Identifying the Average Treatment Effect in Ordered Treatment Models Without Unconfoundedness*

Arthur Lewbel
Boston College

Thomas Tao Yang
Australian National University

original May 2013, revised April 2016

Abstract

We show identification of the Average Treatment Effect (ATE) when treatment is specified by ordered choice in cross section or panel models. Treatment is determined by location of a latent variable (containing a continuous instrument) relative to two or more thresholds. We place no functional form restrictions on latent errors and potential outcomes. Unconfoundedness of treatment doesn't hold and identification at infinity for the treated is not possible. Yet we still show nonparametric point identification and estimation of the ATE. We apply our model to reinvestigate the inverted-U relationship between competition and innovation, and find no inverted-U in US data.

JEL Codes: C14, C21, C26

Keywords: Average treatment effect, Ordered choice model, Special regressor, Semiparametric, Competition and innovation, Identification.

1 Introduction

We consider a model where possible treatments are specified by an ordered choice model. For example, treatment could be determined by an ordered probit. However, unlike probit, we will not specify the distribution of the latent error term. We also do not specify how outcomes are determined as functions of treatment. We place no functional form restrictions on the joint distribution of latent errors and potential outcomes. Unconfoundedness of treatment (either unconditional or conditional on covariates) does not hold, and identification at infinity for the treated is not possible. Yet we still show nonparametric point identification of the Average Treatment Effect (ATE), and we provide an associated estimator, which converges at

*Earlier versions of this paper circulated under the title, "Identifying the Average Treatment Effect in a Two Threshold Model." The authors would like to thank Aamir Hashmi for providing data, and Han Hong, Jim Heckman, Bertan Turhan, Filippo De Marco, Yatfung Wong and four anonymous referees for helpful comments and suggestions. All errors are our own. Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/lewbel/. Thomas Tao Yang, Research School of Economics, Australian National University, ACT 0200, Australia. tao.yang@anu.edu.au.

parametric rates. We describe application of the estimator to both cross section and panel data, though we focus on panel data. The panel model allows for fixed effects in both the treatment and outcome equations. In general, ordered choice panel models with fixed effects suffer from the incidental parameters problem, leading to slow rates of convergence, but we provide conditions under which our estimator does not suffer from the incidental parameters problem.

Suppose an outcome Y is given by

$$Y = Y_0 + (Y_1 - Y_0) D \tag{1.1}$$

where Y_0 and Y_1 are potential outcomes as in Rubin (1974), and D is a binary treatment indicator. Generally, point identification of the average treatment effect (ATE) $E(Y_1 - Y_0)$ requires either i) conditional or unconditional unconfoundedness of treatment, or ii) an instrument for D that can drive D to zero and to one with probability one (i.e., identification at infinity), or iii) functional restrictions on the joint distribution of Y_0, Y_1 and D. In contrast, we provide a novel point identification result, and an associated estimator, for the ATE in a model where none of these conditions hold.

Let V be a continuous instrument that affects the probability of treatment but not the outcome, and let X denote a vector of other covariates. Let D be given by a structure that is identical to one of the choices in an ordered choice model, that is,

$$D = I[\alpha_0(X) \leq V + U \leq \alpha_1(X)] \tag{1.2}$$

where $I(\cdot)$ is the indicator function that equals one if $\cdot$ is true and zero otherwise, U is a latent error term, $\alpha_0(X)$ and $\alpha_1(X)$ are unknown functions, and the coefficient of V is normalized to equal one. The joint distribution of $(U, Y_0, Y_1 \mid X)$ is assumed to be unknown. Later we provide an extension to a model where V in the treatment equation (1.2) is replaced with $\varsigma(V)$ for some unknown function ς.

In the special case of equation (1.2) where $\alpha_0(X)$ and $\alpha_1(X)$ are linear with the same slope, this is equivalent to treatment being given by the more standard looking ordered choice specification

$$D = I(\delta_0 \leq X'\beta_1 + V + U \leq \delta_1)$$

for constants δ_0, δ_1, and β_1. However, we don't impose these linearity restrictions. In addition, unlike standard ordered choice models, we allow the distribution of U to depend on X in completely unknown ways. Equivalently, the covariates X can all be endogenous regressors, with no available associated instruments. The only covariate we require to be exogenous is V.

In the proposed model, treatment and outcomes are confounded (both conditionally and unconditionally), because the unobservable U that affects D can be correlated with Y_0 and Y_1, with or without conditioning on X. No parametric or semiparametric restrictions are placed on the distribution of $(U, Y_0, Y_1 \mid X)$, so treatment effects are not identified by functional form restrictions on the distributions of unobservables. We assume V has large support, but the model is not identified at infinity. This is because both very large and very small values of V drive the probability of treatment close to zero, but no value of V (or of other covariates) drives the probability of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even a local ATE (LATE) is not identified in the usual way (e.g., Imbens and Angrist 1994), because monotonicity of treatment with respect to the instrument cannot hold in the proposed model. Nevertheless, we show that the ATE is identified in our model, using a special regressor argument as in Lewbel (1998, 2000, 2007). Our results include a test of the large support assumption required for this identification. We construct a very simple estimator of the ATE

based on this identification.

To illustrate the model and foreshadow our later empirical application, suppose the outcome Y is a measure of innovation in an industry and $D = 1$ when a latent measure of competitiveness in the industry lies between two estimated thresholds, otherwise $D = 0$. According to the "Inverted-U" theory in Aghion, Bloom, Blundell, Griffith, and Howitt (2005) (hereafter ABBGH), industries with intermediate levels of competitiveness have more innovation than those with low levels or high levels of competition. As in Revenga (1990, 1992), Bertrand (2004), and Hashmi (2013), we use a source-weighted average of industry exchange rates as an instrumental variable for competition, which we take to be our special regressor V. This instrument is computed from the weighted average of the US dollar exchange rate with the currencies of its trading partners. When V is low, products from the U.S. become relatively cheaper, thereby reducing competition by driving out competitors. The treatment effect we estimate is therefore the gains in innovation that result from facing moderate (rather than low or high) levels of competition.

With equations (1.1) and (1.2), one has $D = 0$ if the latent variable is either above the upper threshold or below the lower threshold. In many applications we would want to distinguish between those two cases. We would then have a standard ordered choice model for treatment, that is,

$$D_0 = I\left[V + U < \alpha_0\left(X\right)\right], \quad D_1 = I\left[\alpha_0\left(X\right) \leq V + U < \alpha_1\left(X\right)\right], \quad D_2 = I\left[\alpha_1\left(X\right) \leq V + U\right] \tag{1.3}$$

so an individual receives treatment j for $j = 0, 1, 2$ if $D_j = 1$. Letting W_j denote the potential outcome of an individual who receives treatment j, we would now have

$$Y = D_0 W_0 + D_1 W_1 + D_2 W_2. \tag{1.4}$$

In particular, W_0 is the potential outcome when lying below the lower threshold and W_2 is the potential outcome when lying above the upper threshold. The earlier model of equations (1.1) and (1.2) are the special case of this model where $D = D_1$, $Y_1 = W_1$, and

$$Y_0 = D_0 W_0 + D_2 W_2.$$

In this standard ordered choice model for treatment, the goal would be identification of the means of three potential outcomes, $E\left(W_j\right)$ for $j = 0, 1, 2$, corresponding to low, medium, and high values of the unobserved latent variable that determines selection.

In an extension section, we show identification of this ordered choice model, and identification of a more general model having any number of choices J instead of just the above case of $J = 3$. For $J = 3$, identification and estimation of $E\left(W_1\right)$ is identical to identification and estimation of $E\left(Y_1\right)$. But, unlike identification of $E\left(Y_0\right)$, identification of the separate extreme potential outcomes $E\left(W_0\right)$ and $E\left(W_2\right)$ requires identification at infinity arguments as in Heckman, Urzua and Vytlacil (2006). This means that, unlike estimation of our original treatment effect $E\left(Y_1 - Y_0\right)$, estimation of treatment effects like $E\left(W_1 - W_0\right)$ or $E\left(W_1 - W_2\right)$ will converge at slower than parametric rates. In contrast, identification at infinity is not needed for identification of $E\left(Y_0\right)$ and $E\left(Y_1\right)$, and indeed is not even possible for identification of $E\left(Y_1\right)$.

With this extension, our method can applied to most situations where treatment is defined by ordered response. For example, one might consider returns to education where the three possible treatments correspond to dropping out of school (the low group), completing high school (the middle group), and completing college (the high group). In our later empirical application, this extension will help us to distinguish between competing alternatives to the inverted-U hypothesis.

Even without this extension, our estimator is potentially useful in applications where one wants to assess the impact of a treatment defined as a moderate level of some activity, versus low or high levels. Many such treatments exist. For example, one might want to assess the effects of moderate levels of BMI or of alcohol consumption on a variety of health outcomes (see, e.g., Cao et al. 2014, Koppes et al. 2005, and Solomon et al. 2000). Other examples are the effect of moderate levels of financial development on the growth rates of countries (see Cecchetti and Kharroubi 2012) or the effects of moderate levels of financial regulation on measures of financial instability (see Huang 2015).

Our empirical application uses panel data. We extend our method to show identification of $E\left(Y_{jit}\right)$, and hence of $E\left(Y_{1it}-Y_{0it}\right)$, in the panel data model

$$Y_{it}=\widetilde{a}_i+\widetilde{b}_t+\left(1-D_{it}\right)Y_{0it}+Y_{1it}D_{it}, \tag{1.5}$$

$$D_{it}=I(\alpha_0(x_{it})\leq a_i+b_t+V_{it}+U_{it}\leq\alpha_1(x_{it})), \tag{1.6}$$

where $a_i,\widetilde{a}_i,b_t,\widetilde{b}_t$ are individual and time dummies in the selection and outcome equations. To interpret equation (1.5), define $Y_{it}^*=Y_{it}-\widetilde{a}_i-\widetilde{b}_t$, so Y_{it}^* is the outcome Y_{it} after time and individual specific fixed effects have been removed. Equation (1.5) then becomes $Y_{it}^*=\left(1-D_{it}\right)Y_{0it}+Y_{1it}D_{it}$, so Y_{0it} and Y_{1it} are the potential outcomes, not of Y_{it}, but of Y_{it}^*. With this construction, then the estimand $E\left(Y_{1it}-Y_{0it}\right)$ can be interpreted as a generalization of difference-in-difference (DID) estimation, where unlike standard DID, here D_{it} can be endogenous and hence correlated with the potential outcomes, so unconfoundedness of treatment and outcomes does not hold.

Let $\widetilde{Y}_{0it}$ and $\widetilde{Y}_{1it}$ denote the potential outcomes of Y_{it} itself. Depending on the application, it may be reasonable to add the assumption that $\widetilde{Y}_{dit}=\widetilde{a}_i+\widetilde{b}_t+Y_{dit}$, that is, that potential outcomes $\widetilde{Y}_{0it}$ and $\widetilde{Y}_{1it}$ each have the same fixed effects. If this assumption holds, then $Y_{it}=\left(1-D_{it}\right)\widetilde{Y}_{0it}+\widetilde{Y}_{1it}D_{it}$, and in that case our estimand $E\left(Y_{1it}-Y_{0it}\right)$ equals the standard ATE $E\left(\widetilde{Y}_{1it}-\widetilde{Y}_{0it}\right)$. In what follows, we will identify and estimate $E\left(Y_{1it}-Y_{0it}\right)$, which can always be given the interpretation of the ATE for Y_{it}^*, and with the additional assumption may also equal the ATE for Y_{it} itself.

Under either interpretation, equation (1.5) is itself a generalization of the panel outcome structure in Manski and Pepper's (2013) panel treatment model, [1] and in both ABBGH and Hashmi. Despite the presence of fixed effects in the nonlinear selection equation, and hence the potential for an incidental parameters problem (Neyman and Scott 1948), we attain a rate root nT estimate for the ATE in this panel model. This means our method can be applied with relatively short panels. In an appendix we also consider other panel specifications, including dynamic panels.

The next section is a literature review. In Section 3 we provide formal assumptions of our model, prove identification, and establish the consistency and asymptotic normality of our cross section and panel estimators. Section 4 contains some extensions, including a test of the support assumption regarding V, and identification of the general ordered treatment model, including more choices, and separately identifying the expected potential outcomes of lying below the lowest threshold or above the highest threshold. In Section 5 we empirically apply our estimator to investigate the relationship between competition and innovation. In this section we also implement simulation experiments to evaluate small sample properties of our estimators, using a Monte Carlo design that replicates features of our empirical data. This is followed by conclusions.

The paper additionally includes some appendices. Appendix A provides an evaluation of how the robust-

[1] Manski and Pepper (2013) consider the linear treatment response model $Y_{it}=\alpha_i+\beta D_{it}+\gamma t+\varepsilon_{it}$ where α_i is the individual fixed effect, γt is the time trend, ε_{it} is the random disturbance, and β defines the ATE. Our model generalizes theirs by replacing their fixed ATE β with a random coefficient and replacing the time trend γt with time fixed effects $\widetilde{b}_t$.

ness of our approach compares to more structural models in the presence of measurement errors. Appendix B provides some additional extensions including a generalization of the selection equation, alternative panel data asymptotics under different assumptions, and dynamic panel models. Appendix C provides additional technical assumptions and proofs. Finally, in a supplemental appendix separate from the main paper, we provide more details regarding derivation of the limiting distribution of our estimators, and other technical material.

2 Literature Review

Existing methods for point identifying ATE's are discussed in surveys such as Heckman and Vytlacil (2007a, 2007b) and Imbens and Wooldridge (2009). Early work on the identification of treatment effects (or comparing outcomes) was based either on functional form restrictions as in Haavelmo (1943), Roy (1951), and Heckman (1978), or on randomization and unconfoundedness, as in Neyman (1923), Cox (1958), Cochran and Rubin (1973), Rubin (1974), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), and Heckman and Robb (1984). But in our application (as noted by ABBGH) competition is an endogenous regressor, e.g., both competitiveness and innovation may be affected by business cycles and other variables. Much of what determines both is difficult to observe or even define, making it very unlikely that conditional unconfoundedness could ever hold, regardless of what observable covariates one conditions upon. For similar reasons, one would worry seriously about misspecification and omitted variables in any attempt to gain identification by functional form.

Without unconfoundedness of treatment, instrumental variables have been used in a variety of ways to identify causal effects. Instead of estimating the ATE, Imbens and Angrist (1994) show identification of a local average treatment effect (LATE), which is the ATE for a subpopulation called compliers (the definition of who compliers are, and hence the LATE, depends on the choice of instrument). An assumption for identifying the LATE is that the probability of treatment increase monotonically with the instrument. This assumption does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to decrease. Kitagawa (2009) shows that, with a monotonic instrument, ATE can only be identified at infinity. An implication of his result is that if ATE could be obtained another way, then nonmonotonicity of the instrument would be necessary. Our model possesses this necessary nonmonotonicity, and so constitutes the first example showing that such identification is possible.

Building on Björklund and Moffitt (1987), Heckman and Vytlacil (1999, 2005, 2007a) describe identification of a marginal treatment effect (MTE) as a basis for program evaluation. The MTE is based on having a continuous instrument, as we do. However, identification of the ATE using the MTE requires the assumption that variation in V can drive the probability of treatment to either zero or one, and hence depends on an identification at infinity argument. As we have already noted, identification at infinity is not possible in our model, since no value of V can drive the probability of treatment to one.

A few other papers consider identification of treatment effects in ordered choice models, such as Angrist and Imbens (1995) and Heckman, Urzua, and Vytlacil (2006). However, these papers consider identification of LATE and MTE, respectively, not ATE. Identification could also be achieved by functional form as in earlier Heckman (1979) selection models, e.g., by assuming that the latent error and potential outcomes are jointly normal. We impose no functional form restrictions on this distribution.

The way we achieve identification here is based on special regressor methods. Special regressors were introduced by Lewbel (1998, 2000). The instrumental variable V in our model needs to be continuous, conditionally independent of other variables and have a large support, which are all standard assumptions

for special regressor based estimators. See, e.g., Dong and Lewbel (2015), Lewbel, Dong, and Yang (2012), and Lewbel (2014). Some of the previously discussed papers also implicitly assume a special regressor, notably, Heckman, Urzua, and Vytlacil (2006).

Our proof of identification of $E(Y_1 \mid X)$ uses special regressor machinery based on Lewbel (2007), which exploits a related result to identify a class of semiparametric selection models. The fact that $E(Y_0 \mid X)$ and therefore that the ATE $E(Y_1 - Y_0)$ can be identified in the same way (without identification at infinity arguments) is new and, though very simple to show, is not trivial, since $D = 0$ means the latent variable lies outside thresholds, unlike $D = 1$. All of our panel model results, which comprise the bulk of the paper, including all of our results regarding fixed effects, are new. Also new is our test of the large support assumption and our extension to replacing V with some unknown function $\varsigma(V)$.

In addition to the ATE, our methods can be immediately extended to estimate quantile treatment effects as in Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2005), Bitler, Gelbach, and Hoynes (2006), or Firpo (2006). This is done by replacing Y with $I(Y \leq y)$ in our estimator.

In the panel context of equations (1.5) and (1.6), if unconfoundedness held, making $(Y_{0it}, Y_{1it}) \perp D_{it} \mid X_{it}$, and if in addition a_i and b_t were absent from the selection equation, then one could achieve identification via difference-in-difference methods, as in Ashenfelter (1978), Ashenfelter and Card (1985), Cook and Tauchen (1982, 1984), Card (1990), Meyer, Viscusi, and Durbin (1995), Card and Krueger (1993, 1994) and many others. In contrast, we obtain identification without unconfoundedness, and while allowing for a_i and b_t fixed effects. Analogous to Honore and Lewbel (2002), in panel data our identification and estimation strategy overcomes the incidental parameters problem associated with these fixed effects, and we attain a rate root nT estimate for the ATE.

Chernozhukov et al. (2009) discuss partial identification of marginal effects in nonlinear panel data, while Manski and Pepper (2013) provide partial identification of the (ATE) in a panel data context. Manski and Pepper also consider additional assumptions needed for point identification of the ATE in a panel setting (see their Section 3.1). Our panel data point identification requires some but not all of the assumptions they list as needed, including an average treatment response that is time-invariant, and the instrument exclusion restriction in the outcome equation.

3 The Model

In this section we first prove identification of the ATE in our model. The proof we provide is constructive, and we next describe a corresponding estimator. This is followed by some extensions, in particular, a panel data estimator with fixed effects. The remaining parts of this section then provide limiting distribution theory for the estimators.

3.1 Identification and Estimation

Let Ω_x and f_x denote the support and probability density function of the random variable X, and similarly for other variables.

Assumption 3.1 *We observe realizations of an outcome Y, binary treatment indicator D, a covariate V, and a $k \times 1$ covariate vector X. Assume the outcome Y and treatment indicator D are given by equations (1.1) and (1.2) respectively, where $\alpha_0(X)$ and $\alpha_1(X)$ are unknown threshold functions with $\alpha_0(X) < \alpha_1(X)$, U is an unobserved latent random error, and Y_0 and Y_1 are unobserved random untreated and treated potential outcomes. The joint distribution of (U, Y_0, Y_1), either unconditional or conditional on X, is unknown.*

Assumption 3.2 *Assume* $E(Y_j|X,V,U) = E(Y_j|X,U)$ *for* $j = 0, 1$, *and* $V \perp U \mid X$. *Assume* $V \mid X$ *is continuously distributed with probability density function* $f(V \mid X)$. *For all* $x \in supp(X)$, *the* $supp(V \mid X = x)$ *is an interval on the real line, and the interval* $[\inf supp(\alpha_0(X) - U \mid X = x), \sup supp(\alpha_1(X) - U \mid X = x)]$ *is contained in* $supp(V \mid X = x)$.

Assumption 3.1 defines the model, while Assumption 3.2 says that V is an instrument, in that V affects the probability of treatment but not outcomes (after conditioning on X). The instrument V is also continuously distributed, and has a large enough support so that, for any values U and X may take on, V can be small enough to make $D = 0$ or large enough to make $D = 0$. But no value of V and X will force $D = 1$, so identification at infinity is not possible.[2]

Remark 3.1 For identification, the assumption that $\text{supp}(V \mid X = x)$ equals an interval can be relaxed, as long as this support suitably contains $\alpha_0(x) - U$ and $\alpha_1(x) - U$ for all x. We maintain the single interval support to simplify notation in the identification proofs, and to apply the testing results in Section 4.1.

In this model, obtaining identification by imposing unconfoundedness would be equivalent to assuming that U was independent of $Y_1 - Y_0$, possibly after conditioning on covariates X. However, we do not make any assumption like this, so unconfoundedness does not hold. Alternatively, one might parametrically model the dependence of $Y_1 - Y_0$ on U to identify the model. In contrast we place no restrictions on the joint distribution of (U, Y_0, Y_1), either unconditional or conditioning upon X.

Assumption 3.3 *For some positive constant* τ, *define the trimming function* $I_\tau(v,x) = I[\inf supp(V|X = x)+\tau \leq v \leq \sup supp(V|X = x)-\tau]$. *Assume the interval* $[\inf supp(\alpha_0(X) - U \mid X = x), \sup supp(\alpha_1(X) - U \mid X = x)]$ *is contained in* $\{v : I_\tau(v,x) = 1\}$.

Assumption 3.4 *Assume there exists a positive constant* $\widetilde{\tau} < \tau$ *such that, for all* v,x *having* $I_{\widetilde{\tau}}(v,x) = 1$, *the density* $f(v|x)$ *is bounded away from zero (except possibly on a set of measure zero) and is bounded.*

Assumptions 3.3 and 3.4 will not be not necessary for identification, but will be convenient for simplifying the limiting distribution theory of the estimator we construct based on the identification. To save notation, let $I_\tau \equiv I_\tau(V,X)$. Define the function $\psi(X)$ by

$$\psi(X) \equiv \frac{E[I_\tau DY/f(V \mid X) \mid X]}{E[I_\tau D/f(V \mid X) \mid X]} - \frac{E[I_\tau (1-D) Y/f(V \mid X) \mid X]}{E[I_\tau (1-D)/f(V \mid X) \mid X]} \tag{3.1}$$

Theorem 3.1 *Let Assumptions 3.1, 3.2, 3.3 and 3.4 hold, or let Assumptions 3.1, 3.2 hold and set* $I_\tau \equiv 1$. *Then*

$$\psi(X) = E(Y_1 - Y_0 \mid X)$$

This theorem is proved in Appendix C.

Remark 3.2 This identification result immediately extends to the case of more choices. See the identification of $E(W_j \mid X)$ for $0 < j < J - 1$ in Theorem 4.3 for details.

[2] If instead of the ordered choice $D = I[\alpha_0(X) \leq V + U \leq \alpha_1(X)]$ we had a threshold crossing binary choice $D = I(\alpha_0(X) \leq V + U)$, then Assumption 3.2 would suffice to use "identification at infinity" to identify the treatment effect, by using data where V was arbitrarily low to estimate $E(Y_0 \mid X)$ and data where V was arbitrarily high to estimate $E(Y_1 \mid X)$. However, in our ordered choice model identification at infinity is not possible, since no value of V guarantees with high probability that Y will equal Y_1.

Theorem 3.1 shows identification of the conditional ATE since $\psi(X)$ is defined in terms of moments and densities of observed variables. The second part of the theorem shows that just Assumptions 3.1 and 3.2 are needed for identification. The first part of the theorem, giving identification including the additional Assumptions 3.3 and 3.4, is convenient because inclusion of the trimming term I_τ simplifies the asymptotics of the associated estimator. In particular, these assumptions permit fixed trimming that avoids boundary bias in our kernel estimators. Alternatives to Assumptions 3.3 and 3.4 for estimation could be based on asymptotic trimming arguments. In particular, without Assumption 3.3 we might on estimation replace τ with τ_n where $\tau_n \to 0$ as $n \to \infty$, and Assumption 3.4 might be replaced with another trimming indicator $I(f(v|x) > b_n)$, $b_n \to 0$, as $n \to \infty$.

We now briefly consider estimation with fixed trimming. Let $\widehat{E}(\cdot)$ denote the sample mean of the argument inside, and let $\widehat{f}(\cdot)$ and $\widehat{E}(\cdot|\cdot)$ denote nonparametric Nadayara-Watson kernel density and kernel regression estimators of the corresponding density and conditional mean functions $f(\cdot)$ and $E(\cdot|\cdot)$. It follows immediately from Theorem 3.1 that $\Psi \equiv E[\psi(X)]$ equals the ATE, which is therefore identified and can be consistently estimated by $\widehat{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}(x_i)$ where

$$\widehat{\psi}(x) = \frac{\widehat{E}\left[I_\tau DY/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{E}\left[I_\tau D/\widehat{f}(V \mid X) \mid X = x\right]} - \frac{\widehat{E}\left[I_\tau (1-D) Y/\widehat{f}(V \mid X) \mid X = x\right]}{\widehat{E}\left[I_\tau (1-D)/\widehat{f}(V \mid X) \mid X = x\right]},$$

with uniformly consistent kernel estimators $\widehat{f}$ and $\widehat{E}$.

To provide some intuition for Theorem 3.1, just for now simplify the analysis a bit by assuming that X is empty and that $V \perp (U, Y_0, Y_1)$. Then

$$E(D \mid U, Y_0, Y_1) = E(I[\alpha_0 - U \leq V \leq \alpha_1 - U] \mid U, Y_0, Y_1)$$

$$= \int_{supp(V)} I[\alpha_0 - U \leq v \leq \alpha_1 - U] f(v)\, dv = \int_{\alpha_0 - U}^{\alpha_1 - U} f(v)\, dv = F(\alpha_1 - U) - F(\alpha_0 - U)$$

where F is the cumulative distribution function of V. We have confoundedness because the above expression depends on U, which is correlated with Y_0 and Y_1. However, if V were uniformly distributed, then the above expression would simplify to $E(D \mid U, Y_0, Y_1) = \alpha_1 - \alpha_0$, which is independent of (U, Y_0, Y_1). So if V were uniformly distributed, the model would be unconfounded. Moreover, in that case f would be constant and equation (3.1) would reduce to the standard propensity score weighted estimator of the (unconfounded) average treatment effect. Scaling by the density of V in equation (3.1) is equivalent to converting to a uniform V, and so is equivalent to converting our model into one that is unconfounded. Density weighting is a feature of some special regressor estimators including Lewbel (2000, 2007), and indeed V has the properties of a special regressor, including appearing additively to unobservables in the model, a continuous distribution, large support, and conditional independence.

3.2 Small Extensions

The above identification and associated estimator $\widehat{\psi}(x)$ can be extended to handle random thresholds. In particular, equation (3.1) will still hold replacing the deterministic functions $\alpha_1(X)$ and $\alpha_0(X)$ with with random variables α_1 and α_0 (having distributions that could depend on X), provided that $(\alpha_0, \alpha_1) \perp (U, Y_1, Y_0) \mid X$.

These results also immediately extend to permit estimation of quantile treatment effects. The proof

of Theorem 3.1 shows that the first term in equation (3.1) equals $E(Y_1 \mid X)$ and the second term equals $E(Y_0 \mid X)$. Suppose we strengthen the assumption that $E(Y_j \mid X, V, U) = E(Y_j \mid X, U)$ for $j = 0, 1$ to say that $F_j(Y_j \mid X, V, U) = F_j(Y_j \mid X, U)$, where F_j is the distribution function of Y_j for $j = 0, 1$. Then one can apply Theorem 3.1 replacing Y with $I(Y \leq y)$ for any y, and thereby estimate $E(I(Y_j \leq y) \mid X) = F_j(y \mid X)$. Given this identification and associated estimators for the distributions $F_j(y \mid X)$ of the counterfactuals Y_j, we could then immediately recover quantile treatment effects.

3.3 Panel Data

We now consider a panel data version of the model, allowing for fixed effects. Let the model of treatment be

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(x_{it})), \tag{3.2}$$

and let the outcome equation be

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + (1 - D_{it}) Y_{0it} + Y_{1it} D_{it}, \tag{3.3}$$

where a_i and $\widetilde{a}_i$ equal the coefficients of individual i dummy variables, and where b_t and $\widetilde{b}_t$ equal the coefficients of time dummies in the two equations. For example, b_t is the coefficient of a dummy variable that equals one for all observations in time period t and zero otherwise.

As before, the observables in the model are the outcome Y, treatment D, instrument V, and covariate vector X. We assume that a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ for all i and t are random variables, in that we make some mild assumptions regarding their distribution. However, we interpret a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ as fixed effects, in that their values are not estimated, their distribution is not parameterized or estimated, and they are permitted to correlate with both X and with the unobservables in the model in unknown ways. When we take expectations of functions of a_i, b_t, $\widetilde{a}_i$, or $\widetilde{b}_t$, these should be taken as expectations over their distribution, however, all of our assumptions permit the distribution of a_i, b_t, $\widetilde{a}_i$, or $\widetilde{b}_t$ to be degenerate, and hence these fixed effects can be predetermined constants. Our model therefore encompasses both standard panel random effects and constant fixed effects.

Assumption 3.5 *For all individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables (though with possibly degenerate distribution), and*

$$E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} | X_{it}, V_{it}, a_i, b_t, U_{it}\right) = E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{jit} \middle| X_{it}, a_i, b_t, U_{it}\right),$$

for $j = 0, 1$. $V_{it} \perp a_i, b_t, U_{it} | X_{it}$.

Remark 3.3 We let a_i, b_t, $\widetilde{a}_i$, and $\widetilde{b}_t$ for all i and t be random variables (which can be correlated with X_{it}) to clarify the minimum restrictions we require of them, which is the above conditional independence with V_{it}. Note that the joint distribution of $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t, U_{it}, Y_{0it}, Y_{1it})$ conditional or unconditional on X_{it}, is assumed to be unknown and does not need to be specified or estimated. A similar assumption regarding fixed effects in discrete choice panel models appears in Honore and Lewbel (2002).

Assumption 3.6 *Assumptions 3.3 and 3.4 hold after replacing $supp(\alpha_0(X) - U, \alpha_1(X) - U)$ with $supp(\alpha_0(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it},\ \alpha_1(x_{it}) - \widetilde{a}_i - \widetilde{b}_t - U_{it})$ and replacing $I_\tau(v_i, x_i)$ with $I_{\tau it} \equiv I_\tau(v_{it}, x_{it})$.*

Assumptions 3.5 and 3.6 are essentially the panel data versions of Assumptions 3.2, 3.3, and 3.4.

Theorem 3.2 *Let Assumption 3.1, 3.5, and 3.6 hold for each individual i in each time period t. Let f_{v_t} denote the density of V in time t. Then*

$$\frac{E[I_{\tau it} D_{it} Y_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[I_{\tau it} D_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]} - \frac{E[I_{\tau it}(1-D_{it}) Y_{it} / f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[I_{\tau it}(1-D_{it}) / f_{v_t}(V_{it}|X_{it})|X_{it}]} = E(Y_{1it} - Y_{0it}|X_{it}). \tag{3.4}$$

This theorem is proved in Appendix C. Analogous to Theorem 3.1, the trimming parameter $I_{\tau it}$ and assumptions associated with trimming are not needed for identification, and are only included to simplify the associated asymptotic inference. Note that the expectations in equation (3.4) are taken over both i and t, and the resulting object we identify in this theorem, $E(Y_{1it} - Y_{0it} \mid X_{it})$, does not depend upon the fixed effects. The interpretation of this object as an average treatment effect is as discussed in the Introduction.

In typical panel data models, removing individual specific fixed effects requires some type of differencing over time, and similarly for removing time fixed effects. Moreover, in nonlinear models such differencing is generally not possible and fixed effects need to be estimated, leading to the Neyman and Scott (1948) incidental parameters problem. However, despite the presence of fixed effects in both the linear outcome equation (3.3) and the nonlinear treatment equation (3.2), we have that equation (3.4) is virtually the same as the expression for $\psi(X)$ in equation (3.1). As a result, no differencing or incidental parameters estimation is required. The estimator for panel data, corresponding to equation (3.4) in Theorem 3.2, is essentially identical to the cross section estimator $\widehat{\psi}(x)$ based on Theorem 3.1. We show later that the panel estimator does not experience the slow convergence rates associated with incidental parameters problems.

The intuition for this result is that the same density weighting that eliminates the confounding effects of U in the cross section also happens to remove the nonlinear treatment equation fixed effects a_i and b_t, while the differencing of the two terms that appear in equation (3.4) eliminates the outcome equation fixed effects $\widetilde{a}_i$ and $\widetilde{b}_t$.

As in the cross section case, estimation based on equation (3.4) simply replaces f_{v_t} with a kernel estimator of this density, and replaces the expectations with averages, or nonparametric regressions if elements of X_{it} are continuous. If the distribution of V varies by time then the density of f_{v_t} must be estimated separately in each time period, but averaging or nonparametric regressions is done across all individuals in all time periods. No differencing or other techniques for removing the fixed effects are required.

Identification and estimation based on more general panel models is possible. We present one such extension, allowing for dynamic effects, in Appendix B.

3.4 Asymptotic Normality

Our identification theorems permit fixed trimming, indexed by $I_{\tau i}$ in the cross section and $I_{\tau it}$ in the panel. This trimming allows our limiting distribution derivation to follow standard arguments like those in Newey and McFadden (1994), avoiding the complications associated with kernel estimator bias when V is near the boundary of its support. As a result, we can estimate $\psi(X)$ at the standard nonparametric rate associated with the dimension of X. As noted briefly in Lewbel (2000) and discussed more thoroughly in Khan and Tamer (2010), without fixed trimming obtaining standard convergence rates with inverse density weighted estimators like ours would generally require V to have very thick tails. Our fixed trimming avoids these issues.

For determining limit distributions, we put standard assumptions regarding kernels, bandwidths and smoothness, as well as detailed proofs, in Appendix C. Assumptions that require some discussion are kept in the main text below.

3.4.1 Cross Section Asymptotics

We first derive properties for the cross section version of our estimator. Let x be an interior point in the support of X. Define

$$h_{1i} \equiv \frac{D_i I_{\tau i} Y_i}{f(v_i|x_i)},\ g_{1i} \equiv \frac{D_i I_{\tau i}}{f(v_i|x_i)},\ h_{2i} \equiv \frac{(1-D_i) I_{\tau i} Y_i}{f(v_i|x_i)},\ g_{2i} \equiv \frac{(1-D_i)\, I_{\tau i}}{f(v_i|x_i)},\ \psi_1(x) \equiv \frac{E\left(h_{1i}|\, x\right)}{E\left(g_{1i}|\, x\right)},\ \psi_2(x) \equiv \frac{E\left(h_{2i}|\, x\right)}{E\left(g_{2i}|\, x\right)}$$

From the proof of Theorem 3.1, $\psi_1(x) = E(Y_1|x)$ and $\psi_2(x) = E(Y_0|x)$. We let the sample counterpart estimator of $\psi(x) = \psi_1(x) - \psi_2(x)$ be

$$\widehat{\psi}_1(x) - \widehat{\psi}_2(x) = \frac{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{D_i I_{\tau i} Y_i}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{D_i I_{\tau i}}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)} - \frac{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{(1-D_i) I_{\tau i} Y_i}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n} \frac{(1-D_i) I_{\tau i}}{\widehat{f}(v_i|x_i)} K\left(\frac{x_i - x}{h}\right)}, \tag{3.5}$$

where $\widehat{f}(v_i|x_i) = \widehat{f}_{xv}(x_i, v_i)/\widehat{f}_x(x_i)$ with $\widehat{f}_x(x_i)$ and $\widehat{f}_{xv}(x_i, v_i)$ being the standard leave-one-out nonparametric density estimators

$$\begin{aligned} \widehat{f}_x(x_i) &= \frac{1}{nh^k} \sum_{l=1, l\neq i}^{n} K\left(\frac{x_l - x_i}{h}\right), \\ \widehat{f}_{xv}(x_i, v_i) &= \frac{1}{nh^{k+1}} \sum_{l=1, l\neq i}^{n} K\left(\frac{x_l - x_i}{h}, \frac{v_l - v_i}{h}\right), \end{aligned}$$

where K is a kernel function and h is the bandwidth. For notational convenience, we let h be the same for all covariates.

Assumptions 9.1, 9.2, 9.3 and 9.4 provided in Appendix C, are all standard. Given these assumptions, the asymptotic normality of estimator (3.5) is established as follows.

Theorem 3.3 *Let Assumption 3.1, 3.2, 3.3, 3.4, 9.1, 9.2, 9.3 and 9.4 hold. As $n \to \infty$, let $h \to 0$, $nh^k \to \infty$, and $nh^{k+2p} \to c_0 \in (0, +\infty)$. Then for any point x in the interior of the support of X, we have*

$$\frac{\sqrt{nh^k}}{var\left(q_i\left(x\right)|x\right) \int_{\mathbb{R}^k} K^2\left(u\right) du} \left[\widehat{\psi}_1(x) - \widehat{\psi}_2(x) - E\left(Y_1 - Y_0|x\right) - \mathbb{B}_p\left(x\right)\right] \xrightarrow{d} N\left(0, 1\right),$$

where $q_i\left(x\right)$ and $\mathbb{B}_p\left(x\right)$ are defined in equation (10.6) and (10.7), respectively, in the supplemental Appendix.

The proof is in the supplemental online appendix.

Remark 3.4 In the usual way, the bias term $\mathbb{B}_p\left(x\right)$ becomes asymptotically irrelevant if $h \to 0$ faster than the root mean square minimizing optimal rate. Given Theorem 3.3, the unconditional average treatment effect $E\left(Y_1 - Y_0\right)$ could be estimated as $\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\psi}_1(x_i) - \widehat{\psi}_2(x_i)\right]$. It is generally possible to attain parametric convergence rates for an estimator like this (averages of smooth functions of kernel estimated densities and regressions), though doing so requires dealing with standard boundary bias issues for values of x near the boundary of its support. One method for doing so would be to use boundary bias corrections as in Hickman and Hubbard (2014). Another approach is to employ asymptotic trimming as in Robinson (1988) or Hardle and Stoker (1989).

3.4.2 Panel Data Asymptotics

The panel version of our estimator is essentially identical to averaging our cross section estimator across multiple time periods, because, as noted in the proof of Theorem 3.2, the estimator automatically accounts for fixed effects. Deriving the asymptotic properties of the panel estimator is therefore relatively straightforward but tedious. The main difference from the cross section case comes from allowing the distribution of V to vary over time. However, it is also necessary to keep track of the fixed effects, since they can affect the limiting distribution of the estimator.

To simplify the analysis and to focus on the new issues raised by panel data, assume we have no covariates X. This will be the case for our empirical application. Equations (3.2) and (3.3) then simplify to

$$Y_{it} = a_i + b_t + Y_{0it} + (Y_{1it} - Y_{0it})\, D_{it}, \tag{3.6}$$

$$D_{it} = I\left[0 \leq \widetilde{a}_i + \widetilde{b}_t + V_{it} + U_{it} \leq \alpha\right], \tag{3.7}$$

where $i = 1, 2, ..., n$, $t = 1, 2, ..., T$, and α is an unknown constant. The sample counterpart we estimate is then

$$\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}I_{\tau it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}I_{\tau it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})I_{\tau it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})I_{\tau it}}{\widehat{f}_{v_t}(v_{it})}}. \tag{3.8}$$

If we did have covariates X_{it}, the estimator would then be analogous to equation (3.5) with summation over both i and t, and we would need to combine the asymptotics we do here with those of the previous section.

We consider asymptotics where n goes to infinity faster than T, and obtain a convergence rate of $\sqrt{nT}$. Define ε_{jit} by $Y_{jit} = E(Y_j) + \varepsilon_{jit}$ for $j = 0, 1$, where $E(\varepsilon_{jit}) = 0$. Define

$$\Lambda_{1it} \equiv \frac{\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it}I_{\tau it} - E\left[\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right) D_{it}I_{\tau it}\middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Lambda_{2it} \equiv \frac{\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right)(1 - D_{it})\, I_{\tau it} - E\left[\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right)(1 - D_{it})\, I_{\tau it}\middle| v_{it}\right]}{f_{v_t}(v_{it})},$$

$$\Pi_{1it} \equiv \frac{D_{it}I_{\tau it}}{f_{v_t}(v_{it})},\ \overline{\Pi}_1 \equiv E\left(\frac{D_{it}I_{\tau it}}{f_{v_t}(v_{it})}\right),\ \Pi_{2it} \equiv \frac{(1 - D_{it})\, I_{\tau it}}{f_{v_t}(v_{it})},\ \overline{\Pi}_2 \equiv E\left(\frac{(1 - D_{it})I_{\tau it}}{f_{v_t}(v_{it})}\right).$$

Assumption 3.7 *$n \to \infty, T \to \infty$, and $T = o\left(n^{1-c_T}\right)$, for some $c_T \in (0, 1)$.*

Because $\sqrt{n}$ convergence of $\widehat{f}_{v_t}$ is not attainable, we need $T = o\left(n^{1-c_T}\right)$ to attain the convergence rate $\left(\widehat{f}_{v_t}(v) - f_{v_t}(v)\right)^2 = o_p\left((nT)^{-1/2}\right)$ with appropriate choice of bandwidth and kernel function.

Assumption 3.8 *$a_i, \widetilde{a}_i$ are i.i.d. across i and $b_t, \widetilde{b}_t$ are i.i.d. across t. (Y_{0it}, Y_{1it}) are identically distributed across i, t. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{i't'}, Y_{0i't'}, Y_{1i't'})$ for any $i \neq i'$, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{it'}, Y_{0it'}, Y_{1it'})\,|a_i, \widetilde{a}_i$ for any i, $t \neq t'$. $(U_{it}, Y_{0it}, Y_{1it}) \perp (U_{i't}, Y_{0i't}, Y_{1i't})|\, b_t, \widetilde{b}_t$ for any t, $i \neq i'$.*

The assumption that (Y_{0it}, Y_{1it}) is identically distributed over t as well as over i for each t is made only for convenience, and could be relaxed at the expense of additional notation that would include redefining the estimand to be the average value over time of $E(Y_1 - Y_0|t)$. We could allow heterogeneity (non-identical

distributions) over the time dimension for other variables as well, but we do exploit the i.i.d. assumption across i, conditional on t. These i.i.d. assumptions could also be relaxed to allow for weak dependence, at the cost of requiring more notation and a more general central limit theorem. Variables with the same i or the same t subscript are correlated with each other through individual or time dummies.

In Assumption 3.8, we define $a_i, \widetilde{a}_i, b_t, \widetilde{b}_t$ as random variables, but we estimate the model treating them as one would handle fixed effects, without estimating their values or their distributions and without imposing the kinds of assumptions that would be required for random effects estimation. For example, a_i and b_t are allowed to be correlated with U_{it} and Y_{it} in arbitrary unknown ways.

Remark 3.5 Although they are not estimated, $\widetilde{a}_i$ and $\widetilde{b}_t$ affect our limiting distribution, because the weights on these variables in the first and second components of our estimator are not identical in finite samples. In Lemma 10.7 in the online supplemental appendix, we show that the difference in these components due to $\widetilde{a}_i$ and $\widetilde{b}_t$ is $O_P\left((nT)^{-1/2}\right)$.

Assumption 3.9 *V_{it} are independent across i and t. V_{it} are identically distributed across i given t, with density $f_{v_t}(V_{it})$.*

For each time period t, Assumption 3.9 is equivalent to the cross section special regressor assumption without X. In addition it is assumed that special regressor observations are independent over time, but the distribution of V_{it} is allowed to vary with t. This independence assumption could be relaxed, and it would even be possible to let V_{it} be fixed over time for each i, though this would require dropping the cross section fixed effects from the model.

Assumption 3.10 $E\left(\varepsilon_{0it}|a_i, \widetilde{a}_i\right) = E\left(\varepsilon_{1it}|a_i, \widetilde{a}_i\right)$ *and* $E\left(\varepsilon_{0it}|b_t, \widetilde{b}_t\right) = E\left(\varepsilon_{1it}|b_t, \widetilde{b}_t\right)$.

This condition does not conflict with either interpretation of our estimand $E\left(Y_{1it} - Y_{0it}\right)$ given in the introduction. In particular, it is consistent with the possible assumption that the potential outcomes each have the same fixed effects.

Remark 3.6 Assumption 3.10 is necessary to attain $\sqrt{nT}$-convergence. To see why the assumption is necessary, suppose we could observe the counterfactuals Y_{1it} and Y_{0it}. Then the direct estimator for $E\left(Y_1\right) - E\left(Y_0\right)$ would just be $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Y_{1it} - Y_{0it}\right)$. The random component for this estimator is $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it} - \varepsilon_{0it}\right)$, which is equal to

$$\begin{aligned}&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it} - \varepsilon_{0it} - E\left(\varepsilon_{1it} - \varepsilon_{0it}|a_i, \widetilde{a}_i\right) - E\left(\varepsilon_{1it} - \varepsilon_{0it}|b_t, \widetilde{b}_t\right)\right) \\ &+\frac{1}{n}\sum_{i=1}^{n}E\left(\varepsilon_{1it} - \varepsilon_{0it}|a_i, \widetilde{a}_i\right) + \frac{1}{T}\sum_{t=1}^{T}E\left(\varepsilon_{1it} - \varepsilon_{0it}|b_t, \widetilde{b}_t\right).\end{aligned} \tag{3.9}$$

The first term is $O_P\left((nT)^{-1/2}\right)$, the second term is $O_P\left(n^{-1/2}\right)$, and the third term is $O_P\left(T^{-1/2}\right)$. So the convergence rate of this estimator is $O_P\left(T^{-1/2}\right)$ if $E\left[E\left(\varepsilon_{1it} - \varepsilon_{0it}|b_t, \widetilde{b}_t\right)^2\right] > 0$. So even in the infeasible case where counterfactuals are observable, Assumption 3.10 would be necessary to obtain $\sqrt{nT}$-convergence instead of rate $\sqrt{T}$. All currents results hold in the fixed T case, the case commonly seen in micro-economic data, except that the convergence rate is $\sqrt{n}$. That is because the technical lemmas for this

section, specifically Lemmas 10.7, 10.8, and 10.9 continue to hold in the fixed T case (only the convergence rate is $\sqrt{n}$ instead of $\sqrt{nT}$). The intuition can be seen from equation (3.9) that the first term is $O_P\left(n^{-1/2}\right)$ in the fixed T case. Thus our estimator is practical for short panels.

Additional Assumptions 9.3 and 9.5 provided in the Appendix are standard. Given these assumptions, the rate $\sqrt{nT}$ asymptotic normality of estimator (3.8) is established as follows.

Theorem 3.4 *Let Assumption 3.1, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 9.3, 9.5 hold. Assume that bandwidth* $h = c_0 n^{-c_T/2}$ *in* $\widehat{f}_{v_t}$*, and assume a kernel of order* $p \geq (1 - c_T/2)/c_T$*. Then*

$$\begin{aligned} & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\ = \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}\right) + o_P\left((nT)^{-1/2}\right), \end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}\right) = O_p\left((nT)^{-1/2}\right).$

Remark 3.7 Suppose $\left(a_i, \widetilde{a}_i, b_t, \widetilde{b}_t\right)$ is a series of constants instead of random variables. From the proof of Lemma 10.7, our estimator will still be consistent as long as $\frac{1}{n^2T}\left(\sum_{i=1}^{n}\widetilde{a}_i^2\right) = o(1)$ and $\frac{1}{nT^2}\left(\sum_{t=1}^{T}\widetilde{b}_t^2\right) = o(1)$. The estimator will also, given Assumption 3.10, still converge at rate $\sqrt{nT}$ with the same limiting distribution if $\frac{1}{n}\left(\sum_{i=1}^{n}\widetilde{a}_i^2\right) = O(1)$ and $\frac{1}{T}\left(\sum_{t=1}^{T}\widetilde{b}_t^2\right) = O(1)$. This result allows for limited forms of time trends of unknown form.

Remark 3.8 This theorem gives the influence function $\frac{\Lambda_{1it}}{\overline{\Pi}_1} - \frac{\Lambda_{2it}}{\overline{\Pi}_2}$ for our estimator. The terms in the influence function are identically distributed. From Lemmas 10.7, 10.8, and 10.9 in the supplemental appendix, these terms are dependent through the fixed effects. To establish asymptotic normality, we require that the influence function satisfy a central limit theorem, which means assuming that the dependence induced by these fixed effects is sufficiently weak. Such conditions could affect how $a_i, \widetilde{a}_i, b_t$, and $\widetilde{b}_t$ are drawn (relative to the other variables), and the relative rates of n and T.

Some additional results involving panel data asymptotics are provided in the Appendix B. In particular, we provide limiting distribution theory under some more general conditions, including if Assumption 3.10 does not hold, and a more general model of fixed effects.

4 Extensions

4.1 Testing the Large Support Assumption

The large support assumption on V is crucial for our identification. In this section, we provide a formal test of this assumption.

Suppose $\text{supp}(V) = [-m', m]$, where m' and m are positive constants, f_v is bounded away from zero, and the support of U is a fixed interval on the real line. For simplicity, assume there are no covariates X and that

$$D = I\left(0 \leq V + U \leq \alpha\right). \tag{4.1}$$

The large support assumption we require is that $\text{supp}(-U) \subseteq \text{supp}(V)$ and $\text{supp}(\alpha - U) \subseteq \text{supp}(V)$. Under the model specification and the assumption that the supports of U and V are both fixed intervals on the real line, this large support assumption holds if and only if $P\left(D = 1 | V = m\right) = 0$ and $P\left(D = 1 | V = -m'\right) = 0$. Without loss of generality, we describe testing the former of these two, involving m, with the test for the latter condition involving m' following analogously.

Let $m^* = \sup \text{supp}(\alpha - U)$. The required large support involving m is that $m \geq m^*$. Note that $m < m^*$ if and only if there exists $\varepsilon > 0$ such that $P(D = 1 | V = m) \geq \varepsilon$. Interestingly, if $m = m^*$ then the test statistic converges at rate$\sqrt{n}$ despite $P\left(D = 1 | V = m\right)$ being estimated nonparametrically (see Theorem 4.2 and Remark 4.3 below). However, if $m > m^*$ then the test statistic would degenerate to a constant zero. To avoid this discontinuous behavior, we compromise a bit and instead test

$$\mathbb{H}_0 : P\left(D = 1 | V = m\right) \geq \varepsilon^*,$$

where ε^* is a pre-determined small positive value instead of zero. This compromise means that we do not have power against certain alternatives, in particular, we could reject the null either because large support holds, or because $0 < P\left(D = 1 | V = m\right) < \varepsilon^*$.

With n observations, let $V_n^{(1)} = \max\left\{V_i, i = 1, ..., n\right\}$. Then by Lemma 10.13, $m - V_n^{(1)} = O_P\left(n^{-1}\right)$. We approximate m by $\widehat{m} = V_n^{(1)}$. Denote $G_D\left(v\right) \equiv P\left(D = 1 | V = v\right)$. Let $G'_{D,-}\left(m\right)$ and $G''_{D,-}\left(m\right)$ denote the left first and second derivatives at m respectively. Since we are interested in estimation at the boundary, we employ local linear regression, which has better boundary behavior than standard kernel regression (see, e.g., Fan and Gijbels 1992). This estimation of $G_D\left(v\right)$ is based on

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(I\left(D_i = 1\right) - \beta_0 - \beta_1 \left(\frac{V_i - v}{h} \right) \right) K_h\left(V_i - v\right),$$

where $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)^T$, $K_h\left(V_i - v\right) \equiv \frac{1}{h} K\left(\frac{V_i - v}{h}\right)$. Let $\widehat{\boldsymbol{\beta}}\left(v\right) = \left(\widehat{\beta}_0\left(v\right), \widehat{\beta}_1\left(v\right)\right)^T$ be the estimates from this minimization.

We are interested in estimating $G_D\left(v\right)$ at the boundary point $v = m$. Since we do not know m, we approximate it by $\widehat{m}$. Reorder the data so that $V_n = V_n^{(1)} = \widehat{m}$. We then employ the leave-one-out estimator $\widehat{G}_D\left(\widehat{m}\right) = e_1^T \widehat{\beta}\left(\widehat{m}\right)$ where $e_1 = \left(1, 0\right)^T$ and

$$\begin{aligned} \widehat{\boldsymbol{\beta}}\left(\widehat{m}\right) &= \left[\mathbf{S}_h\left(\widehat{m}\right)\right]^{-1} \frac{1}{n-1} \sum_{i=1}^{n-1} K_h\left(V_i - \widehat{m}\right) \begin{pmatrix} 1 \\ \left(V_i - \widehat{m}\right)/h \end{pmatrix} I\left(D_i = 1\right), \\ \mathbf{S}_h\left(\widehat{m}\right) &\equiv \frac{1}{n-1} \sum_{i=1}^{n-1} K_h\left(V_i - \widehat{m}\right) \begin{pmatrix} 1 \\ \left(V_i - \widehat{m}\right)/h \end{pmatrix} \left(1, \left(V_i - \widehat{m}\right)/h\right). \end{aligned} \tag{4.2}$$

Define $S_{j,-} \equiv \int_{-\infty}^{0} K\left(u\right) u^j du$ for any positive integer j, and $\overline{\mathbf{S}} \equiv \begin{pmatrix} S_{0,-} & S_{1,-} \\ S_{1,-} & S_{2,-} \end{pmatrix}$. Note that $\overline{\mathbf{S}} f_v\left(m\right)$ is the limit of $\mathbf{S}_h\left(\widehat{m}\right)$ in probability. We assume the following technical condition.

Assumption 4.1 *Suppose the model is given by equation (4.1). Assume that $U \perp V$, the supports of U and V are intervals on the real line, the density function f_v for V is continuous, bounded, and bounded away from zero on its support, the density function f_u and cdf F_u for U is continuous differentiable, and f_u and its first derivative are bounded on its support.*

Theorem 4.1 *Suppose Assumption 4.1 holds. Assume the kernel function K satisfies Assumption 9.3. Assume i.i.d. observations. Then under $\mathbb{H}_0$,*

$$\sqrt{nh}\left(\widehat{G}_D(\widehat{m}) - G_D(m) - \mathbb{B}_h\right) \xrightarrow{d} N\left(0, \sigma^2(m)\right),$$

where $\mathbb{B}_h \equiv e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) h^2$, $\sigma^2(m) \equiv e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(m)(1 - G_D(m)) f_v(m)^{-1}$, $\mathbf{Q} \equiv \begin{pmatrix} Q_{0,-} & Q_{1,-} \\ Q_{1,-} & Q_{2,-} \end{pmatrix}$, $Q_{j,-} \equiv \int_{-\infty}^{0} K^2(u) u^j du$.

This theorem is proved in the supplemental appendix.

By Theorem 4.1, we can test $\mathbb{H}_0$ using a standard z-test. The P-value is calculated as $P = \Phi\left(\frac{\widehat{G}_D(\widehat{m}) - \widehat{\mathbb{B}}_h - \varepsilon^*}{\widehat{\sigma}(\widehat{m})}\right)$, where $\widehat{\mathbb{B}}_h, \widehat{\sigma}(\widehat{m})$ can be obtained as in Remark 4.1. We use $\varepsilon^* = 0.05$ in our empirical application.

Remark 4.1 Given the derivation of our test, the optimal bandwidth for estimation of $\widehat{G}_D(\widehat{m})$ is

$$h_{\text{opt}} = n^{-1/5} \left[\left(e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(m)(1 - G_D(m)) f_v(m)^{-1} \right) \Big/ \left(e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) \right)^2 \right]^{1/5}.$$

To get $\widehat{\mathbb{B}}_h$, $\widehat{\sigma}^2(m)$ and h_{opt}, one could estimate $G''_{D,-}(m)$ using local quadratic estimation, and estimate $f_v(m)$ using a kernel estimator with boundary correction as in, e.g., Hardle (1990)[3]. Fan and Gijbels (1992) discuss the good properties and performance of this type of plug-in estimator.

Remark 4.2 The condition in Assumption 4.1 that $f_v(m) > 0$ is used to obtain the asymptotic normality in Theorem 4.1. If this condition does not hold, then the convergence rate of $\widehat{m}$ may be affected (for details, see Lemma 10.13). In our data, the density of V appears to approach zero in the tails of its support. We therefore propose applying our test on a truncated sample, that is, we trim out the top and bottom 1% of our data, thereby making Assumption 4.1 more likely to hold. Having the support of this truncated V sample cover the corresponding support of U or $\alpha - U$ implies that the original large support assumption must hold, so this trimming makes the test stronger than necessary.

We now consider the asymptotic behavior of $\widehat{G}_D(\widehat{m})$ when the support of V covers the support of $\alpha - U$ on the right end.

Theorem 4.2 *Suppose Assumption 4.1 holds. Assume the kernel function K satisfies Assumption 9.3. Assume i.i.d. observations. Assume the support of V covers the support of $\alpha - U$ on the right, and that $h = c_0 n^{-2/5}$, for some $c_0 > 0$. Then*

$$\sqrt{n}\left(\widehat{G}_D(\widehat{m}) - G_D(m)\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2(m)\right),$$

[3]Here we use the modified kernel function: $K(x) \Big/ \int_{-\infty}^{0} K(x)\, dx$.

where $\widetilde{\sigma}^2(m) \equiv e_1^T \overline{\mathbf{S}}^{-1} \widetilde{\mathbf{Q}} \overline{\mathbf{S}}^{-1} e_1 G'_{D,-}(m) f_v(m)^{-1}$, $\widetilde{\mathbf{Q}} \equiv \begin{pmatrix} Q_{1,-} & Q_{2,-} \\ Q_{2,-} & Q_{3,-} \end{pmatrix}$, $Q_{j,-} \equiv \int_{-\infty}^{0} K^2(u) u^j du$.

This theorem is proved in the supplemental appendix.

Remark 4.3 In the case where the support of V strictly covers the support of U, $G_D(v) = G'_{D,-}(v) = 0$ in an interval around the boundary point. In this case, according to the above theorem, $\sigma^2(m) = 0$. The estimates $\widehat{G}_D(\widehat{m})$ will then degenerate to zero in the limit.

4.2 General Ordered Choice Identification

We now extend our results to a general ordered choice model of treatment. We now have J possible treatments, with an individual receiving treatment j, for $j = 0, 1, ...J-1$, if $D_j = 1$. The model is now

$$D_j = I[\alpha_{j-1}(X) \leq V + U < \alpha_j(X)] \tag{4.3}$$

$$Y = \sum\nolimits_{j=0}^{J-1} = D_j W_j \tag{4.4}$$

where $\alpha_{-1}(X) = -\infty$ and $\alpha_{J-1}(X) = \infty$. Here W_j denotes the potential outcome from receiving treatment j, which occurs when $D_j = 1$. Equations (1.3) and (1.4) in the introduction are the special case of this model where $J = 3$. Our main model given by Equations (1.1) and (1.2) in the introduction corresponds to this model with $J = 3$, $D = D_1$, $Y_1 = W_1$, and $Y_0 = D_0 W_0 + D_2 W_2$.

In the model of equations (4.3) and (4.4), identification and estimation of $E(W_j \mid X)$ for middle choices $j = 1, ..., J-2$ is identical to identification and estimation of $E(Y_j \mid X)$, replacing D with D_j. What is new in this section is identification of the separate extreme potential outcomes, $E(W_0 \mid X)$ and $E(W_{J-1} \mid X)$, instead of just the combined outcome $E(Y_0 \mid X)$. As noted by Heckman, Urzua and Vytlacil (2006), without invoking functional form assumptions, identification of $E(W_0 \mid X)$ and $E(W_{J-1} \mid X)$ requires $E(D_0 \mid X, V) \to 1$ and $E(D_{J-1} \mid X, V) \to 1$ for limiting values of one or more covariates. In our case we use this identification at infinity technique to identify $E(W_0 \mid X)$ taking the limit of $E(D_0 \mid X, V)$ as V gets sufficiently small, and identifying $E(W_{J-1} \mid X)$ taking the limit of $E(D_{J-1} \mid X, V)$ as V gets sufficiently large.

This extension permits estimation for a wider set of applications than before, since one is often interested in treatment effects like $E(W_1 - W_0)$ rather than $E(Y_1 - Y_0)$. However, this extension has the disadvantage that estimation of the separate effects of the highest and lowest treatment categories, $E(W_0)$ and $E(W_{J-1})$, requires identification at infinity, and so will result in slower than parametric convergence rates (for middle treatment categories like $E(W_1)$, identification at infinity is not possible). In contrast, with equations (1.1) and (1.2) we obtain parametric convergence rates for both $E(Y_0)$ and $E(Y_1)$.

For the general ordered choice model we impose the following assumptions, which are very similar to our earlier assumptions.

Assumption 4.2 *We observe realizations of* Y *and* D_j *for* $j = 0, 1, ..., J-1$*, a covariate* V*, and a* $k \times 1$ *covariate vector* X*. Assume the treatment indicators* D_j *and outcome* Y *are given by equations (4.3) and (4.4) respectively, where each* $\alpha_j(X)$ *is an unknown threshold function with* $\alpha_{j-1}(X) < \alpha_j(X)$*,* $\alpha_{-1}(X) = -\infty$*,* $\alpha_{J-1}(X) = \infty$*,* U *is an unobserved latent random error, and* W_j *for* $j = 0, 1, ..., J-1$ *are unobserved random potential outcomes, each corresponding to treatment* $D_j = 1$*. The joint distribution of* $(U, W_0, W_1, ...W_{J-1})$*, either unconditional or conditional on* X*, is unknown.*

Assumption 4.3 *Assume* $V \perp (U, W_0, W_1, ...W_{J-1}) \mid X$. *For all* $x \in supp\,(X)$, *the* $supp(V \mid X = x)$ *covers the* $supp\,(\alpha_{J-2}\,(X) - U \mid X = x)$ *on the right and covers* $supp(\alpha_0\,(X) - U \mid X = x)$ *on the left.*

Theorem 4.3 (**Identification**) *Suppose Assumptions 4.2 and 4.3 hold. Assume* $\{\gamma_n\,(X)\}_{n=1}^{\infty}$ *and* $\{\gamma'_n\,(X)\}_{n=1}^{\infty}$ *are increasing series such that* $\lim_{n\to\infty} E\,(D_0 \mid X, V \leq -\gamma_n\,(X)) = 1$ *and* $\lim_{n\to\infty} E\,(D_{J-1} \mid X, V \geq \gamma'_n\,(X)) = 1$. *Then*

$$E\,(W_0 \mid X) = \lim_{n\to\infty} E\,(D_0 Y \mid X, V \leq -\gamma_n\,(X))\,, \quad E\,(W_{J-1} \mid X) = \lim_{n\to\infty} E\,(D_{J-1} Y \mid X, V \geq \gamma'_n\,(X))\,,$$

$$\textit{and} \quad E\,(W_j \mid X) = \frac{E\,[D_j Y / f(V \mid X) \mid X]}{E\,[D_j / f(V \mid X) \mid X]} \quad \textit{for } j = 1, ..., J-2.$$

This theorem is proved in the supplemental Appendix. The tuning parameters $\gamma_n\,(X)$ and $\gamma'_n\,(X)$ determine the set of V values that we average over as the sample size grows. The intuition of this identification at infinity is that the larger in magnitude are $\gamma_n\,(X)$ and $\gamma'_n\,(X)$, the more extreme are the values of V that we average over, and hence the lower is the probability that the confounder U affects D. Eventually, the effect of the confounder vanishes in the limit.

As before, corresponding estimators are immediately obtained by replacing $f(V \mid X)$ with an estimated density, and replacing expectations with nonparametric regressions. When there are no covariates X, these nonparametric regressions further simplify to just equaling sample averages. These estimators come with some caveats that do not apply to our main identification theorem. Due to only being identified at infinity, the estimators $\widehat{E}\,(W_0)$ and $\widehat{E}\,(W_{J-1})$ converge slower than the parametric rate, and these estimates may be sensitive to the choice of tuning parameters $\gamma_n\,(X)$ and $\gamma'_n\,(X)$. In contrast, our main identification results converged at the parametric rates, and did not require associated tuning parameters $\gamma_n\,(X)$ and $\gamma'_n\,(X)$.

5 Competition and Innovation

We apply our model to test the "Inverted-U" theory of ABBGH (Aghion, Bloom, Blundell, Griffith, and Howitt 2005) relating innovation investments to competitiveness in an industry. ABBGH consider two types of oligopoly industries, called Neck-and-Neck (NN) industries, in which firms are technologically close to equal, and Leader-Laggard (LL) industries, where one firm is technologically ahead of others. For these industries there are two opposing effects of competition on innovation. One is the *Schumpeterian effect*, where increased competition reduces profits and thus reduces the incentive to innovate. The second is the *escape-competition effect,* where firms innovate to increase the profits associated with being a leader. For these latter firms, increased competition increases the incentive to innovate. ABBGH argue that the escape-competition effect dominates in NN industries while the Schumpeterian effect dominates in LL industries. This theory results in an inverted-U relationship, because low levels of competition are associated with NN industries and hence with low innovation, by the escape-competition effect, and high levels of competition are associated with LL industries, again leading to low innovation but now by the Schumpeterian effect. In contrast, with an intermediate level of competition, both NN and LL industries innovate to some extent, yielding a higher overall level of innovation in steady state than in either the low or high competition industries.

ABBGH find empirical support for the inverted-U based mainly on UK data. Hashmi (2013) revisits the relationship using a dataset from the US, and finds no inverted-U. Hashmi notes that his finding can be reconciled with the ABBGH model by the assumption that the manufacturing industries in the UK are, on the average, more neck and neck than their counterparts in the US.

For identification and estimation, both the ABBGH and Hashmi empirical results depend heavily on functional form assumptions, by fully parameterizing both the relationship of competitiveness to innovation and the functional form of error distributions. In contrast, we apply our model to test for an inverted-U relationship with minimal restrictions on error distributions.

5.1 Data

Our sample, from Hashmi (2013), consists of US three-digit level industry annual data from 1976 to 2001. There are 116 industries, resulting in 2716 industry-year observations. The analysis is based on three key variables: a measure of industry competitiveness, a measure of industry innovation, and an instrument. ABBGH and Hashmi require a detailed measure of competitiveness, while we only need to classify industries as either moderately competitive or not (corresponding to D equal to one or zero). Hashmi uses an average of industry exchange rates for an instrument, which will serve as our continuous special regressor V. Summary statistics for this data are reported in Table 1. We only applied our estimator to Hashmi's data and not to ABBGH's data, because the latter uses a discrete rather than a continuous instrument.

Hashmi's measure of the level of competition for industry i at time t, denoted c_{it}, is defined by

$$c_{it} = 1 - \frac{1}{n_{it}} \sum\nolimits_{j=1}^{n_{it}} l_{jt}, \tag{5.1}$$

where i indexes firms, l_{jt} is the Lerner index of the price-cost margin of firm j in year t, and n_{it} is the number of firms in industry i in year t. The higher c_{it} is, the higher is the level of competition. The innovation index, denoted y_{it}, is a measure of citation-weighted patent counts, constructed using data from the NBER Patent Data Project. Details regarding the construction of this data can be found in Hashmi (2013).

As ABBGH point out, innovation and competition are endogenous, that is, there are likely to exist unobserved characteristics of each industry i in each time period t that can affect both. To deal with this endogeneity, Hashmi uses a source-weighted average of industry exchange rates as an instrument for competition (ABBGH use a different, events related instrument). Hashmi's instrument, V_{it}, is a weighted average of the US dollar exchange rate with the currencies of trading partners, with weights that vary by industry according to the share of each country in the imports to the US. This instrument has been used in other similar applications, including Revenga (1990, 1992) and Bertrand (2004).

5.2 Model Specifications

Hashmi (2013) adopts a control function approach to deal with endogeneity. In a first stage, c_{it} is regressed on V_{it}, industry dummies and time dummies, so

$$c_{it} = V_{it}\beta + a_i + b_t + w_{it}, \tag{5.2}$$

where a_i and b_t are fixed effects (coefficients of industry and time dummies) and w_{it} is the error from the first stage regression. The fitted residuals $\widehat{w}_{it}$ from this regression are then included as additional regressors in an outcome equation of the form

$$\ln(y_{it}) = \widetilde{a}_i + \widetilde{b}_t + \theta_0 + \theta_1 c_{it} + \theta_2 c_{it}^2 + \delta \widehat{w}_{it} + \varepsilon_{it}, \tag{5.3}$$

where $\widetilde{a}_i$ and $\widetilde{b}_t$ are outcome equation fixed effects (coefficients of industry and time dummies). Hashmi estimates the coefficients in equation (5.3) by maximum likelihood, where the distribution of errors ε_{it} is

determined by assuming that y_{it} has a negative binomial distribution, conditional on c_{it}, industry, and year dummies. This model assumes the relationship of $\ln(y)$ to c is quadratic, with an inverted-U shape if θ_1 is positive and θ_2 negative. The industry and time dummies cannot be differenced out in this model, and so are estimated along with the other parameters.

Possible problems with this model are that the quadratic could be a misspecification, or the endogeneity could take a form that is not completely eliminated by the control function addition of $\widehat{w}$ as a regressor, or the distribution might not be negative binomial. In addition, Hashmi's estimates could suffer the from the incidental parameters problem of Neyman and Scott (1948). This problem here is that the need to estimate industry and time fixed effects results in inconsistent parameter estimates unless both T and n go to infinity. In this application neither T nor n is particularly small, but the presence of the fixed effects still results in over 100 nuisance parameters to estimate, which can lead to imprecision. Our intention is not to criticize Hashmi's or ABBGH's model, but only to point out that there are many reasons why it is desirable to provide a less parametric alternative, to verify that their results are not due to potential model specification or estimation problems.

To apply our estimator, we do not assume that the true level of competitiveness is observed, and instead treat c_{it} as just a rough, possibly mismeasured proxy for competitiveness. We only use c_{it} to divide observations into two groups. Those with moderate levels of c_{it} are assigned $D_{it} = 1$, while the rest are labeled $D_{it} = 0$ (in a later extension we will consider three groups: low, medium and high). To the extent that c_{it} mismeasures competitiveness, even by a substantial amount, a few industries near the boundaries defining these groups may be misclassified. Differences between the ranking of c_{it} versus true competitiveness can only affect D_{it} for the small number industries that happen to be near the c_{it} cutoffs used to define treatment. We show in an appendix that even large differences between c_{it} and true competitiveness results in little bias in our estimator, and far less bias than in Hashmi's model.

Given the treatment indicator D_{it}, we assume observed innovation y_{it} is determined by

$$y_{it} = \widetilde{a}_i + \widetilde{b}_t + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}. \tag{5.4}$$

where $\widetilde{a}_i$ and $\widetilde{b}_t$ are the industry and time dummies respectively, and as discussed in the introduction, Y_{0it}, Y_{1it} are unobserved potential outcomes for $y_{it} - \widetilde{a}_i - \widetilde{b}_t$, that is, potential outcomes for industry i in time t after removing time and industry fixed effects. Unlike the error distribution imposed in equation (5.3), both Y_{1it} and Y_{0it} here are random variables with completely unknown distributions that can be correlated with each other, and with the error term in the D_{it} equation, in completely unknown ways. We will then estimate the ATE $E(Y_{1it} - Y_{0it})$, which equals the average difference in outcomes y (after controlling for fixed effects), between industries with moderate levels of competitiveness, versus industries that have very low or very high levels of competitiveness.

What our model assumes about the treatment indicator D_{it} is

$$D_{it} = I(\alpha_0 \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1), \tag{5.5}$$

where a_i and b_t are industry and time dummies, U_{it} are unobserved, unknown factors that affect competition, and α_0 and α_1 are unknown constants. The way to interpret equation (5.5) is that the latent variable c^*_{it} given by

$$c^*_{it} = a_i + b_t + V_{it} + U_{it} \tag{5.6}$$

is some unobserved true level of competitiveness of industry i in time t. Our model does not require the

observed competitiveness measure c_{it} to equal the true measure c_{it}^*, but if they do happen to be equal then our model is consistent with having Hashmi's equation (5.2) hold. Note when comparing the models for c_{it}^* and c_{it} to each other that replacing c_{it}^* with βc_{it}^* to make equation (5.6) line up with equation (5.5) is a free scale normalization that can be made without loss of generality, because the definition of D_{it} is unaffected by rescaling c_{it}^*.[4]

As in Hashmi's model, our estimator assumes that V_{it} is a valid instrument, affecting competitiveness c_{it}^* and hence the treatment indicator D_{it}, but not directly affecting the outcome y_{it}. We also require that V_{it} has a large support. This appears to be the case in our data, e.g., the exchange rate measure sometimes as much as doubles or halves over time even within a single industry, and varies substantially across industries as well. We later test and do not reject the assumption of large support.

5.3 Measurement Errors in Competitiveness

In our empirical application, we define D_{it} to be one when the observed c_{it} lies between the .25 and .75 quantiles of the empirical c_{it} distribution (we also experiment with other quantiles). This is therefore consistent with equation (5.2) if c_{it} is linear in c_{it}^*. However, our model remains consistent even if c_{it} differs greatly from c_{it}^*, as long as the middle 50% of industry and time periods in the c_{it} distribution corresponds to the middle 50% of industry and time periods in the c_{it}^* distribution.

More generally, suppose c_{it} equals c_{it}^* plus some measurement error. Then the Hashmi model, even if correctly specified, will be consistent only if this measurement error satisfies the conditions necessary for validity of their control function estimator. Some control function estimators remain consistent in models containing measurement errors that are classical, i.e., independent of the true c_{it}^* and of the true model. However, the Hashmi control function estimator would not be consistent even with classical measurement errors, because equation (5.3) is nonlinear in the potentially mismeasured variable c_{it} (this is not intended as a criticism of Hashmi's empirical application, since that work uses control functions only to deal with endogeneity and never made any claims regarding measurement errors). In contrast, our estimator can remain consistent in theory even with measurement errors that are large and nonclassical, as long as c_{it} correctly sorts industries into moderate versus non-moderate levels of competitiveness. However, in practice, measurement error in c_{it} will likely cause some industries to be misclassified, so D_{it} is likely to be mismeasured for a small number of industries that lie close to the .25 and .75 quantile cutoffs.

To summarize: competitiveness is difficult to precisely define and measure, and as a result the impact of measurement errors on this analysis could be large. One advantage of our methodology is that it only depends on sorting industries into two groups (that is, moderate versus extreme levels of competitiveness as indicated by D_{it}). This sorting greatly mitigates measurement error biases, because only a small number of observations of D_{it} are likely to be mismeasured even if most or all of the c_{it} observations are mismeasured to some extent. To verify that this intuition is correct, in an appendix we do a monte carlo analysis that compares the accuracy of our estimator with that of Hashmi's in the presence of measurement errors.

[4] As noted earlier, it is very unlikely that c_{it} perfectly measures true competitiveness in each industry and time period. However, if c_{it} is not mismeasured, then the thresholds used to construct D_{it} from c_{it} would be proportional, up to the scaling of the coefficient of V, to the unknown thresholds $\alpha_1(X)$ and $\alpha_0(X)$ (after accounting for unknown fixed effects a_i and b_t). In theory this information might be usable to increase estimation efficiency, by exploiting the fact that $E(D/f_v|X)$, which we estimate, equals $\alpha_1(X) - \alpha_0(X)$.

5.4 Estimation

Our estimator is quite easy to implement, in part because it does not entail any numerical searches or maximizations. We first estimate the density of V_{it} separately for each year, using a standard kernel density estimator $\widehat{f}_{v_t}(v_{it}) = \frac{1}{n-1}\sum_{j\neq i,,j=1}^{n}\frac{1}{h}K\left(\frac{v_{it}-v_{jt}}{h}\right)$. Note that the density is estimated at each of the data points v_{it}. We employ a Gaussian kernel function K, and choose the bandwidth h using Silverman's rule of thumb. Our estimator involves dividing by these nonparametric density estimates, which can result in outlier observations when $\widehat{f}$ is close to zero. As suggested in Lewbel (2000) and Dong and Lewbel (2015) for other special regressor based estimators, we trim out (i.e., discard from the sample) the 2% of observations with the smallest values of $\widehat{f}_{v_t}$. This defines the trimming function $I_\tau(v)$ from our asymptotic theory.

Given the density estimates $\widehat{f}_{v_t}(v_{it})$, our resulting estimate of the ATE $E(Y_{1it}-Y_{0it})$ is then given by

$$\text{Trim-ATE} = \frac{\sum_i\sum_t I_\tau(v_{it})D_{it}Y_{it}/\widehat{f}_{v_t}(v_{it})}{\sum_i\sum_t I_\tau(v_{it})D_{it}/\widehat{f}_{v_t}(v_{it})} - \frac{\sum_i\sum_t I_\tau(v_{it})(1-D_{it})Y_{it}/\widehat{f}_{v_t}(v_{it})}{\sum_i\sum_t I_\tau(v_{it})(1-D_{it})/\widehat{f}_{v_t}(v_{it})} \tag{5.7}$$

where the i and t sums are over the 98% of observations that were not trimmed out. This model corresponds to the estimator (3.8), which has standard errors that we calculate based on the asymptotic distribution provided in Theorem 3.4. To assess the effect of the trimming on this estimator, we construct a corresponding estimate of ATE that is not trimmed, given by

$$\text{No-Trim-ATE} = \frac{\sum_i\sum_t D_{it}Y_{it}/\widehat{f}_{v_t}(v_{it})}{\sum_i\sum_t D_{it}/\widehat{f}_{v_t}(v_{it})} - \frac{\sum_i\sum_t(1-D_{it})Y_{it}/\widehat{f}_{v_t}(v_{it})}{\sum_i\sum_t(1-D_{it})/\widehat{f}_{v_t}(v_{it})}. \tag{5.8}$$

For comparison, we also calculate a Naive-ATE estimator given by

$$\text{Naive-ATE} = \frac{\sum_i\sum_t D_{it}Y_{it}}{\sum_i\sum_t D_{it}} - \frac{\sum_i\sum_t(1-D_{it})Y_{it}}{\sum_i\sum_t(1-D_{it})}. \tag{5.9}$$

This Naive-ATE just subtracts the average value of Y_{it} when $D_{it}=0$ from the average value of Y_{it} when $D_{it}=1$. This would be a consistent estimator of the ATE if treatment were unconfounded, that is, if low or high competitiveness as indicated by D_{it} was randomly assigned over firms and time periods. One could also consider a LATE estimator such as an instrumental variables regression of Y on D using V as an instrument. However, as noted in the Introduction, LATE requires that the probability of treatment increase monotonically with the instrument. This requirement does not hold in our application, since both increasing or decreasing V sufficiently causes the probability of treatment to decrease.

We also compare our results to a parametric maximum likelihood estimate of the ATE (denoted ML-ATE) assuming a Heckman (1979) type selection model for treatment. This model assumes equations (5.4) and (5.5) hold and that U, Y_0, Y_1 are jointly normally distributed. Let Φ denote the standard normal cumulative distribution function, $\theta_0 = E(Y_0)$, $\theta_1 = E(Y_1)$, and $\Sigma = cov[U, Y_0, Y_1]$ be the three by three covariance

matrix of elements σ_{kl} for $k = 1, 2, 3$ and $l = 1, 2, 3$. Then the ML-ATE is defined by

$$\text{ML-ATE} = \widehat{\theta}_1 - \widehat{\theta}_0 \quad \text{where} \quad \left[\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\Sigma}\right] = \arg\max \sum_i \sum_t$$

$$\left\{ (1 - D_{it}) \log \left(\frac{1}{\sigma_{22}} \phi \left(\frac{Y_{it} - \theta_0}{\sigma_{22}} \right) \left[\Phi \left(\frac{\alpha_0 - v_{it} - \frac{\sigma_{12}}{\sigma_{22}} (Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) + 1 - \Phi \left(\frac{\alpha_1 - v_{it} - \frac{\sigma_{12}}{\sigma_{22}} (Y_{it} - \theta_0)}{\sqrt{\sigma_{11} - \sigma_{12}^2/\sigma_{22}}} \right) \right] \right) \right.$$

$$\left. + D_{it} \log \left(\frac{1}{\sigma_{33}} \phi \left(\frac{Y_{it} - \theta_1}{\sigma_{33}} \right) \left[\Phi \left(\frac{\alpha_1 - v_{it} - \frac{\sigma_{13}}{\sigma_{33}} (Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) - \Phi \left(\frac{\alpha_0 - v_{it} - \frac{\sigma_{13}}{\sigma_{33}} (Y_{it} - \theta_1)}{\sqrt{\sigma_{11} - \sigma_{13}^2/\sigma_{33}}} \right) \right] \right) \right\}.$$

5.5 Empirical Results

Figure 1 shows our kernel density estimates $\widehat{f}_{v_t}$ for each year t. The estimates can be seen to vary quite a bit over time, so we use separate density estimates for each year instead of assuming a constant distribution across years. Figure 2 shows a scatterplot of our competitiveness and innovation data. For illustration, a linear and a quadratic curve are fitted to the data using ordinary least squares regression. The line is slightly downward sloping while the fitted quadratic is only slightly U-shaped. Note that these fitted curves do not deal with the endogeneity issue.

Table 2A shows our main empirical results. The first row of Table 2A provides estimates where D_{it} is defined to equal one for the middle half of the data, that is, D_{it} equals one for firms and years that lie between the 25th and 75th percent quantiles of the observed measure of competition, making half the observations treated and the other half untreated. Other rows of Table 2A report results using different quantiles to define D_{it}. In each row of Table 2A we report four estimates of ATE, as described in the previous section. Standard errors for all the estimates are provided in parentheses.

An inverted-U would imply a positive ATE, but all of our estimates are negative, confirming Hashmi's finding that the inverted-U is not present in US data. For example, our main estimate from the first row of Table 2A is that the Trim-ATE equals -3.9, and is strongly statistically significant. We also find that failure to appropriately control for error correlations between competitiveness and innovation substantially biases the magnitudes of estimated treatment effects. Our semiparametric estimates of the ATE are 50% to 100% larger than both the naive estimates that ignore these correlations, and the maximum likelihood estimates that allow for correlations but require the errors to be jointly normally distributed.

Attempts to find a positive ATE by experimenting with more unusual quantiles for defining D_{it} were for the most part fruitless. An exception, based on examination of Figure 2, was to define the left and right thresholds by 0.62 (10%) and 0.68 (20%) respectively. This implies a heavily skewed inverted U where 80% of firms are in the upper tail. This yields a positive ATE of 8.66, but this model is implausible, since it treats a very narrow spike in Figure 2 as the set of all moderately sized firms. We also experimented with varying the degree of trimming, but we only report results without trimming and with 2% percent trimming because the impacts of other changes in trimming were very small.

The quantiles of c_{it} vary over time, so instead of defining D_{it} based on quantiles of the entire sample of c_{it} observations, one could instead define D_{it} for each year t based on the quantiles of c_{it} just in year t. As a robustness check, results are reported in Table 2B based on estimates calculating D_{it} this alternative way. Comparing Table 2A and 2B, shows that the results are quite similar using either definition.

Hashmi models the mean of innovation using equation (5.3), so the following object constructed from Hashmi's paper can be compared with our ATE estimates

$$E\left(\exp\left(\widetilde{a}_i + \widetilde{b}_t + \theta_0 + \theta_1 c_{it} + \theta_2 c_{it}^2 + \delta w_{it} \right) \middle| D_{it} = 1 \right) - E\left(\exp\left(\widetilde{a}_i + \widetilde{b}_t + \theta_0 + \theta_1 c_{it} + \theta_2 c_{it}^2 + \delta w_{it} \right) \middle| D_{it} = 0 \right).$$

We estimate this by replacing the expectations with D_{it} cell means, and using Hashmi's estimates for the parameters $\widetilde{a}_i$, $\widetilde{b}_t$, θ_0, θ_1, θ_2, and δ.[5] The value of this quantity we find from his model is -1.8, which is about half of our estimated ATE and similar to the ML-ATE and Naive-ATE. Again, we agree with Hashmi's main result regarding signs of effects, but not magnitudes. This discrepancy might come from misspecification of Hashmi's model, sensitivity to measurement error in c_{it} in his model, or imprecision in his estimates of $\widetilde{a}$ and $\widetilde{b}$ due to the incidental parameters problem.

5.6 Monte Carlo Designed for the Empirical Example

To assess how our main estimator works in small samples, we provide two sets of Monte Carlo experiments. We designed these experiments to closely match moments and other features of our empirical data, to see how likely our estimator is to perform well in a controlled setting that mimics our actual application. The number of observations is set to 2716, the same as the number of observations in our empirical dataset. The same four estimators we applied on the actual data, Trim-ATE, No-Trim-ATE, Naive-ATE and ML-ATE, are analyzed in each set of Monte Carlo simulations.

Let e_{1i}, e_{2i}, e_{3i}, and V_i be random variables that are drawn independently of each other. We consider a few different distributions for these variables as described below. The counterfactual outcomes in our simulation are defined by

$$Y_{0i} = \theta_0 + \theta_{01} e_{1i} + \theta_{02} e_{3i} \text{ and } \; Y_{1i} = \theta_1 + \theta_{11} e_{2i} + \theta_{12} e_{3i}.$$

True competitiveness is constructed to equal $V_i + \theta_2 e_{3i}$, and treatment D_i is defined to equal one for observations i that lie between the 25th and 75th quantile of the distribution of $V_i + \theta_2 e_{3i}$. The observed outcome is then constructed as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i.$$

For simplicity, fixed effect type dummies are omitted from the model. Note that e_{3i} appears in D_i, Y_{0i}, and Y_{1i}, and so is the source of confounding in this specification. By construction, the unobserved U_i in our theoretical model is given by $U_i = \theta_2 e_{3i}$. Let θ denote the vector of parameters $(\theta_0, \theta_1, \theta_2, \theta_{01}, \theta_{02}, \theta_{11}, \theta_{12})$. In each Monte Carlo experiment the parameter vector θ is set to match the moments and outcomes of our actual data, specifically, they are set to make the ATE $\theta_1 - \theta_0$ equal our estimate -3.90, and to make the mean and variance of Y_i and D_i, and the covariance between Y_i and D_i, equal the values observed in our data. The variance of V_i is freely normalized (inside the binomial response indicator) to equal one.

The ML-ATE estimator is asymptotically efficient when e_{1i}, e_{2i}, and e_{3i} are normally distributed. In our first experiment we let e_{1i}, e_{2i}, e_{3i}, and V_i each have a standard normal distribution, so the resulting ML-ATE estimates can then serve as an efficient benchmark.

As noted by Khan and Tamer (2010), single threshold crossing model special regressor estimators converge at slow rates when f_v has thin tails, as in the previous design. Although their results are not directly applicable to this paper's two threshold model, it is still sensible to see if our estimator works better with thicker tails, so our second experiment gives e_{1i}, e_{2i}, e_{3i}, and V_i each a uniform distribution on $[-0.5, 0.5]$. Note this is still likely not the best case for our estimator, since Khan and Tamer (2010) note that special regressor methods converge fastest when V has a thick tail and all other variables have thin tails.

Both the normal and uniform designs have symmetric errors, which favors the ML alternative over our

[5] Hashmi only reports θ_1 and θ_2. These are in the fourth column of table 2 in Hashmi (2013). Other parameter estimates can be found in the Stata log file he posts online.

estimator. However, with symmetric errors it is impossible to define a vector θ that matches all the moments of the empirical data, because symmetry prevents matching the empirical covariance between Y and D. Therefore, in both designs we choose values for θ that match all the other moments and come as close as possible to matching this covariance (the required values for θ are given in the footnote of Table 4).

To match the empirical correlation between Y and D along with other moments, we next consider designs that introduce asymmetry into the confounder e_{3i}. In our third experiment, we let e_{1i}, e_{2i}, and V_i be standard normal and let e_{3i} equal a standard normal with probability one half when negative and equal θ_3 times a standard normal with probability one half when non-negative. We then choose θ_3 along with the other elements of θ to match the moments of the empirical data including the covariance of Y with D. This required setting $\theta_3 = 2.65$. Similarly, in a fourth experiment we let e_{1i}, e_{2i}, and V_i be uniform on $[-0.5, 0.5]$ and take e_{3i} to equal a (demeaned) mixed uniform distribution. This mixture was uniform on $[-2, 0]$ with probability one half and uniform on $[0, 5]$ with probability one half, before demeaning.

Each of these four Monte Carlo experiments was replicated 10,000 times, and the results are summarized in Table 4. Panel A in Table 4 is the symmetric normal design. Because of symmetry, all of the estimators in this design are unbiased. ML, being efficient here, has the lowest root mean squared error (RMSE), and the naive estimator is almost as efficient as ML in this case, since it just involves differencing simple covariance estimates. Our Trim-ATE estimator performs reasonably well compared to the efficient estimator, being unbiased and having a RMSE of .43 versus the efficient .30. Trimming improves the RMSE enormously here, as expected because f_v has thin tails, which produces outliers in the denominator of averages weighted by f_v.

Panel B of Table 4 shows that, in the symmetric uniform design, all four estimators are almost identical. This happens because, with V is uniform, $\widehat{f_v}$ is close to a constant, and the estimators for the average effects of the treated and the untreated are close to their sample means.

In the asymmetric designs, given in panels C and D of Table 4, the ML-ATE and Naive-ATE are no longer consistent, and both become substantially downward biased, with an average value of about one half the true value of -3.90. In contrast, our trimmed and untrimmed ATE estimates had far smaller downward biases, resulting in much smaller RMSE, particularly for the Trim-ATE.

The differences in biases between the inconsistent estimators (ML-ATE and NAIVE-ATE) and our proposed estimator in these asymmetric Monte Carlos closely match the observed differences in our empirical application estimates. Specifically, in case 1 of Table 2A the estimated ATE using the ML and Naive estimators is about one half the estimate of -3.90 we obtained using Trim-ATE. This provides evidence that the Monte Carlo results in panels C and D of Table 4 are relevant for assessing the empirical performance of our proposed estimator.

In addition to assessing the quality of estimators we also assess the quality of associated standard error estimates, by providing, in the last column of Table 4, the percentage of times the true ATE fell in the estimated 95% confidence interval (defined as the estimated ATE plus or minus two estimated standard errors). In the symmetric designs all the estimated standard errors for all the estimators were too large, yielding overly conservative inference. In the asymmetric designs the estimated 95% confidence intervals of the inconsistent estimators ML-ATE and NAIVE-ATE were very poor, containing the true value less than 25% of the time. The No-Trim-ATE did much better, but our preferred estimator, Trim-ATE, was by far the best, giving correct 95% coverage in panel C, and conservative 99% coverage in panel D.

5.7 Empirical Results - Extensions

We now give the empirical results of applying our extensions to the Hashmi data. These extensions are the test of the large support assumption and the estimation of the general ordered choice model.

For testing large support in our data we apply Theorem 4.1 (the compromise test), setting $\varepsilon^* = 0.05$. The P-value is calculated as $P = \Phi\left(\frac{\widehat{G}_D(\widehat{m}) - \widehat{\mathbb{B}}_h - \varepsilon^*}{\widehat{\sigma}(\widehat{m})}\right)$ as in Remark 4.1. Following our main empirical results, we define D_{it} with thresholds that are the 25% and 75% quantiles of the entire sample of c_{it} observations. For our first test we use the whole data set to get estimates of $\widehat{G}_D$ at the left and right boundary (minimum and maximum of V respectively). We calculate the optimal bandwidth as described in Remark 4.1. The resulting $\widehat{G}_D$ near both boundaries are very close to zero. The P-values are both zero to three decimal places, which rejects the null hypothesis that $P(D = 1|V = m) \geq 0.05$, thereby supporting $P(D = 1|V = m) < 0.05$.

This application of the test implicitly assumes the supports of V_{it} and U_{it} do not vary over time. In case the supports do vary, we also apply the test separately in each time period. This results in testing both ends of V in each of 26 time periods, for a total of 52 tests. These results are collected in Panel A of Table 3A. The null hypothesis is rejected at the 5% significance level in 36 of these 52 cases. These separate year tests need to be interpreted cautiously, however, since many of these failures to reject could be due to quite small sample sizes for nonparametric estimation. Specifically, each period contains 116 or fewer observations. Following Remark 4.2, we apply our test on the truncated sample that discards observations in the lowest and highest one percent of V values. For the whole truncated sample, the P-values are again both zero to three decimal places. For those small samples in each time period, the test results are also similar to what we have before, and are reported in Panel B of Table 3A.

To summarize, the null hypothesis is strongly rejected for the whole sample in favor of our large support assumption. The null hypothesis is also rejected for most cases when we conduct the test over each time period separately, though these latter tests suffer from small sample sizes. We conclude that the required large support assumption generally appears to hold in our application.

For our final set of estimates, we apply the results from Section 4.2, where identification at infinity is used to extend the identification result to the general ordered choice model of treatment. The model is now

$$Y_{it} = D_{0it}W_{0it} + D_{1it}W_{1it} + D_{2it}W_{2it} \tag{5.10}$$

where W_{jit} is the potential outcome of Y_{it} given treatment j (as discussed in the introduction, the potential outcome of Y_{it} itself given the additional assumption that all potential outcomes have the same fixed effects) and, for treatments $j = 0, 1, 2$,

$$D_{jit} = I\left[\alpha_{j-1} \leq V_{it} + a_i + b_t + U_{it} < \alpha_j\right]. \tag{5.11}$$

with $\alpha_{-1} = -\infty$ and $\alpha_2 = \infty$. Here, relative to Theorem 4.3, we have no covariates X and we have added the fixed effects to the treatment equations.

Here $Y_{1it} = W_{1it}$, so the estimator of $E(W_1)$ equals the estimator of $E(Y_1)$. But unlike estimation of $E(Y_0)$, estimation of $E(W_0)$ and $E(W_2)$ require tuning parameters, and converge at slower than parametric rates, because they are based on identification at infinity. Based on Theorem 4.3, the sample counterpart estimators for $E(W_0)$, $E(W_1)$ and $E(W_2)$ are

$$\widehat{E}(W_0) = \frac{\sum_i \sum_t D_{0it} Y_{it} I(v_{it} \leq -\gamma_{nT})}{\sum_i \sum_t D_{0it} I(v_{it} \leq -\gamma_{nT})},\ \widehat{E}(W_2) = \frac{\sum_i \sum_t D_{2it} Y_{it} I(v_{it} \geq \gamma'_{nT})}{\sum_i \sum_t D_{2it} I(v_{it} \geq \gamma'_{nT})},$$

$$\widehat{E}(W_1) = \frac{\sum_i \sum_t I_\tau(v_{it}) D_{1it} Y_{it} \Big/ \widehat{f}_{v_t}(v_{it})}{\sum_i \sum_t I_\tau(v_{it}) D_{1it} \Big/ \widehat{f}_{v_t}(v_{it})}$$

where γ_{nT} and γ'_{nT} are increasing series such that $\lim_{n,T\to\infty} E(D_{0it} \mid V \leq -\gamma_{nT}) = 0$ and $\lim_{n,T\to\infty} E(D_{2it} \mid V \geq \gamma'_{nT}) = 1$.

The estimator $\widehat{E}(W_1)$ duplicates our previous results, based on Theorem 3.2. To provide standard errors for $\widehat{E}(W_0)$ and $\widehat{E}(W_2)$ we apply the asymptotic theory for models identified at infinity provided by Andrews and Schafgans (1998) and Schafgans (1998). The resulting standard error for $\widehat{E}(W_0)$ is obtained using

$$\frac{\left(\sum_i \sum_t D_{0it} \left(Y_{it} - \widehat{E}(W_0)\right)^2 I(v_{it} \leq -\gamma_{nT})\right)^{1/2}}{\sum_i \sum_t D_{0it} I(v_{it} \leq -\gamma_{nT})}.$$

and similarly for $\widehat{E}(W_2)$. Following Schafgans (1998), we consider various values for γ_{nT} and γ'_{nT}, based on the percentage of uncensored observations (e.g., for $E(W_0)$ uncensored means $D_{0it} = 1$) used in the estimation. Specifically we consider 50%, 40%, 30%, 20%, 10%, and 5% uncensored.

Empirical results are reported in Table 3B. Panel A displays the estimates when we define D_{jit} with thresholds that are the 25% and 75% quantiles of the entire sample of c_{it} observations. As a robustness check, panel B displays the results when we define D_{jit} using the 25% and 75% quantiles of c_{it} defined separately in each year of data. The results do not vary much by year, and are also not very sensitive to the choice of γ_{nT} and γ'_{nT}, especially for $\widehat{E}(W_2)$. Not surprisingly, the standard errors become larger when the tuning parameters are larger, corresponding to averages over fewer observations.

The estimate $\widehat{E}(W_1)$ from the previous section is 4.33. Seen from Table 3B, $\widehat{E}(W_2)$ is slightly (but not significantly) higher than $\widehat{E}(W_1)$, while $\widehat{E}(W_0)$ is much higher than $\widehat{E}(W_1)$ and $\widehat{E}(W_2)$. Therefore we obtain a mostly decreasing relationship between innovation and competition. This pattern is similar to the quadratic least squares estimation of the raw data (see Figure 2). This result is also consistent with Hashmi (2013), who speculates that manufacturing in the US does not appear to have an inverted-U shape because the US is likely dominated by Leader-Laggard industries.

6 Conclusions

In this article, we propose a new method to estimate the average treatment effect in models were treatment is determined by an ordered choice model. In our empirical application, industries are defined as treated if they possess an intermediate level of competition, versus a low or high level of competition.

Unconfoundedness of treatment (either unconditional or conditional on covariates) does not hold in the model, because the unobservables that affect treatment are correlated in unknown ways with potential outcomes, with or without conditioning on other covariates. No parametric or semiparametric restrictions are placed on distributions of treatment and potential outcomes, so treatment effects are not identified by functional form. Our model assumes a continuous instrument V with large support, but treatment effects are not identified at infinity, because both very large and very small values of V drive the probability of intermediate levels of treatment close to zero, while no value of V (or of other covariates) drives the probability of intermediate levels of treatment close to one. So in this framework none of the conditions that are known to permit point identification of the ATE hold. Even the monotonicity conditions generally required for identifying LATE in our base model cannot hold. Nevertheless, we show that the ATE is identified, using a

special regressor argument, and we provide conditions under which the corresponding estimate of the ATE is asymptotically normal and converges at the parametric rate. Root nT consistency is obtained in a panel context with fixed effects, despite nonlinearities that would usually induce an incidental parameters problem in the equation determining probability of treatment. We provide Monte Carlo results that show that our estimator works well in small samples (comparable to the data in our empirical application). We show in an Appendix that our estimator is also relatively robust to substantial measurement errors. We also provide a test of our main identifying assumption, that V has large support, and we provide some extensions and generalizations of our main results and asymptotics.

We use our method to investigate the relationship between competition and innovation. Our estimates using a dataset from Hashmi (2013) confirm Hashmi's finding that an inverted-U is not present in US data. However, although we agree with his sign of the treatment effect, we find both in our data and in a corresponding Monte Carlo that parametric models, including that of ABBGH and Hashmi, and other naive treatment effect estimators, substantially underestimate the magnitude of the treatment effect in this application.

References

[1] Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," Econometrica. Vol. 70, No. 1, 91-117.

[2] Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005), "Competition and Innovation: an Inverted-U Relationship," Quarterly Journal of Economics, 120(2):701-28.

[3] Andrews, D., and M. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," Review of Economic Studies, 65, 497-17.

[4] Angrist, J. D., and G.W. Imbens (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," Journal of the American Statistical Association 90:430, 431-42.

[5] Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics, 60, 47-57.

[6] Ashenfelter, O., and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," Review of Economics and Statistics, 67, 648-660.

[7] Barnow, B. S., G. G. Cain, and A. S. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in Evaluation Studies, Vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, 43–59.

[8] Bertrand, M. (2004), "From the Invisible Handshake to the Invisible Hand? How Import Competition Changes the Employment Relationship," Journal of Labor Economics, 22(4):722-65.

[9] Bitler, M., J. Gelbach, and H. Hoynes (2006), "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," American Economic Review, 96, 4, 988-1012.

[10] Björklund, A., and R. Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models," Review of Economics and Statistics, Vol. LXIX, 42–49.

[11] Cao, S., R. Moineddin, M.L. Urquia, F. Razak, and J.G. Ray (2014), "J-shapedness: an Often Missed, Often Miscalculated Relation: the Example of Weight and Mortality," Journal of Epidemiology and Community Health, 68, 683–690.

[12] Card, D. (1990), "The Impact of the Mariel Boatlift on the Miami Labor Market," Industrial and Labor Relations Review 43, 245-257.

[13] Card, D., and A. Krueger (1993), "Trends in Relative Black-White Earnings Revisited," American Economic Review, vol. 83, no. 2, 85-91.

[14] Card, D., and A. Krueger (1994), "MinimumWages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania," American Economic Review, 84 (4), 772-784.

[15] Cecchetti, S. G., and E. Kharroubi (2012), "Reassessing the Impact of Finance on Growth," working paper.

[16] Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," Econometrica, 73(1), 245-261.

[17] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2009), "Identification and Estimation of Marginal Effects in Nonlinear Panel Models," Technical report, CEMMAP.

[18] Cook, P.J., and G. Tauchen (1982), "The effect of Liquor Taxes on Heavy Drinking," Bell Journal of Economics, 13(2): 379-90.

[19] Cook, P.J., and G. Tauchen (1984), "The Effect of Minimum Drinking age Legislation on Youthful Auto Fatalities, 1970-1977," Journal of Legal Studies, 13(1): 169-90.

[20] Cochran, W., and D. Rubin (1973), "Controlling Bias in Observational Studies: A Review," Sankhyā, 35, 417–446.

[21] Cox, D. R., (1958), Planning of Experiments. New York: Wiley.

[22] Dong Y., and A. Lewbel (2015), "A Simple Estimator for Binary Choice Models with Endogenous Regressors," Econometric Reviews, 34, 82-105.

[23] Fan, J., and I. Gijbels (1992), "Variable Bandwidth and Local Linear Regression Smoothers," The Annals of Statistics, 20(4), 2008-2036.

[24] Firpo, S. (2006), "Efficient Semiparametric Estimation of Quantile Treatment Effects," Econometrica, 75(1), 259-276.

[25] Hardle, W. (1990), Applied Nonparametric Regression. Cambridge University Press.

[26] Hardle, W., and T. M. Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," Journal of the American Statistical Association, 84, 986-995

[27] Hashmi, A.R. (2013), "Competition and Innovation: the Inverted-U Relationship Revisited," Review of Economic Statistics, 95, 5, 1653-1668.

[28] Haavelmo, T. (1943), "The statistical implications of a system of simultaneous equations," Econometrica, 11, 1-12.

[29] Heckman, J. J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica, 46, 931-59.

[30] Heckman, J. (1979), "Sample Selection Bias as a Specification Error," Econometrica, 47(1), 153-162.

[31] Heckman, J. J., and S. Navarro (2007), "Dynamic Discrete Choice and Dynamic Treatment Effects," Journal of Econometrics, 136, (2), 341-396.

[32] Heckman, J., and R. Robb (1984), "Alternative Methods for Evaluating the Impact of Interventions," Longitudinal Analysis of Labor Market Data, ed. by J. Heckman and B. Singer. Cambridge, U.K.: Cambridge University Press, 156–245.

[33] Heckman, J.J., S. Urzua, and E. Vytlacil (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity," Review of Economics and Statistics, 88(3), 389-432.

[34] Heckman, J. J., and E. Vytlacil (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," Proceedings of the National Academy of Science, USA, 96, 4730–4734.

[35] Heckman, J. J., and E. Vytlacil (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica, 73, 669–738.

[36] Heckman, J. J., and E. Vytlacil (2007a), "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Evaluation of Public Policies," Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 70.

[37] Heckman, J. J., and E. Vytlacil (2007b), "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," Handbook of Econometrics, J.J. Heckman and E.E. Leamer (eds.), Vol. 6, North Holland, Chapter 71.

[38] Hickman, B. R., and T. P. Hubbard (2014), "Replacing Sample Trimming with Boundary Correction in Nonparametric Estimation of First-Price Auctions," Journal of Applied Econometrics, forthcoming.

[39] Honore, B., and A. Lewbel (2002), "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," Econometrica, 70, 2053-2063.

[40] Huang, J. (2015), "Banking and Shadow Banking," working paper.

[41] Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," Econometrica, Vol. 61, No. 2, 467-476.

[42] Imbens, G., and J. Wooldridge (2009), "Recent Development in the Econometrics of Program Evaluation," Journal of Economic Literature, 47:1, 5-86.

[43] Khan, S., and E. Tamer (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Econometrica, 6, 2021-2042.

[44] Kitagawa, T. (2009), "Identification Region of the Potential Outcome under Instrument Independence," working paper.

[45] Koppes, L.L., J.M. Dekker, H.F. Hendriks, L.M. Bouter, and R.J. Heine (2005), "Moderate Alcohol Consumption Lowers the Risk of Type 2 Diabetes: a Meta–Analysis of Prospective Observational Studies," Diabetes Care, 28, 719–25.

[46] Lewbel, A. (1998), "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," Econometrica, 66, 105-122.

[47] Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity and Instrumental Variables," Journal of Econometrics, 97, 145-177.

[48] Lewbel, A. (2007), "Endogenous Selection or Treatment Model Estimation," Journal of Econometrics, 141, 777-806.

[49] Lewbel, A. (2014), "An Overview of the Special Regressor Method," in the Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Co-edited by Aman Ullah, Jeffrey Racine, and Liangjun Su, 2014, 38-62, Oxford University Press.

[50] Lewbel, A., Y. Dong, and T.T. Yang (2012), "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," Canadian Journal of Economics, 45, 809-829.

[51] Manski, C. F., and J. V. Pepper (2013), "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections," Journal of Quantitative Criminology 29 (1), 123-141.

[52] Masry, E., 1996, "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency Rates," Journal of Time Series Analysis 17, 571-599.

[53] Meyer, B., K. Viscusi, and D. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," American Economic Review, Vol. 85, No. 3, 322-340.

[54] Newey, W. K., and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[55] Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments." Essay on Principles, Section 9. published in english in (1990) Statistical Science 5(4), 465–472, translated by Dorota M. Dabrowska and Terence P. Speed.

[56] Neyman, J., and E.L. Scott (1948), "Consistent Estimation from Partially Consistent Observations," Econometrica 16, 1-32.

[57] Powell, J.L., J.H. Stock and T.M. Stoker, (1989), "Semiparametric Estimation of Weighted Average Derivatives," Econometrica, 57, 1403-1430.

[58] Revenga, A. (1990), "Essays on Labor Market Adjustment and Open Economics," PhD diss., Harvard University, Economics Department.

[59] Revenga, A. (1992), "Exporting Jobs? The Impact of Import Competition on Employment and Wages in U.S. Manufacturing," Quarterly Journal of Economics, 107, 1255–1284.

[60] Robinson, P. M. (1988), "Root-n-consistent Semiparametric Regression," Econometrica, 56, pp. 931-954.

[61] Rosenbaum, P., and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70, 41–55.

[62] Roy, A. (1951). "Some Thoughts on the Distribution of Earnings," Oxford Economic Papers, 3(2), 135–146.

[63] Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," Journal of Educational Psychology, 66, 688–701.

[64] Schafgans, M.M.A. (1998), "Ethnic Wage Differences in Malaysia: Parametric and Semiparametric Estimation of the Chinese-Malay Wage-Gap," Journal of Applied Econometrics, 13, 481-504.

[65] Solomon C.G., F.B. Hu, M.J. Stampfer, et al. (2000), "Moderate Alcohol Consumption and Risk of Coronary Heart Disease among Women with Type 2 Diabetes Mellitus," Circulation, 102, 494–99.

7 Appendix A: Robustness to Measurement Errors

Observable indices of competitiveness of an industry, like the average Lerner index in equation (5.1), may be relatively crude measures of true competitiveness. In this section we therefore assess the robustness of our estimator, relative to a parametric model estimator like Hashmi's, to measurement error in the index of competitiveness. We find that the bias in our estimator resulting from measurement error is quite small relative to the alternative, even when the parametric alternative model is otherwise correctly specified. As discussed in the text, this is not surprising, since even substantial measurement error in observed competitiveness for every industry only results in a fraction of industries being incorrectly classified by the construction of D. The relative performance of our estimator is even better when the parametric model functional form is not correctly specified.

Consider the case where competitiveness is mismeasured, but a parametric model like ABBGH or Hashmi's is the correct specification in terms of true competitiveness. Dropping fixed effects for simplicity, this model assumes

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \widetilde{e}, \tag{7.1}$$

where $\ln Y$ is logged innovation, c^* is the true level of competitiveness, and $\widetilde{e}$ is an error term. For simplicity we ignore discreteness in $\ln Y$, and we assume c^* can be linearly decomposed into the observable instrument V and an unobserved independent component R, so

$$c^* = V + R. \tag{7.2}$$

Also assume validity of the ABBGH and Hashmi control function type assumption that $\widetilde{e} = \lambda R + e$ where e is independent of R and V, so

$$\ln Y = \theta_0 + \theta_1 c^* + \theta_2 c^{*2} + \lambda R + e \tag{7.3}$$

In this model, if c^* were observed, then control function estimation (first regressing c^* on a constant and V, getting the residuals $\widehat{R}$, and then regressing $\ln Y$ on a constant, c^*, c^{*2}, and $\widehat{R}$) would consistently estimate the θ coefficients and hence any desired treatment effects based on θ.

Now assume the observable competitiveness measure c equals the true measure c^* plus measurement error c_e, so

$$c = c^* + c_e, \tag{7.4}$$

where c_e is the measurement error and independent of c^* and e. To take the best case scenario for the parametric model, assume that the measurement error c_e is classical, having mean zero and is independent of V, R, and e.

Substituting equation (7.4) into equation (7.3) gives

$$\ln Y = \theta_0 + \theta_1 c + \theta_2 c^2 + \lambda R + e^* \tag{7.5}$$

where

$$e^* = \theta_1 c_e - 2\theta_2 c c_e - \theta_2 c_e^2 + e.$$

The error e^* does not have mean zero and correlates with c and c^2, which makes the control function estimator inconsistent. Unlike the case of linear models with independent mean zero measurement errors, the control function estimator becomes inconsistent in the presence of measurement errors because of the nonlinearity of the model.

Now consider applying our nonparametric estimator to this model. The treatment indicator D that we construct is defined as equaling one for firms in the .25 to .75 quantile of c and zero otherwise, while the corresponding indicator D^* based on the true measure of competitiveness equals one for firms in the .25 to .75 quantile of c^* and zero otherwise. Unless the measurement error c_e is extremely large, for the large majority of firms D will equal D^*. This is part of what makes our estimator more robust to measurement error. Even if all firms have c mismeasured to some extent, most will still be correctly classified in terms of D.

To check the relative robustness of these estimators to measurement error, we perform an additional Monte Carlo analysis. As before, we construct simulated data to match moments and the sample size of the empirical data set, and to make what would be the true treatment effect in the model match our empirical estimate of -3.9. We do two simulations, one using normal errors and one based on uniform errors, as before. In both, V and R are scaled to have equal magnitudes, so $V = \delta_0 + \delta_1\varepsilon_1$ and $R = \delta_0 + \delta_1\varepsilon_2$. To match moments in the real data, our normal error simulations set $\delta_0 = 0.375$, $\delta_1 = 0.0733$, and $c_e = \kappa_1\varepsilon_3$ where ε_1, ε_2, and ε_3 are independent standard normals and κ_1 is a constant with values that we vary to obtain different magnitudes of measurement error. The uniform error simulations set $\delta_0 = \delta_1 = 0.25$, and $c_e \sim \kappa_2(\varepsilon_3 - 0.5)$, where now ε_1, ε_2, and ε_3 are independent random variables that are uniformly distributed on $[0, 1]$.

To check for robustness against an alternative specification as well as measurement error, we also generate data replacing the quadratic form in equation (7.1) with the step function

$$\ln Y = \theta_0 + (\theta_1 - \theta_0)D^* + \widetilde{e}, \tag{7.6}$$

where D^*, D, c^*, c, V, R, and e are all defined as above.

The Monte Carlo results, based on 10,000 replications, are reported in Tables 5 and 6. In addition to trying out the four estimators we considered earlier, (Trim-ATE, No-Trim-ATE, Naive-ATE, and ML-ATE) we also apply the control function estimator described above, analogous to Hashmi's estimator.

Our main result is that, with both normal and uniform errors, the greater the magnitude of measurement error is (that is, the larger the κ_1 and κ_2 are), the better our estimator performs relative to other estimators. For the quadratic model without measurement error the control function would be a consistent parametric estimator and so should in that case outperform our semiparametric estimator. We find this also holds with very small measurement error (e.g., $\kappa_1 = .02$ in the left side block of Table 5), however, both the control

function and Trim-ATE estimators perform about equally at $\kappa_1 = .03$, and at the still modest measurement error level of $\kappa_2 = .04$, our Trim-ATE estimate has smaller RMSE (root mean squared error) than all the other estimators, including the control function estimator. Similar results hold for the uniform error model reported in Table 6. Also, in the step function model (shown on the right side of Tables 5 and 6) our Trim-ATE is very close to, or superior to, all the other estimators including control functions at all measurement error levels.

It is worth noting that possible measurement error affects our empirical application only because we defined treatment D in terms of an observed, possibly mismeasured underlying variable, competitiveness. In other applications the treatment indicator may be observed without error even when an underlying latent measure is completely unobserved. For example, suppose an outcome Y is determined in part by an individual's chosen education level, which in turn is determined by an ordered choice specification. The true education level of a student might be unobserved, but a treatment D defined as having graduated high school but not college could still be correctly measured.

8 Appendix B: Additional Extensions

8.1 Identifying an Additive Function of V

In previous sections, we assumed V appears in the selection equation in the form $V + U$. In this section, we consider the generalization where selection depends on $\varsigma(V) + U$ for some unknown function $\varsigma(V)$. This may be more realistic in some applications, since economic theory may not indicate a priori the function $\varsigma(V)$. Given identification and an associated estimator for ς, one could then redefine V as $\varsigma(V)$ and then estimate treatment effects as before. Though not empirically relevant, it is theoretically interesting to note that in the very special case where the function ς equals the cumulative density function of V, the model becomes unconfounded automatically and our proposed estimator then reduces to standard propensity score weighting.

To identify ς, we assume that there is a continuously distributed exogenous covariate Z that affects selection but does not affect the thresholds. Then the selection equation takes the form

$$D = I\left(\widetilde{\alpha}_0(X) \leq \varsigma(V) + \varpi(X, Z) + U \leq \widetilde{\alpha}_1(X)\right). \tag{8.1}$$

Formally, we assume the following.

Assumption 8.1 *Equation (8.1) holds for observed covariates V, Z, and vector X, where $\varsigma, \varpi, \widetilde{\alpha}_0, \widetilde{\alpha}_1$ are unknown functions, ς is differentiable, 0 is in the support of V, $\varsigma(0) = 0$, and $\varsigma'(0) = 1$, and $(V, Z) \perp U \mid X$.*

We could equivalently write equation (8.1) as

$$D = I\left(\alpha_0(X, Z) \leq \varsigma(V) + U \leq \alpha_1(X, Z)\right)$$

for some unknown functions ς, α_0, and α_1 where $\alpha_1(X, Z) - \alpha_0(X, Z) = \delta(X)$ for some function δ. In the standard specification of ordered choice models where $D = I\left(\delta_0 \leq X'\beta_1 + V\beta_2 + U \leq \delta_1\right)$ and X is exogenous, every continuous regressor contained in the vector X could be relabeled as Z and would then satisfy Assumption 8.1. This is much stronger than necessary, since we only the require existence of one such regressor.

The assumptions that zero is in the support of V, that $\varsigma(0) = 0$, and that $\varsigma'(0) = 1$ are all free normalizations that are made without loss of generality. To see this, first note that there must exist some value of v in the support of V for which $\varsigma'(v) \neq 0$, since otherwise $\varsigma(V)$ would be a constant, not a function of V. Redefining V as $V - v$ then ensures that zero is the support of V and that $\varsigma'(0) \neq 0$. Next redefine all of the unknown functions, and U, by dividing them all by $\varsigma'(0)$. After this scale normalization, we will have by construction that $\varsigma'(0) = 1$. Finally, $\varsigma(0) = 0$ is a free location normalization, since if $\varsigma(0) = c \neq 0$ then we can redefine $\varpi(X, Z)$ as $\varpi(X, Z) + c$ to make $\varsigma(0) = 0$.

The following theorem shows identification of the function ς. The proof is constructive, so one could obtain a consistent estimator of ς by mimicking the steps of the proof, using standard kernel based nonparametric regression derivative estimators. After estimating ς, our previous estimators could then be applied by replacing the density of V with the density of $\varsigma(V)$.

Theorem 8.1 *Suppose we observe X, V, Z, D and D follows equation (8.1). Given Assumption 8.1, the functions $\varsigma(V)$ and $\frac{\partial \varpi(X,Z)}{\partial Z}$ are identified.*

This theorem is proved in the online supplemental Appendix.

8.2 Additional Panel Data Asymptotics

We showed earlier that in the panel model, Assumption 3.10 was necessary for obtaining a $\sqrt{nT}$ convergence rate. Here we consider asymptotics when Assumption 3.10 is not imposed. In this case we can also replace Assumption 3.7 with the weaker Assumption 8.2, yielding $\left(\widehat{f}_{v_t}(v) - f_{v_t}(v)\right)^2 = o_p\left(n^{-1/2}\right)$, because the convergence rate of the estimator will now only be $\sqrt{T}$. Similarly, a higher order kernel will no longer be needed.

Assumption 8.2 *$n \to \infty, T \to \infty$, and $T = o(n)$.*

Theorem 8.2 *Let Assumption 3.1, 3.4, 3.5, 3.6, 3.8, 3.9, 8.2, 9.3, and 9.5 hold. Assume that bandwidth $h = c_0 n^{-1/5}$ in $\widehat{f}_{v_t}$ and assume a symmetric kernel of order $p = 2$. Then*

$$\sqrt{T}\left[\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1) + E(\widetilde{a}_i + \widetilde{b}_t + Y_0)\right]$$
$$\xrightarrow{d} N\left(0, \frac{var\left(E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1^2} - \frac{2cov\left(E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right], E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_1\overline{\Pi}_2} + \frac{var\left(E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]\right)}{\overline{\Pi}_2^2}\right).$$

This theorem is proved in the supplemental online appendix.

Remark 8.1 In this $\sqrt{T}$ convergence case, we can allow arbitrary dependence between Y_{jit} and $\left(\widetilde{a}_i, \widetilde{b}_t\right)$, which implies that Y_{jit} can contain some general function of $\widetilde{a}_i$ *and* $\widetilde{b}_t$ as long as $E(Y_{jit}) = E(Y_j)$, $j = 0, 1$. We can similarly allow for more general fixed effects of the form $g\left(\widetilde{a}_i, \widetilde{b}_t\right)$ instead of $\widetilde{a}_i + \widetilde{b}_t$ for some unknown function g, because these fixed effects will now still difference out.

Remark 8.2 Suppose $\left(a_i, \widetilde{a}_i, b_t, \widetilde{b}_t\right)$ is a series of constants instead of random variables. From the proof of Lemma 10.7, the above rate $\sqrt{T}$ limiting distribution will still hold if $\frac{1}{n^2}\left(\sum_{i=1}^{n} \widetilde{a}_i^2\right) = O(1)$ and $\frac{1}{nT}\left(\sum_{t=1}^{T} \widetilde{b}_t^2\right) = O(1)$.

8.3 Dynamic Panels

Our identification can extend to some cases of dynamic panel data models.[6] Define the treatment indicator equation as

$$D_{it} = I(\alpha_0(x_{it}) \leq a_i + b_t + V_{it} + \vartheta\left(D_{it-1}\right) + U_{it} \leq \alpha_1(x_{it})), \tag{8.2}$$

and the outcome equation as

$$Y_{it} = \widetilde{a}_i + \widetilde{b}_t + g\left(Y_{it-1}\right) + Y_{0it} + (Y_{1it} - Y_{0it})D_{it}, \tag{8.3}$$

where the treatment indicator and the outcome variable are now related to those in the last period, and these effects are captured by two potentially unknown functions ϑ and g.

As before, the observables in the model are the outcome Y, treatment D, instrument V, and covariate vector X. Also as before, $(a_i, b_t, \widetilde{a}_i, \widetilde{b}_t)$ are treated like fixed effects, which can correlate with unobservables and with X in unknown ways. We provide two alternative identifying theorems. The first, Theorem 8.3, imposes equation (8.5), which generally requires that $V_{it} \perp V_{it-1}$. We imposed $V_{it} \perp V_{it-1}$ in Assumption 3.9, but there this restriction was imposed for convenience, and could have been relaxed to allow for weak dependence. In contrast, for our first dynamic panel identification theorem, we do not allow dependence between V_{it} and V_{it-1}. However, this restriction can still be relaxed in one of two ways. First, if $\vartheta\left(D_{it-1}\right)$ is identically zero, then we could directly replace $V_{it} \perp V_{it-1}$ with weak dependence in Theorem 8.3. The second result is Theorem 8.4. This permits identification under a weaker set of assumptions that permits some limited dependence among $\{V_{it}\}_{t=1}^{T}$. For example, $\{V_{it}\}_{t=1}^{T}$ can be a Markov chain (see Remark 8.4). However, this result requires modifying the associated estimator as given by equation (8.9), where $f_{v_t}(V_{it}|X_{it})$ is replaced by $f_{v_t}(V_{it}|X_{it}, V_{it-1})$.

Heckman and Navarro (2007) obtain identification of structural dynamic discrete choice models and models of dynamic treatment effects. Although the model specifications in Heckman and Navarro (2007) and our paper are very different, the identification strategies are similar in terms of dealing with dynamics. Both their paper and ours rely on sufficient variation on some covariates. Heckman and Navarro (2007) permit more general serial correlation than we do, however, we allow for feedback and fixed effects in the choice and outcome equations that they do not.

Assumption 8.3 *For individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables.*

$$E\left(g\left(Y_{it-1}\right) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit}|X_{it}, V_{it}, a_i, b_t, U_{it}, D_{it-1}\right) = E\left(g\left(Y_{it-1}\right) + \widetilde{a}_i + \widetilde{b}_t + Y_{jit}\middle| X_{it}, a_i, b_t, U_{it}, D_{it-1}\right), \tag{8.4}$$

for $j = 0, 1$, and

$$V_{it} \perp a_i, b_t, U_{it}, D_{it-1}|X_{it}. \tag{8.5}$$

Remark 8.3 Equation (8.4) is not much more restrictive than the corresponding part of Assumption 3.5.

[6] We thank a referee pointing out these potential extensions.

Assumption 8.4 *Assumption 3.3 holds after replacing supp[$\alpha_0(X)-U, \alpha_1(X)-U$] with supp[$\alpha_0(x_{it})-\widetilde{a}_i-\widetilde{b}_t-U_{it}-\vartheta\left(D_{it-1}\right), \alpha_1(x_{it})-\widetilde{a}_i-\widetilde{b}_t-U_{it}-\vartheta\left(D_{it-1}\right)$].*

Theorem 8.3 *Let Assumption 3.1, 8.3, and 8.4 hold for each individual i in each time period t. Let f_{v_t} denote the density of V in time t. Then*

$$\frac{E[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[D_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]}-\frac{E[(1-D_{it})Y_{it}/f_{v_t}(V_{it}|X_{it})|X_{it}]}{E[(1-D_{it})/f_{v_t}(V_{it}|X_{it})|X_{it}]}=E(Y_{1it}-Y_{0it}|X_{it}). \tag{8.6}$$

The proof of this result is omitted, since it follows from the same logic and steps as the proof of Theorem 3.2.

Assumption 8.5 *For individuals i and time periods t, $a_i, b_t, \widetilde{a}_i, \widetilde{b}_t$ are random variables.*

$$\begin{aligned} & E\left(g\left(Y_{it-1}\right)+\widetilde{a}_i+\widetilde{b}_t+Y_{jit}|X_{it},V_{it},a_i,b_t,U_{it},D_{it-1},V_{it-1}\right) \\ = \; & E\left(g\left(Y_{it-1}\right)+\widetilde{a}_i+\widetilde{b}_t+Y_{jit}\Big|\, X_{it},a_i,b_t,U_{it},D_{it-1},V_{it-1}\right), \end{aligned} \tag{8.7}$$

for $j=0,1$, and

$$V_{it}\perp a_i,b_t,U_{it},D_{it-1}|\left(X_{it},V_{it-1}\right). \tag{8.8}$$

Remark 8.4 Condition (8.8) can be implied by $V_{it}\perp a_i,b_t,\{U_{ij}\}_{j=1}^{t}\Big|\, X_{it}$, and $V_{it}\perp\{V_{ij}\}_{j=1}^{t-2}\Big|\, V_{it-1}$. So $\{V_{it}\}_{t=1}^{T}$ can be a Markov chain under this condition.

Theorem 8.4 *Let Assumption 3.1, 8.4 and 8.5 hold for each individual i in each time period t. Let f_{v_t} denote the density of V in time t. Then*

$$\frac{E[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}{E[D_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}-\frac{E[(1-D_{it})Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}{E[(1-D_{it})/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}]}=E(Y_{1it}-Y_{0it}|X_{it}). \tag{8.9}$$

The proof of this theorem is provided in the supplemental Appendix.

9 Appendix C: Additional Assumptions and Proofs

Proof of Theorem 3.1. We first prove the result under Assumptions 3.1, 3.2, 3.3 and 3.4.

$$\begin{aligned} E\left(\frac{I_\tau DY}{f\left(V\mid X\right)}\mid U,X\right) &= E\left[E\left(\frac{I_\tau DY_1}{f\left(V\mid X\right)}\mid V,U,X\right)\mid U,X\right] \\ &= E\left[\frac{I_\tau I\left[\alpha_0\left(X\right)\le V+U\le\alpha_1\left(X\right)\right]E\left(Y_1\mid V,U,X\right)}{f\left(V\mid X\right)}\mid U,X\right] \\ &= \int_{supp(V|U,X)}\frac{I_\tau I\left[\alpha_0\left(X\right)-U\le v\le\alpha_1\left(X\right)-U\right]E\left(Y_1\mid U,X\right)}{f\left(v\mid X\right)}f\left(v\mid U,X\right)dv \\ &= \int_{\alpha_0(X)-U}^{\alpha_1(X)-U}\frac{E\left(Y_1\mid U,X\right)}{f\left(v\mid X\right)}f\left(v\mid X\right)dv=E\left(Y_1\mid U,X\right)\int_{\alpha_0(X)-U}^{\alpha_1(X)-U}1dv \\ &= \left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]E\left(Y_1\mid U,X\right), \end{aligned}$$

the fourth equality holds by Assumption 3.3. Therefore

$$E\left(\frac{I_\tau DY}{f\left(V\mid X\right)}\mid X\right)=\left[\alpha_1\left(X\right)-\alpha_0\left(X\right)\right]E\left(Y_1\mid X\right).$$

The same analysis dropping Y gives

$$E\left(\frac{I_\tau D}{f\left(V \mid X\right)} \mid X\right) = \alpha_1\left(X\right) - \alpha_0\left(X\right)$$

so

$$E\left(\frac{I_\tau DY}{f\left(V \mid X\right)} \mid X\right) = E\left(Y_1 \mid X\right) E\left(\frac{I_\tau D}{f\left(V \mid X\right)} \mid X\right).$$

Similarly,

$$\begin{aligned}
E\left(\frac{I_\tau\left(1-D\right)Y}{f\left(V \mid X\right)} \mid X\right) &= E\left(\frac{I_\tau\left(1-D\right)Y_0}{f\left(V \mid X\right)} \mid X\right) \\
&= E\left(\frac{I_\tau Y_0}{f\left(V \mid X\right)} \mid X\right) - E\left(\frac{I_\tau DY_0}{f\left(V \mid X\right)} \mid X\right) \\
&= E\left(Y_0 \mid X\right) E\left(\frac{I_\tau}{f\left(V \mid X\right)} \mid X\right) - \left[\alpha_1\left(X\right) - \alpha_0\left(X\right)\right] E\left(Y_0 \mid X\right) \\
&= E\left(Y_0 \mid X\right) E\left(\frac{I_\tau}{f\left(V \mid X\right)} - \left[\alpha_1\left(X\right) - \alpha_0\left(X\right)\right] \mid X\right) \\
&= E\left(Y_0 \mid X\right) E\left(\frac{I_\tau\left(1-D\right)}{f\left(V \mid X\right)} \mid X\right)
\end{aligned}$$

Together these equations prove the result.

Replacing I_τ with one shows that the same result holds with just Assumptions 3.1, 3.2, and $I_\tau = 1$. ■

Proof of Theorem 3.2. This proof is very similar to the proof for Theorem 3.1. First consider

$$\begin{aligned}
& E\left(\frac{I_{\tau it}D_{it}Y_{it}}{f_{v_t}\left(V_{it}|X_{it}\right)}\middle| U_{it}, a_i, b_t, X_{it}\right) \\
= \; & E\left[E\left(\frac{I_{\tau it}D_{it}\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)} \mid V_{it}, U_{it}, a_i, b_t, X_{it}\right) \mid U_{it}, a_i, b_t, X_{it}\right] \\
= \; & E\left[\frac{I_{\tau it}I\left(\alpha_0(X_{it}) \leq a_i + b_t + V_{it} + U_{it} \leq \alpha_1(X_{it})\right) E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid V_{it}, U_{it}, a_i, b_t, X_{it}\right)}{f_{v_t}\left(V_{it}|X_{it}\right)} \mid U_{it}, a_i, b_t, X_{it}\right] \\
= \; & \int_{supp(V_{it}|U_{it},a_i,b_t,X_{it})} \frac{I_{\tau it}I\left(\alpha_0(X_{it}) - a_i - b_t - U_{it} \leq v_{it} \leq \alpha_1(X_{it}) - a_i - b_t - U_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)} \\
& E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) f_{v_t}\left(v_{it} \mid U_{it}, a_i, b_t, X_{it}\right) dv_{it} \\
= \; & \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}}^{\alpha_1(X_{it})-a_i-b_t-U_{it}} \frac{E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right)}{f_{v_t}\left(v_{it}|X_{it}\right)} f_{v_t}\left(v_{it}|X_{it}\right) dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}}^{\alpha_1(X_{it})-a_i-b_t-U_{it}} 1 dv_{it} \\
= \; & E\left(\widetilde{a}_i + \widetilde{b}_t + Y_{1it} \mid U_{it}, a_i, b_t, X_{it}\right) \left[\alpha_1(X_{it}) - \alpha_0(X_{it})\right]
\end{aligned}$$

and therefore

$$\begin{aligned} & E\left[I_{\tau it}D_{it}Y_{it}/f_{v_t}\left(V_{it}|X_{it}\right)|X_{it}\right] \\ = \; & E\left[E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}\mid U_{it},a_i,b_t,X_{it}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]|X_{it}\right] \\ = \; & E\left(Y_{1it}+\widetilde{a}_i+\widetilde{b}_t\middle| X_{it}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]. \end{aligned}$$

Given the above result, the rest of the proof follows similarly as in the proof for Theorem 3.1. ■

We now provide applicable standard assumptions for the asymptotics of our kernel based estimators. Let $\mathbf{m}_k \equiv (m_1, m_2, ..., m_k)$ be a $k \times 1$ vector of non-negative integers. Following Masry (1996), we adopt the notation: $u^{\mathbf{m}_k} \equiv \Pi_{i=1}^k u_i^{m_i}$, $\mathbf{m}_k! \equiv \Pi_{i=1} m_i!$, $|\mathbf{m}_k| \equiv \sum_{i=1}^k m_i$, and $\sum_{|\mathbf{m}_k|=p} \equiv \sum_{m_1=0}^p \cdots \sum_{\substack{m_k=0 \\ m_1+\cdots+m_k=p}}^p$. Let $D^{\mathbf{m}_k} f_x(x) \equiv \partial^{|\mathbf{m}_k|} f_x(x) / \partial^{m_1} x_1 \cdots \partial^{m_k} x_k$. Covariates of dimensions other than k are denoted analogously, e.g. for dimension $k+1$ we have $\mathbf{m}_{k+1}$ and the other terms above are changed accordingly with k replaced by $k+1$.

Assumption 9.1 *Observations are i.i.d. across i.*

Assumption 9.2 *$f_x(x)$, $E(g_{1i}|x)$, and $E(g_{2i}|x)$ are bounded away from zero over the support of X.*

Assumption 9.3 *The kernel functions $K(v)$, $K(x)$, and $K(x,v)$ and their first derivatives have supports that are convex and bounded on $\mathbb{R}^1$, $\mathbb{R}^k$, and $\mathbb{R}^{k+1}$ respectively. Each kernel function integrates to one over its support, is symmetric around zero, and has order p, i.e., for $K(x)$,*

$$\int_{\mathbb{R}^k} x^{\mathbf{m}_k} K(x)dx = 0 \quad for\ |\mathbf{m}_k| < p,$$

$$\int_{\mathbb{R}^k} x^{\mathbf{m}_k} K(x)dx \neq 0 \quad for\ some\ |\mathbf{m}_k| = p,$$

and $\int K(x)^2 dx$, $\int_{\mathbb{R}^k} |x^{\mathbf{m}_k}| K(x)dx$ for $|\mathbf{m}_k| = p$ are finite. This similarly holds for $K(v)$ and $K(x,v)$.

Assumption 9.4 *Let $s_{1i} \equiv \frac{D_i I_{\tau i} Y_i}{f_{xv}(x_i,v_i)}$, $s_{2i} \equiv \frac{D_i I_{\tau i} Y_i f_x(x_i)}{f_{xv}^2(x_i,v_i)}$, $s_{3i} \equiv \frac{D_i I_{\tau i}}{f_{xv}(x_i,v_i)}$, $s_{4i} \equiv \frac{D_i I_{\tau i} f_x(x_i)}{f_{xv}^2(x_i,v_i)}$, $s_{5i} \equiv \frac{(1-D_i) I_{\tau i} Y_i}{f_{xv}(x_i,v_i)}$, $s_{6i} \equiv \frac{(1-D_i) I_{\tau i} Y_i f_x(x_i)}{f_{xv}^2(x_i,v_i)}$, $s_{7i} \equiv \frac{(1-D_i) I_{\tau i}}{f_{xv}(x_i,v_i)}$, $s_{8i} \equiv \frac{(1-D_i) I_{\tau i} f_x(x_i)}{f_{xv}^2(x_i,v_i)}$. Then for each s_{ji}, $j = 1, ..., 8$, $f_x(x)$, $f_{xv}(x,v)$ satisfy the Lipschitz condition that there exists some positive numbers $M_1, ...M_{10}$, such that*

$$|E(s_{ji}|x+e_x) - E(s_{ji}|x)| \leq M_j \|e_x\|, \; j = 1, ..., 8$$

$$\begin{aligned} |f_x(x+e_x) - f_x(x)| &\leq M_9 \|e_x\|, \\ |f_{xv}(x+e_x, v+e_v) - f_x(x,v)| &\leq M_{10} \|(e_x, e_v)\|. \end{aligned}$$

$E(s_{ji}|x_i)$, $j = 1, ..., 8$, f_x, f_{xv} are p-th order differentiable and the p-th order derivatives are bounded. The p-th order derivatives of f_x, f_{xv} also satisfy the Lipschitz condition. The second moment of $q_i(x)$ (defined in equation 10.6) exists.

Assumption 9.5 *$E(D_{it} I_{\tau it} Y_{it}|v)$, $E[(1-D_{it}) I_{\tau it} Y_{it}|v]$, $f_{v_t}(v)$ are p times continuous differentiable in v, and the p-th order derivatives are bounded. Second moments of $\frac{D_{it} I_{\tau it} Y_{it}}{f_{v_t}(v_{it})}$, $\frac{D_{it} I_{\tau it}}{f_{v_t}(v_{it})}$, $\frac{(1-D_{it}) I_{\tau it} Y_{it}}{f_{v_t}(v_{it})}$, and $\frac{(1-D_{it}) I_{\tau it}}{f_{v_t}(v_{it})}$ are bounded.*

Table 1: Summary Statistics of the US Dataset

	MEAN	SD	LQ	MED	UQ
Competition	0.76	0.11	0.70	0.76	0.83
Innovation	5.53	9.98	0.22	1.59	5.77
Source-weighted Interest Rate	0.91	0.23	0.79	0.87	0.99

Note: MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper).

Table 2A: Empirical ATE Estimates

	Right Threshold	Left Threshold	Trim-ATE	No-Trim-ATE	Naive-ATE	ML-ATE
Case 1	25%(0.70)	75%(0.83)	−3.90 (0.61)	−4.25 (0.75)	−1.89 (0.27)	−1.85 (0.39)
Case 2	33%(0.72)	67%(0.80)	−3.27 (0.52)	−3.47 (0.66)	−1.67 (0.26)	−1.69 (0.37)
Case 3	10%(0.63)	90%(0.89)	−2.77 (0.98)	−2.75 (1.10)	−1.95 (0.29)	−4.40 (3.48)
Case 4	20%(0.68)	80%(0.85)	−4.25 (0.71)	−4.62 (0.86)	−2.22 (0.28)	−2.12 (0.43)
Case 5	30%(0.71)	70%(0.82)	−3.54 (0.54)	−3.95 (0.68)	−1.83 (0.26)	−1.81 (0.37)
Case 6	40%(0.74)	60%(0.79)	−2.49 (0.54)	−2.58 (0.67)	−1.18 (0.25)	−1.48 (0.39)

Notes: Right Threshold and Left Threshold are the α_0 and α_1 in equation (5.5) respectively. The first value is the percentage of competition set for the thresholds, with corresponding value of competition in the parenthesis. Four different estimates are reported here, with standard errors in parenthesis. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|D=1) - E(Y_0|D=0)$. ML-ATE is Heckman's selection MLE.

Table 2B: Empirical ATE Estimates - Robustness Checks

	Right Threshold	Left Threshold	Trim-ATE	No-Trim-ATE	Naive-ATE	ML-ATE
Case 1	25%	75%	-4.02 (0.63)	-4.29 (0.80)	-2.04 (0.27)	-2.02 (0.39)
Case 2	33%	67%	-3.46 (0.53)	-4.02 (0.66)	-1.81 (0.26)	-4.46 (0.64)
Case 3	10%	90%	-3.05 (1.06)	-2.98 (1.20)	-2.26 (0.29)	-4.51 (3.00)
Case 4	20%	80%	-4.98 (0.74)	-5.03 (0.93)	-2.75 (0.28)	-2.69 (0.44)
Case 5	30%	70%	-3.62 (0.56)	-3.86 (0.70)	-1.86 (0.26)	-5.95 (0.50)
Case 6	40%	60%	-2.41 (0.57)	-2.99 (0.67)	-0.99 (0.26)	-0.97 (0.44)

Notes: Right Threshold and Left Threshold are the α_0 and α_1 in equation (5.5) respectively. Four different estimates are reported here, with standard errors in parenthesis. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|D=1) - E(Y_0|D=0)$. ML-ATE is Heckman's selection MLE.

Table 3A: P-Values of the Large Support Assumption Tests

Panel A: The whole sample						Panel B: The truncated sample					
Year	Left	Right	Year	Left	Right	Year	Left	Right	Year	Left	Right
1976	0.002***	0.000***	1977	0.029**	0.010***	1976	0.001***	0.006***	1977	0.000***	0.012**
1978	0.000***	0.000***	1979	0.000***	0.000***	1978	1.000	0.000***	1979	0.000***	0.000***
1980	0.242	0.000***	1981	1.000	0.000***	1980	0.000***	0.000***	1981	1.000	0.000***
1982	1.000	1.000	1983	0.722	0.000***	1982	1.000	0.000***	1983	0.971	0.000***
1984	0.000***	0.000***	1985	0.000***	0.000***	1984	0.000***	0.000***	1985	0.000***	0.000***
1986	0.000***	0.000***	1987	0.755	0.010***	1986	0.000***	0.008***	1987	0.172	0.000***
1988	0.927	0.000***	1989	0.000***	0.000***	1988	1.000	0.000***	1989	0.000***	0.000***
1990	0.000***	0.000***	1991	0.000***	0.000***	1990	0.000***	0.000***	1991	0.000***	0.000***
1992	0.000***	0.000***	1993	1.000	0.005***	1992	0.000***	0.000***	1993	0.000***	0.000***
1994	1.000	1.000	1995	0.000***	1.000	1994	0.000***	1.000	1995	0.000***	1.000
1996	0.853	1.000	1997	0.000***	0.826	1996	1.000	1.000	1997	1.000	0.986
1998	0.000***	0.000***	1999	0.000***	0.878	1998	0.000***	0.000***	1999	0.000***	1.000
2000	0.000***	0.316	2001	0.000***	0.000***	2000	0.468	0.092*	2001	0.001***	0.000***

Notes: ***, **, * denote the cases when P-values are less than 0.01, 0.05, 0.10 respectively.

Table 3B: Ordered Choice Estimates (Identification at Infinity)

Trimming Parameter	50%	40%	30%	20%	10%	5%
Panel A: Define treatment from the whole sample						
$E(W_0)$	10.17 (0.97)	11.35 (1.17)	12.78 (1.46)	15.56 (1.98)	17.40 (2.87)	26.94 (4.44)
$E(W_2)$	5.86 (0.50)	5.96 (0.58)	5.94 (0.67)	5.94 (0.88)	4.56 (0.76)	5.06 (0.98)
Panel B: Define treatment separately each year (robustness check)						
$E(W_0)$	9.61 (0.95)	10.55 (1.17)	11.67 (1.41)	15.05 (2.08)	14.85 (2.42)	18.90 (4.53)
$E(W_2)$	6.55 (0.49)	6.39 (0.53)	6.54 (0.62)	6.07 (0.77)	5.79 (0.95)	7.96 (1.76)

Notes: The corresponding estimate of $E(W_1)$ from earlier is 4.33 with standard error 1.14. These estimates are obtained from the identification at infinity, with standard deviation in parentheses. The choice of the trimming parameters is based on the specified percentages of uncensored observations.

Table 4: Monte Carlo results matching the empirical data

	MEAN(−3.9)	SD	LQ	MED	UQ	RMSE	MAE	MDAE	%2SE
Panel A: Symmetric setting with normal errors									
Trim-ATE	−3.90	0.43	−4.19	−3.90	−3.61	0.43	0.34	0.00	1.00
No-Trim-ATE	−3.90	1.22	−4.67	−3.92	−3.12	1.22	0.95	0.02	1.00
Naive-ATE	−3.90	0.32	−4.11	−3.90	−3.68	0.32	0.25	0.00	1.00
ML-ATE	−3.90	0.30	−4.10	−3.90	−3.70	0.30	0.24	0.00	1.00
Panel B: Symmetric setting with uniform errors									
Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
No-Trim-ATE	−3.90	0.38	−4.16	−3.90	−3.64	0.38	0.31	0.00	1.00
Naive-ATE	−3.90	0.38	−4.16	−3.90	−3.65	0.38	0.30	0.00	1.00
ML-ATE	−3.91	0.38	−4.17	−3.90	−3.65	0.38	0.30	0.00	1.00
Panel C: Asymmetric setting with normal errors									
Trim-ATE	−3.21	0.51	−3.55	−3.21	−2.87	0.86	0.73	0.69	0.95
No-Trim-ATE	−3.65	1.33	−4.50	−3.65	−2.81	1.35	1.06	0.25	0.77
Naive-ATE	−1.99	0.34	−2.21	−2.00	−1.77	1.94	1.91	1.90	0.15
ML-ATE	−1.98	0.35	−2.22	−1.98	−1.75	1.95	1.92	1.92	0.15
Panel D: Asymmetric setting with uniform errors									
Trim-ATE	−3.45	0.48	−3.77	−3.45	−3.12	0.66	0.54	0.45	0.99
No-Trim-ATE	−3.76	1.08	−4.47	−3.76	−3.06	1.09	0.86	0.14	0.85
Naive-ATE	−1.84	0.37	−2.08	−1.84	−1.59	2.10	2.06	2.06	0.09
ML-ATE	−2.07	0.39	−2.34	−2.07	−1.81	1.87	1.83	1.83	0.25

Note: True E(Y_1)−E(Y_0) = −3.9. Parameters set (θ_0, θ_1, θ_{01}, θ_{02}, θ_{11}, θ_{12}, θ_2) for the four MC in order are as follows: (6.94 3.04 5.64 8.44 6.71 4.87 1.06), (6.97 3.07 23.67 −24.30 22.62 25.72 1.07), (6.67 2.77 6.57 −2.91 4.51 −5.43 0.43), (7.41 3.51 8.43 −4.27 5.47 −1.47 0.55). Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|D=1) - E(Y_0|D=0)$. ML-ATE is Heckman's selection MLE. All statistics are for the simulation estimates. MEAN = mean. SD = standard errors. LQ = 25% quantile (lower). MED = 50% quantile (median). UQ = 75% quantile (upper). RMSE = root mean square errors. MAE = mean absolute errors. MDAE = median absolute errors. %2SE = percentage of simulations in which the true coefficient was within two estimated standard errors of the estimated coefficient.

Table 5: Robustness checks: Monte Carlo with normal errors

	Quadratic			Step		
	MEAN (≈ -3.9)	SD	RMSE	MEAN (-3.9)	SD	RMSE
Panel A: $\kappa_1 = 0.02$, Noise Ratio = 0.19						
Trim-ATE	-4.23	0.46	0.49	-3.19	0.41	0.82
No-Trim-ATE	-7.79	1.57	4.20	-3.31	1.04	1.20
Naive-ATE	-3.75	0.38	0.39	-3.14	0.34	0.83
ML-ATE	-3.67	0.73	0.76	-3.10	0.66	1.04
Control Function	-3.74	0.24	0.31	-1.38	0.20	2.52
Panel B: $\kappa_1 = 0.03$, Noise Ratio = 0.28						
Trim-ATE	-4.08	0.42	0.42	-2.85	0.42	1.11
No-Trim-ATE	-7.68	1.61	4.11	-2.96	1.11	1.46
Naive-ATE	-3.60	0.37	0.47	-2.79	0.34	1.16
ML-ATE	-3.54	0.74	0.82	-2.74	0.64	1.33
Control Function	-3.59	0.23	0.41	-1.33	0.21	2.58
Panel C: $\kappa_1 = 0.04$, Noise Ratio = 0.36						
Trim-ATE	-3.93	0.48	0.48	-2.55	0.42	1.41
No-Trim-ATE	-7.63	1.64	4.07	-2.66	1.09	1.65
Naive-ATE	-3.40	0.38	0.62	-2.45	0.34	1.49
ML-ATE	-3.33	0.66	0.87	-2.42	0.59	1.60
Control Function	-3.40	0.26	0.59	-1.26	0.20	2.62

Note: True mean value is -3.9. Noise ratio is defined as the ratio of standard deviation of c_e to the standard deviation of $c*$. The first three and last three columns are the results when the true response forms are quadratic and step function respectively. Five different estimators are reported here. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|D=1) - E(Y_0|D=0)$. ML-ATE is Heckman's selection MLE. Control function approach is defined as in the paper. MEAN = mean. SD = standard errors. RMSE = root mean square errors.

Table 6: Robustness checks: Monte Carlo with uniform errors

	Quadratic			Step		
	MEAN (≈ -3.9)	SD	RMSE	MEAN (-3.9)	SD	RMSE
Panel A: $\kappa_2 = 0.06$, Noise Ratio = 0.17						
Trim-ATE	-3.86	0.36	0.36	-3.23	0.34	0.76
No-Trim-ATE	-3.96	0.36	0.36	-3.23	0.34	0.75
Naive-ATE	-3.79	0.34	0.35	-3.24	0.34	0.74
ML-ATE	-3.54	1.76	1.79	-3.23	0.51	0.84
Control Function	-3.71	0.25	0.31	-1.87	0.23	2.04
Panel B: $\kappa_2 = 0.07$, Noise Ratio = 0.19						
Trim-ATE	-3.83	0.35	0.35	-3.13	0.34	0.84
No-Trim-ATE	-3.92	0.35	0.35	-3.14	0.33	0.83
Naive-ATE	-3.76	0.35	0.37	-3.13	0.34	0.84
ML-ATE	-3.46	1.87	1.92	-3.10	0.55	0.97
Control Function	-3.65	0.25	0.35	-1.84	0.23	2.07
Panel C: $\kappa_2 = 0.08$, Noise Ratio = 0.22						
Trim-ATE	-3.79	0.36	0.37	-3.04	0.33	0.91
No-Trim-ATE	-3.88	0.36	0.36	-3.05	0.33	0.91
Naive-ATE	-3.70	0.35	0.39	-3.02	0.33	0.94
ML-ATE	-3.40	1.85	1.91	-3.02	0.53	1.03
Control Function	-3.59	0.25	0.40	-1.82	0.23	2.10

Note: True mean value is -3.9. Noise ratio is defined as the ratio of standard deviation of c_e to the standard deviation of $c*$. The first three and last three columns are the results when the true response forms are quadratic and step function respectively. Five different estimators are reported here. Trim-ATE and No-Trim-ATE are our proposed estimator with and without trimming (2%) respectively. Naive-ATE is an estimate for $E(Y_1|D=1) - E(Y_0|D=0)$. ML-ATE is Heckman's selection MLE. Control function approach is defined as in the paper. MEAN = mean. SD = standard errors. RMSE = root mean square errors.

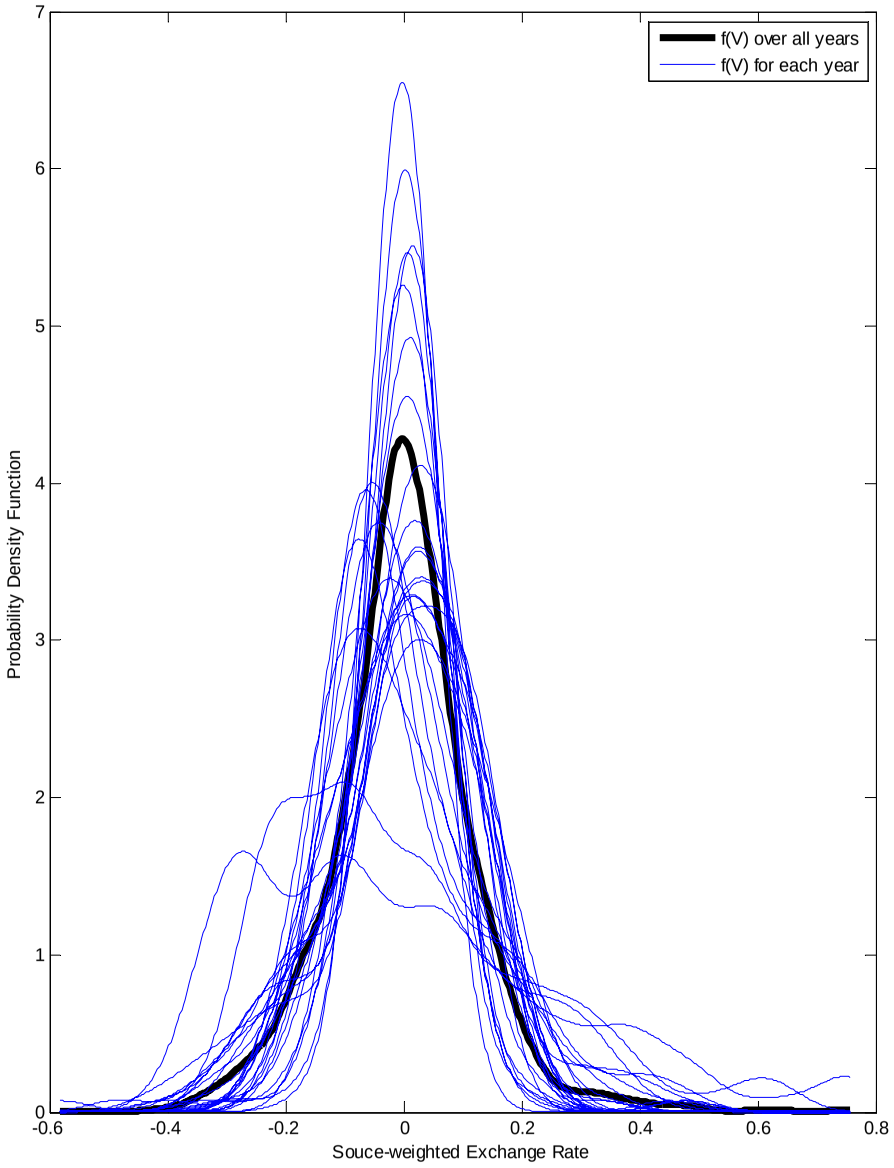

Figure 1: Distribution of V

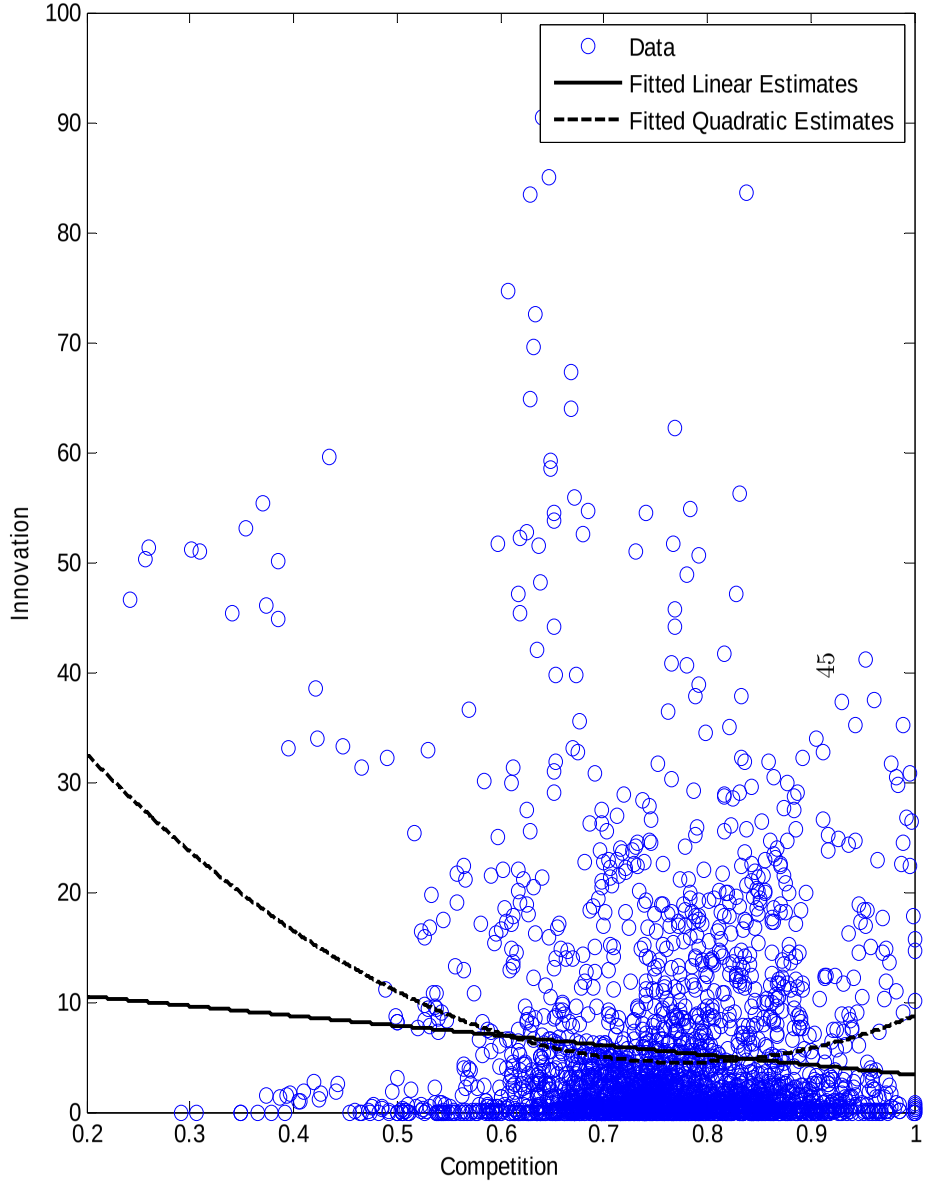

Figure 2: Competition and Innovation

Identifying the Average Treatment Effect in Ordered Treatment Models Without Unconfoundedness - Supplemental Appendix[7]

Arthur Lewbel — Boston College
Thomas Tao Yang — Australian National University

This online appendix provides proofs of Theorem 3.3, 3.4, 8.2 in Section 10.1, proofs of Theorem 4.1, 4.2, 4.3 in Section 10.2, and proofs of Theorem 8.1, 8.4 in Section 10.3.

Remark 10.1 (Uniform Convergence) Based on Silverman (1978), we have the uniform convergence of $\widehat{f}_{xv}(x,v)$ and $\widehat{f}_x(x)$ over a compact set of (V,X) and X respectively. We can apply these results to our estimator (3.5) as follows. For the estimation at x, an interior point of the support of X, we use a kernel function K with bounded support, so those x_i outside of a small interval around x will have zero weights. When h is small enough, all x_i with non-zero weights eventually fall into the compact set for which we have uniform convergence. For the estimation at v, we include a fixed trimming indicator in the associated identification theorem. By selecting a compact set that strictly covers the one where the trimming indicator is nonzero, we can again apply the uniform convergence results for all v_i with nonzero weights.

Bearing the above in mind, to reduce notation in this supplemental appendix we suppress the trimming indicators $I_{\tau i}$ and $I_{\tau it}$. In these proofs, we use R to denote any set of residual terms that are shown to be asymptotically negligible for our derived limiting distributions.

10.1 Proof of Theorem 3.3 and 3.4, and 8.2

Let $\widehat{h}_{1i} = \frac{D_i Y_i}{\widehat{f}(v_i|x_i)}$, $\widehat{g}_{1i} = \frac{D_i}{\widehat{f}(v_i|x_i)}$ where $\widehat{f}(v_i|x_i) = \frac{\widehat{f}_{xv}(x_i,v_i)}{\widehat{f}_x(x_i)}$, and both $\widehat{f}_x(x_i)$ and $\widehat{f}_{xv}(x_i,v_i)$ are standard leave-one-out nonparametric density estimator given by

$$\begin{aligned}\widehat{f}_x(x_i) &= \frac{1}{nh^k}\sum_{l=1,l\neq i}^{n} K\left(\frac{x_l - x_i}{h}\right),\\ \widehat{f}_{xv}(x_i,v_i) &= \frac{1}{nh^{k+1}}\sum_{l=1,l\neq i}^{n} K\left(\frac{x_l - x_i}{h}, \frac{v_l - v_i}{h}\right).\end{aligned}$$

Here h is the bandwidth and K is the kernel function. To simplify notation, without loss of generality we use the same h for each covariate. The kernel function K is defined in Assumption 9.3.

The sample counterpart estimate for $\psi_1(x)$ is then

$$\widehat{\psi}_1(x) = \frac{\widehat{E}\left(\widehat{h}_{1i}\middle| x\right)}{\widehat{E}\left(\widehat{g}_{1i}\middle| x\right)}, \tag{10.1}$$

[7] Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/ lewbel/. Thomas Tao Yang, Research School of Economics, Australian National University, ACT 0200, Australia. tao.yang@anu.edu.au.

where $\widehat{E}$ denotes the standard kernel nonparametric estimation:

$$\begin{aligned}\widehat{E}\left(\widehat{h}_{1i}\middle|x\right) &= \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{h}_{1i}K\left(\frac{x_i-x}{h}\right)\Bigg/\left[\frac{1}{nh^k}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)\right],\\ \widehat{E}\left(\widehat{g}_{1i}|\,x\right) &= \frac{1}{nh^k}\sum_{i=1}^{n}\widehat{g}_{1i}K\left(\frac{x_i-x}{h}\right)\Bigg/\left[\frac{1}{nh^k}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)\right].\end{aligned}$$

For simplicity, we abuse the notation a bit by defining

$$\widetilde{h}_{1i}\equiv h_{1i}f_x(x_i),\ \widetilde{g}_{1i}\equiv g_{1i}f_x(x_i) \tag{10.2}$$

and $\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)$ and $\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)$ are defined as the numerators in $\widehat{E}\left(\widehat{h}_{1i}\middle|x\right)$ and $\widehat{E}\left(\widehat{g}_{1i}|\,x\right)$ respectively:

$$\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)\equiv\frac{1}{nh^k}\sum_{i=1}^{n}\widehat{h}_{1i}K\left(\frac{x_i-x}{h}\right), \tag{10.3}$$

$$\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)\equiv\frac{1}{nh^k}\sum_{i=1}^{n}\widehat{g}_{1i}K\left(\frac{x_i-x}{h}\right). \tag{10.4}$$

It follows from the definition of $\widetilde{h}_{1i}$ and $\widetilde{g}_{1i}$ that

$$E\left(\widetilde{h}_{1i}\middle|x\right)=E\left(h_{1i}|\,x\right)f_x\left(x\right)\ \text{and}\ E\left(\widetilde{g}_{1i}|\,x\right)=E\left(g_{1i}|\,x\right)f_x\left(x\right),$$

and $\widehat{\psi}_1(x)=\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}|x\right)}$.

Replacing the subscript 1 with 2, similarly define $\widehat{\psi}_2(x)$, $\widehat{E}\left(\widehat{h}_{2i}\middle|x\right)$, $\widehat{E}\left(\widehat{g}_{2i}|\,x\right)$, $\widetilde{h}_{2i}$, $\widetilde{g}_{2i}$, $\widehat{\widetilde{h}}_{2i}$, $\widehat{\widetilde{g}}_{2i}$, $\widehat{E}\left(\widehat{\widetilde{h}}_{2i}\middle|x_i\right)$, $\widehat{E}\left(\widehat{\widetilde{g}}_{2i}\middle|x_i\right)$. The resulting estimator is then

$$\begin{aligned}\widehat{\psi}_1(x)-\widehat{\psi}_2(x) &= \frac{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{D_iY_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{D_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)}-\frac{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n}\frac{1-D_i}{\widehat{f}(v_i|x_i)}K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\widehat{E}\left(\widehat{h}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{g}_{1i}|\,x\right)}-\frac{\widehat{E}\left(\widehat{h}_{2i}\middle|x\right)}{\widehat{E}\left(\widehat{g}_{2i}|\,x\right)}=\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)}-\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{2i}\middle|x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{2i}\middle|x\right)}.\end{aligned} \tag{10.5}$$

Define the following term for the influence function of $\widehat{\psi}_1(x_i)-\widehat{\psi}_2(x_i)$:

$$
\begin{aligned}
q_i(x) \equiv & \left(\frac{h_{1i}}{E(\widetilde{g}_{1i}|x)} + \frac{E(h_{1i}|x_i)}{E(\widetilde{g}_{1i}|x)} - \frac{E(h_{1i}|x_i,v_i)}{E(\widetilde{g}_{1i}|x)} - \frac{E\left(\widetilde{h}_{1i}\middle|x\right) g_{1i}}{E(\widetilde{g}_{1i}|x)^2} - \frac{E\left(\widetilde{h}_{1i}\middle|x\right) E(g_{1i}|x_i)}{E(\widetilde{g}_{1i}|x)^2} \right. \\
& \left. + \frac{E\left(\widetilde{h}_{1i}\middle|x\right) E(g_{1i}|x_i,v_i)}{E(\widetilde{g}_{1i}|x)^2} \right) - \left(\frac{h_{2i}}{E(\widetilde{g}_{2i}|x)} + \frac{E(h_{2i}|x_i)}{E(\widetilde{g}_{2i}|x)} - \frac{E(h_{2i}|x_i,v_i)}{E(\widetilde{g}_{2i}|x)} - \frac{E\left(\widetilde{h}_{2i}\middle|x\right) g_{2i}}{E(\widetilde{g}_{2i}|x)^2} \right. \\
& \left. - \frac{E\left(\widetilde{h}_{2i}\middle|x\right) E(g_{2i}|x_i)}{E(\widetilde{g}_{2i}|x)^2} + \frac{E\left(\widetilde{h}_{2i}\middle|x\right) E(g_{2i}|x_i,v_i)}{E(\widetilde{g}_{2i}|x)^2} \right). \qquad (10.6)
\end{aligned}
$$

The bias term resulting from nonparametric regression is given by:

$$
\begin{aligned}
\mathbb{B}_p(x) \equiv & \frac{\mathbb{B}_{1,p}}{E(g_{1i}|x)} - \frac{\mathbb{B}_{2,p}}{E(g_{1i}|x)} - \frac{E(h_{1i}|x)\mathbb{B}_{3,p}}{E(g_{1i}|x)^2} + \frac{E(h_{1i}|x)\mathbb{B}_{4,p}}{E(g_{1i}|x)^2} \qquad (10.7) \\
& - \frac{\mathbb{B}_{5,p}}{E(g_{2i}|x)} + \frac{\mathbb{B}_{6,p}}{E(g_{2i}|x)} + \frac{E(h_{2i}|x)\mathbb{B}_{7,p}}{E(g_{2i}|x)^2} - \frac{E(h_{2i}|x)\mathbb{B}_{8,p}}{E(g_{2i}|x)^2},
\end{aligned}
$$

where $\mathbb{B}_{j,p}$, $j = 1, ..., 8$, are defined as follows.

$$
\begin{aligned}
\mathbb{B}_{1,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[D_i Y_i / f_{xv}(x_i,v_i) D^{\mathbf{m}_k} f_x(x_i)|\, x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K(u_l)\, du_l, \qquad (10.8) \\
\mathbb{B}_{2,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[D_i Y_i f_x(x_i)/f_{xv}^2(x_i,v_i) D^{\mathbf{m}_{k+1}} f_{xv}(x_i,v_i)\middle|\, x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l, \\
\mathbb{B}_{3,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[D_i / f_{xv}(x_i,v_i) D^{\mathbf{m}_k} f_x(x_i)|\, x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K(u_l)\, du_l, \\
\mathbb{B}_{4,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[D_i f_x(x_i)/f_{xv}^2(x_i,v_i) D^{\mathbf{m}_{k+1}} f_{xv}(x_i,v_i)\middle|\, x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l, \\
\mathbb{B}_{5,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[(1-D_i) Y_i / f_{xv}(x_i,v_i) D^{\mathbf{m}_k} f_x(x_i)|\, x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K(u_l)\, du_l, \\
\mathbb{B}_{6,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[(1-D_i) Y_i f_x(x_i)/f_{xv}^2(x_i,v_i) D^{\mathbf{m}_{k+1}} f_{xv}(x_i,v_i)\middle|\, x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l, \\
\mathbb{B}_{7,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[(1-D_i) / f_{xv}(x_i,v_i) D^{\mathbf{m}_k} f_x(x_i)|\, x\right]}{\mathbf{m}_k!} \int_{\mathbb{R}^k} u_l^{\mathbf{m}_k} K(u_l)\, du_l,, \\
\mathbb{B}_{8,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[(1-D_i) f_x(x_i)/f_{xv}^2(x_i,v_i) D^{\mathbf{m}_{k+1}} f_{xv}(x_i,v_i)\middle|\, x\right]}{\mathbf{m}_{k+1}!} \int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}} K(u_l)\, du_l,
\end{aligned}
$$

Lemma 10.1 *Assume we observe* $\underset{(k+1)\times 1}{W_i} = (\underset{k\times 1}{X_i}\ \underset{1\times 1}{V_i})$, $\underset{1\times 1}{s_i}$, $\underset{(k+2)\times 1}{Z_i} = (\underset{(k+1)\times 1}{W_i}\ \underset{1\times 1}{s_i})$,, *which are i.i.d. across i. The density functions* f_x, f_w *for* X *and* W *are bounded.* f_x *and* f_w *are p-th order differentiable, and p-th order derivatives are bounded.* $E(s_i|w_i)$, f_x, f_w *satisfy the Lipschitz condition*

$$
|E(s_i|w_i + e_w) - E(s_i|w_i)| \leq M_1 \|e_w\|,
$$

$$|f_x(x_i + e_x) - f_x(x_i)| \leq M_2 \|e_x\|,$$
$$|f_w(w_i + e_w) - f_w(w_i)| \leq M_3 \|e_w\|$$

for some positive M_1, M_2, M_3. Under the above assumptions, when x is an interior point of X, we have

$$\widehat{E}\left(s_i \widehat{f}_w(w_i)\middle| x\right)\widehat{f}_x(x) \tag{10.9}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^k}s_i\left[\frac{1}{n-1}\sum_{l=1,l\neq i}^{n}\frac{1}{h^{k+1}}K\left(\frac{w_l - w_i}{h}\right)\right]K\left(\frac{x_i - x}{h}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right)f_w(w_i) + \frac{E(s_i|w_i)}{h^k}K\left(\frac{x_i - x}{h}\right)f_w(w_i)\right.$$
$$\left. -2E\left(\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right)f_w(w_i)\right)\right] + E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i - x}{h}\right)K(\frac{w_l - w_i}{h})\right] + o_P\left(\frac{1}{\sqrt{nh^k}}\right) \tag{10.10}$$

and

$$\widehat{E}\left(s_i \widehat{f}_x(x_i)\middle| x\right)\widehat{f}_x(x) \tag{10.11}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^k}s_i\left[\frac{1}{n-1}\sum_{l=1,l\neq i}^{n}\frac{1}{h^{k+1}}K\left(\frac{x_l - x_i}{h}\right)\right]K\left(\frac{x_i - x}{h}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right)f_x(x_i) + \frac{E(s_i|x_i)}{h^k}K\left(\frac{x_i - x}{h}\right)f_x(x_i)\right.$$
$$\left. -2E\left(\frac{s_i}{h^k}K\left(\frac{x_i - x}{h}\right)f_x(x_i)\right)\right] + E\left[\frac{s_i}{h^{2k}}K\left(\frac{x_i - x}{h}\right)K(\frac{x_l - x_i}{h})\right] + o_P\left(\frac{1}{\sqrt{nh^k}}\right). \tag{10.12}$$

Proof of Lemma 10.1. Consider first the following term,

$$\frac{1}{n(n-1)h^{2k+1}}\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}s_iK\left(\frac{x_i - x}{h}\right)K\left(\frac{w_l - w_i}{h}\right) \tag{10.13}$$
$$= \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n}\frac{1}{2}\left[s_iK\left(\frac{x_i - x}{h}\right) + s_lK\left(\frac{x_l - x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l - w_i}{h}\right).$$

Let

$$P_1(z_i, z_l) = \frac{1}{2}\left[s_iK\left(\frac{x_i - x}{h}\right) + s_lK\left(\frac{x_l - x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l - w_i}{h}\right).$$

Then equation (10.13) becomes

$$\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n}P_1(z_i, z_l). \tag{10.14}$$

Following Powell et al. (1989), we first verify that $E\left[P_1(z_i,z_l)^2\right] = o_p(n)$.

$$
\begin{aligned}
& E\left[P_1(z_i,z_l)^2\right] \\
= & \int\int_{\Omega_{w_i,w_l}} E\left\{ \left[\frac{1}{2}\left[s_i K\left(\frac{x_i-x}{h}\right) + s_l K\left(\frac{x_l-x}{h}\right)\right]\frac{1}{h^{2k+1}}K\left(\frac{w_l-w_i}{h}\right)\right]^2 \middle| w_i, w_l \right\} \\
& f_w(w_i) f_w(w_l) \mathrm{d}w_i \mathrm{d}w_l. \\
= & \int\int_{\Omega_{u_i,u_l}} \frac{1}{h^{2k+1}} E\left\{ \left[\frac{1}{2}\left[s_i K\left(u_i\right) + s_l K\left(u_i + h u_l\right)\right] K(u_l)\right]^2 \middle| (x+hu_i, v_i), (x+hu_i+hu_l, v_i+hu_l) \right\} \\
& f_w(x+hu_i, v_i) f_w\left(x+hu_i+hu_l, v_i+hu_l\right) \mathrm{d}u_i \mathrm{d}v_i \mathrm{d}u_l \\
= & O_p\left(\frac{1}{h^{2k+1}}\right) = o_p(n),
\end{aligned}
$$

where the second equality holds by the change of variables $u_l = \frac{w_l - w_i}{h}$, $u_i = \frac{x_i - x}{h}$, the third equality holds by the bounds conditions, and the last equality holds by the assumption that $nh^{2k+1} \to \infty$. According to Lemma 3.2 in Powell et al. (1989), equation (10.14) is equal to

$$
E\left[P_1(z_i,z_l)\right] + \frac{2}{n}\sum_{i=1}^{n}\left\{E\left[P_1(z_i,z_l)|z_i\right] - E\left[P_1(z_i,z_l)\right]\right\} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{10.15}
$$

The term inside the summation in equation (10.15) has the following form:

$$
\begin{aligned}
& E\left[P_1(z_i,z_l)|z_i\right] - E\left[P_1(z_i,z_l)\right] \\
= & \frac{1}{2}E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] + \frac{1}{2}E\left[\frac{s_l}{h^{2k+1}}K\left(\frac{x_l-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
& -E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\right].
\end{aligned}
$$

Since

$$
\begin{aligned}
& E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
= & \int_{\Omega_{w_l}} \frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right) f_w(w_l)\mathrm{d}w_l \\
= & \frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right) f_w(w_i) + \frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\int_{\Omega_{u_l}} K\left(u_l\right)\left[f_w(w_i+hu_l) - f_w(w_i)\right]\mathrm{d}u_l,
\end{aligned}
$$

and similarly

$$
\begin{aligned}
& E\left[\frac{s_l}{h^{2k+1}}K\left(\frac{x_l-x}{h}\right)K\left(\frac{w_i-w_l}{h}\right)\middle| z_i\right] \\
= & \frac{E\left[s_i|w_i\right]}{h^k}K\left(\frac{x_i-x}{h}\right) f_w(w_i) + \frac{1}{h^k}\int_{\Omega_{u_l}}\left[E\left[s_i|w_i+hu_l\right]K\left(\frac{x_i+hu_l-x}{h}\right) f_w(w_i+hu_l)\right. \\
& \left. -E\left[s_i|w_i\right]K\left(\frac{x_i-x}{h}\right) f_w(w_i)\right] K\left(u_l\right)\mathrm{d}u_l,
\end{aligned}
$$

the following holds

$$\begin{aligned} & E\left[P_1(z_i,z_l)|z_i\right]-E\left[P_1(z_i,z_l)\right] \\ = & \frac{1}{2}\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)+\frac{1}{2}\frac{E\left[s_i|w_i\right]}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)-E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w(w_i)\right] \\ & +R_{1i}-E\left(R_{1i}\right), \end{aligned} \tag{10.16}$$

where

$$\begin{aligned} R_{1i} = & \frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\int K\left(u_l\right)\left[f_w(w_i+hu_l)-f_w(w_i)\right]\mathrm{d}u_l \\ & +\frac{1}{h^k}\int_{\Omega_{u_l}}\left[E\left[s_i|w_i+hu_l\right]K\left(\frac{x_i+hu_l-x}{h}\right)f_w(w_i+hu_l)\right. \\ & \left.-E\left[s_i|w_i\right]K\left(\frac{x_i-x}{h}\right)f_w(w_i)K\left(u_l\right)\mathrm{d}u_l\right] \end{aligned} \tag{10.17}$$

and, since $E(s_i|w_i)$, f_x, f_w satisfy the Lipschitz condition, $E\left(R_{1i}^2\right)=o_p\left(\frac{1}{h^k}\right)$. So

$$\frac{1}{n}\sum_{i=1}^{n}\left[R_{1i}-E\left(R_{1i}\right)\right]=o_p\left(\frac{1}{\sqrt{nh^k}}\right). \tag{10.18}$$

By the fact that $p(z_i,z_l)$ is symmetric for z_i,z_l, we have

$$E\left[P_1(z_i,z_l)\right]=E\left[s_iK\left(\frac{x_i-x}{h}\right)K(\frac{w_l-w_i}{h})\right].$$

From equation (10.15) (10.16) and (10.18), we have

$$\begin{aligned} & \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{l=i+1}^{n}P_1(z_i,z_l) \\ = & \frac{1}{n}\sum_{i=1}^{n}\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)+\frac{E\left(s_i|w_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)\right. \\ & \left.-2E\left(\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)f_w\left(w_i\right)\right)\right]+E\left[s_iK\left(\frac{x_i-x}{h}\right)K(\frac{w_l-w_i}{h})\right]+o_P\left(\frac{1}{\sqrt{nh^k}}\right), \end{aligned}$$

which implies the first part of the Theorem.

The second part holds by the same line of analysis after replacing W with X. ■

Lemma 10.2 *Adopt the same notation and assumptions as in Lemma 10.1, and assume D^pf_x and D^pf_w*

also satisfy the Lipschitz condition. Then

$$\begin{aligned} & E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_l-w_i}{h}\right)\right] \\ = \; & E\left[\frac{s_i f_w(w_i)}{h^k}K\left(\frac{x_i-x}{h}\right)\right]+\mathbb{S}_{1,p}f_x(x)+o(h^p) \qquad (10.19) \\ & E\left[\frac{s_i}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{x_l-x_i}{h}\right)\right] \\ = \; & E\left[\frac{s_i f_x(x_i)}{h^k}K\left(\frac{x_i-x}{h}\right)\right]+\mathbb{S}_{2,p}f_x(x)+o(h^p) \qquad (10.20) \end{aligned}$$

where

$$\begin{aligned} \mathbb{S}_{1,p} &\equiv h^p \sum_{|\mathbf{m}_{k+1}|=p} \frac{E\left[s_i D^{\mathbf{m}_{k+1}} f_w(w_i)|\, x\right]}{\mathbf{m}_{k+1}!}\int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}}K(u_l)\,du_l, \\ \mathbb{S}_{2,p} &\equiv h^p \sum_{|\mathbf{m}_k|=p} \frac{E\left[s_i D^{\mathbf{m}_k} f_w(x_i)|\, x\right]}{\mathbf{m}_k!}\int_{\mathbb{R}^k} u_l^{\mathbf{m}_k}K(u_l)\,du_l. \end{aligned}$$

Proof of Lemma 10.2.

$$\begin{aligned} & E\left[\frac{s_i}{h^{2k+1}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_l-w_i}{h}\right)\right] \\ = \; & \int_{\Omega_{\omega_i}}\int_{\Omega_{\omega_l}}\frac{1}{h^{k+1}}K\left(\frac{w_l-w_i}{h}\right)f_w(w_l)\,\mathrm{d}w_l E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\middle|\, w_i\right]f_w(w_i)\,\mathrm{d}w_i \\ = \; & B_1+E\left[\frac{s_i f_w(w_i)}{h^k}K\left(\frac{x_i-x}{h}\right)\right], \end{aligned}$$

where

$$B_1 \equiv \int_{\Omega_{\omega_i}}\int_{\Omega_{\omega_l}}\frac{1}{h^{k+1}}K\left(\frac{w_l-w_i}{h}\right)(f_w(w_l)-f_w(w_i))\,\mathrm{d}w_l E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\middle|\, w_i\right]f_w(w_i)\,\mathrm{d}w_i.$$

Then, doing the standard change of variables transformation $u_l=\frac{w_l-w_i}{h}$, we have

$$B_1 = h^p\sum_{|\mathbf{m}_{k+1}|=p}\int_{\Omega_{\omega_i}}\int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}}K(u_l)\frac{D^{\mathbf{m}_{k+1}}f_w(\widetilde{w}_i)}{\mathbf{m}_{k+1}!}du_l E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)\middle|\, w_i\right]f_w(w_i)\,\mathrm{d}w_i,$$

where $\widetilde{w}_i$ is some value between w_i and w_i+hu_l. Since the kernel has bounded support and $D^p f_w(\widetilde{w}_i)$ satisfies the Lipschitz condition, we have

$$\begin{aligned} B_1 &= h^p\sum_{|\mathbf{m}_{k+1}|=p}\int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}}K(u_l)\,du_l\int_{\Omega_{\omega_i}}\frac{E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)D^{\mathbf{m}_{k+1}}f_w(w_i)\middle|\, w_i\right]}{\mathbf{m}_{k+1}!}f_w(w_i)\,\mathrm{d}w_i+o(h^p) \\ &= h^p\sum_{|\mathbf{m}_{k+1}|=p}\int_{\mathbb{R}^{k+1}} u_l^{\mathbf{m}_{k+1}}K(u_l)\,du_l\frac{E\left[\frac{s_i}{h^k}K\left(\frac{x_i-x}{h}\right)D^{\mathbf{m}_{k+1}}f_w(w_i)\right]}{\mathbf{m}_{k+1}!}+o(h^p). \end{aligned}$$

Substituting this into B_1, we have

$$B_1=\mathbb{S}_{1,p}+o(h^p),$$

which is equation (10.19). The second conclusion can be proved similarly. ■

Corollary 10.3 *Under the same assumptions in Lemma 10.2 and assuming $E(s_i|x)$ and $E(s_i|w)$ are p-th order differentiable with bounded p-th order derivatives, we have*

$$\widehat{E}\left[\frac{s_i}{\widehat{f}(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) - E\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right] f_x(x) = O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right) + O_P\left(\frac{\log(n)}{nh^{k+1}}\right). \tag{10.21}$$

Proof of Corollary 10.3.

$$\begin{aligned}
&\widehat{E}\left[\frac{s_i}{\widehat{f}(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) - E\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right] f_x(x) \\
= \;& \widehat{E}\left[\frac{s_i}{\widehat{f}(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) - \widehat{E}\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) + \widehat{E}\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right]\left[\widehat{f}_x(x) - f_x(x)\right] \\
&+\left\{\widehat{E}\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right] - E\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right]\right\} f_x(x).
\end{aligned}$$

All terms except the first term are readily seen to be $O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right)$. For the first term

$$\begin{aligned}
&\widehat{E}\left[\frac{s_i}{\widehat{f}(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) - \widehat{E}\left[\frac{s_i}{f(v_i|x_i)}\middle| x\right]\widehat{f}_x(x) \\
= \;& \frac{1}{n}\sum_{i=1}^n \frac{s_i\widehat{f}_x(x_i)}{\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right) - \frac{1}{n}\sum_{i=1}^n \frac{s_i f_x(x_i)}{f_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right) \\
= \;& \frac{1}{n}\sum_{i=1}^n \frac{s_i\left[\widehat{f}_x(x_i) - f_x(x_i)\right]}{f_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right) && (10.22) \\
&+\frac{1}{n}\sum_{i=1}^n \frac{s_i f_x(x_i)\left[\widehat{f}_w(w_i) - f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right) && (10.23) \\
&+\frac{1}{n}\sum_{i=1}^n \frac{s_i\left[\widehat{f}_x(x_i) - f_x(x_i)\right]\left[\widehat{f}_w(w_i) - f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right) && (10.24) \\
&+\frac{1}{n}\sum_{i=1}^n \frac{s_i\widehat{f}_x(x_i)\left[\widehat{f}_w(w_i) - f_w(w_i)\right]^2}{f_w^2(w_i)\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i - x}{h}\right). && (10.25)
\end{aligned}$$

According the results in Lemma 10.1 and Lemma 10.2, equation (10.22) and (10.23) are $O_p(h^p) + O_p\left(\frac{1}{\sqrt{nh^k}}\right)$. From Silverman (1978) and Remark 10.1, we have

$$\sup_{K\left(\frac{x_i - x}{h}\right)\neq 0,}\left|\widehat{f}_x(x_i) - f_x(x_i)\right| = O_p\left[\sqrt{\frac{\log(n)}{nh^k}}\right], \tag{10.26}$$

$$\sup_{K\left(\frac{x_i - x}{h}\right)\neq 0,\ I_{\tau i}\neq 0}\left|\widehat{f}_w(w_i) - f_w(w_i)\right| = O_p\left[\sqrt{\frac{\log(n)}{nh^{k+1}}}\right]. \tag{10.27}$$

Then

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]\left[\widehat{f}_w(w_i)-f_w(w_i)\right]}{f_w^2(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\right| \\
\leq & \sup_{K\left(\frac{x_i-x}{h}\right)\neq 0,\ I_{\tau i}\neq 0}\left|\widehat{f}_x(x_i)-f_x(x_i)\right|\left|\widehat{f}_w(w_i)-f_w(w_i)\right|\frac{1}{n}\sum_{i=1}^{n}\left|\frac{s_i}{f_w^2(w_i)h^k}K\left(\frac{x_i-x}{h}\right)\right| \\
= & O_P\left(\frac{\log(n)}{nh^{k+1/2}}\right),
\end{aligned}
$$

and similarly, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{s_i\widehat{f}_x(x_i)\left[\widehat{f}_w(w_i)-f_w(w_i)\right]^2}{f_w^2(w_i)\widehat{f}_w(w_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)=O_P\left(\frac{\log(n)}{nh^{k+1}}\right)
$$

Therefore, we know that equation (10.24) and (10.25) are of the orders $O_P\left(\frac{\log(n)}{nh^{k+1/2}}\right)$ and $O_P\left(\frac{\log(n)}{nh^{k+1}}\right)$ respectively.

Combining the above results then proves the Corollary. ■

Proof of Theorem 3.3. We first derive the properties of $\widehat{\psi}_1(x)$. This can be divided into several components as follows

$$
\begin{aligned}
\widehat{\psi}_1(x)-\psi_1(x) &= \frac{\widehat{E}\left(\widehat{h}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{g}_{1i}|\,x\right)}=\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)} \\
&= \frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{E\left(\widetilde{g}_{1i}|\,x\right)}-\frac{E\left(\widetilde{h}_{1i}\middle|x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)}{E\left(\widetilde{g}_{1i}|\,x\right)^2}+R_2(x), \qquad (10.28)
\end{aligned}
$$

where

$$
R_2(x)\equiv\frac{\left[\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)-E\left(\widetilde{h}_{1i}\middle|x\right)\right]\left[\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)-E\left(\widetilde{g}_{1i}|\,x\right)\right]}{\left[E\left(\widetilde{g}_{1i}|\,x\right)\right]^2}+\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)\left[\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)-E\left(\widetilde{g}_{1i}|\,x\right)\right]^2}{\left[E\left(\widetilde{g}_{1i}|\,x\right)\right]^2}.
$$

According to Corollary 10.3, and the assumption that $\frac{1}{E(\widetilde{g}_{1i}|x)}$ is bounded, $R_2(x)$ is of order $o_P\left(\frac{1}{\sqrt{nh^k}}\right)$. So

$$
\widehat{\psi}_1(x)-\psi_1(x)=\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|x\right)}{E\left(\widetilde{g}_{1i}|\,x\right)}-\frac{E\left(\widetilde{h}_{1i}\middle|x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|x\right)}{\left[E\left(\widetilde{g}_{1i}|\,x\right)\right]^2}+o_P\left(\frac{1}{\sqrt{nh^k}}\right). \qquad (10.29)
$$

Notice that

$$\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|\, x\right)}{E\left(\widetilde{g}_{1i}|\, x\right)} = \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\widehat{f}(v_i|x_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \tag{10.30}$$

$$\begin{aligned} = \quad & \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_iY_i\widehat{f}_x(x_i)}{f_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)\right. \\ & \left.-\frac{D_iY_if_x(x_i)\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]}{f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)+R_{3i}\right\}, \end{aligned}$$

where

$$\begin{aligned} R_{3i} \quad \equiv \quad & \frac{D_iY_i\widehat{f}_x(x_i)\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]^2}{E\left(\widetilde{g}_{1i}|\, x\right)f_{xv}^2(x_i,v_i)\widehat{f}_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\ & -\frac{D_iY_i\left[\widehat{f}_x(x_i)-f_x(x_i)\right]\left[\widehat{f}_{xv}(x_i,v_i)-f_{xv}(x_i,v_i)\right]}{E\left(\widetilde{g}_{1i}|\, x\right)f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right). \end{aligned}$$

Following the same proof as in Corollary 10.3,

$$\frac{1}{n}\sum_{i=1}^{n}R_{3i} = O_p\left(\frac{\log(n)}{nh^{2k+1}}\right) = o_p\left(\frac{1}{\sqrt{nh^k}}\right). \tag{10.31}$$

Applying Lemma 10.1 to the first term in equation (10.30),

$$\frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i\widehat{f}_x(x_i)}{f_{xv}(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \tag{10.32}$$

$$\begin{aligned} = \quad & \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)+\frac{E\left(h_{1i}|x_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)\right. \\ & \left.-2E\left(\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)\right)\right]+E\left[\frac{D_iY_i}{f_{xv}(x_i,v_i)}\frac{1}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{x_l-x_i}{h}\right)\right]. \end{aligned}$$

By the same reasoning, the second component in equation (10.30) is

$$\frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_if_x(x_i)\widehat{f}_{xv}(x_i,v_i)}{f_{xv}^2(x_i,v_i)}\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \tag{10.33}$$

$$\begin{aligned} = \quad & \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)+\frac{E\left(h_{1i}|x_i,v_i\right)}{h^k}K\left(\frac{x_i-x}{h}\right)\right. \\ & \left.-2E\left(\frac{h_{1i}}{h^k}K\left(\frac{x_i-x}{h}\right)\right)\right]+E\left[\frac{D_iY_if_x(x_i)}{f_{xv}^2(x_i,v_i)}\frac{1}{h^{2k}}K\left(\frac{x_i-x}{h}\right)K\left(\frac{w_l-w_i}{h}\right)\right]. \end{aligned}$$

Substituting equation (10.32) and (10.33) back into equation (10.30) and using the results in Lemma 10.2,

we have

$$\begin{aligned}\frac{\widehat{E}\left(\widehat{\widetilde{h}}_{1i}\middle|\, x\right)}{E\left(\widetilde{g}_{1i}|\, x\right)} &= \frac{1}{E\left(\widetilde{g}_{1i}|\, x\right)}\frac{1}{n}\sum_{i=1}^{n}\left[h_{1i}+E\left(h_{1i}|x_i\right)-E\left(h_{1i}|x_i,v_i\right)\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\ &\quad +\frac{\mathbb{B}_{1,p}}{E\left(g_{1i}|\, x\right)}-\frac{\mathbb{B}_{2,p}}{E\left(g_{1i}|\, x\right)}+o_P\left(h^p\right). \end{aligned} \tag{10.34}$$

Applying the same strategy to the next term in equation (10.29), we get

$$\begin{aligned}\frac{E\left(\widetilde{h}_{1i}\middle|\, x\right)\widehat{E}\left(\widehat{\widetilde{g}}_{1i}\middle|\, x\right)}{\left[E\left(\widetilde{g}_{1i}|\, x\right)\right]^2} &= \frac{E\left(\widetilde{h}_{1i}\middle|\, x\right)}{E\left(\widetilde{g}_{1i}|\, x\right)^2}\frac{1}{n}\sum_{i=1}^{n}\left[g_{1i}+E\left(g_{1i}|x_i\right)-E\left(g_{1i}|x_i,v_i\right)\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\ &\quad +\frac{E\left(h_{1i}|\, x\right)\mathbb{B}_{3,p}}{E\left(g_{1i}|\, x\right)^2}-\frac{E\left(h_{1i}|\, x\right)\mathbb{B}_{4,p}}{E\left(g_{1i}|\, x\right)^2}+o_P\left(h^p\right) \end{aligned} \tag{10.35}$$

Substituting equation (10.34) and (10.35) into equation (10.29), we have

$$\begin{aligned}\widehat{\psi}_1(x)-\psi_1(x) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{1i}}{E\left(\widetilde{g}_{1i}|\, x\right)}+\frac{E\left(h_{1i}|x_i\right)}{E\left(\widetilde{g}_{1i}|\, x\right)}-\frac{E\left(h_{1i}|x_i,v_i\right)}{E\left(\widetilde{g}_{1i}|\, x\right)}-\frac{E\left(\widetilde{h}_{1i}\middle|\, x\right)g_{1i}}{E\left(\widetilde{g}_{1i}|\, x\right)^2}\right. \\ &\quad \left.-\frac{E\left(\widetilde{h}_{1i}\middle|\, x\right)E\left(g_{1i}|x_i\right)}{E\left(\widetilde{g}_{1i}|\, x\right)^2}+\frac{E\left(\widetilde{h}_{1i}\middle|\, x\right)E\left(g_{1i}|x_i,v_i\right)}{E\left(\widetilde{g}_{1i}|\, x\right)^2}\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\ &\quad +\frac{\mathbb{B}_{1,p}}{E\left(g_{1i}|\, x\right)}-\frac{\mathbb{B}_{2,p}}{E\left(g_{1i}|\, x\right)}-\frac{E\left(h_{1i}|\, x\right)\mathbb{B}_{3,p}}{E\left(g_{1i}|\, x\right)^2}+\frac{E\left(h_{1i}|\, x\right)\mathbb{B}_{4,p}}{E\left(g_{1i}|\, x\right)^2}+o_P\left(h^p\right)o_p\left(\frac{1}{\sqrt{nh^k}}\right). \end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{\psi}_2(x)-\psi_2(x) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h_{2i}}{E\left(\widetilde{g}_{2i}|\, x\right)}+\frac{E\left(h_{2i}|x_i\right)}{E\left(\widetilde{g}_{2i}|\, x\right)}-\frac{E\left(h_{2i}|x_i,v_i\right)}{E\left(\widetilde{g}_{2i}|\, x\right)}-\frac{E\left(\widetilde{h}_{2i}\middle|\, x\right)g_{2i}}{E\left(\widetilde{g}_{2i}|\, x\right)^2}\right. \\ &\quad \left.-\frac{E\left(\widetilde{h}_{2i}\middle|\, x\right)E\left(g_{2i}|x_i\right)}{E\left(\widetilde{g}_{2i}|\, x\right)^2}+\frac{E\left(\widetilde{h}_{2i}\middle|\, x\right)E\left(g_{2i}|x_i,v_i\right)}{E\left(\widetilde{g}_{2i}|\, x\right)^2}\right]\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right) \\ &\quad +\frac{\mathbb{B}_{5,p}}{E\left(g_{2i}|\, x\right)}-\frac{\mathbb{B}_{6,p}}{E\left(g_{2i}|\, x\right)}-\frac{E\left(h_{2i}|\, x\right)\mathbb{B}_{7,p}}{E\left(g_{2i}|\, x\right)^2}+\frac{E\left(h_{2i}|\, x\right)\mathbb{B}_{8,p}}{E\left(g_{2i}|\, x\right)^2}+o_P\left(h^p\right)+o_p\left(\frac{1}{\sqrt{nh^k}}\right). \end{aligned}$$

Putting these results together gives

$$\widehat{\psi}_1(x)-\widehat{\psi}_2(x)-\left(\psi_1(x)-\psi_2(x)\right)=\frac{1}{n}\sum_{i=1}^{n}q_i\left(x\right)\frac{1}{h^k}K\left(\frac{x_i-x}{h}\right)+\mathbb{B}_p\left(x\right)+o_P\left(h^p\right)+o_p\left(\frac{1}{\sqrt{nh^k}}\right),$$

which implies that

$$\frac{\sqrt{nh^k}}{var\left(q_i\left(x\right)|x\right)\int_{\mathbb{R}^k}K^2\left(u\right)\mathrm{d}u}\left[\widehat{\psi}_1(x)-\widehat{\psi}_2(x)-\left(\psi_1(x)-\psi_2(x)\right)-\mathbb{B}_p\left(x\right)\right]\xrightarrow{d}N\left(0,1\right)$$

■

Proof of Theorem 3.4. The first-order asymptotics of our estimator follow directly from Lemmas 10.4, 10.7, 10.8 and 10.9. The convergence rate of the resulting influence function can be seen from Lemmas 10.7, 10.8 and 10.9. ■

Lemma 10.4 *Let Assumptions 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 9.3, and 9.5 hold. Assume that bandwidth* $h = c_0 n^{-c_T/2}$ *in* $\widehat{f}_{v_t}$*, and assume a kernel of order* $p \geq (1 - c_T/2)/c_T$*. Then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\
= \quad & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_P\left(\frac{1}{\sqrt{nT}}\right).
\end{aligned}$$

Proof of Lemma 10.4. First note that

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1) - E(Y_0)] \\
= \quad & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(Y_{it} - E\left(\widetilde{a}_i + \widetilde{b}_t + Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} \\
& - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(Y_{it} - E\left(\widetilde{a}_i + \widetilde{b}_t + Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})}.
\end{aligned}$$

We first show that

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it} - E(\widetilde{a}_i + \widetilde{b}_t + Y_1)\right)}{\widehat{f}_{v_t}(v_{it})} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it} + o_p(\frac{1}{\sqrt{nT}}).$$

To this end,

$$\begin{aligned} & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{\widehat{f}_{v_t}(v_{it})} \\ = \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}(v_{it})} \\ & -\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right)+R_{nit}, \end{aligned} \tag{10.36}$$

where

$$R_{nit}\equiv\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})\widehat{f}_{v_t}(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right)^2.$$

Again, by the uniform convergence of $\widehat{f}_{v_t}(v_{it})$ (our assumption on p guarantees that the bias term vanishes fast enough),

$$\sup_{I_{\tau it}\neq 0}\left|\widehat{f}_{v_t}\left(v_{it}\right)-f_v\left(v_{it}\right)\right|=O_P\left(\log\left(n\right)\Big/\sqrt{nh}\right)=O_P\left(\log\left(n\right)\Big/n^{1/2-c_T/4}\right)=o_p\left((nT)^{-1/4}\right),$$

such that $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}|R_{nit}|=o_p\left(\frac{1}{\sqrt{nT}}\right)$.

Generalizing Lemma 10.1 a little, we have, $E\left[p\left(z_i,z_j\right)^2\right]=O\left(1/h\right)=o\left(n/T\right)$, and

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{f_{v_t}^2(v_{it})}\left(\widehat{f}_{v_t}(v_{it})-f_{v_t}(v_{it})\right) \\ = \quad & \frac{1}{n}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}\middle|\, v_{it}\right]}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right), \end{aligned}$$

for $t=1,...,T$. Substituting this back into equation (10.36), we get that $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)}{\widehat{f}_{v_t}(v_{it})}$ is equal to

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}-E\left[\left(Y_{it}-E(\widetilde{a}_i+\widetilde{b}_t+Y_1)\right)D_{it}|v_{it}\right]}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right),$$

which is $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}+o_p\left(\frac{1}{\sqrt{nT}}\right)$.

For the same reason

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}=\overline{\Pi}_1+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}-E\left(D_{it}|v_{it}\right)}{f_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right).$$

By the independence assumption on V_{it} across i and t, we know $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}-\frac{D_{it}}{f_{v_t}(v_{it})}\right)=O_p\left(\frac{1}{\sqrt{nT}}\right)$. Therefore

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})}=\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})}+o_p\left(\frac{1}{\sqrt{nT}}\right)\\
=&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(D_{it}\Big/\widehat{f}_{v_t}(v_{it})-D_{it}/f_{v_t}(v_{it})\right)\right)}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}\right)\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\Big/\widehat{f}_{v_t}(v_{it})\right)}+o_p\left(\frac{1}{\sqrt{nT}}\right)\\
=&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}}+o_p\left(\frac{1}{\sqrt{nT}}\right),
\end{aligned}$$

where the last equality holds by $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}=o_P(1)$ and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}-\frac{D_{it}}{f_{v_t}(v_{it})}\right)=O_p\left(\frac{1}{\sqrt{nT}}\right)$. Applying the same analysis to the second component of the estimator finishes the proof. ■

Lemma 10.5 *Let Assumption 3.4, 3.5, 3.6, 3.8 hold, then*

$$\begin{aligned}
\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1&=O_P\left((nT)^{-1/2}\right),\\
\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1-D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_2&=O_P\left((nT)^{-1/2}\right).
\end{aligned}$$

Proof of Lemma 10.5. Here we prove the first equality of the lemma, and the second follows by the same logic. Note that

$$\begin{aligned}
E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|a_i,\widetilde{a}_i\right)&=E\left(E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|a_i,b_t,u_{it}\right)\middle|a_i,\widetilde{a}_i\right)\\
&=E\left(\int\frac{I\left(0\le a_i+b_t+v_{it}+u_{it}\le\alpha\right)}{f_{v_t}(v_{it})}f_{v_t}\left(v_{it}|a_i,\widetilde{a}_i,b_t,u_{it}\right)dv_{it}\middle|a_i,\widetilde{a}_i\right)\\
&=E\left(\int I\left(0\le a_i+b_t+v_{it}+u_{it}\le\alpha\right)dv_{it}\middle|a_i,\widetilde{a}_i\right)\\
&=\alpha=\overline{\Pi}_1.
\end{aligned}$$

Similarly, we have

$$E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle|b_t,\widetilde{b}_t\right)=\alpha=\overline{\Pi}_1.$$

By this result, we have

$$\begin{aligned}
&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1 \\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})}-E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle| a_i,a_i\right)-E\left(\frac{D_{it}}{f_{v_t}(v_{it})}\middle| b_t,\widetilde{b}_t\right)+\overline{\Pi}_1\right).
\end{aligned}$$

By the conditional independence assumption, we know the covariance of the above terms across either a different i or t is zero. So we have

$$\sqrt{nT}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}-\overline{\Pi}_1\right)\xrightarrow{d}N\left(0,var\left(\Pi_{1it}\right)\right).$$

The second part of the theorem follows similarly. ■

Lemma 10.6 *Let Assumption 3.4, 3.5, 3.6, 3.8, for $j=0,1$*

$$\begin{aligned}
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle| a_i,\widetilde{a}_i\right] &= E\left[\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{jit}\middle| a_i,\widetilde{a}_i\right]=0,\\
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle| b_t,\widetilde{b}_t\right] &= E\left[\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{jit}\middle| b_t,\widetilde{b}_t\right]=0,\\
E\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\middle| a_i,\widetilde{a}_i\right] &= E\left[\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\middle| a_i,\widetilde{a}_i\right]=0,\\
E\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\middle| b_t,\widetilde{b}_t\right] &= E\left[\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\middle| b_t,\widetilde{b}_t\right]=0.
\end{aligned}$$

Proof of Lemma 10.6. Note that by the proof of Lemma 10.5

$$\begin{aligned}
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle| a_i,\widetilde{a}_i\right] &= E\left(\varepsilon_{jit}|\, a_i,\widetilde{a}_i\right)-E\left(\varepsilon_{jit}|\, a_i,\widetilde{a}_i\right)=0,\\
E\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{jit}\middle| b_t,\widetilde{b}_t\right] &= E\left(\varepsilon_{jit}|\, b_t,\widetilde{b}_t\right)-E\left(\varepsilon_{jit}|\, b_t,\widetilde{b}_t\right)=0,
\end{aligned}$$

for $j=0,1$. Others follow similarly. ■

Lemma 10.7 *Let Assumption 3.4, 3.5, 3.6, 3.8 hold. Then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\left(\widetilde{a}_i+\widetilde{b}_t-E\left(\widetilde{a}_i+\widetilde{b}_t\right)\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t-E\left(\widetilde{a}_i+\widetilde{b}_t\right)\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})}\\
= \;& \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)\right]+o_P\left((nT)^{-1/2}\right),
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)\right]=O_p\left((nT)^{-1/2}\right).$

Proof of Lemma 10.7.

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} (1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} (1-D_{it})/f_{v_t}(v_{it})} \\
= \quad & \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})} - \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right) \\
& - \left(\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} (1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} (1-D_{it})/ f_{v_t}(v_{it})} - \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right) \right).
\end{aligned}$$

We analyze the first term.

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})} - \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\widetilde{a}_i+\widetilde{b}_t\right) \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i+\widetilde{b}_t\right) - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})-\overline{\Pi}_1\right)}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})\right)\overline{\Pi}_1} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+\left(E\left(\widetilde{a}_i\right)+E\left(\widetilde{b}_t\right)\right)\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right) \\
& - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)\Big/ f_{v_t}(v_{it})\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})-\overline{\Pi}_1\right)}{\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}/f_{v_t}(v_{it})\right)\overline{\Pi}_1} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

So we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}\left(\widetilde{a}_i+\widetilde{b}_t\right)}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})}}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})\left(\widetilde{a}_i+\widetilde{b}_t\right)}{f_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{f_{v_t}(v_{it})}} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\right]\left(\widetilde{a}_i-E\left(\widetilde{a}_i\right)+\widetilde{b}_t-E\left(\widetilde{b}_t\right)\right)+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

The rate of the influence function above can be similarly seen from Lemma 10.6. ■

Lemma 10.8 *Let Assumption 3.4, 3.5, 3.6, 3.8, and 3.9 hold, then*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle|v_{it}\right]\Big/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle|v_{it}\right]}{\overline{\Pi}_1 f_{v_t}(v_{it})}+o_P\left((nT)^{-1/2}\right)
\end{aligned}$$

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle|v_{it}\right]\Big/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle|v_{it}\right]}{\overline{\Pi}_2 f_{v_t}(v_{it})}+o_P\left((nT)^{-1/2}\right)
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_1\right)\right)D_{it}\middle|v_{it}\right]}{\overline{\Pi}_1 f_{v_t}(v_{it})} = O_p\left((nT)^{-1/2}\right)$, $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{E\left[\left(Y_{it}-E\left(\widetilde{a}_i+\widetilde{b}_t+Y_0\right)\right)(1-D_{it})\middle|v_{it}\right]}{\overline{\Pi}_2 f_{v_t}(v_{it})} = O_p\left((nT)^{-1/2}\right)$.

Proof of Lemma 10.8. The first part of this theorem follows the same line of proof as Lemma 10.7. The $\sqrt{nT}$ convergence rate then follows by Assumption 3.9. ■

Lemma 10.9 *Letting Assumption 3.4, 3.5, 3.6, 3.8, and 3.10 hold, we have*

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\varepsilon_{1it}/\,f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\varepsilon_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})} \\
= \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

and $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}\right]=O_P\left((nT)^{-1/2}\right).$

Proof of Lemma 10.9. Following the same proof as in Lemma 10.7, we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}\varepsilon_{1it}/\,f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\varepsilon_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})} \qquad (10.37)\\
=\;&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}-1\right)\varepsilon_{1it}-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}-1\right)\varepsilon_{0it}\\
&+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\varepsilon_{1it}-\varepsilon_{0it}-E\left(\varepsilon_{1it}-\varepsilon_{0it}|a_i,\widetilde{a}_i\right)-E\left(\varepsilon_{1it}-\varepsilon_{0it}|b_t,\widetilde{b}_t\right)\right)\\
&+\frac{1}{n}\sum_{i=1}^{n}E\left(\varepsilon_{1it}-\varepsilon_{0it}|a_i,\widetilde{a}_i\right)+\frac{1}{T}\sum_{t=1}^{T}E\left(\varepsilon_{1it}-\varepsilon_{0it}|b_t,\widetilde{b}_t\right)+o_P\left((nT)^{-1/2}\right),
\end{aligned}$$

where the first three terms are $O_p\left((nT)^{-1/2}\right)$ and last two terms are zero by Assumption 3.10. So we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}Y_{1it}/\,f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}D_{it}/f_{v_t}(v_{it})}-\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{0it}/f_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})/f_{v_t}(v_{it})}-E\left(Y_1-Y_0\right)\\
=\;&\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left[\frac{D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_1}\varepsilon_{1it}-\frac{1-D_{it}}{f_{v_t}(v_{it})\overline{\Pi}_2}\varepsilon_{0it}\right]+o_P\left((nT)^{-1/2}\right).
\end{aligned}$$

■

Lemma 10.10 *Assume a_i, b_t are random vectors that satisfy Assumption 3.8. Each w_{it} is a random vector and $w_{it}\perp w_{it'}|a_i$ for $t\neq t'$, $w_{it}\perp w_{i't}|b_t$ for $i\neq i'$, and $w_{it}\perp w_{i't'}$ for $i\neq i'$, $t\neq t'$. $h(a_i,b_t,w_{it})$ is a real function with first and second moments that exist, and $E\big[h(a_i,b_t,w_{it})^2\big]=o(n)$. $E[h(a_i,b_t,w_{it})]=E[h(a_{i'},b_{t'},w_{i't'})]$ for any i,t,i',t'. $T\to\infty$ as $n\to\infty$. Then*

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}h(a_i,b_t,w_{it})$$

is equal to

$$E[h(a_i,b_t,w_{it})]+\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[E[h(a_i,b_t,w_{it})|a_i]+E[h(a_i,b_t,w_{it})|b_t]-2E[h(a_i,b_t,w_{it})]\right]+o_p\left(\frac{1}{\sqrt{T}}\right).$$

Here w_{it} are heterogeneous across t, but $\mathrm{E}(h(a_i,b_t,w_{it}))$ are assumed the same across t. This would typically be satisfied by having $\mathrm{E}(h(a_i,b_t,w_{it}))=0$ for any i,t.

Proof of Lemma 10.10. Let

$$Q = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[h(a_i,b_t,w_{it}) - \mathrm{E}(h(a_i,b_t,w_{it})|a_i) - \mathrm{E}(h(a_i,b_t,w_{it})|b_t) + \mathrm{E}(h(a_i,b_t,w_{it}))\right], \tag{10.38}$$

To establish that $Q = o_p(\frac{1}{\sqrt{T}})$, begin with

$$\mathrm{E}[Q^2] = \frac{1}{n^2T^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{i'=1}^{n}\sum_{t'=1}^{T}\mathrm{E}\left[\left(h - \mathrm{E}(h|a_i) - \mathrm{E}(h|b_t) + \mathrm{E}(h)\right)\left(h - \mathrm{E}(h|a_{i'}) - \mathrm{E}(h|b_{t'}) + \mathrm{E}(h)\right)\right].$$

For $i \neq i', t \neq t'$, the term inside summation is zero. Now consider the case where only one index is equal to the other one, i.e., $i = i'$or $t \neq t'$. Since

$$\begin{aligned}\mathrm{E}\left[h(a_i,b_t,w_{it})h(a_i,b_{t'},w_{it'})\right] &= \mathrm{E}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})h(a_i,b_{t'},w_{it'})|a_i\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]\mathrm{E}[h(a_i,b_{t'},w_{it'})|a_i]\right],\end{aligned}$$

the term inside summation is zero again. So we can rewrite $\mathrm{E}[Q^2]$ as

$$\mathrm{E}[Q^2] = \frac{1}{n^2T^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathrm{E}\left[\left(h - \mathrm{E}(h|a_i) - \mathrm{E}(h|b_t) + \mathrm{E}(h)\right)^2\right].$$

By assumption $\mathrm{E}\big(h^2\big) = o_p(n)$, so $\mathrm{E}[Q^2] = o_p\left(\frac{1}{T}\right)$, which implies $Q = o_p\left(\frac{1}{\sqrt{T}}\right)$. ■

Lemma 10.11 *Make the same assumptions as in Lemma 10.10 and Assumption 3.7. Further assume* $var(E\left[h(a_i,b_t,w_{it})|a_i\right]) \leq M$*, for all* i*, where* M *is a finite positive number. Then*

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[E\left[h(a_i,b_t,w_{it})|a_i\right] + E\left[h(a_i,b_t,w_{it})|b_t\right] - 2E\left[h(a_i,b_t,w_{it})\right]\right]$$

is equal to $\frac{1}{T}\sum_{t=1}^{T}\left[E\left[h(a_i,b_t,w_{it})|b_t\right] - E\left[h(a_i,b_t,w_{it})\right]\right] + o_p\left(\frac{1}{\sqrt{T}}\right)$.

Proof of Lemma 10.11.

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right] + \mathrm{E}\left[h(a_i,b_t,w_{it})|b_t\right] - 2\mathrm{E}(h(a_i,b_t,w_{it}))\right]$$

First by assumption that $\omega_{it}|b_t$ is i.i.d across i, we know that

$$\mathrm{E}\left[h(a_i,b_t,w_{it})|b_t\right] = \mathrm{E}\left[h(a_{i'},b_t,w_{i't})|b_t\right],$$

which gives

$$\begin{aligned}&\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|b_t\right] - \mathrm{E}(h(a_i,b_t,w_{it}))\right] \qquad (10.39)\\ =\ &\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|b_t\right] - \mathrm{E}(h(a_i,b_t,w_{it}))\right]\end{aligned}$$

For the other part, note that

$$\begin{aligned} & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right] \\ = \quad & \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right], \end{aligned}$$

where $\mathrm{E}[h(a_i,b_t,w_{it})|a_i]$ is independent across i.

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right)^2\right] \\ = \quad & \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left(\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]\right)^2\right]\leq\frac{M}{n}, \end{aligned}$$

by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]=O_p(\frac{1}{\sqrt{n}}),$$

which gives that

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathrm{E}\left[h(a_i,b_t,w_{it})|a_i\right]-\mathrm{E}(h(a_i,b_t,w_{it}))\right]=O_p(\frac{1}{\sqrt{n}}). \tag{10.40}$$

The lemma then follows from combining equation (10.39) and equation (10.40). ■

Lemma 10.12 *Denote $\zeta_n=(A_{1n},B_{1n},A_{2n},B_{2n})'$, a 4-by-1 vector, where $A_{1n},B_{1n},A_{2n},B_{2n}$ are random variables that evolve as n goes to infinity. Assume that ζ_n converge in probability to $\overline{\zeta}=(0,\overline{B}_1,0,\overline{B}_2)'$, where $\overline{B}_1\neq 0,\overline{B}_2\neq 0$, and*

$$\sqrt{n}\left[\zeta_n-\zeta\right]\xrightarrow{d}N(\mathbf{0},\mathbf{\Omega}),$$

where $\mathbf{\Omega}$ is a positive definite matrix

$$\Omega=\begin{pmatrix}\sigma_{A_1}^2 & \sigma_{A_1B_1} & \sigma_{A_1A_2} & \sigma_{A_1B_2}\\ \cdot & \sigma_{B_1}^2 & \sigma_{B_1A_2} & \sigma_{B_1B_2}\\ \cdot & \cdot & \sigma_{A_2}^2 & \sigma_{A_2B_2}\\ \cdot & \cdot & \cdot & \sigma_{B_2}^2\end{pmatrix}.$$

Then

$$\sqrt{n}\left(\frac{A_{1n}}{B_{1n}}-\frac{A_{2n}}{B_{2n}}\right)\xrightarrow{d}N\left(0,\frac{\sigma_{A_1}^2}{\overline{B}_1^2}-\frac{2\sigma_{A_1A_2}}{\overline{B}_1\overline{B}_2}+\frac{\sigma_{A_2}^2}{\overline{B}_2^2}\right).$$

Proof. The Lemma follows immediately from the delta method. ■

Proof of Theorem 8.2. First we have

$$\sup_{I_{\tau it}\neq 0}\left|\widehat{f}_{v_t}\left(v_{it}\right)-f_v\left(v_{it}\right)\right|=O_P\left(\log\left(n\right)\Big/\sqrt{nh}\right)=O_P\left(\log\left(n\right)n^{-2/5}\right).$$

Following the proof of Lemma 10.4, we have[8]

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} D_{it}\Big/\widehat{f}_{v_t}(v_{it})} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})Y_{it}\Big/\widehat{f}_{v_t}(v_{it})}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(1-D_{it})\Big/\widehat{f}_{v_t}(v_{it})} - [E(Y_1)+E(Y_0)] \\
= &\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Applying Lemma 10.11 on this expression, it is equivalent to

$$\frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{T}}\right).$$

Applying Lemma 10.12 to this expression, we have

$$\begin{aligned}
&\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{D_{it}}{\widehat{f}_{v_t}(v_{it})}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})Y_{it}}{\widehat{f}_{v_t}(v_{it})}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{(1-D_{it})}{\widehat{f}_{v_t}(v_{it})}} - E(\widetilde{a}_i+\widetilde{b}_t+Y_1) + E(\widetilde{a}_i+\widetilde{b}_t+Y_0) \\
= &\frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{1it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Lambda_{2it}}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{n}}\right) \\
= &\frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{1it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{1it}} - \frac{\frac{1}{T}\sum_{t=1}^{T} E\left[\Lambda_{2it}|\, b_t, \widetilde{b}_t\right]}{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\Pi_{2it}} + o_p\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}$$

which then gives the conclusion by applying Lemma 10.12. ■

10.2 Proof of Theorem 4.1, 4.2 and 4.3

Lemma 10.13 *Let $[-m', m]$ be the support of V. Suppose that f_v is bounded and bounded away from zero on its support. Then $m - V_n^{(1)} \propto n^{-1}$ in probability.*

[8] Note that the residual here is $o_P\left(\frac{1}{\sqrt{n}}\right)$. We do not need this to be $o_P\left(\frac{1}{\sqrt{nT}}\right)$ due to the slower convergence of our estimator.

Proof. Let $\{a_n\}_{n=1}^{\infty}$ be any series that $a_n \to \infty$ and $a_n = o(n)$. Let $\underline{c}_v = \inf_{v \in \text{supp}(V)} f_v(V)$. Then

$$P\left(V_n^{(1)} < m - \frac{a_n}{n}\right) \leq \left(1 - \underline{c}_v \frac{a_n}{n}\right)^n = \left(\left(1 - \underline{c}_v \frac{a_n}{n}\right)^{\frac{n}{\underline{c}_v a_n}}\right)^{\underline{c}_v a_n} = \left(e\left(1 + o(1)\right)\right)^{-\underline{c}_v a_n} \to 0,$$

where the second equality holds by the fact that $\lim_{x \to 0} (1-x)^{\frac{1}{x}} = e^{-1}$. So we have $m - V_n^{(1)} = O_P\left(n^{-1}\right)$. Let $\overline{c}_v = \sup_{v \in \text{supp}(V)} f_v(V)$. On the other hand, if $a_n \to 0$, then

$$P\left(V_n^{(1)} < m - \frac{a_n}{n}\right) \geq \left(1 - \overline{c}_v \frac{a_n}{n}\right)^n = \left(e\left(1 + o(1)\right)\right)^{-\underline{c}_v a_n} \to 1.$$

So we have in probability $m - V_n^{(1)} \propto n^{-1}$.

If we do not have $f_v(m) > 0$, however, the convergence rate of $V_n^{(1)}$ will be slower and depend on how fast f_v converge to zero towards m. ■

Proof of Theorem 4.1. First,

$$G_D(v) = P(D = 1 | V = v) = F_u(\alpha - v) - F_u(-v),$$

therefore $G_D(v)$ is twice continuous differentiable by Assumption 4.1.

We define the components of the bias term and variance term from the estimates by $\mathbf{B}_h$ and $\mathbf{V}_h$ respectively:

$$\begin{aligned}
\mathbf{B}_h(\widehat{m}) &\equiv \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(V_i - \widehat{m}) \begin{pmatrix} 1 \\ (V_i - \widehat{m})/h \end{pmatrix} \left[G_D(V_i) - G_D(\widehat{m}) - G_D'(\widehat{m})(V_i - V)\right], \\
\mathbf{V}_h(\widehat{m}) &\equiv \frac{1}{n-1} \sum_{i=1}^{n-1} K_h(V_i - \widehat{m}) \begin{pmatrix} 1 \\ (V_i - \widehat{m})/h \end{pmatrix} \left[I(D_i = 1) - G_D(V_i)\right].
\end{aligned}$$

Then $\widehat{G}_D(\widehat{m}) = G_D(\widehat{m}) + e_1^T \left[\mathbf{S}_h(\widehat{m})\right]^{-1} \left(\mathbf{B}_h(\widehat{m}) + \mathbf{V}_h(\widehat{m})\right)$.

Suppose we know the true value of m, we define the following "oracle" estimator $\widehat{G}_D^o(m) = e_1^T \widehat{\boldsymbol{\beta}}^o(m)$ where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^o(m) &= \left[\mathbf{S}_h^o(m)\right]^{-1} \frac{1}{n} \sum_{i=1}^{n} K_h(V_i - m) \begin{pmatrix} 1 \\ (V_i - m)/h \end{pmatrix} I(D_i = 1), \\
\mathbf{S}_h^o(m) &\equiv \frac{1}{n} \sum_{i=1}^{n} K_h(V_i - m) \begin{pmatrix} 1 \\ (V_i - m)/h \end{pmatrix} (1, (V_i - m)/h).
\end{aligned}$$

We similarly define the bias term and variance tem for $\widehat{G}_D^o(m)$:

$$\begin{aligned}
\mathbf{B}_h^o(m) &\equiv \frac{1}{n} \sum_{i=1}^{n} K_h(V_i - m) \begin{pmatrix} 1 \\ (V_i - m)/h \end{pmatrix} \left[G_D(V_i) - G_D(m) - G_{D,-}'(m)(V_i - V)\right], \\
\mathbf{V}_h^o(m) &\equiv \frac{1}{n} \sum_{i=1}^{n} K_h(V_i - m) \begin{pmatrix} 1 \\ (V_i - m)/h \end{pmatrix} \left[I(D_i = 1) - G_D(V_i)\right],
\end{aligned}$$

then $\widehat{G}_D^o(m) = G_D(m) + e_1^T \left[\mathbf{S}_h^o(m)\right]^{-1} \left(\mathbf{B}_h^o(m) + \mathbf{V}_h^o(m)\right)$.

In Lemma 10.14, we show that

$$\begin{aligned} \mathbf{S}_h^o(m) &\xrightarrow{P} \overline{\mathbf{S}} f_v(m), \\ \mathbf{B}_h^o(m) &= h^2 \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) f_v(m) + o_P\left(h^2\right), \end{aligned}$$

and

$$\begin{aligned} E\left[\mathbf{V}_h^o(m)\right] &= 0 \\ E\left[\mathbf{V}_h^o(m) \mathbf{V}_h^o(m)^T\right] &= \frac{1}{nh} \mathbf{Q} G_D(m)\left(1 - G_D(m)\right) f_v(m) + o\left(\frac{1}{nh}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \operatorname{bias}\left(\widehat{G}_D^o(m)\right) &= e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_D(m) h^2 + o\left(h^2\right), \\ \operatorname{var}\left(\widehat{G}_D^o(m)\right) &= \frac{1}{nh} e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(m)\left(1 - G_D(m)\right) f_v(m)^{-1} + o\left(\frac{1}{nh}\right), \end{aligned}$$

where the leading terms in the bias and variance are $\mathbb{B}_h$ and $\sigma^2(m)$ respectively.

By selecting $h = c_0 n^{-1/5}$, easy to verify that the triangular series $\mathbf{V}_h^o(m)$ satisfy the conditions needed for the Liapunov's Central Limit Theorem for Triangular Arrays (by comparing the third moment with the second moment). Therefore, we have

$$\sqrt{nh}\left(\widehat{G}_D^o(m) - G_D(m) - \mathbb{B}_h\right) \xrightarrow{d} N\left(0, \sigma^2(m)\right)$$

By Lemma 10.15 and $h \propto n^{-1/5}$, $\widehat{G}_D^o(m) - \widehat{G}_D(\widehat{m}) = O_P\left(n^{-1}h^{-1}\right) = o_P\left(n^{-1/2}h^{-1/2}\right)$.

Thus

$$\sqrt{nh}\left(\widehat{G}_D(\widehat{m}) - G_D(m) - \mathbb{B}_h\right) \xrightarrow{d} N\left(0, \sigma^2(m)\right),$$

which is the conclusion.

Since $\operatorname{MSE}\left(\widehat{G}_D(\widehat{m})\right) = \left[\operatorname{bias}\left(\widehat{G}_D(\widehat{m})\right)\right]^2 + \operatorname{var}\left(\widehat{G}_D(\widehat{m})\right)$, to minimize mean squared error we can get h_{opt} as

$$h_{\text{opt}} = n^{-1/5} \left[\left(e_1^T \overline{\mathbf{S}}^{-1} \mathbf{Q} \overline{\mathbf{S}}^{-1} e_1 G_D(m)\left(1 - G_D(m)\right) f_v(m)^{-1}\right) \Big/ \left(e_1^T \overline{\mathbf{S}}^{-1} \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m)\right)^2 \right]^{1/5}.$$

■

Lemma 10.14 *Suppose Assumption 4.1 holds. Suppose i.i.d., and $h = c_0 n^{-1/5}$ for some $c_0 > 0$. We have*

$$\begin{aligned} \mathbf{S}_h^o(m) &\xrightarrow{P} \overline{\mathbf{S}} f_v(m), \\ \mathbf{B}_h^o(m) &= h^2 \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) f_v(m) + o_P\left(h^2\right), \\ E\left[\mathbf{V}_h^o(m)\right] &= 0, \\ E\left[\mathbf{V}_h^o(m) \mathbf{V}_h^o(m)^T\right] &= (nh)^{-1} \mathbf{Q} G_D(m)\left(1 - G_D(m)\right) f_v(m) + o\left((nh)^{-1}\right), \end{aligned}$$

where $\mathbf{S}_h^o(m)$, $\mathbf{B}_h^o(m)$ *and* $\mathbf{V}_h^o(m)$ *are defined in the proof of Theorem 4.1.*

Proof of Lemma 10.14. First $\mathbf{S}_h(m) \xrightarrow{P} \overline{\mathbf{S}} f_v(m)$ is standard in nonparametric econometrics (e.g., Theorem 3.2 in Fan and Gijbels 1996).

Now we turn to $\mathbf{B}_h^o(m)$. First note that

$$\mathbf{B}_h^o(m) = E[\mathbf{B}_h^o(m)] + O_P\left(\sqrt{\text{var}(\mathbf{B}_h^o(m))}\right). \tag{10.41}$$

In view of the above fact, and note that we have proved G_D is twice continuously differentiable in the proof of Theorem 4.1,

$$\begin{aligned} E[\mathbf{B}_h^o(m)] &= h^2 \int_{-\infty}^0 \begin{pmatrix} K(u)u^2 \\ K(u)u^3 \end{pmatrix} G_D''(\widetilde{m}) f_v(m+uh)\,du, \\ &= \begin{pmatrix} \int_{-\infty}^0 K(u)u^2 du \\ \int_{-\infty}^0 K(u)u^3 du \end{pmatrix} G_{D,-}''(m) f_v(m) h^2 (1+o(1)), \end{aligned}$$

where $\widetilde{m}$ is some value very close to $\widehat{m}$ (we have assumed K with bounded support). By the conditional independence, for the first element of $\mathbf{B}_h^o(m)$

$$\text{var}\left(\mathbf{B}_{h,1}^o(m)\right) = n^{-1}h^3 \int_{-\infty}^0 K^2(u)u^4 du G_{D,-}''(m)^2 f_v(m)(1+o(1)) = O\left(n^{-1}h^3\right).$$

By the same reason $\text{var}\left(\mathbf{B}_{h,2}^o(m)\right) = O\left(n^{-1}h^3\right)$.

By $h \propto n^{-1/5}$, $\left(n^{-1}h^3\right)^{1/2} \propto n^{-4/5} \propto h^4$. From equation (10.41) and the calculation followed,

$$\mathbf{B}_h^o(m) = \begin{pmatrix} \int_{-\infty}^0 K(u)u^2 du \\ \int_{-\infty}^0 K(u)u^3 du \end{pmatrix} G_{D,-}''(m) f_v(m) h^2 + o_P\left(h^2\right).$$

Now we turn to $\mathbf{V}_h^o(m)$. Since by the definition of G_D, $E[I(D_i=1) - G_D(V_i)|V_i] = 0$,

$$E[\mathbf{V}_h^o(m)] = E\{E[\mathbf{V}_h^o(m)|(V_1,...,V_n)]\} = 0.$$

Then by i.i.d,

$$\begin{aligned} & E\left[\mathbf{V}_h^o(m)\mathbf{V}_h^o(m)^T\right] \\ = & \frac{1}{n}E\left[K_h(V_i-m)^2 \begin{pmatrix} 1 & (V_i-m)/h \\ (V_i-m)/h & (V_i-m)^2/h^2 \end{pmatrix} [I(D_i=1) - G_D(V_i)]^2\right] \\ = & \frac{1}{nh}\begin{pmatrix} \int_{-\infty}^0 K(u)^2 du & \int_{-\infty}^0 K(u)^2 u du \\ \int_{-\infty}^0 K(u)^2 u du & \int_{-\infty}^0 K(u)^2 u^2 du \end{pmatrix} G_D(m)(1-G_D(m)) f_v(m)(1+o(1)) \\ = & (nh)^{-1}\mathbf{Q}G_D(m)(1-G_D(m)) f_v(m)(1+o(1)) = (nh)^{-1}\mathbf{Q}G_D(m)(1-G_D(m)) f_v(m) + o\left(n^{-1}h^{-1}\right), \end{aligned}$$

where the second line holds for by the standard variable transformation. Therefore

$$E\left[\mathbf{V}_h^o(m)\mathbf{V}_h^o(m)^T\right] = (nh)^{-1}\mathbf{Q}G_D(m)(1-G_D(m)) f_v(m) + o\left((nh)^{-1}\right).$$

■

Lemma 10.15 *Under the same assumptions for Theorem 4.1, we have*

$$\widehat{G}_D(\widehat{m}) - \widehat{G}_D^o(m) = O_p\left((nh)^{-1}\right).$$

Proof. By noting that

$$\begin{aligned}
\widehat{G}_D(\widehat{m}) &= G_D(\widehat{m}) + e_1^T\left[\mathbf{S}_h(\widehat{m})\right]^{-1}\left(\mathbf{B}_h(\widehat{m}) + \mathbf{V}_h(\widehat{m})\right), \\
\widehat{G}_D^o(m) &= G_D(m) + e_1^T\left[\mathbf{S}_h^o(m)\right]^{-1}\left(\mathbf{B}_h^o(m) + \mathbf{V}_h^o(m)\right),
\end{aligned}$$

we compare the components in $\widehat{G}_D(\widehat{m})$ and $\widehat{G}_D^o(m)$ one by one.

By Lemma 10.13 and finiteness of $G_D'(\cdot)$, $G_D(\widehat{m}) - G_D(m) = G_D'(\widetilde{m})(\widehat{m} - m) = O_P\left(n^{-1}\right) = o_p\left((nh)^{-1}\right)$, where $\widetilde{m}$ is some value between m and $\widehat{m}$.

For $\mathbf{S}_h(\widehat{m}) - \mathbf{S}_h^o(m)$,

$$\begin{aligned}
\mathbf{S}_h(\widehat{m}) - \mathbf{S}_h^o(m) &= \frac{1}{n-1}\sum_{i=1}^{n-1}\left[K_h(V_i - \widehat{m})\begin{pmatrix} 1 \\ (V_i - \widehat{m})/h \end{pmatrix}(1, (V_i - \widehat{m})/h)\right. \\
&\quad \left. - K_h(V_i - m)\begin{pmatrix} 1 \\ (V_i - m)/h \end{pmatrix}(1, (V_i - m)/h)\right] + \frac{\mathbf{S}_h^o(m)}{n(n-1)} \\
&\quad + \frac{n-2}{n(n-1)}K_h(V_n - m)\begin{pmatrix} 1 \\ (V_n - m)/h \end{pmatrix}(1, (V_n - m)/h) \\
&\equiv \mathcal{J}_{n1} + \mathcal{J}_{n2} + \mathcal{J}_{n3}.
\end{aligned}$$

Using the fact that the support of $K(\cdot)$ is bounded, we know $(V_i - m)/h$ is bounded when $K(\cdot)$ is nonzero, so by Lemma 10.13

$$\mathcal{J}_{n1} = O_P\left((m - \widehat{m})/h\right) = O_P\left(n^{-1}h^{-1}\right).$$

$\mathcal{J}_{n2} = O_P\left(n^{-2}\right)$ because $\mathbf{S}_h^o(m) = O_P(1)$. $\mathcal{J}_{n3} = O_P\left(n^{-1}h^{-1}\right)$, because $(V_n - m)/h$ is finite when $K(\cdot)$ is nonzero. In sum, we have

$$\mathbf{S}_h(\widehat{m}) - \mathbf{S}_h^o(m) = O_P\left(n^{-1}h^{-1}\right).$$

For $\mathbf{B}_h(\widehat{m}) - \mathbf{B}_h^o(m)$, note that each element of $\mathbf{B}_h(\widehat{m}) - \mathbf{B}_h^o(m)$ has very similar structure to the elements in $\mathbf{S}_h(\widehat{m}) - \mathbf{S}_h^o(m)$. Because G_D and G_D' are finite, by the same reason as above

$$\mathbf{B}_h(\widehat{m}) - \mathbf{B}_h^o(m) = O_P\left(n^{-1}h^{-1}\right).$$

For $\mathbf{V}_h(\widehat{m}) - \mathbf{V}_h^o(m)$, we similarly have

$$\mathbf{V}_h(\widehat{m}) - \mathbf{V}_h^o(m) = O_P\left(n^{-1}h^{-1}\right).$$

Finally, note that for any B, C of the dimension marked below

$$\underset{1\times 2}{e_1^T}\underset{2\times 2}{\widehat{B}^{-1}}\underset{2\times 1}{\widehat{C}} - \underset{1\times 2}{e_1^T}\underset{2\times 2}{B^{-1}}\underset{2\times 1}{C} = e_1^T\widehat{B}^{-1}\left(B - \widehat{B}\right)B^{-1}\widehat{C} + e_1^T B^{-1}\left(\widehat{C} - C\right).$$

Apply this formula to $\widehat{G}_D(\widehat{m}) - \widehat{G}_D^o(m)$, by that $\mathbf{S}_h(\widehat{m})$ are uniformly bounded away from zero in probability

because of the consistency of $\mathbf{S}_h(\widehat{m})$, and with the results on each component, we have

$$\widehat{G}_D(\widehat{m}) - \widehat{G}_D^o(m) = O_P\left(n^{-1}h^{-1}\right).$$

■

Proof of Theorem 4.2. Most of the proof of this theorem is standard, except that $\widehat{G}_D(\widehat{m})$ converges at the $\sqrt{n}$ rate, while in the typical case the convergence rate is $\sqrt{nh}$. The intuition for this result is that in $\mathbf{V}_h^o(m)\ E\left[\left(I(D_i=1) - G_D(V_i)\right)^2 \middle| V_i = m-h\right] \propto h$ (by Assumption 4.1, there is no mass point in the distribution of U), when the right end of the support of V is equal to the right end of the support of $\alpha - U$, and we put the most weight on the observations around m within ch during estimation, for some $c > 0$. The variance for those observations is of the order h, resulting in a faster rate of convergence.

As shown in Lemma 10.16,

$$E\left[\mathbf{V}_h(\widehat{m})\mathbf{V}_h(\widehat{m})^T\right] = \frac{1}{n}\widetilde{\mathbf{Q}}G'_{D,-}(m)f_v(m) + \frac{\widehat{m}-m}{nh}\widetilde{\mathbf{Q}}G'_{D,-}(m)f_v(m) + o_P\left(\frac{1}{n} + \frac{\widehat{m}-m}{nh}\right),$$

where the leading term in $E\left[\mathbf{V}_h(\widehat{m})\mathbf{V}_h(\widehat{m})^T\right]$ in Theorem 4.1 becomes zero here.

Again from Lemma 10.16, when $h \propto n^{-2/5}$

$$\begin{aligned}
\text{bias}\left(\widehat{G}_D(\widehat{m})\right) &= h^2 e_1^T \overline{\mathbf{S}}^{-1}\begin{pmatrix} S_{2,-} \\ S_{3,-}\end{pmatrix} G''_{D,-}(m) + o_P\left(h^2 + n^{-1/2}h^{3/2}\right) = \mathbb{B}_h + o_P\left(h^2\right), \\
\text{var}\left(\widehat{G}_D(\widehat{m})\right) &= \frac{1}{n}e_1^T\overline{\mathbf{S}}^{-1}\widetilde{\mathbf{Q}}\overline{\mathbf{S}}^{-1}e_1 G'_{D,-}(m)f_v(m)^{-1} + \frac{\widehat{m}-m}{nh}e_1^T\overline{\mathbf{S}}^{-1}\widetilde{\mathbf{Q}}\overline{\mathbf{S}}^{-1}e_1 G'_{D,-}(m)f_v(m)^{-1} \\
&\quad + o_P\left(\frac{1}{n} + \frac{\widehat{m}-m}{nh}\right) \\
&= n^{-1}\widetilde{\sigma}^2(m) + o_P\left(n^{-1}\right),
\end{aligned}$$

here the leading terms of the bias and variance terms are $\mathbb{B}_h$ and $\widetilde{\sigma}^2(m)$, respectively.

Since $\text{MSE}\left(\widehat{G}_D(\widehat{m})\right) = \left[\text{bias}\left(\widehat{G}_D(\widehat{m})\right)\right]^2 + \text{var}\left(\widehat{G}_D(\widehat{m})\right)$, minimizing the mean squared error we can get h_{opt} as

$$h_{\text{opt}} = \left(\frac{\widehat{m}-m}{n}\right)^{1/5}\left[e_1^T\overline{\mathbf{S}}^{-1}\widetilde{\mathbf{Q}}\overline{\mathbf{S}}^{-1}e_1 G'_{D,-}(m)f_v(m)^{-1} \Big/ \left(e_1^T\overline{\mathbf{S}}^{-1}\begin{pmatrix} S_{2,-} \\ S_{3,-}\end{pmatrix} G''_{D,-}(m)\right)^2\right]^{1/5}.$$

By Lemma 10.13 $\widehat{m} - m \propto n^{-1}$ in probability, so $h_{\text{opt}} \propto n^{-2/5}$. In this case, the bias term is asymptotically negligible, so is the second term in $\text{var}\left(\widehat{G}_D(\widehat{m})\right)$.

Following the same analysis as above and by essentially the same argument in the proof of Theorem 4.1,

$$\sqrt{n}\left(\widehat{G}_D^o(m) - G_D(m)\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2(m)\right).$$

From Lemma 10.15, $\widehat{G}_D(\widehat{m}) - G_D^o(m) = o_P\left((nh)^{-1}\right) = o_P\left(n^{-3/5}\right)$ by $h \propto n^{-2/5}$. Thus

$$\sqrt{n}\left(\widehat{G}_D(\widehat{m}) - G_D(m)\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2(m)\right),$$

which is the conclusion. ■

Lemma 10.16 *Suppose Assumption 4.1 holds. Suppose i.i.d., the support of V covers the support of $\alpha - U$ on the right and $h \to 0$. We have*

$$\begin{aligned}
\mathbf{B}_h(\widehat{m}) &= h^2 \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) f_v(m) + o_P\left(h^2 + n^{-1/2}h^{3/2}\right) \\
E\left[\mathbf{V}_h(\widehat{m})\mathbf{V}_h(\widehat{m})^T\right] &= \frac{1}{n}\widetilde{\mathbf{Q}}G'_{D,-}(m) f_v(m) + \frac{\widehat{m}-m}{nh}\widetilde{\mathbf{Q}}G'_{D,-}(m) f_v(m) + o_P\left(\frac{1}{n}\right) + o_P\left(\frac{\widehat{m}-m}{nh}\right),
\end{aligned}$$

where $\mathbf{B}_h(\cdot)$, and $\mathbf{V}_h(\cdot)$ are defined in the proof of Theorem 4.1.

Proof of Lemma 10.16. The following is essential for our proof. Conditional on $V_n^{(1)}$, the only restriction of V_i, $i = 1, ..., n-1$, is $V_i < V_n^{(1)}$. By i.i.d. of the original sample, we have the conditional independence $V_i \perp V_j | V_n^{(1)}$, for $i \neq j, i, j \in \{1, ..., n-1\}$, and the conditional distribution of $V_i \left| V_n^{(1)} \right.$ is:

$$f_{v\left|V_n^{(1)}=v_n\right.}(s) = \begin{cases} f_v(s)/F_v(v_n) & \text{if } s \leq v_n, \\ 0 & \text{otherwise.} \end{cases}$$

Now we turn to $\mathbf{B}_h(\widehat{m})$. First note that

$$\mathbf{B}_h(\widehat{m}) = E\left[\mathbf{B}_h(\widehat{m}) \middle| V_n^{(1)} = \widehat{m}\right] + O_P\left(\sqrt{\operatorname{var}\left(\mathbf{B}_h(\widehat{m}) \middle| V_n^{(1)} = \widehat{m}\right)}\right). \tag{10.42}$$

In view of the above fact, conditional on $V_n^{(1)} = \widehat{m}$, by some standard calculation (note we have proved G_D is twice continuously differentiable in the proof of Theorem 4.1),

$$\begin{aligned}
E\left[\mathbf{B}_h(\widehat{m}) \middle| V_n^{(1)} = \widehat{m}\right] &= h^2 \int_{-\infty}^{0} \begin{pmatrix} K(u)u^2 \\ K(u)u^3 \end{pmatrix} G''_D(\widetilde{m}) \left[f_v(\widehat{m}+uh) F_v(\widehat{m}+uh)^{-1}\right] du, \\
&= \begin{pmatrix} \int_{-\infty}^{0} K(u)u^2 du \\ \int_{-\infty}^{0} K(u)u^3 du \end{pmatrix} \left[G''_{D,-}(\widehat{m}) f_v(\widehat{m}) F_v(\widehat{m})^{-1}\right] h^2 (1 + o(1)),
\end{aligned}$$

where $\widetilde{m}$ is some value very close to $\widehat{m}$ (we have assumed K with bounded support). By the conditional independence, for the first element of $\mathbf{B}_h(\widehat{m})$

$$\operatorname{var}\left(\mathbf{B}_{h,1}(\widehat{m}) \middle| V_n^{(1)} = \widehat{m}\right) = n^{-1}h^3 \int_{-\infty}^{0} K^2(u) u^4 du G''_{D,-}(\widehat{m})^2 f_v(\widehat{m}) F_v(\widehat{m})^{-1} (1 + o(1)) = O\left(n^{-1}h^3\right).$$

By the same reason $\operatorname{var}\left(\mathbf{B}_{h,2}(\widehat{m}) \middle| V_n^{(1)} = \widehat{m}\right) = O\left(n^{-1}h^3\right)$.

From equation (10.42) and the calculation followed,

$$\mathbf{B}_h(\widehat{m}) = \begin{pmatrix} \int_{-\infty}^{0} K(u)u^2 du \\ \int_{-\infty}^{0} K(u)u^3 du \end{pmatrix} \left[G''_{D,-}(\widehat{m}) f_v(\widehat{m}) F_v(\widehat{m})^{-1}\right] h^2 + o_P\left(h^2 + n^{-1/2}h^{3/2}\right).$$

By Lemma 10.13 that $\widehat{m} \xrightarrow{P} m$, so $F_v(\widehat{m}) \xrightarrow{P} 1$. By the continuity of $G''_{D,-}$ and f_v,

$$\mathbf{B}_h(\widehat{m}) = h^2 \begin{pmatrix} S_{2,-} \\ S_{3,-} \end{pmatrix} G''_{D,-}(m) f_v(m) + o_P\left(h^2 + n^{-1/2}h^{3/2}\right).$$

Now we turn to $\mathbf{V}_h(\widehat{m})$. Since $E[I(D_i = 1) - G_D(V_i)|V_i, \widehat{m}] = 0$ $\left(U_i \perp V_i \text{ and so by i.i.d. } U_i \perp V_n^{(1)}\right)$

$$E[\mathbf{V}_h(\widehat{m})] = E\{E[\mathbf{V}_h(\widehat{m})|V_i, \widehat{m}]\} = 0.$$

We inspect the first entry in $\mathbf{V}_h(\widehat{m})$. By $V_i \perp V_j | V_n^{(1)}$, for $i \neq j, i, j \in \{1, ..., n-1\}$

$$\begin{aligned}
& E\left[\left(\frac{1}{n-1}\sum_{i=1}^{n-1} K_h(V_i - \widehat{m})[I(D_i = 1) - G_D(V_i)]\right)^2 \middle| V_n^{(1)} = \widehat{m}\right] \\
= & \frac{1}{(n-1)h}\int_{-\infty}^{0} K^2(u) G_D(\widehat{m} + hu)[1 - G_D(\widehat{m} + hu)] f_v(\widehat{m} + hu) F_v(\widehat{m} + hu)^{-1} du \\
= & \frac{1}{(n-1)h}\int_{-\infty}^{0} K^2(u) G_D(m)[1 - G_D(m)] f_v(m) F_v(m)^{-1} du \\
& + \frac{1}{(n-1)h}\int_{-\infty}^{0} K^2(u)\Big\{G_D(m)[1 - G_D(m)] F_v(m)^{-2} f_v(m) + G_D(m)[1 - G_D(m)] f'_v(m) F_v(m)^{-1} \\
& + G'_{D,-}(m)[1 - G_D(m)] f_v(m) F_v(m)^{-1} - G_D(m) G'_{D,-}(m) f_v(m) F_v(m)^{-1}\Big\}(hu + \widehat{m} - m) du \\
& + O_p\left(\frac{(h + \widehat{m} - m)^2}{nh}\right) \\
= & \frac{1}{n-1}\int_{-\infty}^{0} K^2(u) u du G'_{D,-}(m) f_v(m) + \frac{\widehat{m} - m}{(n-1)h}\int_{-\infty}^{0} K^2(u) u du G'_{D,-}(m) f_v(m) + o_P\left(\frac{1}{n} + \frac{\widehat{m} - m}{nh}\right),
\end{aligned}$$

where the second equality holds by the Taylor expansion around m, the third equality holds by substituting $G_D(m) = 0$ and $F_v(m) = 1$, and the last line holds because $h \to 0$ and $\widehat{m} - m \to 0$. Do this for each element of $E\left[\mathbf{V}_h(\widehat{m})\mathbf{V}_h(\widehat{m})^T \middle| V_n^{(1)} = \widehat{m}\right]$, we get

$$E\left[\mathbf{V}_h(\widehat{m})\mathbf{V}_h(\widehat{m})^T \middle| V_n^{(1)} = \widehat{m}\right] = \frac{1}{n}\widetilde{\mathbf{Q}} G'_{D,-}(m) f_v(m) + \frac{\widehat{m} - m}{nh} e_1^T \widetilde{\mathbf{Q}} e_1 G'_{D,-}(m) f_v(m) + o_P\left(\frac{1}{n}\right) + o_P\left(\frac{\widehat{m} - m}{nh}\right).$$

■

Proof of Theorem 4.3. By Assumption 4.3, $E(W_0|X) = E(W_0|X, V \leq -\gamma_n(X))$. The first part of the theorem follows from

$$\begin{aligned}
& \lim_{n\to\infty} E(D_0 Y|X, V \leq -\gamma_n(X)) - E(W_0|X, V \leq -\gamma_n(X)) \\
= & \lim_{n\to\infty} E(D_0 W_0|X, V \leq -\gamma_n(X)) - E(W_0|X, V \leq -\gamma_n(X)) \\
= & \lim_{n\to\infty} E[(D_0 - 1) W_0|X, V \leq -\gamma_n(X)] = 0,
\end{aligned}$$

where the last equality holds by $\lim_{n\to\infty} E(D_0|X, V \leq -\gamma_n(X)) = 1$. This yields the expression for $E(W_0|X)$, and the equality for $E(W_{J-1}|X)$ is obtained in the same way. The part for $E(W_j|X)$, $j = 1, ..., J-2$, follows immediately from Theorem 3.1. ■

10.3 Proof of Theorem 8.1 and 8.4

Proof of Theorem 8.1. First the following is identified:

$$\begin{aligned}
& E\left(D|V=v,Z=z,X=x\right) \\
= \; & F_{U|X}\left(\alpha_1\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right)-F_{U|X}\left(\alpha_0\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right), \\
& \frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial v} \\
= \; & -\left[f_{U|X}\left(\alpha_1\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right)-f_{U|X}\left(\alpha_0\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right)\right]\frac{d\varsigma\left(v\right)}{dv}, \\
& \frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial z} \\
= \; & -\left[f_{U|X}\left(\alpha_1\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right)-f_{U|X}\left(\alpha_0\left(x\right)-\varsigma\left(v\right)-\varpi\left(x,z\right)|x\right)\right]\frac{\partial\varpi\left(x,z\right)}{\partial z}.
\end{aligned}$$

The ratio $\frac{d\varsigma(v)}{dv}\Big/\frac{\partial\varpi(x,z)}{\partial z}$ is identified by

$$\frac{d\varsigma\left(v\right)}{dv}\Big/\frac{\partial\varpi\left(x,z\right)}{\partial z}=\frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial v}\Big/\frac{\partial E\left(D|V=v,Z=z,X=x\right)}{\partial z}. \tag{10.43}$$

Then fix $V=0$, by $\varsigma'\left(0\right)=1$, and $\frac{\partial\varpi(x,z)}{\partial z}$ is identified by varying $\left(X,Z\right)$. Fix X,Z at some point, and then by knowing $\frac{\partial\varpi(x,z)}{\partial z}$, $\varsigma'\left(v\right)$ is identified. Finally, $\varsigma\left(V\right)$ is identified by

$$\varsigma\left(v\right)=\varsigma\left(0\right)+\int_0^v\varsigma'\left(s\right)ds.$$

■

Proof of Theorem 8.4. The proof here is very similar to the proof of Theorem 3.2.

Start by looking at

$$\begin{aligned}
& E\left(\frac{D_{it}Y_{it}}{f_{v_t}(V_{it}|X_{it},V_{it-1})}\middle|U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right) \\
= \; & E\left[E\left(\frac{D_{it}\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\right)}{f_{v_t}(V_{it}|X_{it},V_{it-1})}\mid V_{it},U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right] \\
= \; & E\left[\frac{I\left(\alpha_0(X_{it})\leq a_i+b_t+V_{it}+\vartheta\left(D_{it-1}\right)+U_{it}\leq\alpha_1(X_{it})\right)}{f_{v_t}(V_{it}|X_{it},V_{it-1})}\right. \\
& \left.E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid V_{it},U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right] \\
= \; & \int_{supp(V_{it}|U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1})}\frac{I\left(\alpha_0(X_{it})\leq a_i+b_t+V_{it}+\vartheta\left(D_{it-1}\right)+U_{it}\leq\alpha_1(X_{it})\right)}{f_{v_t}(v_{it}|X_{it},V_{it-1})} \\
& E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)f_{v_t}\left(v_{it}\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)dv_{it} \\
= \; & \int_{\alpha_0(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}^{\alpha_1(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)dv_{it} \\
= \; & E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\int_{\alpha_0(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}^{\alpha_1(X_{it})-a_i-b_t-U_{it}-\vartheta(D_{it-1})}1dv_{it} \\
= \; & E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]
\end{aligned}$$

and therefore

$$\begin{aligned} & E\left[D_{it}Y_{it}/f_{v_t}(V_{it}|X_{it},V_{it-1})|X_{it}\right] \\ = \; & E\left[E\left(\widetilde{a}_i+\widetilde{b}_t+Y_{1it}+g\left(Y_{it-1}\right)\mid U_{it},a_i,b_t,X_{it},D_{it-1},V_{it-1}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]|X_{it}\right] \\ = \; & E\left(Y_{1it}+\widetilde{a}_i+\widetilde{b}_t+g\left(Y_{it-1}\right)\Big| X_{it}\right)\left[\alpha_1(X_{it})-\alpha_0(X_{it})\right]. \end{aligned}$$

Given the above result, the rest of the proof follows from the same logic as the proof for Theorem 3.1. ■

References

[1] Powell, J. L., J. H. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coeffients," Econometrica, 57, 1403-1430.

[2] Silverman, B. W. (1978), "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and its Derivatives," Annals of Statistics, 6, 177-184.