{d} \frac{\partial^2}{\partial x_\ell \partial x_{\ell'}} F(y|x) \int \boldsymbol{r}(u)u_\ell u_{\ell'} w(u)\mathbf{1}\{x+h_1u \in \mathscr{X}\}du \\
+ & \frac{h_2^2}{2}\Omega(x,h_1)^{-1} \frac{\partial^2}{\partial y^2} F(y|x) \int v^2 k(v)dv \int \boldsymbol{r}(u)w(u)\mathbf{1}\{x+h_1u \in \mathscr{X}\}du + o(h_1^2+h_2^2).
\end{aligned}$$

Then the first claim of the theorem follows.

When $x \in \mathring{\mathscr{X}}_{h_1}$, $x+h_1u \in \mathscr{X}$ for all $u \in [-1,1]^d$. In that case, $\Omega(x,h_1)$ becomes the identity matrix because w_ℓ is symmetric and has variance one. Then the second claim of the theorem follows.

□

Proof of Lemma 2. Let $M > 0$ be the uniform bound of $\mathscr{G}$. Notice that each $\mathscr{G}_n$ is a uniformly bounded (by M) Euclidean class with the same coefficients (A, v). Denote

$$\Delta_n^o = \sup_{f \in \mathscr{G}_n} \left| \sum_{i=1}^n Rad_i f(X_i) \right|,$$

where $Rad_i, 1 \leq i \leq n$, is a sequence of iid Rademacher variables. By Proposition 2.1 in Giné and Guillou [2001], we have for all $n \geq 1$,

$$\mathbb{E}\Delta_n^o \leq C \left(vM \log(AM/\sigma_n) + \sqrt{v}\sqrt{n\sigma_n^2 \log(AM/\sigma_n)} \right) = O\left(\sqrt{n\sigma_n^2 |\log \sigma_n|} + |\log \sigma_n| \right).$$

By the symmetrization result in, for example, Lemma 2.3.1 of van der Vaart and Wellner [1996b], we know that $\mathbb{E}\Delta_n \leq 2\mathbb{E}\Delta_n^o = O\left(\sqrt{n\sigma_n^2 |\log \sigma_n|} + |\log \sigma_n| \right)$. Then the claimed result follows from the Chebyshev inequality. □

Proof of Theorem 2. We proceed with two steps. Recall the expression of $H_1 \hat{\boldsymbol{\beta}}$ in Equation (1). In Step 1, we derive the uniform convergence rate of the numerator $\hat{\boldsymbol{\upsilon}}(y, x, h_1, h_2)$. In Step 2, we derive the uniform convergence rate of the denominator $\hat{\Xi}(x, h_1)$.

Step 1. To avoid repetition in the proof, we consider a generic element of the vector $\hat{\boldsymbol{\upsilon}}(y, x, h_1, h_2)$:

$$\begin{aligned} \hat{\boldsymbol{\upsilon}}_\pi(y, x, h_1, h_2) &= \frac{1}{nh_1^d} \sum_{i=1}^n ((X_i - x)/h_1)^\pi K\left(\frac{y - Y_i}{h_2}\right) w\left(\frac{X_i - x}{h_1}\right), \\ &= \frac{1}{nh_1^d} \sum_{i=1}^n K\left(\frac{y - Y_i}{h_2}\right) \prod_{\ell=1}^d ((X_{i\ell} - x_\ell)/h_1)^{\pi_\ell} w_\ell \left(\frac{X_{i\ell} - x_\ell}{h_1}\right), \end{aligned} \tag{B.4}$$

where $\pi = (\pi_1, \cdots, \pi_d), \pi_\ell \in \{0, 1\}, \sum \pi_\ell \leq 1$. We want to derive the following uniform convergence rate of $\hat{\boldsymbol{\upsilon}}_\ell(y, x, h_1, h_2)$:

$$\sup_{y \in \mathbb{R}, x \in \mathscr{X}} \left| \hat{\boldsymbol{\upsilon}}_\ell(y, x, h_1, h_2) - \mathbb{E}\hat{\boldsymbol{\upsilon}}_\ell(y, x, h_1, h_2) \right| = O_p\left(\sqrt{|\log h_1| / (nh_1^d)} \right). \tag{B.5}$$

By defining

$$\psi_n(Y,X;y,x) = K\left(\frac{y-Y_i}{h_2}\right)\prod_{\ell=1}^{d}((X_{i\ell}-x_\ell)/h_1)^{\pi_\ell} w_\ell\left(\frac{X_{i\ell}-x_\ell}{h_1}\right)$$

and $\Psi_n = \{\psi_n(\cdot,\cdot;y,x) : y \in \mathbb{R}, x \in \mathscr{X}\}$, we can write the LHS of (B.5) as

$$\sup_{\psi_n \in \Psi_n}\left|\frac{1}{nh_1^d}\sum_{i=1}^{n}(\psi_n(Y_i,X_i;y,t) - \mathbb{E}\psi_n(Y_i,X_i;y,t))\right|,$$

which can be studied with the empirical process theory introduced previously. Notice that ψ_n and Ψ_n depend on n through the bandwidth h_1 and h_2.

Consider a larger class Ψ that does not depend on n defined by the following product:

$$\Psi = \Psi_Y \Psi_{X_1} \Psi_{X_2} \cdots \Psi_{X_d},$$

where

$$\begin{aligned}
\Psi_Y &= \{(Y,X) \mapsto K\big((y-Y)/h\big) : y \in \mathbb{R}, h > 0\},\\
\Psi_{X_\ell} &= \{(Y,X) \mapsto \big((X_\ell - x_\ell)/h\big)^{\pi_\ell} w_\ell\big((X_\ell - x_\ell)/h\big) : x \in \mathscr{X}, h > 0\}, \ell = 1,\cdots,d.
\end{aligned}$$

For all $n \geq 1$, Ψ_n is a subset of the product class Ψ. Then we want to show that Ψ is uniformly bounded and Euclidean. If that is true, then we can appeal to Lemma 2.

In view of Lemma B.5, we only need to show that Ψ_Y and Ψ_{X_ℓ} are uniformly bounded and Euclidean. The class Ψ_Y is uniformly bounded by 1. The function K is of bounded variation on $\mathbb{R}$ since it is the integral of the integrable function k (Corollary 3.33 in Folland [1999]). Then by Lemma B.1, we know that Ψ_Y is Euclidean. The class Ψ_{X_ℓ} is uniformly bounded by $\|w_\ell\|_\infty$. This is because w_ℓ is support on $[-1,1]$ and hence the term in front of w_ℓ, $(X_\ell - x_\ell)/h$, cannot exceed one in magnitude. To show that Ψ_{X_ℓ} is Euclidean, notice that the function $u_\ell \mapsto u_\ell^{\pi_\ell} w_\ell(u_\ell)$ is of bounded variation. This is because on the support of w_ℓ, $[-1,1]$, both $u_\ell \mapsto u_\ell^{\pi_\ell}$ and w_ℓ are of bounded variation. Then their product is also of bounded variation (Theorem 6.9, Apostol [1974]). Then we know Ψ_{X_ℓ} is Euclidean by appealing to Lemma B.1.

Next, we want to derive a uniform variance bound for each Ψ_n. By the standard change of variables, we know that $\sup_{\psi_n \in \Psi_n} \mathbb{E}[\psi_n(Y,X;y,x)^2]$ is bounded by

$$\sup_{x\in\mathscr{X}} \mathbb{E}\left[w\left(\frac{X-x}{h_1}\right)^2\right] \leq \sup_{x\in\mathscr{X}} h_1^d \int w(u)^2 f_X(x+h_1u)du \leq h_1^d \|f_X\|_\infty \prod_{\ell=1}^{d} \|w_\ell\|_\infty,$$

where we have used the fact that $K \in [0,1]$, and w_ℓ is supported on $[-1,1]$ and integrates to 1. Therefore, we can define $\sigma_{\Psi_n}^2 = h_1^d \|f_X\|_\infty \prod_{\ell=1}^d \|w_\ell\|_\infty$ as a uniform variance bound for Ψ_n. Under the assumption that $nh_1^d/|\log h_1| \to \infty$, we can apply Lemma 2 to the sequence Ψ_n and obtain that

$$\begin{aligned}\sup_{\psi_n\in\Psi_n}\left|\frac{1}{nh_1^d}\sum_{i=1}^{n}(\psi_n(Y_i,X_i;y,x)-\mathbb{E}\psi_n(Y_i,X_i;y,x)\right| &= O_p\left(\frac{\sqrt{n\sigma_{\Psi_n}^2|\log\sigma_{\Psi_n}|}}{nh_1^d}\right)\\ &= O_p\left(\sqrt{\frac{|\log h_1|}{nh_1^d}}\right),\end{aligned}$$

which is the desired result specified in Equation (B.5).

Step 2. Following the same procedure as in Step 1, we can show that the uniform convergence rate for each element of the matrix $\hat{\Xi}(x,h_1)$ is also $\sqrt{|\log h_1|/(nh_1^d)}$. We omit the details for brevity. Then by Lemma 1, we know that with probability approaching one, the eigenvalues of $\hat{\Xi}(x,h_1)$ is in $[1/C,C]$. In particular, with probability approaching one, the inverse matrix $\hat{\Xi}(x,h_1)^{-1}$ is well-defined, and its induced 2-norm $\left\|\hat{\Xi}(x,h_1)^{-1}\right\|_2$ is bounded. Then applying

Lemma 1 once again, we have

$$\begin{aligned}\sup_{x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}-\Xi(x,h_1)^{-1}\right\|_2 &= \sup_{x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}(\Xi(x,h_1)-\hat{\Xi}(x,h_1))\Xi(x,h_1)^{-1}\right\|_2\\ &\le \sup_{x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}\right\|_2\left\|\Xi(x,h_1)-\hat{\Xi}(x,h_1)\right\|_2\left\|\Xi(x,h_1)^{-1}\right\|_2\\ &= O_p\left(\sqrt{|\log h_1|/(nh_1^d)}\right),\end{aligned}$$

where the second line follows from the submultiplicativity of the induced 2-norm. Combing the above result with Step 1, we obtain

$$\begin{aligned}&\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}\hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2)-\Xi(x,h_1)^{-1}\boldsymbol{\upsilon}(y,x,h_1,h_2)\right\|_2\\ \le&\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}-\Xi(x,h_1)^{-1}\right\|_2\left\|\hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2)\right\|_2\\ +&\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2)-\boldsymbol{\upsilon}(y,x,h_1,h_2)\right\|_2\left\|\Xi(x,h_1)^{-1}\right\|_2\\ =&O_p\left(\sqrt{|\log h_1|/(nh_1^d)}\right),\end{aligned}$$

where the last line uses the fact that $\hat{v}$ is uniformly bounded. This proves Equation (5).

□

Proof of Theorem 3. For the unsmoothed estimator, we can now define the pseudo-true value by replacing the term $K((y-Y_i)/h_2)$ with the term $\mathbf{1}\{Y_i \le y\}$ for the minimization problem defined in (3) in the main text. To derive the bias term, we can follow the proof of Theorem 1 and replace $\tilde{F}$ with F in the definition of $\boldsymbol{\upsilon}$. Therefore, the bias term in this case can be controlled by using (B.1) and (B.2) without (B.3), which means that we no longer require the differentiability of F with respect to y. The bias term associated with h_2 (as in Theorem 1) no longer exist. The remaining bias is $O(h_1^2)$. The stochastic term can be dealt with by using the proof of Theorem 2. We replace the class Ψ_Y by the class of indicator functions $\mathbf{1}\{Y_i \le y\}, y \in \mathbb{R}$, which is also uniformly bounded by 1 and is Euclidean. The other parts of the proof remain the same. □

Proof of Theorem 4. Notice that we can write $H_1\left(\hat{\boldsymbol{\beta}}(y,x,h_1,h_2)-\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)$ as

$$\begin{aligned}
&\hat{\Xi}(x,h_1)^{-1}\left(\hat{\boldsymbol{\upsilon}}(y,x,h_1,h_2)-\hat{\Xi}(x,h_1)H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)\\
=&\hat{\Xi}(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\\
+&\hat{\Xi}(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)\left(\tilde{F}(y|X_i)-\boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)^\top H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)w\left(\frac{X_i-x}{h_1}\right)\\
=&\Xi(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)+err_1(y,x)+err_2(y,x)
\end{aligned}$$

where

$$\begin{aligned}
err_1(y,x)&=\hat{\Xi}(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)\\
&\qquad\times\left(\tilde{F}(y|X_i)-\boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)^\top H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)w\left(\frac{X_i-x}{h_1}\right),\\
err_2(y,x)&=\left(\hat{\Xi}(x,h_1)^{-1}-\Xi(x,h_1)^{-1}\right)\frac{1}{nh_1^d}\sum_{i=1}^n \boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2).
\end{aligned}$$

We use the empirical process theory to derive the uniform convergence rates of err_1 and err_2 respectively in the following Step 1 and Step 2.

Step 1. Define a sequence of function classes $\Phi_n=\{\phi_n(\cdot,\cdot;y,x):y\in\mathbb{R},x\in\mathscr{X}\}$, where

$$\phi_n(Y,X;y,x)=\left(\tilde{F}(y|X)-\boldsymbol{r}\left(\frac{X-x}{h_1}\right)^\top H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)\prod_{\ell=1}^d\left(\frac{X_\ell-x_\ell}{h_1}\right)^{\pi_\ell}w_\ell\left(\frac{X_\ell-x_\ell}{h_1}\right)$$

with $\sum\pi_\ell\leq 1$ as before. We want to derive the convergence rate of

$$\sup_{\phi_n\in\Phi_n}\left|\frac{1}{nh_1^d}\sum_{i=1}^n\phi_n(Y_i,X_i;y,x)\right|.$$

Notice that ϕ_n is already centered, that is, $\mathbb{E}\phi_n(Y,X;y,x)=0$, by the first-order condition of (2). Define a larger product class Φ that does not vary with n by $\Phi=\Phi_Y\Psi_{X_1}\Psi_{X_2}\cdots\Psi_{X_d}$, where Ψ_{X_ℓ}

is defined in the proof of Theorem 2 and

$$\begin{aligned}\Phi_Y = \{(Y,X) \mapsto (\mathbb{E}[K((y-Y)/h)|X] - \boldsymbol{r}(X-x)^\top \boldsymbol{\beta} \\ \times \mathbf{1}\{|X_\ell - x_\ell| \leq 1, 1 \leq \ell \leq d\} : y \in \mathbb{R}, x \in \mathscr{X}, h > 0, \left\|\boldsymbol{\beta}\right\|_2 \leq C\}.\end{aligned}$$

To understand the expression of Φ_Y, recall that by definition $\tilde{F}(y|X) = \mathbb{E}[K((y-Y)/h_2) \mid X]$. The term $\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)$ is replaced by a general $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ with a bounded norm. This can be done as both the numerator and denominator of $\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)$ is bounded. The indicator term $\mathbf{1}\{|X_\ell - x_\ell| \leq 1, 1 \leq \ell \leq d\}$ comes from the support of w. This indicator term is needed for deriving the uniform boundedness.

For each n, we have $\Phi_n \in \Phi$. We want to show that Φ is a uniformly bounded Euclidean class. Since Ψ_{X_ℓ} is proven to be uniformly bounded and Euclidean in Theorem 2, we only need to focus on the class Ψ_Y. First notice that the class

$$\left\{(Y,x) \mapsto \boldsymbol{r}(X-x)^\top \boldsymbol{\beta} \mathbf{1}\{|X_\ell - x_\ell| \leq 1, 1 \leq \ell \leq d\} : y \in \mathbb{R}, x \in \mathscr{X}, h > 0, \left\|\boldsymbol{\beta}\right\|_2 \leq C\right\}$$

is uniformly bounded and Euclidean in view of Lemma B.2. By Lemma B.3, we know that the following class is uniformly bounded and Euclidean:

$$\left\{(Y,X) \mapsto \mathbb{E}[K((y-Y)/h) \mid X] \mathbf{1}\{|X_\ell - x_\ell| \leq 1, 1 \leq \ell \leq d\} : y \in \mathbb{R}, h > 0\right\}$$

Then by Lemma B.4, we know that Φ_Y is uniformly bounded and Euclidean. Hence, Φ is uniformly bounded and Euclidean.

Then we want to derive a variance bound for each Φ_n. By the usual change of variables, we have for any $y \in \mathbb{R}$ and $x \in \mathscr{X}$,

$$\mathbb{E}[\phi_n(Y,x;y,x)^2] \leq h_1^d \int \left(\tilde{F}_{Y|X}(y|x+h_1u) - \boldsymbol{r}(u) H_1 \bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right)^2 w(u)^2 f_X(x+h_1u) du.$$

From the uniform bias expansion results in Theorem 1, we have

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left|\tilde{F}_{Y|X}(y|x+h_1u)-\boldsymbol{r}(u)H_1\bar{\boldsymbol{\beta}}(y,x,h_1,h_2)\right|=O(h_1^2+h_2^2)=O(h_1^2).$$

Therefore, we can construct a uniform variance bound $\sigma^2_{\Phi_n}=O(h_1^{d+4})$ for the class Φ_n. Then by Lemma 2, we can show that

$$\sup_{\phi_n\in\Phi_n}\left|\frac{1}{nh_1^d}\sum_{i=1}^n\phi_n(Y_i,X_i;y,x)\right|=O_p\left(\left(\sqrt{nh_1^{d+4}|\log h_1|}+|\log h_1|\right)/(nh_1^d)\right)=O_p\left(\frac{|\log h_1|}{nh_1^d}\right),$$

where the second line follows from the assumption that $nh_1^{d+4}/|\log h_1|\le C$. Therefore,

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}}\|err_1(y,x)\|_2=\sup_{x\in\mathscr{X}}\left\|\hat{\Xi}(x,h_1)^{-1}\right\|_2O_p\left(|\log h_1|/(nh_1^d)\right)=O_p\left(|\log h_1|/(nh_1^d)\right).$$

Step 2. Similar as before, we can show that

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\frac{1}{nh_1^d}\sum_{i=1}^n\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\right\|_2=O_p\left(\sqrt{|\log h_1|/(nh_1^d)}\right).$$

It is straightforward to see that the summand is centered, and the relevant function classes are uniformly bounded and Euclidean. For the variance bound, we can simply bound the term $\left(K\left((y-Y_i)/h_2\right)-\tilde{F}(y|X_i)\right)^2$ by 1. We omit the details of the derivation. Then by the uniform convergence rate of $\hat{\Xi}(x,h_1)^{-1}$ derived in the proof of Theorem 2, we have

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}}|err_2(y,x,h_1,h_2)|=O_p\left(|\log h_1|/(nh_1^d)\right).$$

Therefore, we have shown that both the terms $err_1(y,x)$ and $err_2(y,x)$ are $O_p\left(|\log h_1|/(nh_1^d)\right)$ uniformly. Then the desired result follows.

□

Proof of Corollary 2. By the asymptotic linear representation in Theorem 4 and the mean value theorem, we have

$$
\begin{aligned}
&\sup_{|y_1-y_2|\le\delta_n, x\in\mathscr{X}} \left|\hat{F}(y_1|x)-F(y_1|x)-(\hat{F}(y_2|x)-F(y_2|x))\right| \\
=&\sup_{|y_1-y_2|\le\delta_n, x\in\mathscr{X}} \left|\Xi(x,h_1)^{-1}\frac{1}{nh_1^d}\sum_{i=1}^n(\boldsymbol{s}(Y_i,X_i;y_1,x,h_1,h_2)-\boldsymbol{s}(Y_i,X_i;y_2,x,h_1,h_2))\right| \\
&+O_p\left(\frac{|\log h_1|}{nh_1^d}\right) \\
\le& C\delta_n\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\frac{1}{nh_1^d}\sum_{i=1}^n\frac{\partial}{\partial y}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\right\|_2+O_p\left(\frac{|\log h_1|}{nh_1^d}\right),
\end{aligned}
$$

where we have used the fact that $\|\Xi(x,h_1)^{-1}\|_2$ is bounded (Lemma 1). The partial derivative $\frac{\partial}{\partial y}\boldsymbol{s}$ is equal to

$$
\frac{\partial}{\partial y}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)=\frac{1}{h_2}\boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)\left(k\left(\frac{y-Y_i}{h_2}\right)-\mathbb{E}\left[k\left(\frac{y-Y_i}{h_2}\right)\mid X_i\right]\right)w\left(\frac{X_i-x}{h_1}\right).
$$

Similar as before, we can use Lemma 2 to show that

$$
\begin{aligned}
&\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\frac{1}{nh_1^d}\sum_{i=1}^n\boldsymbol{r}\left(\frac{X_i-x}{h_1}\right)\left(k\left((y-Y_i)/h_2\right)-\mathbb{E}\left[k\left((y-Y_i)/h_2\right)\mid X_i\right]\right)w\left(\frac{X_i-x}{h_1}\right)\right\|_2 \\
=&O_p\left(\sqrt{\frac{|\log h_1|}{nh_1^d}}\right).
\end{aligned}
$$

We omit the details here. It then follows that

$$
\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\frac{1}{nh_1^d}\sum_{i=1}^n\frac{\partial}{\partial y}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\right\|_2=O_p\left(\sqrt{\frac{|\log h_1|}{nh_1^d}}\frac{1}{h_2}\right).
$$

This proves the corollary. □

Proof of Corollary 3. Consider the following bias-variance decomposition of $\hat{\theta}-\theta$:

$$\underbrace{\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\left(\bar{\boldsymbol{\beta}}_0(y,x,h_1,h_2)-\boldsymbol{\beta}_0^*(y,x,h_1,h_2)\right)dxdy}_{\text{bias term}}$$
$$+\underbrace{\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\left(\hat{\boldsymbol{\beta}}_0(y,x,h_1,h_2)-\bar{\boldsymbol{\beta}}_0(y,x,h_1,h_2)\right)dxdy}_{\text{stochastic term}}$$

By Theorem 1 and the assumption that $\sqrt{n}h_1^2=o(1)$, we know that the bias term is $o(n^{-1/2})$. For the stochastic term, we first want to take care of the matrix $\Xi(x,h_1)$. Recall that when $x\in\mathring{\mathscr{X}}_h=[\underline{x}+h,\bar{x}-h]$, $\Xi(x,h_1)$ is equal to the identity matrix $\boldsymbol{I}$. In the proof of Theorem 4, we have shown that

$$\sup_{y\in\mathbb{R},x\in\mathscr{X}}\left\|\frac{1}{nh_1}\sum_{i=1}^{n}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\right\|_2=O_p\left(\sqrt{|\log h_1|/(nh_1)}\right).$$

Therefore, we have

$$\left\|\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\left(\Xi(x,h_1)-\boldsymbol{I}\right)\frac{1}{nh_1}\sum_{i=1}^{n}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)\right\|_2=o_p(1/\sqrt{n}).$$

By Theorem 4 and the assumption that $\sqrt{n}h_1/|\log h_1|\to\infty$, we can write the stochastic term as

$$\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\left(\hat{\boldsymbol{\beta}}_0(y,x,h_1,h_2)-\bar{\boldsymbol{\beta}}_0(y,x,h_1,h_2)\right)dxdy=\frac{1}{n}\sum_{i=1}^{n}Z_i+o_p(1/\sqrt{n})$$

where

$$\begin{aligned}Z_i&=\frac{1}{h_1}\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\boldsymbol{e}_0^\top\boldsymbol{I}\boldsymbol{s}(Y_i,X_i;y,x,h_1,h_2)dxdy\\&=\frac{1}{h_1}\int_{\underline{y}}^{\bar{y}}\int_{\underline{x}}^{\bar{x}}\left(K\left(\frac{y-Y_i}{h_2}\right)-\tilde{F}(y|X_i)\right)w\left(\frac{X_i-x}{h_1}\right)dxdy.\end{aligned}$$

By the standard change of variables, we can write Z_i as

$$Z_i = \int_{\underline{y}}^{\bar{y}} \int_{(X_i-\underline{x})/h_1}^{(X_i-\bar{x})/h_1} \left(K\left(\frac{y-Y_i}{h_2}\right) - \tilde{F}(y|X_i) \right) w(u)\, dudy$$

The random variables $\{Z_i : 1 \leq i \leq n\}$ forms an iid triangular array. Each Z_i is centered, that is, $\mathbb{E}[Z_i] = 0$. Denote the variance of Z_i as V_n, which can be calculated based on change of variables:

$$\begin{aligned} V_n &= \mathbb{E}[Z_i^2] \\ &= \int \left(\int_{\underline{y}}^{\bar{y}} \int_{(t-\underline{x})/h_1}^{(t-\bar{x})/h_1} \left(K\left(\frac{y-s}{h_2}\right) - \tilde{F}(y|t) \right) w(u)\, dudy \right)^2 f(s,t) dtds \\ &= \int \left(\int_{\underline{y}}^{\bar{y}} \int_{-1}^{1} \left(K\left(\frac{y-s}{h_2}\right) - \tilde{F}(y|t) \right) \mathbf{1}_{[(t-\underline{x})/h_1,(t-\bar{x})/h_1]}(u) w(u)\, dudy \right)^2 f(s,t) dtds \end{aligned}$$

As $n \to \infty$, we have the pointwise convergence results $K(\frac{y-s}{h_2}) \to \mathbf{1}\{s \leq y\}, \tilde{F}(y|t) \to F(y|t)$, and $\mathbf{1}_{[(t-\underline{x})/h_1,(t-\bar{x})/h_1]}(u) \to 1, u \in [-1,1]$. We know that these functions are bounded and the support of (Y,X) is compact. Then by the dominated convergence theorem, we have

$$\begin{aligned} V_n &\sim \int \left(\int_{\underline{y}}^{\bar{y}} \int_{-1}^{1} \left(\mathbf{1}\{s \leq y\} - F(y|t) \right) w(u)\, dudy \right)^2 f(s,t) dtds \\ &= \int \left(\int \left(\mathbf{1}\{s \leq y\} - F(y|t) \right) dy \right)^2 f(s,t) dtds = V, \end{aligned}$$

where the second line follows from the fact that w integrates to 1. It is straightforward to see that Z_i is bounded, and hence any moment of $|Z_i|$ is finite. Then we can apply the Lyapnov central limit theorem (for example, Theorem 5.11 in White [2001]) to obtain that $\sum Z_i/\sqrt{n}$ converges in distribution to $N(0,V)$. The desired result is thus proved.

□

B.2 Preliminary Results in Empirical Process Theory

Lemma B.1. *Let $K : \mathbb{R} \to \mathbb{R}$ be a function of bounded variation. Then the following class is Euclidean:*

$$\left\{K\left((\cdot - x)/h\right) : x \in \mathbb{R}, h > 0\right\}.$$

Proof of Lemma B.1. This is a direct application of Lemma 22(i) in Nolan and Pollard [1987]. □

Lemma B.2. *Any uniformly bounded and finite-dimensional vector space of functions is Euclidean.*

Proof of Lemma B.2. This follows from Lemma 2.6.15 and Theorem 2.6.7 in van der Vaart and Wellner [1996b]. □

Lemma B.3. *Let $\mathscr{G}$ be a uniformly bounded Euclidean class with coefficients (A, v). Then the class $\{\mathbb{E}[g(\cdot) \mid X] : g \in \mathscr{G}\}$ is also uniformly bounded and Euclidean with coefficients (A, v).*

Proof of Lemma B.3. This follows from the fact that the conditional expectation is a projection in the Hilbert space $L_2(P)$ and hence reduces the norm. □

Lemma B.4. *Let $\mathscr{G}_1$ and $\mathscr{G}_2$ be two classes of functions that are uniformly bounded and Euclidean with coefficients (A_1, v_1) and (A_2, v_2) respectively. Then the class $\mathscr{G}_1 \oplus \mathscr{G}_2 = \{g_1 + g_2 : g_1 \in \mathscr{G}_1, g_2 \in \mathscr{G}_2\}$ is also uniformly bounded and Euclidean with coefficients $(A_1A_2A_1A_22^{v_1+v_2}, v_1 + v_2)$.*

Proof. By Inequalities (A.4) in Andrews [1994], we have

$$\begin{aligned} N(\mathscr{G}_1 \oplus \mathscr{G}_2, L_2(P), \varepsilon) &\leq N(\mathscr{G}_1, L_2(P), \varepsilon/2) N(\mathscr{G}_2, L_2(P), \varepsilon/2) \\ &\leq A_1(2/\varepsilon)^{v_1} A_2(2/\varepsilon)^{v_2} = A_1A_22^{v_1+v_2}/\varepsilon^{v_1+v_2}. \end{aligned}$$

□

Lemma B.5. *Let $\mathscr{G}_1$ be a class of functions that is uniformly bounded by M_1 and Euclidean with coefficients (A_1, v_1) and $\mathscr{G}_2$ a class of functions that is uniformly bounded by M_2 and Euclidean with coefficients (A_2, v_2). Then the class $\mathscr{G}_1\mathscr{G}_2 = \{g_1 \cdot g_2 : g_1 \in \mathscr{G}_1, g_2 \in \mathscr{G}_2\}$ is uniformly bounded by M_1M_2 and Euclidean with coefficients $(A_1A_2(M_1+M_2)^{v_1+v_2}, v_1+v_2)$.*

Proof of Lemma B.5. The proof is similar to that of Theorem 3 in Andrews [1994]. By definition, for every measure P and every $\varepsilon \in (0,1]$, $N(\mathscr{G}_1, P, \varepsilon) \leq A_1/\varepsilon^{v_1}$ and $N(\mathscr{G}_2, P, \varepsilon) \leq A_2/\varepsilon^{v_2}$. We can construct $\{\tilde{g}_{1,j_1} : 1 \leq j_1 \leq J_1\}$ and $\{\tilde{g}_{2,j_2} : 1 \leq j_2 \leq J_2\}$ to be the ε-covering of $\mathscr{G}_1$ and $\mathscr{G}_2$, respectively, where $J_1 = N(\mathscr{G}_1, P, \varepsilon)$ and $N(\mathscr{G}_1, P, \varepsilon)$. For any $g_1 \in \mathscr{G}$ and $g_2 \in \mathscr{G}_2$, suppose g_1 is in the ε-neighborhood of $\tilde{g}_{1,j_{1,*}}$ and g_2 is in the ε-neighborhood of $\tilde{g}_{2,j_{2,*}}$. Then the $L_2(P)$ distance between g_1g_2 and $\tilde{g}_{1,j_{1,*}}\tilde{g}_{2,j_{2,*}}$ is

$$\begin{aligned}\left\|g_1g_2 - \tilde{g}_{1,j_{1,*}}\tilde{g}_{2,j_{2,*}}\right\|_{L_2(P)} &\leq \left\|g_1g_2 - g_1\tilde{g}_{2,j_{2,*}}\right\|_{L_2(P)} + \left\|g_1\tilde{g}_{2,j_{2,*}} - \tilde{g}_{1,j_{1,*}}\tilde{g}_{2,j_{2,*}}\right\|_{L_2(P)} \\ &\leq M_1 \left\|g_2 - \tilde{g}_{2,j_{2,*}}\right\|_{L_2(P)} + M_2 \left\|g_1 - \tilde{g}_{1,j_{1,*}}\right\|_{L_2(P)} \leq (M_1+M_2)\varepsilon.\end{aligned}$$

This means that $\{\tilde{g}_{1,j_1}\tilde{g}_{2,j_2} : 1 \leq j_1 \leq J_1, 1 \leq j_2 \leq J_2\}$ forms a $(M_1+M_2)\varepsilon$-cover of $\mathscr{G}_1\mathscr{G}_2$. Therefore,

$$\begin{aligned}N(\mathscr{G}_1\mathscr{G}_2, L_2(P), \varepsilon) &\leq N(\mathscr{G}_1, L_2(P), \varepsilon/(M_1+M_2))N(\mathscr{G}_2, L_2(P), \varepsilon/(M_1+M_2)) \\ &\leq A_1A_2(M_1+M_2)^{v_1+v_2}/\varepsilon^{v_1+v_2}.\end{aligned}$$

This proves the result. □

Appendix C

Appendix for Chapter 3

Appendix C.1 contains the proofs for theorems and propositions stated in the main text. Appendix C.2 studies the efficient estimation of parameters implicitly defined by possibly non-smooth and overidentifying moment restrictions.

C.1 Technical Proofs

We assume that Assumptions 3.1 and 3.2 hold throughout this section. The following two lemmas are helpful for proving the identification results.

C.1.1 Proof of the Identification Results

Lemma C.1. *$S \perp Z \mid X$ and $t \in \mathscr{T}$, $Y_t \perp T \mid S, X$.*

Proof of Lemma C.1. The first statement follows from the definition of S and the fact that Z is independent of the vector $(T_{z_1}, \cdots, T_{z_{N_Z}})$ conditioning on X. For the second statement, T is entirely determined by (S, Z, X). Hence, given S and X, T is independent of Y_t since Z is independent of $(Y_{t_1}, \cdots, Y_{t_{N_T}})$ conditional on X. □

Lemma C.2. *For each $t \in \mathscr{T}$ and $k = 1, \cdots, N_Z$, the following identification results hold.*

(i) $\mathbb{P}(S \in \Sigma_{t,k} \mid X) = b_{t,k} P_t(X)$ *a.s.*

(ii) $\mathbb{E}\left[Y_t \mid S \in \Sigma_{t,k}, X\right] = (b_{t,k} Q_t(X))/(b_{t,k} P_t(X))$ *a.s.*

Proof of Lemma C.2. This is Theorem T-6 in Heckman and Pinto [2018a]. The conditioning is explicitly presented. □

Proof of Theorem 3.1. The first statement follows from applying the law of iterated expectation to Lemma C.2(i). For the second statement, we can apply Bayes rule to Lemma C.2 and obtain that

$$
\begin{aligned}
\mathbb{E}\left[Y_t \mid S \in \Sigma_{t,k}\right] &= \int \mathbb{E}\left[Y_t \mid S \in \Sigma_{t,k}, X = x\right] f_{X \mid S \in \Sigma_{t,k}}(x) dx \\
&= \int \mathbb{E}\left[Y_t \mid S \in \Sigma_{t,k}, X = x\right] \frac{\mathbb{P}(S \in \Sigma_{t,k} \mid X = x)}{\mathbb{P}(S \in \Sigma_{t,k})} f_X(x) dx \\
&= \mathbb{E}\left[b_{t,k} Q_t(X)\right] / p_{t,k},
\end{aligned}
$$

where $f_{X \mid S \in \Sigma_{t,k}}$ denotes the conditional density function of X given type $S \in \Sigma_{t,k}$. □

Proof of Theorem 3.2. By Lemma L-16 of Heckman and Pinto [2018b], we know that under the unordered monotonicity assumption, $B_t[\cdot, i] = B_t[\cdot, i']$ for all $s_i, s_{i'} \in \Sigma_{t,k}$. Thus, the set $\mathscr{Z}_{t,k}$ always exists. For the first statement, we have

$$
\begin{aligned}
\mathbb{P}\left(T = t, S \in \Sigma_{t,k}\right) &= \mathbb{P}\left(Z \in \mathscr{Z}_{t,k}, S \in \Sigma_{t,k}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(Z \in \mathscr{Z}_{t,k}, S \in \Sigma_{t,k} \mid X\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(Z \in \mathscr{Z}_{t,k} \mid X\right) \mathbb{P}\left(S \in \Sigma_{t,k} \mid X\right)\right] \\
&= \mathbb{E}\left[b_{t,k} P_t(X) \pi_{t,k}(X)\right],
\end{aligned}
$$

where the second equality follows from the law of iterated expectations and the third equality follows from the fact that $Z \perp S \mid X$ (Lemma C.1). For the second statement, notice that

$$\begin{aligned}\mathbb{P}(T=t, S\in\Sigma_{t,k} \mid X=x) &= \mathbb{P}(T=t \mid S\in\Sigma_{t,k}, X=x)\mathbb{P}(S\in\Sigma_{t,k} \mid X=x)\\ &= \mathbb{P}(Z\in\mathscr{Z}_{t,k} \mid X)\mathbb{P}(S\in\Sigma_{t,k} \mid X=x)\\ &= \pi_{t,k}(X) b_{t,k} P_t(X).\end{aligned}$$

By Lemma C.1, we know that

$$\mathbb{E}\left[Y_t \mid T=t, S\in\Sigma_{t,k}, X=x\right] = \mathbb{E}\left[Y_t \mid S\in\Sigma_{t,k}, X=x\right].$$

Therefore, we can apply Bayes rule and obtain that

$$\begin{aligned}&\mathbb{E}\left[Y_t \mid T=t, S\in\Sigma_{t,k}\right]\\ =&\int \mathbb{E}\left[Y_t \mid T=t, S\in\Sigma_{t,k}, X=x\right] f_{X\mid T=t, S\in\Sigma_{t,k}}(x)dx\\ =&\int \mathbb{E}\left[Y_t \mid S\in\Sigma_{t,k}, X=x\right] \frac{P(T=t, S\in\Sigma_{t,k} \mid X=x)}{P(T=t, S\in\Sigma_{t,k})} f_X(x)dx\\ =&\int \frac{b_{t,k}Q_t(X)}{b_{t,k}P_t(X)} \times \frac{\pi_{t,k}(X)b_{t,k}P_t(X)}{q_{t,k}} f_X(x)dx\\ =&\mathbb{E}\left[b_{t,k}Q_t(X)\pi_{t,k}(X)\right]/q_{t,k}.\end{aligned}$$

□

C.1.2 Semiparametric Efficiency Calculations

We follow the method developed by Newey [1990]. The likelihood of the GLATE model can be specified as

$$\mathscr{L}(Y,T,Z,X) = f_X(X)\prod_{z\in\mathscr{Z}}\left(f_z(Y,T \mid X)\pi_z(X)\right)^{\mathbf{1}\{Z=z\}},$$

where $f_z(\cdot,\cdot \mid X)$ denotes the conditional density of Y,T given $Z=z$ and X. In a regular parametric submodel, where the true underlying probability measure P is indexed by θ^o, we use the following notations to represent the score functions:

$$\begin{aligned}
&s_z(Y,Z \mid X;\theta) = \frac{\partial}{\partial\theta}\log\left(f_z(Y,T \mid X;\theta)\right),\\
&s_\pi(Z \mid X;\theta) = \sum_{z\in\mathscr{Z}} \mathbf{1}\{Z=z\}\frac{\partial}{\partial\theta}\log\left(\pi_z(X;\theta)\right),\\
&s_X(X;\theta) = \frac{\partial}{\partial\theta}\log\left(f_X(X;\theta)\right).
\end{aligned}$$

The score in a regular parametric submodel is

$$s_{\theta^o}(Y,T,Z,X) = \sum_{z\in\mathscr{Z}} \mathbf{1}\{Z=z\}s_z\left(Y,T \mid X;\theta^o\right) + s_\pi(Z \mid X;\theta^o) + s_X(X;\theta^o).$$

Hence, the tangent space of the model is

$$\begin{aligned}
\mathscr{S} = \Big\{s \in L_0^2 : s(Y,T,Z,X) = \sum_{z\in\mathscr{Z}} \mathbf{1}\{Z=z\}s_z\left(Y,T \mid X\right) + s_\pi(Z \mid X) + s_X(X)\\
\text{for some } s_z, s_\pi, s_X \text{ such that } \int s_z(y,t \mid X) f_z(y,t \mid X)dydt \equiv 0, \forall z;\\
\sum_{z\in\mathscr{Z}} s_\pi(z \mid X)\pi_z(X) \equiv 0, \text{ and } \int s_X(x) f_X(x)dx = 0\Big\},
\end{aligned}$$

where L_0^2 is a subspace of L^2 that contains the mean zero functions.

Proof of Theorem 3.3. We only prove statements (i) and (ii) since (iii) and (iv) are easier cases that can be proved along the way. We start with the first statement. The path-wise differentiability of the parameter $\beta_{t,k}$ can be verified in the following way: in any parametric submodel, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\beta_{t,k}(\theta)\Big|_{\theta=\theta^o}\\
=&\frac{\partial}{\partial\theta}(b_{t,k}\mathbb{E}_\theta\left[Q_t(X)\right]/p_{t,k})|_{\theta=\theta^o}\\
=&\frac{1}{p_{t,k}}\left((\partial b_{t,k}\mathbb{E}_\theta\left[Q_t(X)\right]/\partial\theta)|_{\theta=\theta^o}-(b_{t,k}\mathbb{E}_\theta\left[Q_t(X)\right]/p_{t,k})(\partial p_{t,k}/\partial\theta)|_{\theta=\theta^o}\right)\\
=&\frac{1}{p_{t,k}}b_{t,k}\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[Q_t(X)\right]|_{\theta=\theta^o}-\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[P_t(X)\right]|_{\theta=\theta^o}\beta_{t,k}\right),
\end{aligned}$$

where $\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[Q_t(X)\right]|_{\theta=\theta^o}$ and $\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[P_t(X)\right]|_{\theta=\theta^o}$ are $N_Z\times 1$ random vectors whose typical element can be represented respectively by

$$\begin{aligned}
&\int y\mathbf{1}\{\tau=t\}s_z(y,\tau\mid x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\
+&\int y\mathbf{1}\{\tau=t\}s_X(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx
\end{aligned}$$

and

$$\begin{aligned}
&\int \mathbf{1}\{\tau=t\}s_z(y,\tau\mid x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\
+&\int \mathbf{1}\{\tau=t\}s_X(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx,
\end{aligned}$$

respectively, for $z\in\mathscr{Z}$. The EIF is characterized by the condition that

$$\frac{\partial}{\partial\theta}\beta_{t,k}(\theta)\Big|_{\theta=\theta^o}=\mathbb{E}\left[\psi_{\beta_{t,k}}s_{\theta^o}\right],\text{ and }\psi_{\beta_{t,k}}\in\mathscr{S}.$$

The expression of $\psi_{\beta_{t,k}}$ given in Equation (3.2) meets the above requirements. In particular, the correspondence between terms in the EIF and path-wise derivative appears exactly as in Lemma 1 of Hong and Nekipelov [2010b].

For the second statement, the path-wise derivative of $\gamma_{t,k}$ can be computed similarly.

$$\begin{aligned}\frac{\partial}{\partial\theta}\gamma_{t,k}(\theta)\Big|_{\theta=\theta^o} = &\frac{1}{q_{t,k}}b_{t,k}\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[Q_t(X)\pi_{t,k}(X)\right]\Big|_{\theta=\theta^o}\\ &-\frac{\gamma_{t,k}}{q_{t,k}}b_{t,k}\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[P_t(X)\pi_{t,k}(X)\right]\Big|_{\theta=\theta^o},\end{aligned}$$

where $\frac{\partial}{\partial\theta}\mathbb{E}_\theta[Q_t(X)\pi_{W_{t,k}}(X)]|_{\theta=\theta^o}$ and $\frac{\partial}{\partial\theta}\mathbb{E}_\theta[P_t(X)\pi_{W_{t,k}}(X)]|_{\theta=\theta^o}$ are $N_Z\times 1$ random vectors whose typical element can be represented by

$$\begin{aligned}&\int y\mathbf{1}\{\tau=t\}s_z(y,\tau\mid x;\theta^o)\pi_{W_{t,k}}(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\ +&\int y\mathbf{1}\{\tau=t\}s_X(x;\theta^o)\pi_{W_{t,k}}(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\ +&\int y\mathbf{1}\{\tau=t\}\left(\frac{\partial}{\partial\theta}\pi_{t,k}(X;\theta)\Big|_{\theta=\theta^o}\right)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx,\end{aligned}$$

and

$$\begin{aligned}&\int \mathbf{1}\{\tau=t\}s_z(y,\tau\mid x;\theta^o)\pi_{W_{t,k}}(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\ +&\int \mathbf{1}\{\tau=t\}s_X(x;\theta^o)\pi_{W_{t,k}}(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\ +&\int \mathbf{1}\{\tau=t\}\left(\frac{\partial}{\partial\theta}\pi_{t,k}(X;\theta)\Big|_{\theta=\theta^o}\right)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx,\end{aligned}$$

respectively, for $z\in\mathscr{Z}$. The main difference appears when dealing with the last terms in the above two expressions, which can be matched with terms in the efficient influence function of the following two forms

$$\begin{aligned}&\mathbb{E}\left[Y\mathbf{1}\{T=t\}\mid Z=z,X\right]\left(\mathbf{1}\{Z\in\mathscr{Z}_{t,k}\}-\pi_{t,k}(X)\right),\text{ and}\\ &\mathbb{E}\left[\mathbf{1}\{T=t\}\mid Z=z,X\right]\left(\mathbf{1}\{Z\in\mathscr{Z}_{t,k}\}-\pi_{t,k}(X)\right).\end{aligned}$$

Take the latter one as an example. Notice that

$$\mathbf{1}\{Z \in \mathscr{Z}_{t,k}\} - \pi_{t,k}(X) = \sum_{z \in \mathscr{Z}_{t,k}} \left(\mathbf{1}\{Z = z\} - \pi_z(X)\right),$$

and

$$\left(\mathbf{1}\{Z = z\} - \pi_z(X)\right) s_\pi(Z \mid X; \theta^o) = \frac{\mathbf{1}\{Z = z\}}{\pi_z(X)} \frac{\partial}{\partial \theta} \pi_z(X; \theta)\big|_{\theta=\theta^o} - \pi_z(X) s_\pi(Z \mid X; \theta^o).$$

By the law of iterated expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{T = t\} \mid Z = z, X\right] \left(\mathbf{1}\{Z = z\} - \pi_z(X)\right) s_\pi(Z \mid X; \theta^o)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{T = t\} \mid Z = z, X\right] \mathbb{E}\left[\mathbf{1}\{Z = z\}/\pi_z(X) \mid X\right] \frac{\partial}{\partial \theta} \pi_z(X; \theta)\big|_{\theta=\theta^o}\right] \\
-&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{T = t\} \mid Z = z, X\right] \pi_z(X) \mathbb{E}\left[s_\pi(Z \mid X; \theta^o) \mid X\right]\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{T = t\} \mid Z = z, X\right] \frac{\partial}{\partial \theta} \pi_z(X; \theta)\big|_{\theta=\theta^o}\right] \\
=&\int \mathbf{1}\{\tau = t\} \left(\frac{\partial}{\partial \theta} \pi_z(X; \theta)\big|_{\theta=\theta^o}\right) f_z(y, \tau \mid x; \theta^o) f_X(x; \theta^o) dy d\tau dx.
\end{aligned}$$

□

Proof of Proposition 3.1. This proof is based on Section 4 in Newey [1994]. We focus on the case of $\beta_{t,k}$. The other cases are similar. To ease notation, let $h_t = \left(h_{Y,t,Z}, h_{t,Z}, \pi\right)'$. The estimator $\hat{\beta}_{t,k}$ is defined by the moment condition

$$\mathbb{E}[M\left(X, \beta_{t,k}, h_t\right)] = 0,$$

where

$$M\left(X, \beta_{t,k}, h_t\right) \equiv b_{t,k} \left(\frac{h_{Y,t,z_1}(X)}{\pi_{z_1}(X)}, \cdots, \frac{h_{Y,t,z_{N_Z}}(X)}{\pi_{z_{N_Z}}(X)}\right)' - \beta_{t,k} b_{t,k} \left(\frac{h_{t,z_1}(X)}{\pi_{z_1}(X)}, \cdots, \frac{h_{t,z_{N_Z}}(X)}{\pi_{z_{N_Z}}(X)}\right)'.$$

We then compute the derivatives of M with respect to the parameters:

$$\begin{aligned}
\mathbb{E}\left[\partial M/\partial\beta_{t,k}\right] &= -b_{t,k}\mathbb{E}\left[P_t(X)\right] = -p^o_{t,k} \\
\partial M/\partial h_{Y,t,z_i}|_{h_t=h_t^o} &= b_{t,k}[i]/\pi^o_{z_i}(X) \equiv \delta_{Y,t,z_i}(X) \\
\partial/\partial h_{t,z_i}M|_{h_t=h_t^o} &= -(\beta_{t,k}b_{t,k}[i])/\pi^o_{z_i}(X) \equiv \delta_{t,z_i}(X) \\
\partial M/\partial \pi_{z_i}|_{h_t=h_t^o} &= -(b_{t,k}[i]Q^o_{t,z_i}(X))/\pi^o_{z_i}(X) + (\beta_{t,k}b_{t,k}[i]P^o_{t,z_i}(X))/\pi^o_{z_i}(X) \equiv \delta_{\pi,z_i}(X),
\end{aligned}$$

where $b_{t,k}[i]$ denotes the ith element of the vector $b_{t,k}$. Define

$$\begin{aligned}
\alpha\left(Y,T,Z,X\right) \equiv \sum_{z\in\mathscr{Z}} &\delta_{Y,t,z}(X)\left(\mathbf{1}\{Z=z\}Y\mathbf{1}\{T=t\} - h^o_{Y,t,z}(X)\right) \\
&+ \sum_{z\in\mathscr{Z}} \delta_{t,z}(X)\left(\mathbf{1}\{Z=z\}\mathbf{1}\{T=t\} - h^o_{t,z}(X)\right) \\
&+ \sum_{z\in\mathscr{Z}} \delta_{\pi,z}(X)\left(\mathbf{1}\{Z=z\} - \pi^o_z(X)\right).
\end{aligned}$$

We have

$$\begin{aligned}
\alpha\left(Y,T,Z,X\right) = b_{t,k}&\zeta(Z,X,\pi^o)\left(\imath(Y\mathbf{1}\{T=t\}) - Q^o_t(X)\right) \\
&- \beta^o_{t,k}b_{t,k}\zeta(Z,X,\pi^o)\left(\imath\mathbf{1}\{T=t\} - P^o_t(X)\right).
\end{aligned}$$

Then Newey's (1994) Proposition 4 suggests that the influence function of the estimator $\hat{\beta}_{t,k}$ is $(M+\alpha)/p_{t,k}$ which is equal to the EIF $\psi^{\beta_{t,k}}$.

□

C.1.3 Proof of Robustness Results

Proof of Proposition 3.2. We prove the case for $\psi^{p_{t,k}}$, the other cases can be dealt with analogously. First assume $\pi = \pi^o$, then

$$\mathbb{E}\left[\mathbf{1}\{Z=z\}/\pi_z^o(X) \mid X\right] = 1,$$

which implies that $\mathbb{E}\left[\zeta(Z,X,\pi^o) \mid X\right]$ is almost surly equal to the identity matrix $\mathbf{I}$. By the law of total expectations, we have

$$\mathbb{E}\left[\mathbf{1}\{T=t\}\mathbf{1}\{Z=z\}/\pi_z^o(X) \mid X\right] = \mathbb{E}\left[\mathbf{1}\{T=t\} \mid Z=z,X\right] = P_{t,z}^o(X),$$

which implies that $\mathbb{E}\left[\zeta(Z,X,\pi^o)\iota\mathbf{1}\{T=t\}\right] = \mathbb{E}\left[P_t^o(X)\right]$. Therefore,

$$\begin{aligned}&b_{t,k}\mathbb{E}[\zeta(Z,X,\pi^o)\left(\iota(\mathbf{1}\{T=t\}) - P_t(X)\right) + P_t(X)]\\ =&b_{t,k}\mathbb{E}\left[\zeta(Z,X,\pi^o)\iota\mathbf{1}\{T=t\}\right] + b_{t,k}\mathbb{E}\left[(\mathbf{I}-\zeta(Z,X,\pi^o))P_t(X)\right] = b_{t,k}\mathbb{E}\left[P_t^o(X)\right] = p_{t,k}^o.\end{aligned}$$

Now suppose that $P_t = P_t^o$. Then by the law of total expectation, we have

$$\begin{aligned}&\mathbb{E}[\mathbf{1}\{Z=z\}(\mathbf{1}\{T=t\} - P_{t,z}^o(X)) \mid X]\\ =&\pi_z(X)\mathbb{E}[\mathbb{E}[\mathbf{1}\{T=t\} \mid Z=z,X] - P_{t,z}^o(X) \mid X] = 0.\end{aligned}$$

This implies that $\mathbb{E}[\zeta(Z,X,\pi)(\iota(\mathbf{1}\{T=t\}) - P_t^o(X))] = 0$. Hence,

$$b_{t,k}\mathbb{E}\left[\zeta(Z,X,\pi)\left(\iota(\mathbf{1}\{T=t\}) - P_t^o(X)\right) + P_t^o(X)\right] = b_{t,k}\mathbb{E}\left[P_t^o(X)\right] = p_{t,k}^o.$$

This proves the proposition. □

Proof of Proposition 3.3. Since $b_{t,k}$ is a finite vector, it suffices to verify the Neyman orthogonality condition for ψ_z, which is defined by

$$
\begin{aligned}
&\psi_z(Y,T,Z,X,\beta_{t,k},Q_t,P_t,\pi_z)\\
\equiv&\Big((\mathbf{1}\{Z=z\}/\pi_z(X))\left(\mathbf{1}\{T=t\}-P_{t,z}(X)\right)+P_{t,z}(X)\Big)\beta_{t,k}\\
&-(\mathbf{1}\{Z=z\}/\pi_z(X))\left(Y\mathbf{1}\{T=t\}-Q_{t,z}(X)\right)-Q_{t,z}(X).
\end{aligned}
$$

We want to show that

$$
\frac{d}{dr}\mathbb{E}\left[\psi_z(Y,T,Z,X,\beta_{t,k},Q_t^r,P_t^r,\pi_z^r)\right]\Big|_{r=0}=0,
$$

where $Q_t^r=Q_t^o+r(Q_t-Q_t^o)$, $P_t^r=P_t^o+r(P_t-P_t^o)$, and $\pi_z^r=\pi_z^o+r(\pi_z-\pi_z^o)$. In fact,

$$
\begin{aligned}
&\frac{d}{dr}\mathbb{E}\left[\psi_z(Y,T,Z,X,\beta_{t,k},Q_t^r,P_t^r,\pi_z^r)\right]\big|_{r=0}\\
=&\mathbb{E}\Bigg[\frac{-\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}\Big(\mathbf{1}\{T=t\}-P_{t,z}^r(X)\Big)\left(\pi_z(X)-\pi_z^o(X)\right)\beta_{t,k}\\
&+\Bigg(P_{t,z}(X)-P_{t,z}^o(X)-\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}\Big(P_{t,z}(X)-P_{t,z}^o(X)\Big)\Bigg)\beta_{t,k}\\
&+\frac{\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}\Big(Y\mathbf{1}\{T=t\}-Q_{t,z}^r(X)\Big)\left(\pi_z(X)-\pi_z^o(X)\right)\\
&-(Q_{t,z}(X)-Q_t^o(X))+\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}\Big(Q_{t,z}(X)-Q_{t,z}^o(X)\Big)\Bigg]\Bigg|_{r=0}\\
=&\mathbb{E}\Bigg[\frac{-\mathbf{1}\{Z=z\}}{(\pi_z^o(X))^2}\Big(\mathbf{1}\{T=t\}-P_{t,z}^o(X)\Big)\left(\pi_z(X)-\pi_z^o(X)\right)\beta_{t,k}\\
&+\Bigg(P_{t,z}(X)-P_{t,z}^o(X)-\frac{\mathbf{1}\{Z=z\}}{\pi_z^o(X)}\Big(P_{t,z}(X)-P_{t,z}^o(X)\Big)\Bigg)\beta_{t,k}\\
&+\frac{\mathbf{1}\{Z=z\}}{(\pi_z^o(X))^2}\Big(Y\mathbf{1}\{T=t\}-Q_{t,z}^o(X)\Big)\left(\pi_z(X)-\pi_z^o(X)\right)\\
&-(Q_{t,z}(X)-Q_{t,z}^o(X))+\frac{\mathbf{1}\{Z=z\}}{\pi_z^o(X)}\Big(Q_{t,z}(X)-Q_{t,z}^o(X)\Big)\Bigg],
\end{aligned}
$$

which equals zero because of the following three identities:

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{Z=z\}/\pi_z^o(X) \mid X] &= 1,\\ \mathbb{E}[\mathbf{1}\{Z=z\}/\pi_z^o(X)(\mathbf{1}\{T=t\}-P_{t,z}^o(X)) \mid X] &= 0,\\ \mathbb{E}[\mathbf{1}\{Z=z\}/\pi_z^o(X)(Y\mathbf{1}\{T=t\}-Q_{t,z}^o(X)) \mid X] &= 0.\end{aligned}$$

□

Proof of Theorem 3.4. The asserted claims follow from Theorem 3.1, Theorem 3.2, and Corollary 3.2 of Chernozhukov et al. [2018] (henceforth referred to as the DML paper). We want to verify their Assumption 3.1 and 3.2. Adopting the notation from the DML paper, we let

$$\psi^a(T,Z,X,P_t,\pi) = -b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota\mathbf{1}\{T=t\}-P_t(X)\right)+P_t(X)\right)$$

and

$$\psi^b(Y,T,Z,X,Q_t,\pi) = b_{t,k}\left(\zeta(Z,X,\pi)\left(\iota(Y\mathbf{1}\{T=t\})-Q_t(X)\right)+Q_t(X)\right)$$

so that the linearity of the moment condition (with respect to $\beta_{t,k}$) is verified by the fact that $\psi = \psi^a\beta^{t,k}+\psi^b$. Define[1]

$$\varepsilon_n = \max_{z\in\mathscr{Z}}\left(\|\hat{Q}_{t,z}-Q_{t,z}^o\|_2 \vee \|\hat{P}_{t,z}-P_t^o\|_2 \vee \left\|\hat{\pi}_z-\pi_z^o\right\|_2\right).$$

By assumption on the convergence rates of the nonparametric estimators, we have $\varepsilon_n = o(n^{-1/4})$. Define $C_\varepsilon = C_{\varepsilon,1}\vee C_{\varepsilon,2}\vee C_{\varepsilon,3}\vee C_{\varepsilon,4}$, where $C_{\varepsilon,1}, C_{\varepsilon,2}, C_{\varepsilon,3}$, and $C_{\varepsilon,4}$ are positive constant that only depends on C and ε and are specified later in the proof. Let δ_n be a sequence of positive constants approaching zero and satisfies that $\delta_n \geq C_\varepsilon\left(\varepsilon_n^2\sqrt{n}\vee n^{-1/4}\vee n^{-(1-2/q)}\right)$. Such construction is

[1] For simplicity, we drop the superscript l in the nonparametric estimators.

possible since $\sqrt{n}\varepsilon_n^2 = o(1)$. We set the nuisance realization set N_n (denoted by $\mathscr{T}_N$ in the DML paper) to be the set of all vector functions $(Q_t, P_t, \pi_z : z \in \mathscr{Z})$ consisting of square-integrable functions $Q_{t,z}, P_{t,z}$, and π_z such that for all $z \in \mathscr{Z}$:

$$\begin{aligned}
&\left\|Q_{t,z}\right\|_q \leq C, P_{t,z} \in [0,1], \pi_z \in [\varepsilon, 1], z \in \mathscr{Z}, \\
&\|Q_{t,z} - Q_{t,z}^o\|_q \vee \|P_{t,z} - P_{t,z}^o\|_q \vee \left\|\pi_z - \pi_z^o\right\|_q \leq \varepsilon_n, \\
&\left\|\pi_z - \pi_z^o\right\|_2 \times \left(\|Q_{t,z} - Q_{t,z}^o\|_2 + \|P_{t,z} - P_{t,z}^o\|_2\right) \leq \varepsilon_n^2.
\end{aligned}$$

Consider Assumption 3.1 in the DML paper. Assumption 3.1(d), the Neyman orthogonality condition, is verified by Proposition 3.3, where the validity of the differentiation under the integral operation is verified later in the proof. Assumption 3.1(e), the identification condition, is verified by the condition that $p_{t,k}^o \in [\varepsilon, 1]$. The remaining conditions of Assumption 3.1 in the DML paper are trivially verified.

Next, we consider Assumption 3.2 in the DML paper. Note that Assumption 3.2(a) holds by the construction of N_n and ε_n and our assumptions on the nuisance estimates. Assumption 3.2(d) is verified by our assumption that the semiparametric efficiency bound of $\beta_{t,k}$ is above ε. The remaining task is to verify Assumption 3.2(b) and 3.2(c) in the DML paper. To do that, we choose n sufficiently large and let $(Q_{t,z}, P_{t,z}, \pi_z : z \in \mathscr{Z})$ be an arbitrary element of the nuisance realization set N_n. We keep the above notations throughout the remaining part of the proof. Define

$$\psi_z^a(T, Z, X, P_t, \pi_z) = \frac{\mathbf{1}\{Z = z\}}{\pi_z(X)}(\mathbf{1}\{T = t\} - P_{t,z}(X)) + P_{t,z}(X)$$

and

$$\psi_z^b(Y, T, Z, X, Q_t, \pi_z) = \frac{\mathbf{1}\{Z = z\}}{\pi_z(X)}(Y\mathbf{1}\{T = t\} - Q_{t,z}(X)) + Q_{t,z}(X).$$

Since ψ^a is a linear combination of $\psi_z^a, z \in \mathscr{Z}$ and ψ^b is a linear combination of $\psi_z^b, z \in \mathscr{Z}$, we only need $\left\|\psi_z^a(T,Z,X,P_t,\pi_z)\right\|_q$ and $\left\|\psi_z^b(Y,T,Z,X,Q_t,\pi_z)\right\|_q$ to be uniformly bounded (i.e., the bounds do not depend on n) for $z \in \mathscr{Z}$ in order to verify Assumption 3.2(b) in the DML paper. In fact,

$$\begin{aligned}\left\|\psi_z^b(Y,T,Z,X,P_t,\pi_z)\right\|_q &\le \left\|\mathbf{1}\{Z=z\}/\pi_z(X)\big|Y\mathbf{1}\{T=t\}-Q_{t,z}(X)\big|\right\|_q + \left\|Q_{t,z}(X)\right\|_q \\ &\le \frac{1}{\varepsilon}\left(\left\|Y\mathbf{1}\{T=t\}\right\|_q + \left\|Q_{t,z}(X)\right\|_q\right) + \left\|Q_{t,z}(X)\right\|_q \le 2C/\varepsilon + C,\end{aligned}$$

where we have used the assumption that $\pi_z \ge \varepsilon$, $\left\|Y\mathbf{1}\{T=t\}\right\|_q \le C$, and $\left\|Q_t(X)\right\|_q \le C$. Similarly, we have

$$\begin{aligned}\left\|\psi_z^a(T,Z,X,P_t,\pi_z)\right\|_q &\le \left\|\mathbf{1}\{Z=z\}/\pi_z(X)\big|\mathbf{1}\{T=t\}-P_{t,z}(X)\big|\right\|_q + \left\|P_{t,z}(X)\right\|_q \\ &\le \frac{1}{\varepsilon}\left(1+\left\|P_{t,z}(X)\right\|_q\right) + \left\|P_{t,z}(X)\right\|_q \le 2/\varepsilon + 1,\end{aligned}$$

where we have used the assumption that $\pi_z \ge \varepsilon$ and $P_t \in [0,1]$. Thus, Assumption 3.2(b) in the DML paper is verified.

To verify Assumption 3.2(c) in the DML paper, we again only need to verify the corresponding conditions for ψ_z^a and ψ_z^b, respectively. For ψ_z^a, we have

$$\begin{aligned}&\left\|\psi_z^a(T,Z,X,P_t,\pi_z) - \psi_z^a(T,Z,X,P_t^o,\pi_z^o)\right\|_2 \\ \le&\left\|\frac{\pi_z(X)-\pi_z^o(X)}{\pi_z(X)\pi_z^o(X)}\right\|_2 + \left\|\frac{P_{t,z}(X)}{\pi_z(X)} - \frac{P_{t,z}^o(X)}{\pi_z^o(X)}\right\|_2 + \left\|P_{t,z}(X)-P_{t,z}^o(X)\right\|_2 \\ \le&\frac{1}{\varepsilon^2}\left\|\pi_z(X)-\pi_z^o(X)\right\|_2 + \frac{1}{\varepsilon^2}\left\|(P_{t,z}(X)-P_{t,z}^o(X))\pi_z^o(X) + P_{t,z}^o(X)(\pi_z^o(X)-\pi_z(X))\right\|_2 \\ &+\left\|P_{t,z}(X)-P_{t,z}^o(X)\right\|_2 \\ \le&\frac{2}{\varepsilon^2}\left\|\pi_z(X)-\pi_z^o(X)\right\|_2 + \left(1/\varepsilon^2+1\right)\left\|P_{t,z}(X)-P_{t,z}^o(X)\right\|_2 \le C_{\varepsilon,1}\varepsilon_n \le \delta_n,\end{aligned}$$

where the second to last inequality follows from the fact that $P^o_{t,z}, \pi^o_z \in [0,1]$. For ψ^b_z, we have

$$\begin{aligned}
&\left\|\psi^b_z(Y,T,Z,X,Q_t,\pi_z) - \psi^b_z(Y,T,Z,X,Q^o_t,\pi^o_z)\right\|_2 \\
\leq &\frac{1}{\varepsilon^2}\left\|\pi^o_z(X)(Y\mathbf{1}\{T=t\} - Q_{t,z}(X)) - \pi_z(X)(Y\mathbf{1}\{T=t\} - Q^o_{t,z}(X))\right\|_2 \\
&+\left\|Q_{t,z}(X) - Q^o_{t,z}(X)\right\|_2 \\
= &\frac{1}{\varepsilon^2}\left\|(Y\mathbf{1}\{T=t\} - Q^o_{t,z}(X))(\pi^o_z(X) - \pi_z(X)) + \pi^o_z(X)(Q^o_{t,z}(X) - Q_{t,z}(X))\right\|_2 \\
&+\left\|Q_{t,z}(X) - Q^o_{t,z}(X)\right\|_2 \\
\leq &\frac{1}{\varepsilon^2}\left\|(Y\mathbf{1}\{T=t\} - Q^o_{t,z}(X))(\pi^o_z(X) - \pi_z(X))\right\|_2 + \left\|\pi^o_z(X)(Q^o_{t,z}(X) - Q_{t,z}(X))\right\|_2 \\
&+\left\|Q_{t,z}(X) - Q^o_{t,z}(X)\right\|_2 \\
\leq &\frac{C}{\varepsilon^2}\left\|\pi^o_z(X) - \pi_z(X)\right\|_2 + \left(\frac{1}{\varepsilon^2}+1\right)\left\|Q^o_{t,z}(X) - Q_{t,z}(X)\right\|_2 \leq C_{\varepsilon,2}\varepsilon_n \leq \delta_n,
\end{aligned}$$

where the last inequality follows from our assumption that $|Y\mathbf{1}\{T=t\} - Q^o_t(X)| \leq C$ and the fact that $\pi^o_z \in [\varepsilon, 1]$. Combining the above two inequality results, we can verify the first two conditions of Assumption 3.2(c) in the DML paper.

For the last condition of Assumption 3.2(c) in the DML paper, which bounds the second-order Gateaux derivative, we again consider ψ^a_z and ψ^b_z separately. For $r \in [0,1)$, recall that $Q^r_{t,z} = Q^o_{t,z} + r(Q_{t,z} - Q^o_{t,z})$, $P^r_{t,z} = P^o_{t,z} + r(P_{t,z} - P^o_{t,z})$, and $\pi^r_z = \pi^o_z + r(\pi_z - \pi^o_z)$. Clearly, $P^r_{t,z}, \pi^r_z \in$

$[0,1]$. With differentiation under the integral, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial r^2}\mathbb{E}\left[\psi_z^a(T,Z,X,P_t^r,\pi_z^r)\right]\\
=&\frac{\partial}{\partial r}\mathbb{E}\Bigg[\frac{-\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}\Big(\mathbf{1}\{T=t\}-P_{t,z}^r(X)\Big)\Big(\pi_z(X)-\pi_z^o(X)\Big)\\
&+P_{t,z}(X)-P_{t,z}^o(X)-\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}\Big(P_{t,z}(X)-P_{t,z}^o(X)\Big)\Bigg]\\
=&\mathbb{E}\left[\frac{2\times\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^3}(\pi_z(X)-\pi_z^o(X))^2(\mathbf{1}\{T=t\}-P_{t,z}^r(X))\right]\\
&+\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}(\pi_z(X)-\pi_z^o(X))(P_{t,z}(X)-P_{t,z}^o)\right]\\
&+\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}(\pi_z(X)-\pi_z^o(X))(\mathbf{1}\{T=t\}-P_{t,z}^r(X))(P_{t,z}(X)-P_{t,z}^o)\right]\\
&-\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}(\mathbf{1}\{T=t\}-P_{t,z}^r(X))(P_{t,z}(X)-P_{t,z}^o)^2\right].
\end{aligned}$$

Using the fact that $|\mathbf{1}\{T=t\}-P_t^r(X)|\leq 1$ and $\pi_z^r\geq\varepsilon$, we can bound the above derivative by

$$\begin{aligned}
\Big|\frac{\partial^2}{\partial r^2}\mathbb{E}\left[\psi_z^a(T,Z,X,P_t^r,\pi_z^r)\right]\Big| &\leq C_\varepsilon\Big(\big\|\pi_z(X)-\pi_z^o(X)\big\|_2^2+\big\|P_{t,z}(X)-P_{t,z}^o(X)\big\|_2^2\Big)\\
&\quad+C_\varepsilon\big\|\pi_z(X)-\pi_z^o(X)\big\|_2\times\|P_{t,z}(X)-P_{t,z}^o(X)\|_2\\
&\leq C_{\varepsilon,3}\varepsilon_n^2\leq\delta_n/\sqrt{n}.
\end{aligned}$$

By bounding the first and second derivative uniformly with respect to r, we know that the differentiation under the integral operation is valid. So the Neyman orthogonality condition is verified. Analogously, we can show that

$$
\begin{aligned}
&\frac{\partial^2}{\partial r^2}\mathbb{E}\left[\psi_z^b(Y,T,Z,X,Q_t^r,\pi_z^r)\right]\\
=&\mathbb{E}\left[\frac{2\times\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^3}(\pi_z(X)-\pi_z^o(X))^2(Y\mathbf{1}\{T=t\}-Q_{t,z}^r(X))\right]\\
&+\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}(\pi_z(X)-\pi_z^o(X))(Q_{t,z}(X)-Q_{t,z}^o)\right]\\
&-\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{(\pi_z^r(X))^2}(\pi_z(X)-\pi_z^o(X))(Y\mathbf{1}\{T=t\}-Q_{t,z}^r(X))(Q_{t,z}(X)-Q_{t,z}^o)\right]\\
&-\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}(Y\mathbf{1}\{T=t\}-Q_{t,z}^r(X))(Q_{t,z}(X)-Q_{t,z}^o)^2\right].
\end{aligned}
$$

Under the assumption $|Y\mathbf{1}\{T=t\}-Q_{t,z}^o(X)|\leq C$, we have

$$
|Y\mathbf{1}\{T=t\}-Q_{t,z}^r(X)|\leq|Y\mathbf{1}\{T=t\}-Q_{t,z}^o(X)|+r|Q_{t,z}(X)-Q_{t,z}^o|\leq C+1,
$$

for all $r\in[0,1]$ and n large enough. Then we can bound the above derivative by

$$
\begin{aligned}
\left|\frac{\partial^2}{\partial r^2}\mathbb{E}\left[\psi_z^b(Y,T,Z,X,Q_t^r,\pi_z^r)\right]\right|\leq&C_\varepsilon\Big(\left\|\pi_z(X)-\pi_z^o(X)\right\|_2^2+\left\|Q_{t,z}(X)-Q_{t,z}^o(X)\right\|_2^2\Big)\\
&+C_\varepsilon\left\|\pi_z(X)-\pi_z^o(X)\right\|_2\times\left\|Q_{t,z}(X)-Q_{t,z}^o(X)\right\|_2\\
\leq&C_{\varepsilon,4}\varepsilon_n^2\leq\delta_n/\sqrt{n}.
\end{aligned}
$$

Therefore, we have verified the last condition of Assumption 3.2(c) in the DML paper.

Lastly, we need to verify the condition on δ_n in Theorem 3.1 and 3.2 in the DML paper, that is, $\delta_n\geq n^{-[(1-2/q)\wedge(1/2)]}$. This directly follows from the construction of δ_n. □

C.1.4 Proof of Weak IV Inference Results

Proof of Theorem 3.5. We first prove part (i). Consider applying the DML method to the moment condition (3.8) to estimate the parameter $\upsilon-\beta_0 p$ and obtain the standard error. We want to show the convergence in distribution of

$$\check{\sigma}_{\psi}^{-1}\sqrt{n}\left[(\check{\upsilon}-\beta_0\check{p})-(\upsilon-\beta_0 p)\right]=\check{\rho}-\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi} \tag{C.1}$$

to the standard normal distribution uniformly over the DGPs in $\mathscr{P}^{\mathrm{WI}}(c_0,c_1)$. To do that, we need to verify Assumptions 3.1 and 3.2 in the DML paper regarding the above moment condition. Assumptions 3.1(a)-(c) hold trivially. Assumption 3.1(d), the Neyman orthogonality condition, is verified by Proposition 3.3. That is, the Gateaux derivatives with respect to the nuisance parameters are zero regardless of the value of β. Assumption 3.1(e), the identification condition, is verified since the Jacobian of the parameter in the moment condition is 1. Assumption 3.2 in the DML paper can be verified in the same way as in the proof of Theorem 3.4. For brevity, we do not repeat the verification here.

For DGPs in $\mathscr{P}_{\beta_0}^{\mathrm{WI}}(c_0,c_1)$, (C.1) is equal to $\check{\rho}$. Therefore, the uniform convergence in distribution of $|\check{\rho}|$ is established in the null space, and the size of the test is uniformly controlled accordingly. For DGPs in $\mathscr{P}_{\beta}^{\mathrm{WI}}(c_0,c_1)$, where $\beta>\beta_0$, we have

$$\begin{aligned}\check{\rho}&=\left(\check{\rho}-\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi}\right)+\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi}\\&=\left(\check{\rho}-\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi}\right)+\sqrt{n}(\beta-\beta_0)p/\check{\sigma}_{\psi}.\end{aligned}$$

The first term on the RHS of the last equality converges in distribution to $N(0,1)$. In contrast, the second term diverges to infinity since $\check{\sigma}_{\psi}$ converges in probability to $\sigma_{\psi}\geq\sqrt{c_0}$ by Theorem 3.2 in the DML paper. Therefore, the probability of $|\check{\rho}|$ exceeding any finite number converges to 1. The case where $\beta<\beta_0$ is essentially the same.

To prove part (ii) of the theorem, notice that $(\beta-\beta_0)p\leq 0$ for any DGP in the null space $\bigcup_{\beta\leq\beta_0}\mathscr{P}_{\beta}^{\mathrm{WI}}(c_0,c_1)$, which implies that $\check{\rho}\leq\check{\rho}-\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi}$. Therefore,

$$\sup_P \mathbb{P}_P\Big(\check{\rho}>\mathscr{N}_{1-\alpha}\Big)\leq\sup_P \mathbb{P}_P\left(\check{\rho}-\sqrt{n}(\upsilon-\beta_0 p)/\check{\sigma}_{\psi}>\mathscr{N}_{1-\alpha}\right)\to\alpha,$$

where the supremum is taken over $P \in \bigcup_{\beta \leq \beta_0} \mathscr{P}_\beta^{\mathrm{WI}}(c_0, c_1)$. Consistency can be derived in the same way as part (i). □

C.2 Implicitly Defined Parameters

This section studies general parameters defined implicitly through moment conditions. We allow the moment conditions to be non-smooth, which is the case when the parameter of interest is the quantile. We also allow the moment conditions to be overidentifying, which could be the result of imposing the underlying economic theory on multiple levels of treatment and instrument.

To facilitate the exposition, we define a random variable $Y_{t,k}^*$ such that the marginal distribution of $Y_{t,k}^*$ is equal to the conditional distribution of Y_t given $S \in \Sigma_{t,k}$. The joint distribution of the $Y_{t,k}^*$'s is irrelevant and hence left unspecified. For convenience, we use a single index $j \in J$ rather than (t,k) for labeling. That is, we collect the $Y_{t,k}^*$'s into the vector $Y^* \equiv (Y_1^*, \cdots, Y_J^*)$. Let t_j be the treatment level associated with Y_j^*. The quantities p_j and b_j are analogously defined.[2]

Let the parameter of interest be η, which lies in the parameter space $\Lambda \subset \mathbb{R}^{d_\eta}$, $d_\eta \leq J$. The true value of the parameter η_0 satisfies the moment condition

$$\mathbb{E}\left[m(Y^*, \eta^o)\right] = 0,$$

where $m : \mathscr{Y}^J \times \mathbb{R}^{d_\eta} \to \mathbb{R}^J$ is a vector of functions:

$$m(Y^*, \eta) \equiv \big(m_1(Y_1^*, \eta), \cdots, m_J(Y_J^*, \eta)\big)'$$

Since the vector η appears in each m_j, restrictions are allowed both within and across different subpopulations. Another interesting feature of this specification is that the moment conditions

[2] We can further extend the vector Y^* to include variables whose marginal distributions are the same as the conditional distributions of Y_t given $T = t, S \in \Sigma_{t,k}$. Efficient estimation in this more general case is similar and hence omitted for brevity.

are defined for the random variables that are not observed. But their marginal distributions can be identified similar to Theorem 3.1.

Let $\bar{m} \equiv (\bar{m}_1', \cdots, \bar{m}_J')'$, where

$$\bar{m}_j(X,\eta) = \Big(\bar{m}_{j,z_1}(X,\eta), \cdots, \bar{m}_{j,z_{N_Z}}(X,\eta)\Big)'$$

and

$$\bar{m}_{j,z}(X,\eta) = \mathbb{E}\left[m_j(Y,\eta)\mathbf{1}\{T = t_j\} \mid Z = z, X\right].$$

The functions $\bar{m}_{j,z}$ are identified from the data. Similar to Theorem 3.1, we can show that the parameter η is identified by the moment conditions:

$$b_j\mathbb{E}\left[\bar{m}_j(X,\eta)\right] = 0, 1 \leq j \leq J \iff \eta = \eta^o.$$

The following theorem gives the SPEB for the estimation of η.

Theorem C.1. *Assume the following conditions hold.*

(i) $\mathbb{E}\Big[m(Y^*,\eta)^2\Big] < \infty, \eta \in \Lambda.$

(ii) For each j and z, $m_{j,t_j,z}$ is continuously differentiable in its second argument. Let Γ be the $J \times d_\eta$ matrix whose jth row is $b_j\frac{d}{d\eta}\mathbb{E}\left[\bar{m}_j(X,\eta)\right]\big|_{\eta=\eta^o}'$, and assume Γ has full column rank.

Then for the estimation of η, the EIF is

$$-\Big(\Gamma'V^{-1}\Gamma\Big)^{-1}\Gamma'V^{-1}\psi^\eta(Y,T,Z,X,\eta^o,\pi^o,\bar{m}^o), \tag{C.2}$$

where

$$V=\mathbb{E}\left[\psi^{\eta}(Y,T,Z,X,\eta,\pi,\bar{m})\psi^{\eta}(Y,T,Z,X,\eta,\pi,\bar{m})'\right]$$

and $\psi^{\eta}(Y,T,Z,X,\eta,\pi,\bar{m})$ *is a* $J\times 1$ *random vector whose jth element is*

$$b_j\left(\zeta(Z,X,\pi)\left(\iota(m_j(Y,\eta)\mathbf{1}\{T=t_j\})-\bar{m}_j(X,\eta)\right)+\bar{m}_j(X,\eta)\right) \tag{C.3}$$

In particular, the semiparametric efficiency bound is $\left(\Gamma'V^{-1}\Gamma\right)^{-1}$.

Proof of Theorem C.1. The proof is based on the approach described in section 3.6 of Hong and Nekipelov [2010a] and the proof of Theorem 1 in Cattaneo [2010]. We use a constant $d_\eta\times d_m$ matrix A to transform the overidentified vector of moments into an exactly identified system of equations $A\left(b_j\mathbb{E}\left[\bar{m}_j(X,\eta)\right]\right)_{j=1}^{J}=0$, find the A-dependent EIF for the exactly-identified parameter, and choose the optimal A. In a parametric submodel, the implicit function theorem gives that

$$\frac{\partial}{\partial\theta}\eta\big|_{\theta=\theta^o}=-\left(A\Gamma\right)^{-1}A\frac{\partial}{\partial\theta}\left(b_j\mathbb{E}_\theta\left[\bar{m}_j(X,\eta^o)\right]\right)_{j=1}^{J}\big|_{\theta=\theta^o},$$

where $\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\bar{m}_j(X,\eta^o)\right]\big|_{\theta=\theta^o}$ is an $N_Z\times 1$ random vector whose typical element can be represented by

$$\begin{aligned}&\int m_j(y,\eta^o)\mathbf{1}\{\tau=t_j\}s_z(y,\tau\mid x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx\\&+\int m_j(y,\eta^o)\mathbf{1}\{\tau=t_j\}s_X(x;\theta^o)f_z(y,\tau\mid x;\theta^o)f_X(x;\theta^o)dyd\tau dx,\end{aligned}$$

for $z\in\mathscr{Z}$. So the EIF for this exactly-identified parameter is

$$\psi^A(Y,T,Z,X,\eta^o,\pi^o,\bar{m}^o)=-\left(A\Gamma\right)^{-1}A\Psi^{\eta}(Y,T,Z,X,\eta^o,\pi^o,\bar{m}^o),$$

where ψ^{η} is defined by Equation (C.3). It is straightforward to verify that ψ^{A} satisfies $\frac{\partial}{\partial\theta}\eta\big|_{\theta=\theta^o} = \mathbb{E}\left[\psi^A s'_{\theta^o}\right]$, and $\psi^A \in \mathscr{S}$. The optimal A is chosen by minimizing the sandwich matrix $\mathbb{E}\left[\psi^A(\psi^A)'\right] = (A\Gamma)^{-1} A\mathbb{E}\left[\psi^{\eta}(\psi^{\eta})'\right] A' \left(\Gamma' A'\right)^{-1}$. Thus, the EIF for the over-identified parameter is obtained when $A = \Gamma' V^{-1}$. Plugging this expression into ψ^A, we obtain Equation (C.2). □

Note that, for example, $m_j(Y_j^*, \eta) = Y_j^* - \eta$, then $\eta = \beta_j$, and the efficiency bound shown above reduces to the one computed in Theorem 3.3. If $T = Z$, that is, the treatment satisfies the unconfounded, then the Theorem C.1 reduces to Theorem 1 in Cattaneo [2010].

For estimation, we use the EIFs to generate moment conditions and propose a three-step semiparametric GMM procedure. The criterion function is

$$\Psi_n^{\eta}(\eta, \pi, m) = \frac{1}{n}\sum_{i=1}^{n} \psi^{\eta}(Y_i, T_i, Z_i, X_i, \eta, \pi, \bar{m}). \tag{C.4}$$

Its probability limit is denoted as

$$\Psi^{\eta}(\eta, \pi, m_Z) = \mathbb{E}\left[\psi^{\eta}(Y, T, Z, X, \eta, \pi, \bar{m})\right], \tag{C.5}$$

where the expectation is taken with respect to the true parameters $(\pi^o, \bar{m}^o)$. The implementation procedure is as follows. Assume that we have nonparametric estimators $\hat{\pi}$ and $\hat{m}$ that consistently estimate π^o and $\bar{m}^o$, respectively. We first find a consistent GMM estimator $\tilde{\eta}$ using the identity matrix as the weighting matrix, that is,

$$\left\|\Psi_n^{\eta}(\tilde{\eta}, \hat{\pi}, \hat{m})\right\|_2 \leq \inf_{\eta\in\Lambda} \left\|\Psi_n^{\eta}(\eta, \hat{\pi}, \hat{m})\right\|_2 + o_p(1). \tag{C.6}$$

Next, we use this estimate to form a consistent estimator $\hat{V}$ of the covariance matrix V, where

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} \psi^{\eta}(Y_i, T_i, Z_i, X_i, \tilde{\eta}, \hat{\pi}, \hat{m}) \psi^{\eta}(Y_i, T_i, Z_i, X_i, \tilde{\eta}, \hat{\pi}, \hat{m})'.$$

Then we let $\hat{\eta}$ be the optimally-weighted GMM estimator:

$$\begin{aligned}&\Psi_n^{\eta}(\hat{\eta},\hat{\pi},\hat{m}_Z)V_n(\tilde{\eta},\hat{\pi},\hat{m}_Z)^{-1}\Psi_n^{\eta}(\hat{\eta},\hat{\pi},\hat{m}_Z)'\\ &\leq \inf_{\eta\in\Lambda}\Psi_n^{\eta}(\eta,\hat{\pi},\hat{m}_Z)V_n(\tilde{\eta},\hat{\pi},\hat{m}_Z)^{-1}\Psi_n^{\eta}(\eta,\hat{\pi},\hat{m}_Z)'+o_p\left(n^{-1/2}\right).\end{aligned}$$

To conduct inference, we estimate Γ using the estimator $\hat{\Gamma}$ whose elements are defined as

$$\hat{\Gamma}_{jl}=\frac{1}{n}\sum_{i=1}^{n}b_j\frac{\partial}{\partial\eta}\hat{m}_j(X_i,\eta)\Big|_{\eta=\hat{\eta}},$$

where we have implicitly assumed that the estimator $\hat{m}_j$ is differentiable in its second argument.

In the following theorem, we derive the asymptotic properties of the GMM estimators. The main theoretical difficulty is that the random criterion function $\Psi_n(\cdot,\hat{\pi},\hat{m})$ could potentially be discontinuous because we allow $m(Y^*,\cdot)$ to be discontinuous. We use the theory developed in Chen et al. [2003] to overcome this problem.[3] Let Π_z be the function class that contains π_z^o. Let $\mathscr{M}_{j,z}$ be the function class that contains $\bar{m}_{j,z}^o$.

Theorem C.2. *Let the assumptions in Theorem C.1 hold. Assume the following conditions hold.*

(i) The parameter space Λ is compact. The true parameter η^o is in the interior of Λ.

(ii) For any j,z and $\bar{m}_{j,z}\in\mathscr{M}_{j,z}$, there exists $C>0$ such that for $\delta>0$ sufficiently small,

$$\sup_{|\eta'-\eta|\leq\delta}\mathbb{E}\left|\bar{m}_{j,z}(X,\eta')-\bar{m}_{j,z}(X,\eta)\right|^2\leq C\delta^2.$$

(iii) Donsker properties:

$$\int_0^{\infty}\log N(\varepsilon,\Pi_z,\|\cdot\|_{\infty})d\varepsilon,\int_0^{\infty}\log N(\varepsilon,\mathscr{M}_{j,z},\|\cdot\|_{\infty})d\varepsilon<\infty,$$

[3] Cattaneo [2010] instead uses the theory from Pakes and Pollard [1989]. However, the general theory of Chen et al. [2003] is more straightforward to apply in this case since they explicitly assume the presence of infinite-dimensional nuisance parameters, which can depend on the parameters to be estimated.

where $N(\varepsilon, \mathscr{F}, \|\cdot\|)$ denotes the covering number of the space $(\mathscr{F}, \|\cdot\|)$.

(iv) Convergence rates of the nonparametric estimators:

$$\left\|\hat{\pi}_z - \pi_z^o\right\|_\infty, \|\hat{m}_{j,z} - \bar{m}_{j,z}^o\|_\infty = o_p(n^{-1/4}).$$

(v) The function $\sup_{\eta\in\Lambda}\left|\frac{\partial}{\partial\eta}\bar{m}_j^o(\cdot,\eta)\right|$ is integrable. The estimator $\frac{\partial}{\partial\eta}\hat{m}_j$ is consistent uniformly in its second argument, that is,

$$\left\|\frac{\partial}{\partial\eta}\hat{m}_j(x,\eta) - \frac{\partial}{\partial\eta}\bar{m}_j^o(x,\eta)\right\|_\infty = o_p(1), \forall x.$$

Then $\tilde{\eta} = \eta^o + o_p(1)$, $\hat{V} = V + o_p(1)$, $\hat{\Gamma} = \Gamma + o_p(1)$, and

$$\sqrt{n}(\hat{\eta} - \eta^o) \implies N\left(\mathbf{0}, (\Gamma' V^{-1}\Gamma)^{-1}\right),$$

where $\mathbf{0}$ denotes a vector of zeros.

The following lemma is helpful for proving Theorem C.2.

Lemma C.3. *Under the assumptions of Theorem C.1, the class*

$$\mathscr{F} \equiv \left\{\psi^\eta(Y,T,Z,X,\eta,\pi,\bar{m}) : \pi \in \Pi_z, \bar{m}_{j,z} \in \mathscr{M}_{j,z}, 1 \le j \le J, z \in \mathscr{Z}\right\}$$

is Donsker with a finite integrable envelope. The following stochastic equicontinuity condition hold: for any positive sequence $\delta_n = o(1)$,

$$\sup\Big\{\Psi_n^\eta(\eta,\pi,\bar{m}) - \Psi^\eta(\eta,\pi,\bar{m}) - \Psi_n^\eta(\eta^o,\pi^o,m_Z^o) : \\ \|\eta - \eta^o\|_2 \vee \|\pi - \pi^o\|_\infty \vee \|\bar{m} - \bar{m}^o\|_\infty \le \delta_n\Big\} = o_p\left(n^{-1/2}\right),$$

where the supremum is taken over $\eta \in \Lambda$, $\pi_z \in \Pi_z$, and $\bar{m}_{j,z} \in \mathscr{M}_{j,z}$.

Proof of Lemma C.3. We first verify that the moment condition ψ^{η} satisfies Condition (3.2) of Theorem 3 in Chen et al. [2003] (hereafter CLK). In fact, when $\|\bar{m}'_{j,z} - \bar{m}_{j,z}\|_{\infty} \vee \left\|\eta' - \eta\right\|_{\infty} \leq \delta$, the triangle inequality gives that

$$\begin{aligned}
&\mathbb{E}\left|\bar{m}'_{j,z}(X,\eta') - \bar{m}_{j,z}(X,\eta)\right|^2 \\
\leq &2\mathbb{E}\left|\bar{m}'_{j,z}(X,\eta') - \bar{m}'_{j,z}(X,\eta)\right|^2 + 2\mathbb{E}\left|\bar{m}'_{j,z}(X,\eta) - \bar{m}_{j,z}(X,\eta)\right|^2 \\
\leq &const \times \delta^2,
\end{aligned}$$

where we use the notation *const* to denote a generic constant that may have different values at each appearance. The last inequality follows from the assumption (ii). Similarly, we can verify that the remaining terms in ψ^{η} also satisfy the same condition. Therefore, ψ^{η} is locally uniformly L_2-continuous, that is,

$$\begin{aligned}
\mathbb{E}\Big[\sup\Big\{&\left|\psi^{\eta}(Y,T,Z,X,\eta',\pi',\bar{m}') - \psi^{\eta}(Y,T,Z,X,\eta,\pi,\bar{m})\right| : \\
&\left\|\eta' - \eta\right\| \vee \left\|\pi' - \pi\right\|_{\infty} \vee \left\|\bar{m}' - \bar{m}\right\|_{\infty} \leq \delta\Big\}\Big] \leq const. \times \delta^2.
\end{aligned}$$

Following the same steps as in the proof of Theorem 3 in CLK (p. 1607), we can show that the bracketing number of $\mathscr{F}$ is bounded by

$$\begin{aligned}
&N_{[]}\left(\varepsilon, \mathscr{F}, \|\cdot\|_{L_2}\right) \\
\leq &N(\varepsilon/const, \Lambda, \|\cdot\|) \times \prod_{z} N(\varepsilon/const, \Pi_z, \|\cdot\|) \times \prod_{j,z} N(\varepsilon/const, \mathscr{M}_{j,z}, \|\cdot\|).
\end{aligned}$$

Therefore, the bracketing entropy of class $\mathscr{F}$ is bounded by

$$\begin{aligned}
&\log N_{[]}\left(\varepsilon,\mathscr{F},\|\cdot\|_{L_2}\right)\\
\leq &const\times\Bigg(\log N(\varepsilon/const,\Lambda,\|\cdot\|)\vee\max_{z}\log N(\varepsilon/const,\Pi_z,\|\cdot\|)\\
&\vee\max_{j,z}\log N(\varepsilon/const,\mathscr{M}_{j,z},\|\cdot\|)\Bigg).
\end{aligned}$$

Under the assumption that Λ is compact and

$$\int_0^\infty \log N(\varepsilon,\Pi_z,\|\cdot\|)d\varepsilon,\int_0^\infty \log N(\varepsilon,\mathscr{M}_{j,z},\|\cdot\|)d\varepsilon<\infty,\forall j,z,$$

we have that

$$\int_0^\infty \log N_{[]}\left(\varepsilon,\mathscr{F},\|\cdot\|_{L_2}\right)d\varepsilon<\infty.$$

This implies that $\mathscr{F}$ is Donsker with a finite integrable envelope. Lastly, as stated in Lemma 1 of CLK, the asserted stochastic equicontinuity condition is implied by the fact that $\mathscr{F}$ is Donsker and ψ^η is L_2-continuous. □

Proof of Theorem C.2. We follow the large sample theory in CLK and set $\theta=\eta$, $h=(\pi,\bar{m})$, $M(\theta,h)=\Psi^\eta(\eta,\pi,\bar{m})$, and $M_n(\theta,h)=\Psi_n^\eta(\eta,\pi,\bar{m})$.

We first use Theorem 1 in CLK to show the consistency of $\tilde{\eta}$. Condition (1.2) in CLK is satisfied because Λ is compact, and $\Psi^\eta(\eta,\pi^o,\bar{m}^o)$ has a unique zero and is continuous by our second condition in Theorem C.1. As for Condition (1.3) of CLK, we can easily see from the expression of Ψ that it is continuous with respect to $\bar{m}_{j,z}$ and π_z (since π_z is bounded away from zero), and the uniformity in η follows from the fact that $\mathbb{E}\left[m(Y^*,\eta)\right]$ is bounded as a function of η. Condition (1.4) of CLK is satisfied by the assumption of Theorem C.2. The uniform stochastic equicontinuity condition (1.5) of CLK is implied by Lemma C.3. Therefore, $\tilde{\eta}=\eta^o+o_p(1)$.

We use Corollary 1 (which is based on Theorem 2) in CLK to show the consistency of $\hat{V}$ and the asymptotic normality of $\hat{\eta}$. Condition (2.2) in CLK is verified by the assumptions of Theorem C.1. Similar to the proof of Proposition 3.3, we can show that the moment condition Ψ^{η}, based on the EIF, satisfies the Neyman orthogonality condition for the nuisance parameters π and m_Z. In fact, for any j and z, we let $\pi_z^r = \pi_z^o(X) + r(\pi_z(X) - \pi_z^o(X))$ and $\bar{m}_{j,z}^r(X,\eta) = \bar{m}_{j,z}^o(X,\eta) + r\Big(\bar{m}_{j,z}(X,\eta) - \bar{m}_{j,z}^o(X,\eta)\Big)$. Then we have

$$\begin{aligned}
&\frac{d}{dr}\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{\pi_z^r(X)}\Big(m_j(Y,\eta)\mathbf{1}\{T=t_j\} - \bar{m}_{j,z}^r(X,\eta)\Big) + \bar{m}_{j,z}^r(X,\eta)\right]\Bigg|_{r=0} \\
&= \mathbb{E}\Bigg[-\frac{\mathbf{1}\{Z=z\}}{\big(\pi_z^o(X)\big)^2}\left(\pi_z(X) - \pi_z^o(X)\right)\Big(m_j(Y,\eta)\mathbf{1}\{T=t_j\} - \bar{m}_{j,z}^o(X,\eta)\Big) \\
&\quad + \Big(\bar{m}_{j,z}^o(X,\eta) - \bar{m}_{j,z}(X,\eta)\Big)\left(\frac{\mathbf{1}\{Z=z\}}{\pi_z^o(X)} - 1\right)\Bigg] = 0,
\end{aligned}$$

where we have applied the law of iterated expectations and used the fact that

$$\mathbb{E}\left[\frac{\mathbf{1}\{Z=z\}}{\pi_z^o(X)}\Big(m_j(Y,\eta)\mathbf{1}\{T=t_j\} - \bar{m}_{j,z}^o(X,\eta)\Big)\Big|X\right] = 0.$$

Thus, the path-wise derivative of Ψ^{η} with respect to $h = (\pi, \bar{m})$ is zero in any direction. Hence, Condition (2.3) of CLK is verified. Condition (2.4) in CLK directly follows from our assumptions of Theorem C.2. The stochastic equicontinuity condition (condition (2.6) in CLK) follows from Lemma C.3. Lastly, condition (2.6) in CLK is verified using the central limit theorem since the path-wise derivative is zero. Due to the presence of $\hat{V}$, we also need the uniform convergence condition in Corollary 1 of CLK, which can be verified by using Lemma C.3 and an application of Theorem 2.10.14 of van der Vaart and Wellner [1996a].

Lastly, to show the consistency of $\hat{\Gamma}$, we only need to show that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}\hat{m}_{j,t_j,z}(X_i,\hat{\eta}) \xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial\eta}\hat{m}_{j,z}(X,\eta^o)\right] = \frac{\partial}{\partial\eta}\mathbb{E}\left[\hat{m}_{j,z}(X,\eta^o)\right],$$

where the inequality follows from the differentiation under integral operation which holds under the last assumption of the theorem. The convergence in probability follows from the uniform convergence of $\frac{\partial}{\partial \eta} \hat{m}_{j,z}$ and the consistency of $\hat{\eta}$. Therefore, the desired convergence results follow. □

Bibliography

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, 81(3):919–943, 2014.

Douglas Almond, Jr. Doyle, Joseph J., Amanda E. Kowalski, and Heidi Williams. Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns*. *The Quarterly Journal of Economics*, 125(2):591–634, 05 2010. ISSN 0033-5533. doi: 10.1162/qjec.2010.125.2.591. URL https://doi.org/10.1162/qjec.2010.125.2.591.

Donald W.K. Andrews. Chapter 37 empirical process methods in econometrics. volume 4 of *Handbook of Econometrics*, pages 2247–2294. Elsevier, 1994. doi: https://doi.org/10.1016/S1573-4412(05)80006-6. URL https://www.sciencedirect.com/science/article/pii/S1573441205800066.

Donald W.K. Andrews, Xu Cheng, and Patrik Guggenberger. Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics*, 218(2):496–531, 2020.

Isaiah Andrews and Timothy B Armstrong. Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, 8(2):479–503, 2017.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Tom M. Apostol. *Mathematical Analysis: A Modern Approach to Advanced Calculus*. Pearson, New York, NY, 2 edition, 1974. ISBN 9780201002881.

Alan I. Barreca, Melanie Guldi, Jason M. Lindo, and Glen R. Waddell. Saving Babies? Revisiting the effect of very low birth weight classification*. *The Quarterly Journal of Economics*, 126 (4):2117–2123, 10 2011. ISSN 0033-5533. doi: 10.1093/qje/qjr042. URL https://doi.org/10.1093/qje/qjr042.

Guglielmo Beccuti and Silvana Pannain. Sleep and obesity. *Current opinion in clinical nutrition and metabolic care*, 14(4):402, 2011.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, , and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, New York, 1993.

Christine Blandhol, John Bonney, Magne Mogstad, and Alexander Torgovitsky. When is tsls actually late? Technical report, National Bureau of Economic Research, 2022.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Carolina Caetano, Gregorio Caetano, and Juan Carlos Escanciano. Regression discontinuity design with multivalued treatments. *arXiv preprint arXiv:2007.00185*, 2020.

David Card, David S. Lee, Zhuan Pei, and Andrea Weber. Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483, 2015. doi: https://doi.org/10.3982/ECTA11224. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11224.

Matias D Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

Matias D. Cattaneo and Juan Carlos Escanciano. *Regression discontinuity designs: Theory and applications*. Advances in Econometrics. Emerald Group Publishing, 2017.

Matias D. Cattaneo and Rocio Titiunik. Regression discontinuity designs, 2021.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455, 2020.

Xiaohong Chen and Andres Santos. Overidentification in regular models. *Econometrica*, 86(5): 1771–1817, 2018.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1555514.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences*, 110(32):12936–12941, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1300018110. URL https://www.pnas.org/content/110/32/12936.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005. doi: https://doi.org/10.1111/j.1468-0262.2005.00570.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00570.x.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/40664520.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Srinjoy Das and Dimitris N. Politis. Nonparametric estimation of the conditional distribution at regression boundary points. *The American Statistician*, 74(3):233–242, 2020. doi: 10.1080/00031305.2018.1558109. URL https://doi.org/10.1080/00031305.2018.1558109.

Xavier D'Haultfœuille and Philippe Février. Identification of nonseparable triangular models with discrete instruments. *Econometrica*, 83(3):1199–1210, 2015. doi: https://doi.org/10.3982/ECTA10038. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10038.

Yingying Dong. Alternative assumptions to identify late in fuzzy regression discontinuity designs. *Oxford Bulletin of Economics and Statistics*, 80(5):1020–1027, 2018a. doi: https://doi.org/10.1111/obes.12249. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12249.

Yingying Dong. Jump or kink? regression probability jump and kink design for treatment effect evaluation. Working paper, UC Irvine, 2018b.

Yingying Dong and Arthur Lewbel. Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *The Review of Economics and Statistics*, 97(5):1081–1092, 12 2015. ISSN 0034-6535. doi: 10.1162/REST_a_00510. URL https://doi.org/10.1162/REST_a_00510.

Yingying Dong, Ying-Ying Lee, and Michael Gou. Regression discontinuity designs with a continuous treatment. *Journal of the American Statistical Association*, 0(ja):1–31, 2021. doi: 10.1080/01621459.2021.1923509. URL https://doi.org/10.1080/01621459.2021.1923509.

Avraham Ebenstein, Maoyong Fan, Michael Greenstone, Guojun He, and Maigeng Zhou. New evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences*, 114(39):10384–10389, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1616784114. URL https://www.pnas.org/content/114/39/10384.

Yangin Fan and Emmanuel Guerre. *Multivariate Local Polynomial Estimators: Uniform Boundary Properties and Asymptotic Linear Representation*, volume 36, pages 489–537. Emerald Group Publishing Limited, 2021/10/18 2016. ISBN 978-1-78560-786-8, 978-1-78560-787-5. doi: 10.1108/S0731-905320160000036023. URL https://doi.org/10.1108/S0731-905320160000036023.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. Integrated public use microdata series, current population survey: Version 7.0 [dataset]. https://doi.org/10.18128/D030.V7.0, 2020. Minneapolis, MN: IPUMS.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, New York, NY, 2 edition, 1999.

Markus Frölich. Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007.

CAMILA Galindo. Empirical challenges of multivalued treatment effects. Technical report, Job market paper, 2020.

Evarist Giné and Armelle Guillou. Laws of the Iterated Logarithm for Censored Data. *The Annals of Probability*, 27(4):2042 – 2067, 1999. doi: 10.1214/aop/1022874828. URL https://doi.org/10.1214/aop/1022874828.

Evarist Giné and Armelle Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 37(4):503–522, 2001. ISSN 0246-0203. doi: https:

//doi.org/10.1016/S0246-0203(01)01081-0. URL https://www.sciencedirect.com/science/article/pii/S0246020301010810.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6): 907–921, 2002. ISSN 0246-0203. doi: https://doi.org/10.1016/S0246-0203(02)01128-7. URL https://www.sciencedirect.com/science/article/pii/S0246020302011287.

Osea Giuntella and Fabrizio Mazzonna. Sunset time and the economic effects of social jetlag: evidence from us time zone borders. *Journal of Health Economics*, 65:210–226, 2019. ISSN 0167-6296. doi: https://doi.org/10.1016/j.jhealeco.2019.03.007. URL https://www.sciencedirect.com/science/article/pii/S0167629618309718.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001. doi: https://doi.org/10.1111/1468-0262.00183. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00183.

Kristen G Hairston, Michael Bryer-Ash, Jill M Norris, Steven Haffner, Donald W Bowden, and Lynne E Wagenknecht. Sleep duration and five-year abdominal fat accumulation in a minority cohort: the iras family study. *Sleep*, 33(3):289–295, 2010.

Peter Hall, Rodney C. L. Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999. doi: 10.1080/01621459.1999.10473832. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473832.

Bruce E Hansen. Nonparametric estimation of smooth conditional distributions. *Unpublished paper: Department of Economics, University of Wisconsin*, 2004.

Bruce E Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

James J Heckman and Rodrigo Pinto. Unordered monotonicity. *Econometrica*, 86(1):1–35, 2018a.

James J Heckman and Rodrigo Pinto. Web appendix for unordered monotonicity. *Econometrica*, 86(1):1–35, 2018b.

James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111, 2001.

Stefan Hoderlein and Enno Mammen. Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518, 2007.

Stefan Hoderlein and Enno Mammen. Identification and estimation of local average derivatives in non-separable models without monotonicity. *The Econometrics Journal*, 12(1):1–25, 2009.

Sandra L. Hofferth, Sarah M. Flood, Matthew Sobek, and Daniel Backman. American time use survey data extract builder: Version 2.8 [dataset]. https://doi.org/10.18128/D060.V2.8, 2020. College Park, MD: University of Maryland and Minneapolis, MN: IPUMS.

Han Hong and Denis Nekipelov. Semiparametric efficiency in nonlinear late models. *Quantitative Economics*, 1(2):279–304, 2010a.

Han Hong and Denis Nekipelov. Supplement to "semiparametric efficiency in nonlinear late models". *Quantitative Economics*, 1(2):279–304, 2010b.

Hidehiko Ichimura and Whitney K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2007.05.001. URL https://www.sciencedirect.com/science/article/pii/S0304407607001091. The regression discontinuity design: Theory and applications.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009. doi: https://doi.org/10.3982/ECTA7108. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7108.

Takuya Ishihara. Partial identification of nonseparable models using binary instruments. *Econometric Theory*, 37(4):817–848, 2021.

Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015.

Patrick Kline and Christopher R Walters. Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics*, 131(4):1795–1848, 2016.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery*

Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008, volume 2033. Springer Science & Business Media, 2011.

Efang Kong, Oliver Linton, and Yingcun Xia. Uniform bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory*, 26(5):1529–1564, 2010.

David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 6 2010. doi: 10.1257/jel.48.2.281. URL https://www.aeaweb.org/articles?id=10.1257/jel.48.2.281.

Sokbae Lee and Bernard Salanié. Identifying effects of multivalued treatments. *Econometrica*, 86(6):1939–1963, 2018.

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

Elias Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.

Matthew A Masten and Alexander Torgovitsky. Identification of instrumental variable correlated random coefficients models. *Review of Economics and Statistics*, 98(5):1001–1005, 2016.

Rosa L. Matzkin. Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375, 2003. doi: https://doi.org/10.1111/1468-0262.00452. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00452.

Anna Mikusheva. Uniform inference in autoregressive models. *Econometrica*, 75(5):1411–1452, 2007.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2): 99–135, 1990.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Deborah Nolan and David Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987. ISSN 00905364. URL http://www.jstor.org/stable/2241339.

Ryo Okui, Dylan S Small, Zhiqiang Tan, and James M Robins. Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205, 2012.

Ariel Pakes and David Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057, 1989.

Rodrigo Pinto. Beyond intention to treat: Using the incentives in moving to opportunity to identify neighborhood effects. UCLA Working paper, 2021.

Jack Porter. Estimation in the regression discontinuity model. Working paper, University of Wisconsin at Madison, 2003.

Zhongjun Qu and Jungmo Yoon. Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics*, 185(1):1–19, 2015. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2014.10.008. URL https://www.sciencedirect.com/science/article/pii/S0304407614002462.

Yuya Sasaki. What do quantile regressions identify for general structural functions? *Econometric Theory*, 31(5):1102–1116, 2015.

Liangjun Su, Takuya Ura, and Yichong Zhang. Non-separable models with high-dimensional data. *Journal of Econometrics*, 212(2):646–677, 2019.

Yixiao Sun. Adaptive estimation of the regression discontinuity model. Working paper, UC San Diego, 2005.

Zhenting Sun. Instrument validity for heterogeneous causal effects. *arXiv preprint arXiv:2009.01995*, 2021.

Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

Alexander Torgovitsky. Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3):1185–1197, 2015. doi: https://doi.org/10.3982/ECTA9984. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9984.

Alexander Torgovitsky. Minimum distance from independence estimation of nonseparable instrumental variables models. *Journal of Econometrics*, 199(1):35–48, 2017. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2017.01.009. URL https://www.sciencedirect.com/science/article/pii/S0304407617300441.

Aad W van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 1998.

Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer: New York, 1996a.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, NY, 1996b. ISBN 978-1-4757-2545-2.

Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Halbert White. *Asymptotic theory for econometricians*. Academic press, 2001.

Keming Yu. *Smooth regression quantile estimation*. PhD thesis, The Open University, 1997. URL http://oro.open.ac.uk/57655/.

Keming Yu and M. C. Jones. Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237, 1998. ISSN 01621459. URL http://www.jstor.org/stable/2669619.