Doubly Robust Identification of Causal Effects of a Continuous Treatment using Discrete Instruments

Yingying Dong and Ying-Ying Lee*

February 2024

Abstract

Many empirical applications estimate causal effects of a continuous endogenous variable (treatment) using a binary instrument. Estimation is typically done through linear 2SLS. This approach requires a mean treatment change and causal interpretation requires the LATE-type monotonicity in the first stage. An alternative approach is to explore distributional changes in the treatment, where the first-stage restriction is treatment rank similarity. We propose causal estimands that are doubly robust in that they are valid under either of these two restrictions. We apply the doubly robust estimation to estimate the impacts of sleep on well-being. Our new estimates corroborate the usual 2SLS estimates.

JEL codes: C14, C21, I30
Keywords: Doubly Robust Identification, Non-separable Model, Treatment effect heterogeneity, Continuous treatment, Monotonicity, Rank similarity, Sleep time, Well-being

1 Introduction

Many empirical applications estimate causal effects of a continuously distributed endogenous variable (treatment), such as air pollution concentration,

*Yingying Dong and Ying-Ying Lee, Department of Economics, University of California Irvine, yyd@uci.edu and yingying.lee@uci.edu.

poverty rate, income, price, birth weight, and time use, etc. A common approach is to apply two-stage least squares (2SLS) estimation, using a binary or discrete instrumental variable (IV). See, for recent examples, Chay and Greenstone (2005), Kling et al. (2007), Goda et al. (2011), Angrist et al. (2000), Maruyama and Heinesen (2020), Giaccherini et al. (2021), Aggeborn and Ohman (2021), and Bessone et al. (2021). In the case of a binary instrument, the 2SLS estimator essentially estiamtes the Wald ratio. In its basic form without considering covariates, the Wald ratio estimand is given by

$$\tau^{Wald} := \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]}, \tag{1}$$

where Y is the outcome of interest, Z is the binary instrument and T is the treatment. The above estimand τ^{Wald} requires a mean change in the treatment variable, as the denominator cannot be zero. In addition, when treatment effect is heterogenous and individuals select treatment intensity based on idiosyncratic gains, causal interpretation of the Wald ratio relies on a monotonicity assumption, which restricts treatment to change in one direction when the IV changes. This monotonicity assumption is originally proposed in Imbens and Angrist (1994) to show that in the case of a binary treatment and a binary IV, τ^{Wald} identifies a local average treatment effect (LATE).

One drawback of the above approach is that causal identification may be weak or may even fail. Frequently, policy instruments aim to shift one or two tails of the treatment distribution or change other features, such as the variance, of the treatment. As a result, treatment changes may concentrate at some selected quantiles, say lower or upper quantiles. By solely focusing on the mean treatment change, one may miss where the true changes are in the treatment distribution. Examples of such policies include minimum wage, minimum capital requirements, and the pollution ceiling set by Environmental Protection Agency (EPA). In this paper, we consider an alternative approach, which explores the distributional change in the first stage for causal identification. This idea has been proposed and explored in the non-separable IV model literature. The commonly employed restriction in the first stage is

treatment rank invariance or more generally treatment rank similarity. Rank invariance in the first stage is typically stated as the condition that the treatment function is monotonic in a scalar disturbance. See, e.g., Imbens and Newey (2002), and Chesher (2001, 2002) for early papers exploring this condition in non-separable models. Note that both the LATE-type monotonicity and treatment rank similarity are restrictions on the first-stage instrument effect heterogeneity. In general neither assumption implies the other. Also, these assumptions are not verifiable, in the sense that one may at best test their testable implications, which are necessary but not sufficient conditions of these assumptions.[1]

This paper takes a novel nonparametric doubly robust (DR) identification approach to identify causal effects of a continuous treatment using a binary or discrete instrument. We consider the two alternative restrictions on the first-stage instrument effect heterogeneity: the LATE-type monotonicity or treatment rank similarity. Either of these assumptions can be consistent with certain treatment choice behaviors and has been used extensively to identify causal treatment effects. See, e.g., Imbens and Newey (2009) and the reference therein for justifications of rank invariance (a stronger version of rank similarity) and Angrist and Imbens (1995) for a justification of monotonicity when treatment is multi-valued. We focus on discrete instruments since discrete instruments are widely used.

The parameters of interest include 1) the average effect at a given treatment quantile, which captures treatment effect heterogeneity at different treatment intensities and 2) weighted average effects for the largest subpopulation that respond to the IV change. Identification of the former parameter requires treatment rank similarity to hold. In contrast, for the latter parameter, we develop doubly robust estimands that are valid under either monotonicity or treatment rank similarity. When monotonicity holds, these estimands reduce to the LATE-type estimands and individuals who respond to the IV change

[1]See, e.g., Angrist and Imbens (1995) and Fiorini and Stevens (2021) for the discussion of the testable implication of the LATE-type monotonicity when treatment is multi-valued. See Dong and Shu (2018) and Frandsen and Lefgren (2018) for tests of the testable implication of rank similarity.

can be labeled as compliers, since they change treatment in a monotonic way (similar to compliers in the classic LATE model with a binary treatment and binary IV); otherwise, these estimands continue to be valid under treatment rank similarity, i.e., they continue to identify weighted averages of the average treatment effects for all individuals that change their treatment values when the IV changes, even though these individuals no longer respond in a monotonic way. Instead, their treament changes are subject to the rank restriction, i.e., the probability distribution of their treatment ranks stay the same, which is a slight generalization of requiring that their treatment rank to be exactly the same. Since these first-stage restrictions are not verifiable, our doubly robust estimands identify causal effects for the largest subpopulation while allowing either of these two assumptions holds true.

Our identification is nonparametric in that we consider non-separable models for both the first-stage and the outcome equation. Non-separable models allow for treatment effect heterogeneity and individuals self-selection of different treatment levels based on idiosyncratic gains, both of which are important features of the data as supported by economic theory and empirical evidence. For estimation, we opt for convenient semiparametric estimators to avoid cumbersome fully nonparametric estimation. We establish consistency and asymptotic normality of our proposed estimators. Lastly we apply our proposed approach to estimate the impacts of sleep time on individuals' well-being using data from a recent field experiment by Bessone et al. (2021). We show that our doubly robust approach can serve as a valuable tool to corroborate the IV/2SLS estimates.

This paper's identification approach builds upon two strands of literature - the LATE literature and the non-separable IV model literature. The LATE model is proposed in the seminal work of Imbens and Angrist (1994) and is further extended in Angrist and Imbens (1995), Angrist, Imbens and Rubin (1996), Angrist, Graddy and Imbens (2000), Abadie (2003), Frölich (2007), de Chaisemartin (2017), Dahl, Huber, and Mellace (2023), etc. The LATE model relies on the monotonicity assumption mentioned previously or some weaker versions of it for causal identification. Many studies in the nonseparable IV

model literature explore rank invariance or rank similarity in the first stage for causal identification. See, e.g., Chesher (2001, 2002, 2003, 2005), Imbens and Newey (2002, 2009), Florens et al. (2008) and more recently Torgovitsky (2015), and D'Haultfoeuille and Février (2015), among others. In particular, Torgovitsky (2015), and D'Haultfoeuille and Février (2015) provide detailed discussions of the identifying power of rank restrictions in the treatment and/or in the outcome equation. In addition, Masten and Torgovitsky (2016) consider a random correlated coefficients model and utilize treatment rank invariance to identify the average partial effect of continuous treatment variables, using binary or discrete instruments. For the DR identification approach, a few existing studies take this approach, see, e.g., Dong, Lee, and Gou (2023) and Arkhangelsky and Imbens (2022). Both papers are set in different frameworks than the current one. Dong, Lee, and Gou (2023) study the regression discontinuity design, while Arkhangelsky and Imbens (2022) investigate the panel data model.[2]

The rest of the paper proceeds as follows. Section 2 presents the DR identification results for the basic setup with a binary IV and without covariates. Section 3 extends the identification results to the general setup with covariates. Section 4 proposes convenient partial linear estimators and establishes their consistency and asymptotic normality. Section 5 discusses extensions to the case with a multi-valued IV or a vector of discrete IVs, with or without covariates; Section 6 presents our empirical analysis. Short concluding remarks are provided in Section 7.

2 Doubly Robust Identification in the Basic Setup

Let $Y \in \mathcal{Y} \subset \mathcal{R}$ be the outcome of interest, e.g., a measure of well-being. Y can be continuous or discrete. Let $T \in \mathcal{T} \subset \mathcal{R}$ be a continuous treatment

[2]The current paper extends the regression discontinuity setup of Dong, Lee, and Gou (2023) in multiple directions, including allowing the IV independence and treatment rank similarity to hold conditional on a vector of continuous and/or discrete covariates, allowing for a multi-valued IV or a vector of discrete IVs and completely different estimation and inference procedures.

variable, e.g., sleep time. Let $Z \in \{0,1\}$ be a binary IV for T, e.g., an indicator for being randomly assigned to a group receiving encouragement or financial incentives to increase night sleep.

To present the core ideas, we subsume all the covariates in this section. The general setup with covariates is presented in the next section. Assume that Y and T are generated as

$$Y = g(T, \varepsilon), \tag{2}$$
$$T = h(Z, U), \tag{3}$$

where ε captures all the other factors other than T that affect Y, and similarly U captures all the other reduced-form factors other than Z that affect T. The outcome disturbance $\varepsilon \in \mathcal{E} \subset \mathcal{R}^{d_\varepsilon}$ is allowed to be of arbitrary dimension, so d_ε does not need to be finite. Without loss of generality, rewrite eq. (3) as

$$T = ZT_1(U_1) + (1 - Z)\, T_0(U_0), \tag{4}$$

where $T_z(\cdot), z = 0, 1$ are some unknown functions, and the reduced-form disturbance $U_z \in \mathcal{U}_z \subset \mathcal{R}$, $z = 0, 1$. Later we impose an assumption that essentially requires $T_z(\cdot)$ to be the quantile functions and $\mathcal{U}_z$ to be the rank variables. Note by construction $U = ZU_1 + (1 - Z)\, U_0$.

Define $Y_t := g(t, \varepsilon)$ as the potential outcome when T is exogenously set to be t. Further define $T_z := T_z(U_z)$, $z = 0, 1$, as the potential treatment when Z is exogenously set to be z. Denote the support of T_z as $\mathcal{T}_z$. The observed treatment is then $T = ZT_1 + (1 - Z)\, T_0$. We use $F_{\cdot}(\cdot)$ and $F_{\cdot|\cdot}(\cdot|\cdot)$ to denote the unconditional cumulative distribution function (CDF) and conditional CDF, respectively.

Assumption 1 (Treatment quantile representation)**.** *$T_z(u)$ is strictly increasing in u, and $U_z \sim Unif(0,1)$, $z = 0, 1$.*

Assumption 1 requires that the potential treatment T_z is continuous with a strictly increasing CDF. The condition $U_z \sim Unif(0,1)$ involves a normalization. This kind of normalization is necessary, since the identification

results hold up to a monotonic transformation of U_z, as long as U_z is continuous with a strictly increasing CDF. See discussions in Matzkin (2003) and more recently Torgovitsky (2015). By Assumption 1, $T_z(u)$ is the u quantile of T_z, and $U_z = F_{T_z}(T_z)$ is the rank of the potential treatment. Further, $U = ZU_1 + (1 - Z)\,U_0$ is the observed treatment rank.

Assumption 2 (Independence)**.** $Z \perp (U_z, \varepsilon)$, $z = 0, 1$.

Assumption 2 essentially requires Z to be randomly assigned. More generally, we can allow the independence condition to hold only after conditioning on relevant pre-determined covariates, which we will discuss in the next session. Assumptions 1 and 2 imply $U \perp Z$, because for $z = 0, 1$, $\Pr(U \leq \tau | Z = z) = \Pr(U_z \leq \tau | Z = z) = \Pr(U_z \leq \tau) = \tau$, where the last equality follows the condition $Z \perp U_z$ as implied by Assumption 2.

Assumption 3 (First-stage)**.** $T_1(u) \neq T_0(u)$ *for at least some* $u \in (0, 1)$.

Assumption 3 requires that the distribution of T changes with Z. Assumption 3 is strictly weaker than the standard first-stage assumption of the LATE model, which requires $\mathbb{E}[T_1] \neq \mathbb{E}[T_0]$. For example, when the policy instrument Z affects the variance or shifts the tails of the treatment distribution but otherwise leaves the average treatment level unaffected, we have the standard LATE first-stage assumption fails, but the above Assumption 3 holds.

Assumption 4 (Monotonicity)**.** $\Pr(T_1 \geq T_0) = 1$.

Assumption 4 requires that treatment can only change in one direction when Z changes - without loss of generality, we normalize it to be non-decreasing. For example, this assumption holds in the usual linear regression model of T with a constant coefficient on Z.

Assumption 4 can not be tested directly, but it has testable implications. It implies $T_1(u) - T_0(u) \geq 0$ for all $u \in (0, 1)$, i.e., T_1 stochastically dominates T_0. Since stochastic dominance is a necessary but not sufficient condition for Assumption 4, rejecting stochastic dominance could mean monotonicity does not hold, but failing to reject does not necessarily mean that monotonicity

holds. That is, in a given empirical scenario, even we see that the quantile curve of T_1 does not cross the quantile curve of T_0, it does not necessarily mean that Assumption 4 holds. Assumption 4 is essentially not verifiable. Assumptions 2 - 4 together imply $\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0] > 0$.

For convenience of exposition, we generalize the standard definition of compliers, which is defined for a binary treatment (Angrist, Imbens, and Rubin, 1996). Let $\mathcal{T}_c = \{(t_0, t_1) \in \mathcal{T}_0 \times \mathcal{T}_1 : t_1 - t_0 > 0\}$ be the set of compliers. Define $LATE(t_0, t_1) := \mathbb{E}\left[\frac{Y_{t_1} - Y_{t_0}}{t_1 - t_0} | T_1 = t_1, T_0 = t_0\right]$ for any $(t_0, t_1) \in \mathcal{T}_c$. $LATE(t_0, t_1)$ is the local average treatment effect for complier type $(t_0, t_1) \in \mathcal{T}_c$. For example, in the case of a binary treatment, $\tau^{Wald} = LATE(0, 1)$. More generally when treatment is continuous as in our setup, τ^{Wald} is a weighted average of $LATE(t_0, t_1)$ for all $(t_0, t_1) \in \mathcal{T}_c$. We formalize this result in the following lemma.

Lemma 1. *Let Assumptions 1-4 hold. Then*

$$\tau^{Wald} = \iint_{\mathcal{T}_c} w_{t_0,t_1} LATE(t_0, t_1) F_{T_0,T_1}(dt_0, dt_1)$$

where $w_{t_0,t_1} = (t_1 - t_0) / \iint_{\mathcal{T}_c} (t_1 - t_0) F_{T_0,T_1}(dt_0, dt_1)$.

The above lemma states that under Assumptions *1-4*, τ^{Wald} in eq. (1) identifies a weighted average of the average treatment effects for different compliers, where the weights are proportional to their treatment intensity change $(t_1 - t_0)$. Frölich (2007) gives a comparable expression when treatment is multi-valued. Angrist and Imbens (1995) provide a slightly different expression than that of Frölich (2007), but as pointed out by Frölich (2007), these expressions are equivalent.[3]

When $g(T, \varepsilon)$ is continuously differentiable in its first argument, the identified causal parameter can be further expressed as a weighted average derivative of Y w.r.t. T, following Angrist et al. (2000, Theorem 1). The exact form of the weighted average derivative is provided in the proof of Lemma 1 in the

[3]Unlike in Frölich (2007) and here, Angrist and Imbens (1995) present the weighted average effect in terms of overlapping subpopulations.

online supplementary appendix.[4]

In the following, we provide an alternative assumption to Assumption 4, which allows us to identify causal effects at different treatment quantiles and further a convexly weighted average effect. This convexly weighted average effect is in contrast to τ^{Wald}, which is also a convexly weighted average effect under Assumption 4 monotonicity.

Assumption 5 (Treatment Rank Similarity)**.** $U_0|\varepsilon \sim U_1|\varepsilon$.

Assumption 5 assumes that conditional on ε, U_0 and U_1 follow the same distribution. Without conditioning on ε, U_0 and U_1 both follow a uniform distribution over the unit interval due to normalization, so $F_{U_0}(u) = F_{U_1}(u)$ by construction. Assumption 5 implies $\varepsilon|U_0 = u \sim \varepsilon|U_1 = u$ by Bayes' theorem, so ε has the same distribution at the same rank of the potential treatment.

A slightly stronger assumption is rank invariance, which is the condition $U_0 = U_1$. Rank invariance essentially requires that the joint distribution of T_0 and T_1 are degenerate. Rank invariance holds trivially when the treatment model is additively separable in a scalar disturbance, but this assumption does not require additive separability in general. Rank invariance is frequently imposed in the non-separable IV literature. For example, Imbens and Newey (2009) propose a control variable approach to identify various causal parameters for the non-separable IV model. They assume that in the treatment model $T = h(Z, U)$, U is a scalar unobservable, and that $h(Z, u)$ is strictly increasing in u with probability 1. Monotonicity in a scalar disturbance implies rank invariance, because under this assumption, $U_z{:=}F_{T_z}(T_z) = F_U(u)$ for any z in the support of Z. When $Z \in \{0, 1\}$, it means $U_0 = U_1$. In addition to rank invariance, Imbens and Newey (2009) assume $Z \perp (U, \varepsilon)$, which is equivalent to Assumption 2 when rank invariance holds.

Rank similarity in Assumption 5 relaxes rank invariance - instead of assuming the ranks of the potential treatments to be the same, it only assumes that they have the same conditional probability distribution for any given ε, and

[4]Our weighted average derivative appears to be different than that of Angrist et al. (2000). Similar to the point made in Frölich (2007), both are equivalent and the difference lies in that they express their weighted average derivative in terms of overlapping subpopulations.

thereby permits random deviations from the common rank level between the potential treatments. For example, if the common rank level for night sleep (the actual time one is in sleep as measured by actigraphy) is determined by individuals' biological clock (possibly after conditioning on observable covariates as disscussed in our general setup), which does not change with Z, then rank similarity permits that the increase in night sleep is subject to some random factors. Rank similarity was proposed by Chernozhukov and Hansen (2005, 2006) to identify quantile treatment effects in IV models. Note that they impose rank similarity on the ranks of potential outcomes, instead of ranks of potential treatments.

Lemma 2. *Under Assumptions 1, 2 and 5,* $T \perp \varepsilon | U$.

Lemma 2 suggests that U is a control variable as defined by Imbens and Newey (2009), i.e., conditional on the observed treatment rank U, T is exogenous to Y. Intuitively, under Assumptions 2 and 5 and holding U fixed, the only variation in T is the exogenous variation induced by Z.[5]

Based on Lemma 2, one may condition on U in the outcome equation to estimate the causal effect of T on Y. Let $q_z(u) = F^{-1}_{T|Z}(u|z)$ be the conditional u quantile of T given $Z = z$, and further $\Delta q(u) = q_1(u) - q_0(u)$. In addition, let $\mathcal{U} = \{u \in (0,1) \colon \Delta q(u) \neq 0\}$. By conditioning on $U = u$, for any $u \in \mathcal{U}$, the resulting IV estimand can be written as

$$\tau(u) := \frac{\mathbb{E}[Y|Z=1, U=u] - \mathbb{E}[Y|Z=0, U=u]}{\mathbb{E}[T|Z=1, U=u] - \mathbb{E}[T|Z=0, U=u]}. \tag{5}$$

The numerator captures the reduced-form effect of Z on Y given $U = u$, while the denominator captures the first-stage treatment change given $U = u$. The corresponding estimator (by replacing the population means and ranks by their sample analogues) is analogous to the indirect least square estimator in the linear IV model setting.

[5]This result is closely related to Theorem 1 of Imbens and Newey (2009), except that we assume rank similarity instead of rank invariance and that we focus on a binary IV instead of an IV that may have a large support. The large support is required to identify structural parameters, like the average structural function (Blundell and Powell, 2003), when the outcome disturbance is of arbitrary dimension.

Intuitively, conditional on $U = u$, with a binary instrument, T potentially can take two values $T_0(u)$ and $T_1(u)$. When T changes exogenously from $T_0(u)$ and $T_1(u)$, the corresponding average effect on the outcome $\mathbb{E}\left[Y_{T_1(u)} - Y_{T_0(u)}|U = u\right]$ can be identified, as we show in the following Theorem 1. For notational convenience, let $\Delta T(u) = T_1(u) - T_0(u)$.

Theorem 1. *Let Assumptions 1-3 and Assumption 5 hold. Then for any* $u \in \mathcal{U}$,

$$\begin{aligned} \tau(u) &= \mathbb{E}\left[\frac{Y_{T_1(u)} - Y_{T_0(u)}}{\Delta T(u)}|U = u\right] \quad (6) \\ &= \int \{g(T_1(u), e) - g(T_0(u), e)\} \frac{1}{\Delta T(u)} F_{\varepsilon|U}(de|u). \quad (7) \end{aligned}$$

To see the above results, note

$$\begin{aligned} \mathbb{E}[Y|Z = 1, U = u] &= \mathbb{E}[g(T_1(u), \varepsilon)|Z = 1, U = u] \\ &= \mathbb{E}[g(T_1(u), \varepsilon)|U = u] \\ &= \mathbb{E}\left[Y_{T_1(u)}|U = u\right] \end{aligned}$$

where the first equality follows from the models of Y and T, (2) and (4), respectively, the second equality follows from the condition $Z \perp \varepsilon|U$ as shown in the proof of Lemma 2, and the last equality is by the definition of the potential outcome. One can similarly show $\mathbb{E}[Y|Z = 0, U = u] = \mathbb{E}\left[Y_{T_0(u)}|U = u\right]$. That is, we can identify $\mathbb{E}\left[Y_{T_z(u)}|U = u\right]$ for $z = 0, 1$ and $u \in \mathcal{U}$. Ideally one may wish to recover $\mathbb{E}[Y_t]$ for any $t \in \mathcal{T}$, which is known as the average dose-response function or the average structural function. However, identifying $\mathbb{E}[Y_t]$ for any $t \in \mathcal{T}$ is not possible in our setup, because we have a binary instrument and we do not restrict the dimensionality of the outcome disturbance, i.e., we do not impose rank invariance in the outcome model.

Theorem 1 shows that $\tau(u)$ identifies an average (per unit) treatment effect at the u quantile of the treatment. $\tau(u)$ measures treatment effect heterogeneity at different treatment intensities, which can be useful. The denominator in

eq. (6) reflects the fact that $T_z(u) \notin \{0,1\}$ in general. Inside of the integral in eq. (7), T exogenously changes from $T_0(u)$ to $T_1(u)$ while holding ε fixed at e, so $\tau(u)$ is causal from a *ceteris paribus* point of view.

By eq. (4) and Assumption 1, U and T follow a one-to-one mapping given $Z = z$, i.e., conditioning on $U = u$ is the same as conditioning on $T = T_z(u)$.[6] Further by Assumption 2, $T_z(u) = q_z(u)$. Let the conditional mean function of Y given Z and T be $m_z(t) = \mathbb{E}[Y|Z = z, T = t]$, $z = 0, 1$. Then $\tau(u)$ in eq. (5) can be re-written as

$$\tau(u) = \frac{m_1(q_1(u)) - m_0(q_0(u))}{q_1(u) - q_0(u)}. \tag{8}$$

Later our estimation is directly based on eq. (8).

Oftentimes, researchers or policy makers are interested in some summary measure of the overall treatment effect. With $\tau(u)$, one can further identify and estimate a weighted average of $\tau(u)$, i.e.,

$$\tau^{RS}(w) := \int_{\mathcal{U}} \tau(u) w(u) du$$

for any known or estimable weighting function $w(u)$ such that $w(u) \geq 0$ and $\int_{\mathcal{U}} w(u)du = 1$. The weighting function $w(u)$ is required to be non-negative; otherwise, $\tau^{RS}(w)$ can be a weighted difference of the average treatment effects for units. For example, if $\mathcal{U} = (0,1)$, and one chooses $w(u) = 1$, then $\tau^{RS}(w) = \mathbb{E}[\tau(U)]$.

$\tau^{RS}(w)$ is a weighted average of the average treatment effects at all treatment quantiles where treatment changes under Assumption 5, treatment rank similarity. By Lemma 1, τ^{Wald} is a weighted average of the average treatment effects for all compliers under Assumption 4, monotonicity. Note that both assumptions impose restrictions on the first-stage IV effect heterogeneity - monotonicity imposes a sign restriction, while treatment rank similarity imposes a rank restriction. Neither assumption implies the other. Neither assumption is verifiable. In practice, it is not ideal to have to choose one versus

[6] The σ-algebra is the same.

the other estimand based on some pre-testing results. We therefore consider a weighting function that leads to a DR property of the resulting estimand, i.e., the estimand is valid under either of the two alternative identifying assumptions.

Proposition 1. *Let Assumptions 1-3 hold. If either Assumption 4 or Assumption 5 holds, then* $\tau^{DR} := \int_{\mathcal{U}} \tau(u) w^{DR}(u) du$ *for* $w^{DR}(u) = |\Delta q(u)| / \int_{\mathcal{U}} |\Delta q(u)| du$ *identifies a weighted average of the average treatment effects among units for which* $T_1 \neq T_0$.

Proposition 1 combines the results of Lemma 1 and Theorem 1. It shows that under either first-stage restriction on the IV effect heterogeneity, τ^{DR} identifies a weighted average of the average effects for all the units that respond to the IV change. These units represent the largest subpopulation one can identify causal effects for without any further restrictions. The two alternative first-stage assumptions specify exactly how these units respond - either they change treatment in a monotonic way or they change treatment such that the probability distribution of their treatment ranks remains the same.

When Assumption 4 monotonicity holds, $w^{DR}(u) = \Delta q(u) / \int_{\mathcal{U}} \Delta q(u) du$. Then

$$\begin{aligned} \tau^{DR} &= \frac{\int_0^1 \{\mathbb{E}[Y|Z=1, U=u] - \mathbb{E}[Y|Z=0, U=u]\} du}{\int_0^1 \Delta q(u) du} \\ &= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1] - \mathbb{E}[T|Z=0]} \\ &= \tau^{Wald} \end{aligned}$$

That is, τ^{DR} reduces to the standard LATE estimand τ^{Wald} given by eq. (1) when monotonicity holds. By Lemma 1, in this case τ^{DR} identifies a weighted average of the average treatment effects for different compliers. Otherwise, when Assumption 4 monotonicity does not hold, but Assumption 5 rank similarity holds, τ^{DR} is a weighted average of $\tau(u)$ for $u \in \mathcal{U}$, and by Theorem 1, $\tau(u)$ captures the average treatment effects at the u quantile of treatment. Either way, τ^{DR} identifies a weighted average of the average treatment effects for

all the units that change their treatment levels in response to the IV changes. The weights are proportional to the magnitude of their treatment changes.

The weighting function in Proposition 1 allows $\Delta q(u)$ to change signs, which in fact indicates that the LATE monotonicity condition does not hold. As a result, τ^{DR} may average over two different types of units, those who increase their treatment levels and those who decrease their treatment levels when the IV changes.[7] Ideally one may want to separately consider these two types of units. However, individual types are not point identified, so point identification of causal effects over different individual types is not possible. As a mitigation measure, if desired, one may separately consider treatment quantiles where $\Delta q(u) > 0$ and those where $\Delta q(u) < 0$.

Let $\mathcal{U}_+ = \{u \in \mathcal{U} : \Delta q(u) > 0\}$. Define $\tau_+^{DR} := \int_{\mathcal{U}_+} \tau(u) w_+(u) du$, where $w_+(u) = \Delta q(u) / \int_{\mathcal{U}_+} \Delta q(u)\, du$. τ_+^{DR} can be rewritten as the ratio of the mean outcome difference over $u \in \mathcal{U}_+$ to the mean treatment difference over $u \in \mathcal{U}_+$, i.e.,

$$\tau_+^{DR} = \frac{\int_{\mathcal{U}_+} \int \{g(T_1(u), e) - g(T_0(u), e)\} F_{\varepsilon|U}(de|u)\, du}{\int_{\mathcal{U}_+} \Delta T(u) du}.$$

τ_+^{DR} carries a similar interpretation as that of τ^{DR} but is only for the subset of treatment quantiles $u \in \mathcal{U}_+$.[8] In particular, when either monotonicity or treatment rank similarity holds over $\mathcal{U}_+$, τ_+^{DR} identifies a weighted average of the average treatment effects for all the responding (to IV changes) units associated with this subset of quantiles. Similarly, one can define $\tau_-^{DR} := \int_{\mathcal{U}_-} \tau(u) w_-(u) du$, where $\mathcal{U}_- = \{u \in \mathcal{U} : \Delta q(u) < 0\}$ and $w_-(u) = \Delta T(u) / \int_{\mathcal{U}_-} \Delta q(u)\, du$.

So far, we have focused our discussion on (weighted) average effects. One may extend the above identification results to identify distributional effects at a given treatment quantile $u \in \mathcal{U}$. In particular, under Assumptions 1-3

[7] This issue is not unique to our setting. This issue would arise whenever a research estimates some average effects, but does not assume that the treatment change cannot switch signs.

[8] Under treatment rank invariance, monotonicity holds automatically if treatment quantile changes do not switch signs. This is not true in general. Intuitively under rank invariance, $U_1 = u$ implies $U_0 = u$ and vise versa; however, under treatment rank similarity (but not rank invariance), individuals counterfactual treatment rank is not point identified, and hence monotonic treatment quantile changes do not gurantee individual level monotonicity.

and Assumption 5, for any $u \in \mathcal{U}$, $F_{Y_{T_z(u)}|U}(y|u) = \mathbb{E}[\mathbf{1}(Y \leq y)|U = u, Z = z]$, $z = 0, 1$. Let the conditional quantile function of $Y_{T_z(u)}$ given $U = u$ be $F^{-1}_{Y_{T_z(u)}|U=u}(\widetilde{u})$ for $\widetilde{u} \in (0, 1)$ and $z = 0, 1$. The corresponding reduced-form quantile treatment effect is given by $F^{-1}_{Y_{T_1(u)}|U=u}(\widetilde{u}) - F^{-1}_{Y_{T_0(u)}|U=u}(\widetilde{u})$ for any $\widetilde{u} \in (0, 1)$ and $u \in \mathcal{U}$. Replacing Y by $\mathbf{1}(Y \leq y)$ in Theorem 1 and further in Proposition 1 leads to the DR estimand for the weighted average effect of T on $\mathbf{1}(Y \leq y)$ for all $y \in \mathcal{Y}$. It is worth emphasizing that our goal is to develop robust identification results in the presence of both treatment effect heterogeneity and instrument effect heterogeneity. Whether any of these proposed weighted averages are of interest depends on empirical scenarios.

3 Doubly Robust Identification with Covariates

The previous section presents our core idea without considering covariates. IV independence and rank similarity may be more plausible when conditioning on relevant pre-determined covariates. For this claim on rank similarity, see e.g., discussion in Chernozhukov and Hansen (2005, 2006). If covariates enter the non-separable models for Y and T, i.e., (2) and (3), and all the previous assumptions hold conditional on covariates, then it follows readily that all the previous results hold conditional on covariates. However, such conditional results may not be very useful in practice, as it can be unwieldy to present all the conditional results if there are many covariates and worse many continuous covariates. In this section, we seek to directly identify unconditional weighted average effects as before while allowing for covariates.

Let $X \in \mathcal{X} \subset \mathcal{R}^{d_X}$ denote the vector of covariates. We do not require X to be exogenous. We consider the following models for Y and T:

$$\begin{aligned} Y &= G(T, X, \epsilon), && (9) \\ T &= H(Z, X, V) \\ &= ZT_1(X, V_1) + (1 - Z)T_0(X, V_0), && (10) \end{aligned}$$

where by construction $V = V_1 Z + V_0(1 - Z)$.

As before, $T_z{:=}T_z(X, V_z)$ is the potential treatment Z is exogenously set to be $z \in \{0, 1\}$ and $Y_t{:=}G(t, X, \epsilon)$ is the potential outcome when T is exogenously set to be $t \in \mathcal{T} \subset \mathcal{R}$. We extend Assumptions 1, 2, 3 and 5 to condition on covariates X as follows.[9]

Assumption C1 (Conditional Treatment Quantile). *For any $x \in \mathcal{X}$, $T_z(x, v)$, $z = 0, 1$, is strictly increasing in v, and $V_z \sim Unif(0, 1)$.*

By Assumption C1, $T_z(x, v)$ is the conditional quantile function of T_z given X, and $V_z = F_{T_z|X}(T_z|X)$ is the conditional rank of T_z given X.

Assumption C2 (Conditional Independence). $Z \perp (V_z, \epsilon)\,|X$, $z = 0, 1$.

Assumption C3 (Conditional First-stage). $T_z(x, v) \neq T_z(x, v)$ *for at least some $x \in \mathcal{X}$ and $v \in (0, 1)$.*

Assumption C4 (Common Support). $\Pr(Z = 1|X = x) \in (0, 1)$ *for any $x \in \mathcal{X}$.*

Assumption C5 (Conditional Treatment Rank Similarity). $V_1|(\epsilon, X) \sim V_0|(\epsilon, X)$.

Assumption C2 requires Z to be unconfounded, instead of being randomly assigned as required by Assumption 2. Assumption C4 is a common support assumption to ensure our parameters are well-defined. In addition, Assumption C5 requires that treatment rank similarity holds only among the subgroup of units with the same observed covariate values, which is weaker than Assumption 5.[10]. The following Lemma extends Lemma 2 to allow for covariates.

[9]We do not extend Assumption 4 monotonicity, as little will be changed from the identification perspective. For example, one way to relax Assumption 4 is to assume that either $\Pr(T_1 \geq T_0|X = x) = 1$ or $\Pr(T_1 \leq T_0|X = x) = 1$ for any $x \in \mathcal{X}$. Since the sign of the first-stage change is identified from the data given the assumptions here, for those covariate values at which treatment change is negative when Z changes from 0 to 1 (consistent with $\Pr(T_1 \leq T_0|X = x) = 1$), one may change the observed value $Z = 1$ to $Z = 0$ and similarly $Z = 0$ to $Z = 1$, so that after the switch, the condition $\Pr(T_1 \geq T_0|X = x) = 1$ for any $x \in \mathcal{X}$ holds, which then is essentially the same as having $\Pr(T_1 \geq T_0) = 1$. Our identification results in this section would go through with this rearranging values of Z.

[10]Note that V_z is defined conditionally on X, while U_z is defined unconditionally. Given that X are determinants of Y, one can let X be an observable subvector of ε in $Y = g(T, \varepsilon)$. That is, $\varepsilon = (X, \epsilon)$. Assumption 5 $U_1|\varepsilon \sim U_0|\varepsilon$

Lemma 3. *Under Assumptions C1-C3 and C5,* $T \perp \epsilon | (V, X)$.

Lemma 3 is a conditional (on X) version of Lemma 2. So V is a control variable given X. Let $q_z(v, x) = F^{-1}_{T|Z,X}(v|z, x)$ be the conditional v quantile of T given $Z = z$ and $X = x$. Let $\Delta q(x, v) = q_1(v, x) - q_0(v, x)$. Assumptions C1 and C4 ensure that $\Delta q(x, v)$ is well defined for all $x \in \mathcal{X}$ and $v \in (0, 1)$. Further let $\mathcal{S} = \{(x, v) \in \mathcal{X} \times (0, 1) \colon \Delta q(x, v) \neq 0\}$. The resulting IV estimand by conditioning on $X = x$ and $V = v$ can be defined as

$$\pi(x, v) := \frac{\mathbb{E}[Y|Z = 1, X = x, V = v] - \mathbb{E}[Y|Z = 0, X = x, V = v]}{\mathbb{E}[T|Z = 1, X = x, V = v] - \mathbb{E}[T|Z = 0, X = x, V = v]} \tag{11}$$

for any $(x, v) \in \mathcal{S}$. By eq. (10) and Assumption C1, T and V follow a one-to-one mapping given $Z = z$ and $X = x$, i.e., conditioning on $V = v$ is the same as conditioning on $T = T_z(x, v)$ in (11). Further by Assumption C2, $T_z(x, v) = q_z(x, v)$. Then $\pi(x, v)$ can be re-written as

$$\pi(x, v) = \frac{m_1(x, q_1(x, v)) - m_0(x, q_0(x, v))}{q_1(x, v) - q_0(x, v)},$$

where $m_z(x, t) = \mathbb{E}[Y|Z = z, X = x, T = t]$.

Let $\Delta T(x, v) = T_1(x, v) - T_0(x, v)$. We have the following Theorem 2, which extends Theorem 1.

Theorem 2. *Let Assumptions C1-C5 hold. Then for any* $(x, v) \in \mathcal{S}$,

$$\pi(x, v) = \mathbb{E}\left[\frac{Y_{T_1(x,v)} - Y_{T_0(x,v)}}{\Delta T(x, v)} | X = x, V = v\right] \tag{12}$$

$$= \int \{G(T_1(x, v), x, e) - G(T_0(x, v), x, e)\} \frac{F_{\epsilon|X,V}(de|x, v)}{\Delta T(x, v)}. \tag{13}$$

implies $U_1|X \sim U_0|X$, so $F_{U_1|X}(u|x) = F_{U_0|X}(u|x)$ for any $u \in (0, 1)$ and $x \in \mathcal{X}$. It follows that $F_{V_0|X,\epsilon}(v|X = x, \epsilon = e) = \mathbb{E}[1(V_0 \leq v)|X = x, \epsilon = e] = \mathbb{E}\left[1\left(F_{U_0|X}(U_0|x) \leq v\right)|X = x, \epsilon = e\right] = \mathbb{E}\left[1\left(F_{U_1|X}(U_1|x) \leq v\right)|X = x, \epsilon = e\right] = F_{V_1|X,\epsilon}(v|X = x, \epsilon = e)$ for any v, x, and e in their support, where the second equality follows from $V_z = F_{Tz|X}(T_z|X)$ by Assumption C1, which can be further written as $V_z = F_{Uz|X}(U_z|X)$, $z = 0, 1$, since T_z and U_z follow a one-to-one mapping by Assumption 1. Therefore, $V_0|X, \epsilon \sim V_1|X, \epsilon$.

By Theorem 2, $\pi(x,v)$ identifies a conditional weighted average treatment effect at the conditional v quantile of the treatment given $X=x$. By eq. (13), it is clear that $\pi(x,v)$ represents the causal effect of an exogenous change in treatment from $T_0(x,v)$ to $T_1(x,v)$, while holding X and ϵ fixed at x and e.

If desired, one may average $\pi(x,v)$ over the distribution of X to obtain a weighted average effect at the conditional v quantile of the treatment. For notational convenience, in the following, we assume $\pi(x,v)=0$ when $\Delta q(x,v)=0$, so that $\pi(x,v)$ is defined for all $(x,v)\in\mathcal{X}\times(0,1)$. For example, for any $v\in(0,1)$ such that $\Pr(\Delta q(X,v)\neq 0)>0$, one can define

$$\pi(v):=\int_{\mathcal{X}}\pi(x,v)w_v(x)dx,$$

where $w_v(x)=|\Delta q(x,v)|f_X(x)/\int_{\mathcal{X}}|\Delta q(x,v)|f_X(x)dx$. $\pi(v)$ identifies a weighted average effect at the conditional v quantile of the treatment. In contrast, $\tau(u)$ identifies an average effect at the unconditional u quantile of the treatment. $\pi(v)$ can be useful in investigating treatment effect heterogeneity at the conditional v quantile of the treatment.

Consider now constructing a DR estimand for the overall unconditional weighted average effect based on $\pi(x,v)$. Since Z is valid only after conditioning on pre-determined covariates, τ^{Wald} is no longer a valid causal estimand. Define

$$\tau^{Wald\_X}:=\frac{\int_{\mathcal{X}}\{\mathbb{E}[Y|Z=1,X=x]-\mathbb{E}[Y|Z=0,X=x]\}f_X(x)\,dx}{\int_{\mathcal{X}}\{\mathbb{E}[T|Z=1,X=x]-\mathbb{E}[T|Z=0,X=x]\}f_X(x)dx}. \quad (14)$$

The numerator of eq. (14) does not reduce to $\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]$ and similarly the denominator of eq. (14) does not reduce to $\mathbb{E}[T|Z=1,X=x]-\mathbb{E}[T|Z=0,X=x]$, as X is not required to be independent of Z, and hence the distribution of X given $Z=0$ and that given $Z=1$ are different in general. Nevertheless, the following lemma shows that τ^{Wald_X} identifies the same unconditional effect as what τ^{Wald} would if Z were valid without conditioning on covariates.

Lemma 4. *Let Assumptions C1 - C4 and further Assumption 4 hold. Then*

$$\tau^{Wald\_X} = \iint_{\mathcal{T}_c} w_{t_0,t_1} LATE\left(t_0,t_1\right) F_{T_0,T_1}\left(dt_0,dt_1\right)$$

where $w_{t_0,t_1} = \left(t_1 - t_0\right) / \iint_{\mathcal{T}_c} \left(t_1 - t_0\right) F_{T_0,T_1}\left(dt_0,dt_1\right)$.

Frölich (2007) presents a comparable result for a binary or discrete treatment. Our DR estimand below incorporates τ^{Wald_X} (instead of its invalid counterpart τ^{Wald}) as a special case.

Proposition 2. *Let Assumptions C1 - C4 hold. When further either 4 or C5 holds,*

$$\pi^{DR} := \iint_{\mathcal{S}} \pi(x,v) w\left(x,v\right) dvdx$$

for $w(x,v) = |\Delta q(x,v)| f_X(x) / \iint_{\mathcal{S}} |\Delta q(x,v)| f_X(x) dvdx$ *identifies a weighted average of the average treatment effects among all the units for which* $T_1 \neq T_0$.

When Assumption C5 conditional rank similarity holds, π^{DR} is a weighted average of $\pi\left(x,v\right)$ for $(x,v) \in \mathcal{S}$, which by Theorem 2, is a causal estimand; Otherwise, when Assumption 4 monotonicity holds, $\pi^{DR} = \tau^{Wald\_X}$, which we show in Lemma 4 identifies a weighted average of $LATE(t_0,t_1)$ for $(t_0,t_1) \in \mathcal{T}_c$. Either way, π^{DR} identifies a weighted average of the average treatment effects for all the units responding to the IV change, the largest subpopulation one can identify treatment effects without further assumptions. The weights are proportional to both the magnitude of the treatment change and the density of X.

Let $\mathcal{S}_+ = \{(x,v) \in \mathcal{X} \times (0,1)\text{: } \Delta q(x,v) > 0\}$ and $\mathcal{S}_- = \{(x,v) \in \mathcal{X} \times (0,1)\text{: } \Delta q(x,v) < 0\}$. Define

$$\pi_+^{DR} := \iint_{\mathcal{S}_+} \pi(x,v) w_+(x,v) dvdx, \tag{15}$$

where $w_+(x,v) = \Delta q(x,v) f(x) / \iint_{\mathcal{S}_+} \Delta q(x,v) f(x) dvdx$. π_+^{DR} identifies a weighted average of the average treatment effects for all the responding units with $(x,v) \in S_+$, when either monotonicity or conditional treatment rank similar-

ity holds for $\mathcal{S}_+$. π_-^{DR} can be analogously defined by replacing $w_+(x,v)$ with $w_-(x,v)$ and $\mathcal{S}_+$ with $\mathcal{S}_-$ in eq. (15) respectively. π_-^{DR} identifies a weighted average of the average treatment effects for units experiencing negative treatment changes, regardless of whether they stay at the same treatment rank or not.

4 Estimation and Inference

For estimation and inference, we focus on the general setup with covariates. To avoid cumbersome fully nonparametric estimation, we assume that covariates enter linearly and propose convenient semi-parametric estimation. We briefly discuss the practical implications of the additional functional form assumptions required in our estimation toward the end of this section. The estimation without covariates can be seen as a special case of that with covariates.

4.1 Estimation

We assume a linear quantile regression model for the conditional v quantile of T given $Z = z$ and $X = x$, i.e., $q_z(x,v) = a_0(v) + x'a_1(v) + za_2(v) + zx'a_3(v)$; we further assume a partially linear model for the conditional mean function of Y given Z, X and T, i.e., $m_z(x,t) = x'b_0 + g_0(t) + zx'b_1 + zg_1(t)$, where g_z, $z = 0, 1$, are some unknown functions. Given a sample of $i.i.d.$ observations $\{(Y_i, T_i, X_i, Z_i)\}_{i=1}^n$ for (Y, T, X, Z), we propose the following estimation procedure.

Step 1. Estimate the first-stage conditional treatment quantiles $q_z(x,v)$:

- $\widehat{q}_z(x,v) = \widehat{a}_0(v) + x'\widehat{a}_1(v) + z\hat{a}_2(v) + zx'\widehat{a}_3(v)$
  for $v \in V^{(l)}$, where $V^{(l)} = \{v_1, v_2, ..., v_l\}$ is the set of equally spaced quantiles over $(0,1)$.
  Then $\Delta\widehat{q}(x,v) = \hat{a}_2(v) + x'\widehat{a}_3(v)$.

Step 2. Estimate the conditional mean function $m_z(x,t)$ by a partially linear series estimator:

- $\widehat{m}_z(x,t) = x'\widehat{b}_0 + \widehat{g}_0(t) + zx'\widehat{b}_1 + z\widehat{g}_1(t)$
  Let $\Delta\hat{m}(X_i,v) = \widehat{m}_1(X_i,\widehat{q}_1(X_i,v)) - \widehat{m}_0(X_i,\widehat{q}_0(X_i,v))$.

Step 3. Assume that the trimming parameter ϱ_n is a positive sequence that goes to $\varrho = 0$ as $n \to \infty$. For $v \in V^{(l)}$ and $i = 1, ..., n$, the plug-in estimator of $\pi(X_i, v)$ is $\hat{\pi}(X_i, v) = \Delta\hat{m}(X_i, v)/\Delta\widehat{q}(X_i, v)$ when $|\Delta\widehat{q}(X_i, v)| \geq \varrho_n$. Let $\widehat{\pi}(X_i, v) = 0$ when $|\Delta\widehat{q}(X_i, v)| < \varrho_n$.

- Estimate $\pi(v)$ for $v \in V^{(l)}$ such that $\sum_i 1\big(|\Delta\widehat{q}(X_i,v)| \geq \varrho_n\big) \neq 0$:
  $\widehat{\pi}(v) = \sum_i \widehat{\pi}(X_i, v)\widehat{w}_v(X_i)$, where $\widehat{w}_v(X_i) = \frac{|\Delta\widehat{q}(X_i,v)|1(|\Delta\widehat{q}(X_i,v)|\geq\varrho_n)}{\sum_i |\Delta\widehat{q}(X_i,v)|1(|\Delta\widehat{q}(X_i,v)|\geq\varrho_n)}$
- Estimate π^{DR}: $\widehat{\pi}^{DR} = \sum_{v\in V^{(l)}} \sum_i \widehat{\pi}(X_i, v)\widehat{w}(X_i, v)$,
  where $\widehat{w}(X_i, v) = \frac{|\Delta\widehat{q}(X_i,v)|1(|\Delta\widehat{q}(X_i,v)|\geq\varrho_n)}{\sum_{v\in V^{(l)}} \sum_i |\Delta\widehat{q}(X_i,v)|1(|\Delta\widehat{q}(X_i,v)|\geq\varrho_n)}$.

One may estimate π_+^{DR} or π_-^{DR} analogously by replacing $|\Delta\widehat{q}(X_i, v)|$ with $\Delta\widehat{q}(X_i, v)$ or $-\Delta\widehat{q}(X_i, v)$, respectively. The following provides details on the partial linear series estimator in Step 2. Let $\{\psi_{J1}, ..., \psi_{JJ}\}$ be a collection of basis functions of t for approximating the nonparametric component $g_z(t)$. Let $\psi^J(x,t,z) = \big(x', \psi_{J1}(t), ..., \psi_{JJ}(t), zx', z\psi_{J1}(t), ..., z\psi_{JJ}(t)\big)'$, a $2(d_x+J)\times 1$ vector. Let $\Psi = (\psi^J(X_1,T_1,Z_1), ..., \psi^J(X_n,T_n,Z_n))'$, a $n \times 2(d_x + J)$ matrix. Then the series coefficient estimate is $\hat{c} = [\Psi'^{-}\Psi'(Y_1, ..., Y_n)'$, and a series least squares estimator of $m_z(x,t)$ is $\hat{m}_z(x,t) = \psi^J(x,t,z)'\hat{c}$.

For the trimming parameter ϱ_n, one may choose $\varrho_n = 1.96\times\min_{v\in V^{(l)},\{X_i\}_{i=1}^n} se(\Delta\widehat{q}(X_i, v))/\log(n)$. This ϱ_n satisfies the rate condition required by our asymptotic theory as that given in Theorems 3 and 4 in Section 4.2.

4.2 Asymptotic Theory

This section presents inference results for $\pi(v)$ and π^{DR}. Inference results for the other parameters $\pi(x, v)$ and $\pi_\pm^{DR}$ are presented in Section S.2 and Section S.3, respectively, in the online supplementary appendix.

We derive the asymptotic theory based on the literature of quantile regression and sieve estimation. The main complication here is that we need to account for the variation from the Step 1 quantile regression and Step 2 sieve estimation, as well as the trimming function. Let $a(v) = (a_0(v), a_1'(v), a_2(v), a_3'(v))'$

be the quantile coefficients in Step 1. For the quantile regression estimator $\hat{a}(v)$, we apply the results of Angrist, Chernozhukov, and Fernández-Val (2006). They show that $\hat{a}(v)$ converges uniformly over v in a closed subset of $(0,1)$ to a zero mean Gaussian process indexed by v. For the partially linear estimation in Step 2, we apply the results of Chen and Christensen (2018). They establish uniform inference for nonlinear functionals of nonparametric IV regression. We apply their results for a special case of exogenous regressors and linear functionals. Our assumptions for asymptotics collect the assumptions in these two papers.

Assumption A1 collects the conditions in Theorem 3 in Angrist, Chernozhukov, and Fernández-Val (2006).

Assumption A1. *The conditional density $f_{T|X,Z}(t|x,z)$ is bounded and uniformly continuous in t, uniformly for $x \in \mathcal{X}$, $z = 0, 1$. $\mathbb{E}[\|X\|^3] < \infty$. Let $\vartheta(v) := \mathbb{E}[f_{T|X,Z}(S'a(v)|X,Z)SS']$, where $S := (1, X', Z, ZX')'$, be positive definite for all $v \in \mathcal{V}$ which is a closed subset of $(0,1)$.*

Let $e = Y - \mathbb{E}[Y|Z,X,T]$. Let $G = \mathbb{E}\left[\psi^J(X,T,Z)\psi^J(X,T,Z)'\right] = \mathbb{E}[\Psi'\Psi/n]$ be positive definite for each J. Let $\Omega = \mathbb{E}\left[e^2\psi^J(X,T,Z)\psi^J(X,T,Z)'\right]$ and $\mho = G^{-1}\Omega G^{-1}$.

Let $L^\infty(T)$ denote the set of all bounded measurable functions $g : \mathcal{T} \to \mathcal{R}$ endowed with the sup-norm $\|g\|_\infty = \sup_t |g(t)|$. Let $\|\cdot\|_{\ell^q}$ denote the vector ℓ^q-norm when applied to vectors and the operator norm induced by the vector ℓ^q-norm when applied to matrices. If $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, then we say $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n < \infty$.

Consider a collection of linear functionals $\{L_\ell : \ell \in \mathcal{L}\}$ with an index set $\mathcal{L}$. For example, for the conditional mean function $m_z(x,t)$, one can let $L_\ell(m_z) = m_z(x,t)$ with $\ell = (x,t) \in \mathcal{L} = \mathcal{X} \times \mathcal{T}$, for $z = 0, 1$. Assumptions A2 and A3 below collect the assumptions in Chen and Christensen (2018).

Assumption A2. *1. (i) (X,T) have compact rectangular support $\mathcal{XT} \subset \mathcal{R}^{d_x+1}$ and the density of (X,T) is uniformly bounded away from 0 and ∞ on $\mathcal{XT}$.*

(ii) *For* $z = 0, 1$, $m_z \in \mathcal{H} \subset L^\infty(X,T)$. *The sieve space for* (X,T) *is the closed linear span* $\Psi_J = clsp\{\psi_{J1}, ..., \psi_{JJ}\} \subset L^2(X,T)$, *and* $\cup_J \Psi_J$ *is dense in* $(\mathcal{H}, \|\cdot\|_{L^\infty(X,T)})$.

2. *(i)* $\mathbb{E}\big[|e_i|^{2+\delta}\big] < \infty$ *for some* $\delta > 0$.

   (ii) $\mathbb{E}\big[|e_i|^3 | Z_i = z, X_i = x, T_i = t\big] < \infty$ *and* $\mathbb{E}\big[e_i^2 | Z_i = z, X_i = x, T_i = t\big] \in [\underline{\sigma}^2, \bar{\sigma}^2]$ *for some finite and positive constants* $(\underline{\sigma}^2, \bar{\sigma}^2)$, *uniformly for* $(x,t) \in \mathcal{XT}$, *for* $z = 0, 1$.

3. *(i)* Ψ_J *is Hölder continuous: there exist finite constants* $C \geq 0$, $\tilde{C} > 0$ *such that* $\|G^{-1/2}\{\psi^J(x,t,z) - \psi^J(\tilde{x},\tilde{t},z)\}\|_{\ell^2} \lesssim J^C \|(x,t) - (\tilde{x},\tilde{t})\|_{\ell^2}^{\tilde{C}}$ *for* $t, \tilde{t} \in \mathcal{T}$, $x, \tilde{x} \in \mathcal{X}$, $z = 0, 1$.

   (ii) *Let* $\zeta := \sup_{x,t,z} \|G^{-1/2}\psi^J(x,t,z)\|_{\ell^2}$ *satisfy* $\zeta^2/\sqrt{n} = O(1)$ *and* $\zeta^{(2+\delta)/\delta} \sqrt{(\log n)/n} = o(1)$.

4. *(i)* *Let* $\sigma_n^2(L_\ell) = L_\ell(\psi^J)' \mho L_\ell(\psi^J) \nearrow +\infty$ *as* $n \to \infty$ *for each* $\ell \in \mathcal{L}$. *Let* η_n *be a sequence of nonnegative numbers such that* $\eta_n = o(1)$. *Let* $\tilde{m}_z(x,t) = \psi^J(x,t,z)'\tilde{c}$ *where* $\tilde{c} = (\Psi'\Psi)^- \Psi'\big(m_{Z_1}(X_1,T_1), ..., m_{Z_n}(X_n,T_n)\big)'$ *and* $\sup_{\ell \in \mathcal{L}} \sqrt{n}|L_\ell(\tilde{m}_z(x,t)) - L_\ell(m_z(x,t))|/\sigma_n(L_\ell) = O_p(\eta_n)$.

   (ii) *Let* $u_n(L_\ell)(X_i,T_i,Z_i) = \psi^J(X_i,T_i,Z_i)'^{-1} L_\ell(\psi^J)/\sigma_n(L_\ell)$ *be the normalized sieve Riesz representer. Let* $d_n(\ell_1,\ell_2) = \big(\mathbb{E}\big[(u_n(L_{\ell_1})(X_i,T_i,Z_i) - u_n(L_{\ell_2})(X_i,T_i,Z_i))^2\big]\big)^{1/2}$ *be the semimetric on* $\mathcal{L}$. *Let* $N(\mathcal{XT}, d_n, \varsigma)$ *be the* ς*-covering number of* $\mathcal{XT}$ *with respect to* d_n. *There is a sequence of finite constant* $c_n \gtrsim 1$ *that could grow to infinity such that* $1 + \int_0^\infty \sqrt{\log N(\mathcal{XT}, d_n, \varsigma)} d\varsigma = O(c_n)$.

   (iii) *Let* $\delta_{m,n}$ *be a sequence of positive constants such that* $\|\hat{m}_z - m_z\|_\infty = O_p(\delta_{m,n}) = o_p(1)$. *Define* $\delta_{V,n} := \big(\zeta^{(2+\delta)/\delta}\sqrt{(\log J)/n}\big)^{\delta/(1+\delta)} + \delta_{m,n} + \zeta\sqrt{(\log J)/n}$. *There is a sequence of constant* $r_n > 0$ *decreasing to zero slowly such that (a)* $r_n c_n \lesssim 1$ *and* $\zeta J^2/(r_n^3\sqrt{n}) = o(1)$, *(b)* $\zeta\sqrt{(J \log J)/n} + \eta_n + \delta_{V,n} c_n = o(r_n)$.

Assumption A3. *Let* $J\sqrt{(J \log J)/n} = o(1)$. *Let* $B^p_{\infty,\infty}$ *denote the Hölder space of smoothness* $p > 0$ *and* $\|\cdot\|_{B^p_{\infty,\infty}}$ *denote its norm. Let* $B_\infty(p, L) =$

$\{m \in B^p_{\infty,\infty} : \|m\|_{B^p_{\infty,\infty}} \leq L\}$ *denote a Hölder ball of smoothness* $p > 1$ *and radius* $L \in (0, \infty)$. *Let* $m \in B_\infty(p, L)$ *and* Ψ_J *be spanned by a B-spline basis of order* $\gamma > p$ *or a CDV wavelet basis of regularity* $\gamma > p$.

Assumption A3 ensures the uniform consistency of $\partial_t \hat{m}_z(x,t) = \partial \hat{m}_z(x,t)/\partial t$, which is used to account for the Step 1 estimation error.

We show in Theorem 3 below that under Assumptions A1, A2, and A3, the influence function of $\hat{\pi}(v)$ is given by $R_i(v)/B(v) = (R_{1i}(v) + R_{2i}(v) + R_{3i}(v))/B(v)$, where $R_{1i}(v)$ captures the impact of Step 1, $R_{2i}(v)$ captures the impact of Step 2, $R_{3i}(v)$ is the influence function for the sample analogue estimator of $\hat{\pi}(v)$ (without accounting for the step 1 and step 2 estimation errors) in Step 3, and $B(v)$ is from the normalization in the weighting function. The exact formulas of $R_{ki}(v)$, $k = 1, 2, 3$, are given in (S.9) in the online supplementary appendix. Let $\sigma_n^2(v) = \mathbb{E}\left[R_i(v)^2\right]/B(v)^2$, which is the sieve variance of $\sqrt{n}\hat{\pi}(v)$. Further let $\hat{\sigma}^2(v)$ be a uniformly consistent estimator of $\sigma_n^2(v)$ in the sense that $\sup_{v \in \mathcal{V}_\varrho} |\sigma_n(v)/\hat{\sigma}(v) - 1| = o_p(1)$ for a closed set $\mathcal{V}_\varrho = \{v \in \mathcal{V} : \Pr(|\Delta q(X, v)| > \varrho) > 0\}$. For example, $\hat{\sigma}^2(v)$ can be estimated by the sample analogue plug-in estimator, i.e., $\hat{\sigma}^2(v) = n^{-1} \sum_{i=1}^n \widehat{R}_i(v)^2/\widehat{B}(v)^2$, where $\widehat{R}_i(v)$ and $\widehat{B}(v)$ are uniformly consistent estimators of $R_i(v)$ and $B(v)$, respectively. We give the estimation detail of $\hat{\sigma}^2(v)$ in Section S.4 in the online supplementary Appendix.

Theorem 3. *Let Assumptions A1, A2, and A3 hold. Let* $\sqrt{n}(\varrho_n - \varrho) = o(1)$. *Then* $\sqrt{n}\big(\hat{\pi}(v) - \pi(v)\big)/\hat{\sigma}(v) = n^{-1/2} \sum_{i=1}^n R_i(v)/(B(v)\sigma_n(v)) + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)$ *uniformly for* $v \in \mathcal{V}_\varrho$.

A $100(1-\alpha)\%$ confidence interval for $\pi(v)$ can be constructed as $\big[\hat{\pi}(v) - z^*_{1-\alpha}\hat{\sigma}(v)/\sqrt{n}, \hat{\pi}(v) + z^*_{1-\alpha}\hat{\sigma}(v)/\sqrt{n}\big]$, where $z^*_{1-\alpha} = \Phi^{-1}(1-\alpha/2)$ is the $1 - \alpha/2$ quantile of the standard normal distribution, based on the asymptotically normal approximation.

Similarly Theorem 4 shows that under Assumptions A1, A2, and A3, the influence function of $\hat{\pi}^{DR}$ is given by $R_i/B = (R_{1i} + R_{2i} + R_{3i})/B$. The exact formulas of R_{ki}, $k = 1, 2, 3$, are given in (S.8) in the online supplementary

appendix. Let $\sigma_n^2 = \mathbb{E}\left[R_i^2\right]/B^2$, which is the sieve variance of $\sqrt{n}\hat{\pi}^{DR}$. Further let $\hat{\sigma}^2$ be a consistent estimator of σ_n^2 such that $|\sigma_n/\hat{\sigma} - 1| = o_p(1)$.

Theorem 4. *Let Assumptions A1, A2, and A3 hold. Let* $\sqrt{n}(\varrho_n - \varrho) = o(1)$ *and* $\sqrt{n}l^{-1} = o(1)$. *Then* $\sqrt{n}\big(\hat{\pi}^{DR} - \pi^{DR}\big)/\hat{\sigma} = n^{-1/2}\sum_{i=1}^n R_i/(B\sigma_n) + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)$.

Based on Theorem 4, a $100(1-\alpha)\%$ confidence interval for π^{DR} can be constructed as $\left[\hat{\pi}^{DR} - z_{1-\alpha}^*\hat{\sigma}/\sqrt{n}, \hat{\pi}^{DR} + z_{1-\alpha}^*\hat{\sigma}/\sqrt{n}\right]$.

Note that our semiparametric estimation imposes certain functional form assumptions. Causal interpretation of the estimated parameters require these additional functional forms to hold. In theory, fully nonparametric estimation and inference is possible. For example, in Step 1, one can use the nonparametric QR series in Belloni et al. (2009), and in Step 2, one can follow Chen and Christensen (2018) to estimate a fully nonparametric mean regression. Our asymptotic theory for $\hat{\pi}(v)$ and $\hat{\pi}$ can then be extended to the corresponding nonparametric estimators at the cost of more complicated notations and stronger regularity conditions.

When monotonicity (along with other identifying assumptions) holds, $\pi^{DR} = \tau^{Wald\_X}$. So if the assumed semiparametric functional forms are true or if both are non-parametrically estimated, the two estimators converge to the same causal parameter and hence the corresponding estimates should be similar in large samples. Seeing the estimates very different may suggest that monotonicity does not hold (assuming other identifying assumptions hold). When monotonicity does not hold, the usual Wald ratio estimator, even when it is well-defined (or the denominator is not zero), is not consistent, while our estimator can be consistent for a well-defined causal parameter.

5 Extensions to a Multi-valued IV or Multiple IVs

In this section we briefly discuss extensions of identification, estimation and inference to the case of a multi-valued IV or a vector of discrete IVs.[11] We

[11] Mogstad et al. (2021) show that the LATE monotonicity may not be plausible with multiple IVs for a binary treatment. This conclusion is generalizable to a continuous treat-

first consider the basic setup without covariates and then discuss the general setup with covariates.

Assume $T = g(T, \varepsilon)$ and $T = h(Z, U)$ as in Section 2. Denote the support of Z as $\mathcal{Z} = \{z_0, z_1, ..., z_K\}$. So e.g., if $Z = (Z_1, Z_2)$, where $Z_1 \in \{0, 1\}$ and $Z_2 \in \{0, 1\}$, then one can let $z_0 = (0, 0)$, $z_1 = (0, 1)$, $z_2 = (1, 0)$, and $z_3 = (1, 1)$. Let $U_k = F_{T_{z_k}}(T_{z_k})$ be the rank of the potential treatment T_{z_k} if Z is exogenously set to be z_k. The observed rank can be written as $U = \sum_{k=1}^{K} 1(Z = z_k) U_k$. Let $T_{z_k}(u)$ be the u quantile of the potential treatment T_{z_k}. Further let $r_k = \Pr(Z = z_k)$, $p(Z) = \mathbb{E}[T|Z]$, $p_k = \mathbb{E}[T|Z = z_k]$, and $\overline{p} = \mathbb{E}[T]$. Without loss of generality, assume that the $K + 1$ values of Z is ordered such that $p_k \geq p_{k-1}$ for $k = 1, ..., K$, which may involve rearranging and is verifiable from the data.

We continue to use the same sets of assumptions when we consider either the basic setup without covariates or the general setup with covariates, except that the relevant assumptions need to be modified to accommodate the greater support of Z, which is $\mathcal{Z} = \{z_0, z_1, ..., z_K\}$. For example, Assumption 1 now requires that $T_{z_k}(u)$ is strictly monotonic in u for any $z_k \in \mathcal{Z}$, and that $U_k \sim Unif(0, 1)$ for $k = 0, ..., K$, and Assumption 2 independence now requires $Z \perp (U_k, \varepsilon)$ for $k = 0, ..., K$. The same holds true for Assumptions C1 and C2. Further Assumptions 3 and 4, and 5, and similarly Assumptions C3 and C5 need to hold for each pair of IV values z_k and z_{k-1} for $k = 1, ..., K$. That is, Assumption 3 now requires that $T_{z_k}(u) \neq T_{z_{k-1}}(u)$ for $k = 1, ..., K$ and at least some $u \in (0, 1)$. Assumption 4 monotonicity now states that $\Pr\left(T_{z_k} \geq T_{z_{k-1}}\right) = 1$, $k = 1, ..., K$. Assumption 5 now requires that $U_k|\varepsilon \sim U_{k-1}|\varepsilon$, $k = 1, ..., K$. The same holds true for Assumption C3 and Assumption C5. In addition, Assumption C4 common support now requires $\Pr(Z = z_k|X = x) \in (0, 1)$ for $k = 0, ..., K$ and any $x \in \mathcal{X}$.

Define the following estimand for each pair of the IV values $\{z_{k-1}, z_k\}$,

ment. While they seek to provide a causal interpretation for the usual two stage least square (2SLS) estimand under a weaker partial monotonicity condition (i.e., monotonicity holds with one IV while holding other IVs fixed), we provide an estimand that is robust to the failure of the LATE monotonicity assumption.

$k = 1, ...K$,

$$\tau_k(u) := \frac{\mathbb{E}[Y|Z = z_k, U = u] - \mathbb{E}[Y|Z = z_{k-1}, U = u]}{\mathbb{E}[T|Z = z_k, U = u] - \mathbb{E}[T|Z = z_{k-1}, U = u]}$$

if the denominator is not zero; otherwise, define $\tau_k(u) := 0$. Like before, T and U follow a one-to-one mapping given $Z = z_k$, so conditioning on $U = u$ is the same as conditioning on $T = T_{z_k}(u)$. Further given $Z \perp (U_k, \varepsilon)$, we have $T_{z_k}(u) = q_k(u)$, where $q_k(u) = F^{-1}_{T|Z}(u|z_k)$ is the conditional u quantile of T given $Z = z_k$. Then $\tau_k(u)$ can be re-written as

$$\tau_k(u) = \frac{\mathbb{E}[Y|Z = z_k, T = q_k(u)] - \mathbb{E}[Y|Z = z_{k-1}, T = q_{k-1}(u)]}{q_k(u) - q_{k-1}(u)}.$$

Following Theorem 1, $\tau_k(u)$ identifies an average treatment effect at the u quantile of treatment for units responding to the IV change from z_{k-1} to z_k.

Analogous to Proposition 1, define a DR estimand for each pair of IV values. In particular, let $\Delta q_k(u) = q_k(u) - q_{k-1}(u)$, $k = 1, ..., K$. The corresponding DR estimand is given by

$$\tau_k^{DR} := \int_0^1 \tau_k(u) w_k(u)\, du,$$

where $w_k(u) = \frac{|\Delta q_k(u)|}{\int_0^1 |\Delta q_k(u)| du}$. τ_k^{DR} identifies a weighted average of the average treatment effect for all units that respond to the IV change from z_{k-1} to z_k, under either monotonicity or rank similarity. Construct an aggregated DR estimand as

$$\tau^{DR,K} := \sum_{k=1}^{K} \lambda_k \tau_k^{DR}, \tag{16}$$

where $\lambda_k := \frac{(p_k - p_{k-1}) \sum_{l=k}^{K} r_l (p_l - \overline{p})}{\sum_{k=1}^{K} (p_k - p_{k-1}) \sum_{l=k}^{K} r_l (p_l - \overline{p})}$. The weights λ_k follow from Theorem 2 of Imbens and Angrist (1994).

Note that $\lambda_k \geq 0$ and $\sum_{k=1}^{K} \lambda_k = 1$, because the IV values are ordered such that $p_k \geq p_{k-1}$. Therefore, $\tau^{DR,K}$ is a convex combination of τ_k^{DR}, $k = 1, ..., K$,

and hence has the DR property as well.[12] In particular, when monotonicity holds, τ_k^{DR} reduces to the LATE Wald ratio $\tau_k^{Wald}:=\frac{\mathbb{E}[Y|Z=z_k]-\mathbb{E}[Y|Z=z_{k-1}]}{\mathbb{E}[T|Z=z_k]-\mathbb{E}[T|Z=z_{k-1}]}$, and hence $\tau^{DR,K}=\sum_{k=1}^{K}\lambda_k\tau_k^{Wald}$. Further by Theorem 2 of Imbens and Angrist (1994), $\sum_{k=1}^{K}\lambda_k\tau_k^{Wald}=\frac{Cov(Y,p(Z))}{Cov(T,p(Z))}$. Notice that τ_k^{Wald} in this case identifies a weighted average of LATEs for $Z\in\{z_{k-1},z_k\}$ under monotonicity. Therefore, if monotonicity holds, $\tau^{DR,K}$ identifies a doubly weighted average of LATEs, averaging over different compliers for a given pair of IV values and over different pairs of IV values; otherwise, when rank similarity holds, $\tau^{DR,K}$ identifies a doubly weighted average of the average treatment effects at different treatment quantiles - the first averaging is over different treatment quantiles for a given pair of IV values and the second is over different pairs of IV values. Either way, $\tau^{DR,K}$ identifies a doubly weighted average of the average treatment effects for all the units responding to IV changes.

Now consider the general setup where the IV independence and treatment rank similarity are valid only conditional on covariates. One can incorporate covariates as before for each pair of IV values. In particular for $k=1,...,K$, define the following estimand

$$\pi_k(x,v):=\frac{\mathbb{E}\left[Y|Z=z_k,X=x,V=v\right]-\mathbb{E}\left[Y|Z=z_{k-1},X=x,V=v\right]}{\mathbb{E}\left[T|Z=z_k,X=x,V=v\right]-\mathbb{E}\left[T|Z=z_{k-1},X=x,V=v\right]}$$

when the denominator is not zero; define $\pi_k(x,v):=0$, otherwise. Following Theorem 2, $\pi_k\left(x,v\right)$ identifies an average treatment effect at the conditional (on $X=x$) v quantile of treatment.

Further analogous to Proposition 2, define the DR estimand for each pair of IV values, z_{k-1} and z_k, as

$$\pi_k^{DR}:=\iint_{(0,1)\times\mathcal{X}}\pi_k(x,v)w_k(x,v)dvdx,$$

where $w_k(x,v)=\frac{|\Delta q_k(x,v)|f(x)}{\iint_{(0,1)\times\mathcal{X}}|\Delta q_k(x,v)|f(x)dvdx}$, and $\Delta q_k\left(x,v\right)=q_k\left(x,v\right)-q_{k-1}\left(x,v\right)$,

[12] In theory, any convex combination of $\tau_{z_k,z_{k-1}}^{DR}$, $k=1,...,K$, would have the DR property. Here our goal is to incorporate the 2SLS or LATE-type estimand given by $\frac{Cov(Y,p(Z))}{Cov(T,p(Z))}$ as a special case, which leads to the particular choice of λ_k.

and $q_k(x,v) = F^{-1}_{T|Z,X}(v|z_k,x)$ is the conditional v quantile of T given $Z = z_k$ and $X = x$.

Then define the aggregated DR estimand as

$$\pi^{DR,K} := \sum_{k=1}^{K} \lambda_k \pi_k^{DR},$$

where λ_k is defined as in (16). When monotonicity holds, $\pi^{DR,K}$ identifies a doubly weighted average of LATEs; otherwise when rank similarity holds, $\pi^{DR,K}$ identifies a doubly weighted average of the average treatment effects at different conditional treatment quantiles. Note that the identified parameter in this case is still the unconditional doubly weighted average, even though the instrument validity holds only conditional on covariates.

One can estimate $\pi^{DR,K}$ by $\hat{\pi}^{DR,K} = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\pi}_k^{DR}$ given an $i.i.d.$ sample $\{(Y_i, T_i, X_i, Z_i)\}_{i=1}^n$, where $\hat{\pi}_k^{DR}$ is an estimator of π_k^{DR} and $\hat{\lambda}_k$ is an estimator of λ_k. $\hat{\pi}_k^{DR}$ can be obtained similar to $\widehat{\pi}^{DR}$ proposed for a binary IV. $\hat{\lambda}_k$ can be estimated by a simple sample analogue plug-in estimator. Let $D^k = 1(Z = z_k)$. One can estimate $p_k = \mathbb{E}[T|Z = z_k]$ by $\widehat{p}_k = \sum_{i=1}^n T_i D_i^k / \sum_{i=1}^n D_i^k$ for $k = 0, 1, ..., K$, and estimate $\overline{p}$ by $\widehat{\overline{p}} = n^{-1} \sum_{i=1}^n T_i$. One can further estimate r_k by $\widehat{r}_k = n^{-1} \sum_{i=1}^n D_i^k$ for $k = 1, ..., K$. Then the plug-in estimator for λ_k is $\hat{\lambda}_k = \frac{(\widehat{p}_k - \widehat{p}_{k-1}) \sum_{l=k}^K \widehat{r}_l \left(\widehat{p}_l - \widehat{\overline{p}}\right)}{\sum_{k=1}^K (\widehat{p}_k - \widehat{p}_{k-1}) \sum_{l=k}^K \widehat{r}_l \left(\widehat{p}_l - \widehat{\overline{p}}\right)}$, $k = 1, ..., K$

We provide the influence function for $\hat{\pi}^{DR,K}$, denoted as R_{Ki}, in eq. (S.13) in the online supplementary appendix. The influence function given in Theorem 4 is now indexed by k, i.e., R_i/B defined in (S.8) is now R_i^k/B^k. Together with the influence function of $\hat{\lambda}_k$, we can derive the influence function of $\hat{\pi}^{DR,K}$. Define the sieve variance of $\sqrt{n}\hat{\pi}^{DR,K}$ as $\sigma_{Kn}^2 = \mathbb{E}\left[{R_{Ki}}^2\right]$. Let $\hat{\sigma}_K^2$ be a consistent estimator of σ_{Kn}^2, such that $|\sigma_{Kn}/\hat{\sigma}_K - 1| = o_p(1)$. We have the following asymptotics result for $\hat{\pi}^{DR,K}$.

Theorem 5. *Let the conditions in Theorem 4 hold. Then* $\sqrt{n}\left(\hat{\pi}^{DR,K} - \pi^{DR,K}\right)/\hat{\sigma}_K = n^{-1/2} \sum_{i=1}^n R_{Ki}/\sigma_{Kn} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)$.

6 Empirical Analysis

In this section, we apply our doubly robust approach to estimate the effects of night sleep on physical and psychological well-being using data from a recent field experiment (Bessone et al. 2021). 452 adults in Chennai, India participated in the experiment for a period of twenty eight days. Baseline data were collected for the first eight days. Then participants were randomized into three groups - a control group, a group who were provided with (a) devices to improve their home-sleep environments, and (b) information and verbal encouragement to increase their night sleep (the Encouragement group) and a group who were provided with (a), (b) and additional financial incentives to increase night sleep (the Encouragement + Incentives group). The three groups were further cross-randomized with a nap assignment that offered participants the opportunity for a daily half-hour afternoon nap at their workplace. So all together, there are six groups - control, encouragement, encouragement + incentives, naps, encouragement and naps, encouragement + incentives and naps. Details on the study design can be found in Bessone et al. (2021).

We use data from the first three non-nap assignment groups and take night sleep as our treatment variable, i.e., $T =$ night sleep in hours, for two reasons. First, night sleep is a primary form of sleep for most people. Second, the control group has zero hours of nap, while our treatment variable has to be absolutely continuous. For simplicity, we use as our outcome the well-being index, a summary measure of physical and phycological well-being, so $Y =$ well-being index.[13,14] Some of the individual outcomes, like labor supply etc.

[13]Bessone et al. (2021) focus on the reduced-form impacts of the night sleep and nap treatment assignments on a variety of work, well-being, cognition, and economic preferences outcomes.

[14]The well-being index is constructed as a weighted average of a wide range of standardized measures of psychological and physical well-being. Each constituent measure is standardized by the control group's mean and standard deviation. The weights are the inverse of the covariance matrix to ensure that highly correlated measures receive less weights in the aggregation. The measures of psychological well-being are happiness, sense of life possibilities (Cantril Scale), life satisfaction, stress, and depression. The measures of physical well-being are performance in a stationary biking task, reported days of illness, self-reported pain, activities of daily living, and blood pressure.

were recorded on a daily basis during the experimental period. Analysis of these outcomes would require dealing with the panel structure of the data. The well-being index is standardized by the baseline control group's mean and standard deviation (std. dev.) as in Bessone et al. (2021), so the unit of measurement is standard deviations. Following Bessone et al. (2021), our analysis controls for baseline measures of well-being and night sleep. In a subset of analysis we additionally control for participants' gender and age in four quartiles.

Table 1: Sample summary statistics

	(1)	(2)	(3)	(2) − (1)	(3) − (1)
Baseline well-being	0.00 (0.46)	0.03 (0.40)	0.09 (0.41)	0.03 (0.07)	0.19 (0.07)
Baseline night sleep	5.51 (0.90)	5.60 (0.84)	5.65 (0.79)	0.09 (0.14)	0.14 (0.14)
Age in 1st quartile	0.23 (0.43)	0.25 (0.44)	0.31 (0.47)	0.02 (0.07)	0.08 (0.07)
Age in 2nd quartile	0.27 (0.45)	0.27 (0.45)	0.20 (0.40)	-0.01 (0.07)	-0.07 (0.07)
Age in 3rd quartile	0.23 (0.43)	0.27 (0.45)	0.34 (0.48)	0.03 (0.07)	0.10 (0.07)
Female	0.68 (0.47)	0.64 (0.48)	0.64 (0.48)	-0.04 (0.08)	-0.04 (0.08)
Night sleep	5.62 (0.80)	5.99 (0.85)	6.22 (0.95)	0.37 (0.14)	0.60 (0.14)
Well-being	-0.00 (0.41)	0.14 (0.37)	0.10 (0.37)	0.15 (0.06)	0.10 (0.06)
Participants	77	75	74		

Note: Columns 1 - 3 report sample means and standard deviations (in parentheses) of the three groups: (1) Control, (2) Encouragement, (3) Encouragement + Incentives ; Columns 4 and 5 report the mean differences and their standard errors.

Our sample consists of 226 observations, including 77 from the control group, 75 from the Encouragement group and 74 from the Encouragement + Incentives group. Sample summary statistics are presented in Table 1. The three experimental groups are well-balanced across all of the covariates. Consistent with the results in Bessone et al. (2021), being assigned to either the Encouragement group or the Encouragement + Incentives group significantly increases night sleep on average. The increase in the Encouragement + Incentives group is larger, as expected. Interestingly, these simple mean comparisons show that being assigned to the Encouragement group significantly increases well-being (by 0.15 std. dev.), while being assigned to the Encouragement + Incentives group has no significant impacts on well-being, even though it leads to a larger increase in the average sleep time (0.60 vs. 0.37 hours).

Given the three experimental groups, we perform three sets of analysis. Let Z_1 be an indicator for whether one is assigned to the Encouragement group, and Z_2 be an indicator for whether one is assigned to the Encouragement + Incentives group. Define the three IV values based on the values of (Z_1, Z_2), i.e., z_0:=$(0,0)$, z_1:=$(1,0)$ and z_2:=$(0,1)$. In the first set of analysis, we look at the Encouragement group and the control group so that the IV is $Z = Z_1$. In the second set of analysis, we look at the Encouragement + Incentives group and the control group, so the IV is $Z = Z_2$. Note that these single IV analyses condition on the other IV being zero, which is important (see, discussion in Mogstad et al., 2021). In our third set of analysis, we use data from all three groups, so that the IV is $Z = (Z_1, Z_2)$ for $Z \in \{z_0, z_1, z_2\}$. The first two sets of analysis illustrate our proposed approach with a single binary IV, while the the third analysis illustrates our extended result with a multi-valued IV.

For the first set of analysis, the monotonicity assumption requires (a) everyone is likely to increase their night sleep if they are assigned to the Encouragement group instead of the control group; for the second set of analysis, it requires (b) everyone is likely to increase their night sleep if they are assigned to the Encouragement + Incentives group instead of the control group. For the third set of analysis, the monotonicity assumption requires (a) and additionally that everyone is likely to further increase their night sleep if they are assigned to the Encouragement + Incentives group instead of the Encouragement only group. Although we think these conditions are plausible, they are not verifiable in principle.[15] It is therefore useful to apply our doubly robust approach. For comparison purposes, we also implement (i) the usual linear 2SLS estimator, (ii) an estimator of τ^{Wald_X} in eq. (14), where all the condi-

[15]The one-sided Kolmogorov-Smirnov (KS) test fails to reject first-order stochastic dominance of the Encouragement group (or the Encouragement + Incentives group) treatment over the Control group treatment at the 10% significance level. It also fails to reject the dominance of the Encouragement + Incentives group treatment over the Encouragement only group treatment at the 10% significance level. When inspecting the empirical treatment quantile curves for each comparison, we do not find quantile crossing. However, as mentioned, stochastic dominance, $T_1(u) - T_0(u) \geq 0$ for all $u \in (0,1)$, is a necessary but not sufficient condition for monotonicity $\Pr(T_1 - T_0 \geq 0) = 1$, and the KS test is known to have low power for small samples, so we cannot conclude that monotonicity holds in this case.

tional means are assumed to be linear in covariates and fully interacted with the relevant binary IV, as well as (iii) a multi-valued IV extension of τ^{Wald_X}, i.e., $\tau^{Wald\_X,K}:=\sum_{k=1}^{K}\lambda_k\tau_k^{Wald\_X}$ for $K=2$, where λ_k is defined as in (16) and $\tau_k^{Wald\_X}$ is defined analogously to τ^{Wald_X} for the pair of IV values, z_{k-1} and z_k for $k=1,2$.

Table 2: Effects of per hour night sleep on well-being

	2SLS	Wald	DR	DR-2
		IV: Encouragement vs. Control (Z_1)		
(I)	0.427 (0.195)**	0.427 (0.236)*	0.390 (0.225)*	0.382 (0.220)*
(II)	0.408 (0.187)**	0.407 (0.236)*	0.233 (0.128)*	0.234 (0.127)*
		IV: Incentives vs. Control (Z_2)		
(I)	0.130 (0.109)	0.131 (0.121)	0.123 (0.133)	0.122 (0.102)
(II)	0.111 (0.107)	0.111 (0.122)	0.077 (0.103)	0.078 (0.132)
		Two IVs: (Z_1, Z_2)		
(I)	0.151 (0.107)	0.174 (0.229)	0.165 (0.158)	0.157 (0.156)
(II)	0.144 (0.104)	0.149 (0.226)	0.099 (0.122)	0.097 (0.121)

Note: (I) controls for baseline measure of well-being and that of night sleep, and (II) additionally controls for participants gender and age in four quartiles. 2SLS - linear 2SLS estimate; Wald - estimates of τ^{Wald_X} in eq. (14) or a multiple IV extension of it, where the conditional mean functions of Y and T are assumed to be linear in covariates and fully interacted with IV Z (see details in the main text); DR - doubly robust IV estimates based on the estimator in Section 4; DR-2 - doubly robust IV estimates, where the trimming parameter is set to be 3 times the baseline value. Standard errors are in the parenthesis. ** Significant 5%; * Significant at 10%.

Table 2 reports estimates from the three sets of analysis in three panels. Column 1 reports estimates by the usual linear 2SLS estimator. Column 2 reports estimates of τ^{Wald_X} (top and middle panels) and estimates of $\tau^{Wald\_X,2}$ (bottom panel). The linear 2SLS estimator is a special case of the estimator of τ^{Wald_X} (or $\tau^{Wald\_X,2}$). Both estimators similarly require monotonicity for causal interpretation but the former additionally assumes homogeneity of the instrument and treatment effects in covariates. Column 3 reports estimates by our doubly robust estimation proposed in Section 4.1 (top and middle panels) and estimates by the multi-valued IV extension of the doubly robust estimation discussed in Section 5 (bottom panel). Lastly, Column 4 reports similar doubly robust estimates to those reported in Column 3. The difference is that in

Column 3 the trimming parameter ϱ_n is set to be the baseline value specified in Section 4.1, while in Column 4 ϱ_n is set to be three times of the baseline value. In all the doubly robust estimation, the polynomial order of the power series of T is chosen to be one, considering the relatively small sample sizes. We report bootstrapped standard errors based on 200 bootstrap replications for estimates in Columns 2-4, since bootstrapping is straightforward and is computationally convenient.

Note that each instrument is associated with a different group of individuals responding to it. We found interesting treatment effect heterogeneity across the different groups responding to the two instruments. The local estimates using the encouragement assignment Z_1 as an IV range from 0.38 to 0.43 std. dev. when controlling for baseline sleep and baseline well-being, which are significant at the 5% or 10% level. So for the group that respond to the encouragement instrument, increased night sleep has marginally significant impacts on well-being. These estimates reduce to 0.23-0.41 std. dev., which are still significant at the 5% or 10% level, when additionally controlling for participants' gender and age. In contrast, the local estimates using the encouragement + incentives assignment Z_2 as an IV are smaller (yet still positive) but are not statistically significant, meaning that for the groups that respond to the encouragement + incentives instrument, increased night sleep does not translate into better well-being.

Table 3: Effects of per hour night sleep on well-being: breakdown of the combined two IV estimates

	(I)	(II)		(I)	(II)
π_1^{DR}	0.399 (0.187)**	0.233 (0.157)	$\tau_1^{Wald\_X}$	0.440 (0.243)**	0.404 (0.248)
π_2^{DR}	-0.297 (0.189)	-0.167 (0.120)	$\tau_2^{Wald\_X}$	-0.354 (0.357)	-0.355 (0.485)
Wt. avg.	0.165 (0.158)	0.099 (0.122)	Wt. avg.	0.174 (0.229)	0.149 (0.226)

Note: (I) controls for baseline well-being and baseline night sleep; (II) additionally controls for participants' gender and age in four quartiles. π_1^{DR} and$\tau_1^{Wald_X}$ compares the Encouragement group with Control; π_1^{DR} and $\tau_2^{Wald\_X}$ compares the Encouragement + Incentives group with the Encouragement only group. Wt. avg. is the weighted average of π_1^{DR} and π_1^{DR} (or $\tau_1^{Wald\_X}$ and $\tau_2^{Wald\_X}$), where the weights $\lambda_1 = 0.664$ (std. err. $= 0.239$) and $\lambda_2 = 0.336$ (std. err.$= 0.239$). ** Significant 5%.

The estimates using the two instruments Z_1 and Z_2 jointly lie between the estimates using each of the two instruments separately, which are not statistically significant. Recall that the doubly robust estimator estimates $\pi^{DR,2}:=\sum_{k=1}^{2}\lambda_k\pi_k^{DR}$ and the Wald estimator estimates $\tau^{Wald\_X,2}:=\sum_{k=1}^{2}\lambda_k\tau_k^{Wald\_X}$, where π_1^{DR} and $\tau_1^{Wald\_X}$ utilize the IV variation from z_0 to z_1 (comparing the Encouragement group to the control) while π_2^{DR} and $\tau_2^{Wald\_X}$ utilize the IV variation from z_1 to z_2 (comparing the Encouragement + Incentives group to the Encouragement group). Table 3 reports a detailed breakdown of the joint IV estimates. Estimates of π_1^{DR} and $\tau_1^{Wald\_X}$ are positive, while estimates of π_2^{DR} and $\tau_2^{Wald\_X}$ are always negative, even though they are not statistically significant. Consistent with the above analysis, these results once again suggest that those who respond to the additional financial incentives do not experience improved well-being.

Across all analysis, our doubly robust estimates are similar to the estimates of τ^{Wald_X} or $\tau^{Wald\_X,2}$ and the 2SLS estimates. The doubly robust estimates come with slightly inflated standard errors compared with the 2SLS estimates. The inflated standard errors reflect partly the tradeoff between robustness and efficiency. The similarity of the point estimates between our doubly robust estimator and the 2SLS estimator is reassuring. In this case, the doubly robust approach serves as a valuable tool to corroborate the usual 2SLS estimates, so that the 2SLS estimates can be relied upon with greater confidence.

Compared with the analysis in Bessone et al. (2021), which focuses on reduced-form analysis and uses different IVs jointly in one regression, we analyze each IV separately and when using the two IVs jointly, we give a detailed breakdown of the overall estimates. Our analysis yields the new finding that those individuals who slept longer due to the better sleep environment and verbal encouragement experienced improved mental and physical well-being, while those who slept more due to the financial incentives did not retain such benefits. This result is largely in line with the reduced-form estimates in Bessone et al. (2021, see, e.g., Table III).

7 Conclusion

Many empirical applications feature a continuous endogenous variable (treatment) and a binary or discrete IV. In this paper, we propose nonparametric doubly robust identification of the causal effects of a continuous treatment with a binary or discrete instrument.

We consider the two commonly imposed restrictions on the first-stage instrument effect heterogeneity: the LATE-type monotonicity vs. treatment rank similarity. Both assumptions can be used to identify causal effects of treatment in non-separable models, which accommodate arbitrary treatment effect heterogeneity and individuals sell-selection of different treatment levels. These assumptions are not nested. Both assumptions are not verifiable. We first show that with a continuous treatment, both can yield weighted average effects for the units that respond to the instrument change. In practice, it is not ideal to choose estimands based on, say, some pre-testing results. We further develop doubly robust estimands that are robust to failure of either one, so that one does not have to rely on pre-testing. When the LATE-type monotonicity holds, they reduce to the LATE-type estimands; otherwise, they continue to be valid under treatment rank similarity. Further, when treatment rank similarity holds, we can identify treatment effect heterogeneity at different (conditional) treatment quantiles. Based on our nonparametric identification results, we propose convenient semiparametric estimators and establish consistency and asymptotic normality of the proposed estimators. While our primary focus is on a binary instrument, we extend all of the identification, estimation and asymptotic results to the case with a multi-valued IV or a vector of discrete IVs, with or without covariates.

The usefulness of our proposed approach is illustrated in an empirical analysis estimating the impacts of night sleep on well-being, using data from a recent field experiment. We show that the group of individuals who increased night sleep due to information and verbal encouragement had improved psychological and physical well-being, while those who slept more due to the additional financial incentives did not experience such positive effects. In this

case, our doubly robust estimation yields estimates that are similar to the usual linear 2SLS estimates across different sets of analysis, which further establishes the credibility of the IV/2SLS estimates.

It is worth mentioning that we seek robust identification of some unconditional weighted average effects. When monotonicity fails, such weighted average effects average over units experiencing positive treatment changes and those experiencing negative treatment changes, which may not be ideal. However, it is well-known that individual types are not identified; therefore, point identification of (weighted) average effects separately for each individual type is not possible without further assumptions. An interesting direct of future research is then to develop partial identification results.

References

[1] Aggeborn, L. and M. Öhman (2021): "The effects of Fluoride in drinking water," Journal of Political Economy, 129 (2), 465-491.

[2] Arkhangelsky, D. and G. Imbens (2022): "Doubly robust identification for causal panel data models," The Econometrics Journal, 25(3), 649-674.

[3] Abadie, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," Journal of Econometrics, 113, 231–263.

[4] Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the U.S. wage structure," Econometrica, 74(2), 539-63.

[5] Angrist, J., Graddy, K., and G. Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," Review of Economic Studies, 67, 499–527.

[6] Angrist, J. and G. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," Journal of American Statistical Association, 90, 431-442.

[7] Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of causal effects using instrumental variables," Journal of American Statistical Association, 91, 444–472.

[8] Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021): "The economic consequences of increasing sleep among the urban poor," The Quarterly Journal of Economics, 136 (3), 1887-1941.

[9] Blundell, R., and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," in Advances in Economics and Econometrics, Vol. II, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press, 312-357.

[10] Chay, K. Y. and M. Greenstone (2005): "Does air quality matter? Evidence from the housing market," Journal of Political Economy, 113(2), 376-424.

[11] Chen, X. and T. Christensen (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV Regression," Quantitative Economics, 9(1), 39-85.

[12] Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects, " Econometrica, 73, 245-261.

[13] Chernozhukov, V. and C. Hansen (2006): "Instrumental quantile regression Inference for structural and treatment effect Models," Journal of Econometrics, 132, 491-525.

[14] Chesher, A. (2001): "Quantile driven identification of structural derivatives," Cemmap working paper CWP08/01.

[15] Chesher, A. (2002): "Local identification in nonseparable models," Cemmap working paper CWP05/02.

[16] Chesher, A. (2003): "Identification in nonseparable models," Econometrica, 71, 1405-1441.

[17] Chesher, A. (2005): "Nonparametric identification under discrete variation," Econometrica, 73, 1525-1550.

[18] de Chaisemartin, C. (2017): "Tolerating defiance? Local average treatment effects without monotonicity," Quantitative Economics, 8(2), 367-396.

[19] Dahl, C. M., M. Huber, and G. Mellace (2023): "It's never too LATE: A new look at local average treatment effects with or without defiers," Econometrics Journal, 26, 378-404.

[20] D'haultfoeuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," Econometrica, 83 (3), 1199-1210.

[21] Dong, Y., Y.-Y. Lee, and M. Gou (2023): "Regression discontinuity designs with a continuous treatment," Journal of the American Statistical Association, 118:541, 208-22.

[22] Dong, Y. and S. Shen (2018): "Testing for Rank Invariance or Similarity in Program Evaluation," The Review of Economics and Statistics, 100 (1), 78-85.

[23] Fiorini, M. and K. Stevens (2021): "Scrutinizing the Monotonicity Assumption in IV and fuzzy RD designs," Oxford Bulletin and Economics and Statistics, 83 (6), 1475-1526.

[24] Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," Econometrica, 76, 1191-1206.

[25] Frandsen B. R. and L. J. Lefgren (2018): "Testing Rank Similarity," The Review of Economics and Statistics, 100 (1), 86-91.

[26] Frölich, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," Journal of Econometrics, 139, 35-75.

[27] Goda G. S., E. Golberstein, and D. C. Grabowski (2021): "Income and the utilization of long-term care services: evidence from the social security benefit notch," Journal of Health Economics, 30 (4), 719-729.

[28] Giaccherini, M., J. Kopinska, and A. Palma (2021): "When particulate matter strikes cities: Social disparities and health costs of air pollution," Journal of Health Economics, 78, 102478.

[29] Imbens, G. and J. Angrist (1994): "Identification and estimation of local average treatment effects," Econometrica, 62 (92), 467-475.

[30] Imbens, G. and W. Newey (2002): "Identification and estimation of triangular simultaneous equations models without additivity," Technical Working Paper 285, National Bureau of Economic Research.

[31] Imbens, G. and W. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," Econometrica, 77, 1481-1512.

[32] Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental analysis of neighborhood effects," Econometrica, 75 (1), 83-119.

[33] Maruyama, S. and E. Heinesen (2020): "Another look at returns to birthweight," Journal of Health Economics, 70, 102269.

[34] Masten, M. and A. Torgovitsky (2016): "Identification of instrumental variable correlated random coefficients models," The Review of Economics and Statistics, 98 (5), 1001-1005.

[35] Matzkin, R. L. (2003): "Nonparametric estimation of nonadditive random functions," Econometrica, 71, 1339-1375.

[36] Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The causal interpretation of two-stage least squares with multiple instrumental variables," American Economic Review, 111 (11), 3663-98.

[37] Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," Econometrica, 83 (3), 1185-1197.

Online supplementary appendix for Nonparametric Doubly Robust Identification of Causal Effects of a Continuous Treatment using Discrete Instruments

Yingying Dong and Ying-Ying Lee†

In this supplementary appendix, Section S.1 provides proofs for the identification results presented in Sections 2 and 3. Section S.2 presents the inference theory for $\pi(x,v)$. Section S.3 presents the proofs of the inference results presented in Section 4.2. Section S.4 provides more details on computing the standard errors.

S.1 Proofs: Identification

Proof of Lemma 1: By definition,

$$\begin{aligned}
\tau^{Wald} &= \frac{\mathbb{E}\left[g\left(T_1,\varepsilon\right)|Z=1\right]-\mathbb{E}\left[g\left(T_0,\varepsilon\right)|Z=0\right]}{\mathbb{E}\left[T_1|Z=1\right]-\mathbb{E}\left[T_0|Z=0\right]} \\
&= \frac{\mathbb{E}\left[g\left(T_1,\varepsilon\right)-g\left(T_0,\varepsilon\right)\right]}{\mathbb{E}\left[T_1-T_0\right]} \\
&= \frac{\mathbb{E}\left[\left\{g\left(T_1,\varepsilon\right)-g\left(T_0,\varepsilon\right)\right\}\cdot 1\left(T_1-T_0>0\right)\right]}{\mathbb{E}\left[\left\{T_1-T_0\right\}\cdot 1\left(T_1-T_0>0\right)\right]} \\
&= \frac{\iint_{\mathcal{T}_c}\int\left\{g\left(t_1,e\right)-g\left(t_0,e\right)\right\}F_{\varepsilon|T_0,T_1}\left(de|t_0,t_1\right)F_{T_0,T_1}\left(dt_0,dt_1\right)}{\iint_{\mathcal{T}_c}\left\{t_1-t_0\right\}F_{T_0,T_1}\left(dt_0,dt_1\right)} \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1}\left\{\int\frac{g\left(t_1,e\right)-g\left(t_0,e\right)}{t_1-t_0}F_{\varepsilon|T_0,T_1}\left(de|t_0,t_1\right)\right\}F_{T_0,T_1}\left(dt_0,dt_1\right) \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1}\mathbb{E}\left[\frac{Y_{t_1}-Y_{t_0}}{t_1-t_0}|T_0=t_0,T_1=t_1\right]F_{T_0,T_1}\left(dt_0,dt_1\right) \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1}LATE(t_0,t_1)F_{T_0,T_1}\left(dt_0,dt_1\right),
\end{aligned}$$

where the first equality follows from the models for Y and T without covariates as specified in eq.s (2) and (4), respectively, the second equality follows from

†Yingying Dong and Ying-Ying Lee, Department of Economics, University of California Irvine, yyd@uci.edu and yingying.lee@uci.edu.

Assumption 2, the third equality follows from Assumption 4, the fourth equality follows from the law of iterated expectations, and the fifth to the last equalities follow from rearranging and our notation $w_{t_0,t_1} = \frac{t_1-t_0}{\iint_{\mathcal{T}_c}(t_1-t_0)F_{T_0,T_1}(dt_0,dt_1)}$ and $\mathcal{T}_c = \{(t_0,t_1) \in \mathcal{T}_0 \times \mathcal{T}_1 : t_1 - t_0 > 0\}$. Under monotonicity, $w_{t_0,t_1} \geq 0$ and $\iint_{\mathcal{T}_c} w_{t_0,t_1} F_{T_0,T_1}(dt_0,dt_1) = 1$, so τ^{Wald} identifies a weighted average of $LATE(t_0,t_1) := \mathbb{E}\left[\frac{Y_{t_1}-Y_{t_0}}{t_1-t_0} | T_0 = t_0, T_1 = t_1\right]$ for $(t_0,t_1) \in \mathcal{T}_c$.

Further, when $g(T,\varepsilon)$ is continuously differentiable in T,

$$\begin{aligned}
\tau^{Wald} &= \frac{\mathbb{E}\left[\int_{T_0}^{T_1} \frac{\partial g(t,\varepsilon)}{\partial t} dt\right]}{\mathbb{E}\left[\int_{T_0}^{T_1} 1 dt\right]} \\
&= \frac{\mathbb{E}\left[\int_{\mathcal{T}} \frac{\partial g(t,\varepsilon)}{\partial t} 1\left(T_0 \leq t \leq T_1\right) dt\right]}{\mathbb{E}\left[\int 1\left(T_0 \leq t \leq T_1\right) dt\right]} \\
&= \frac{\int_{\mathcal{T}} \mathbb{E}\left[\frac{\partial g(t,\varepsilon)}{\partial t} | T_0 \leq t \leq T_1\right] \Pr\left(T_0 \leq t \leq T_1\right) dt}{\int \Pr\left(T_0 \leq t \leq T_1\right) dt} \\
&= \int_{\mathcal{T}} \mathbb{E}\left[\left.\frac{\partial g\left(t,\varepsilon\right)}{\partial t}\right| T_0 \leq t \leq T_1\right] \widetilde{w} dt,
\end{aligned}$$

where $\widetilde{w} = \frac{\Pr(T_0 \leq t \leq T_1)}{\int_{\mathcal{T}} \Pr(T_0 \leq t \leq T_1) dt}$, the first equality follows from Assumption 4 and differentiability of $g(T,\varepsilon)$ in T, the second to the last equalities follow from the law of iterated expectations and interchanging the order of integration when standard regularity conditions hold.

Proof of Lemmas 2 and 3: By $Z \perp (U_z, \varepsilon)$ specified in Assumption 2, $Z \perp U_z|\varepsilon$. That is, $U_0|\varepsilon \sim U_0|(\varepsilon, Z=0)$ and $U_1|\varepsilon \sim U_1|(\varepsilon, Z=1)$. Further by Assumption 5, $U_0|\varepsilon \sim U_1|\varepsilon$. Together they imply $U_0|(\varepsilon, Z=0) \sim U_1|(\varepsilon, Z=1)$, i.e., $U|(\varepsilon, Z=1) \sim U|(\varepsilon, Z=0)$, so that $U \perp Z|\varepsilon$. Further by Assumption 2, $Z \perp \varepsilon$. Therefore, $Z \perp (U, \varepsilon)$, and hence $Z \perp \varepsilon|U$. It further implies $T \perp \varepsilon|U$, since $T = h(Z,U)$.

Replacing the above proof of Lemma 2 by conditioning on X in each step proves Lemma 3.

Proof of Theorem 2: Similar to the derivation of Lemma 2, one can show $Z \perp \epsilon | (V, X)$ under Assumptions C2 and C5. In particular, Assumption C2 states $Z \perp (V_z, \epsilon) | X$, which implies $Z \perp V_z | (X, \epsilon)$, i.e., $V_z | (X, \epsilon, Z = z) \sim V_z | (X, \epsilon)$, and hence $V | (X, \epsilon, Z = z) \sim V_z | (X, \epsilon)$. In addition, Assumption C5 states $V_1 | (X, \epsilon) \sim V_0 | (X, \epsilon)$. Then, $V | (X, \epsilon, Z = 0) \sim V | (X, \epsilon, Z = 1)$, i.e., $Z \perp V | (X, \epsilon)$. Further by Assumption C2, $Z \perp \epsilon | X$. Therefore, $Z \perp (V, \epsilon) | X$, and hence $Z \perp \epsilon | (V, X)$.

Consider now the two terms in the numerator of $\pi(x, v)$:

$$\begin{aligned}
\mathbb{E}[Y|Z = z, X = x, V = v] &= \mathbb{E}[G(T_z(x, v), x, \epsilon) | Z = z, X = x, V = v] \\
&= \mathbb{E}[G(T_z(x, v), x, \epsilon) | X = x, V = v] \\
&= \mathbb{E}\left[Y_{T_z(x,v)} | X = x, V = v\right] \\
&= \int G(T_z(x, v), x, e) F_{\epsilon|X,V}(de|x, v),
\end{aligned}$$

where the first equality follows from our models (9) and (10), the second equality follows from the condition $Z \perp \epsilon | (V, X)$ shown above, and the third equality follows from the definition of potential outcomes.

Consider next the two terms in the denominator of $\pi(x, v)$. By eq. (10),

$$\mathbb{E}[T|Z = z, X = x, V = v] = T_z(x, v).$$

Together they prove the theorem.

Proof of Lemma 4: First notice

$$\begin{aligned}
& \mathbb{E}[Y|Z = 1, X = x] - \mathbb{E}[Y|Z = 0, X = x] \\
= \; & \mathbb{E}[G(T_1, X, \epsilon) | Z = 1, X = x] - \mathbb{E}[G(T_0, X, \epsilon) | Z = 0, X = x] \\
= \; & \mathbb{E}[G(T_1, X, \epsilon) | X = x] - \mathbb{E}[G(T_0, X, \epsilon) | X = x],
\end{aligned}$$

where the first equality follows from our models of Y and T, equations (9) and (10), respectively, while the second equality follows from Assumption C2.

Consider now the numerator of τ^{LATE_X}:

$$\begin{aligned}
&\int_{\mathcal{X}} \{\mathbb{E}[Y|Z=1,X=x] - \mathbb{E}[Y|Z=0,X=x]\} f_X(x)dx \\
= &\int_{\mathcal{X}} \{\mathbb{E}[G(T_1,X,\epsilon)|X=x] - \mathbb{E}[G(T_0,X,\epsilon)|X=x]\} f_X(x)dx \\
= &\mathbb{E}[(G(T_1,X,\epsilon) - G(T_0,X,\epsilon))] \\
= &\mathbb{E}[G(T_1,X,\epsilon) - G(T_0,X,\epsilon)\cdot 1(T_1 - T_0 < 0)] \\
= &\iint_{\mathcal{T}_c} \iint \{G(t_1,x,e) - G(t_0,x,e)\} F_{X,\epsilon|T_0,T_1}(dx,de|t_0,t_1) F_{T_0,T_1}(dt_0,dt_1),
\end{aligned}$$

where the first equality follows from the derivation above, the second equality follows from the law of total expectation, and the third equality follows from Assumption 4, and the last equality follows from iterated expectations.

Similarly, the numerator of τ^{Wald_X} can be derived as follows

$$\begin{aligned}
&\int_{\mathcal{X}} \{\mathbb{E}[T_1|Z=1,X=x] - \mathbb{E}[T_0|Z=0,X=x]\} f_X(x)dx \\
= &\mathbb{E}[(T_1 - T_0)\cdot 1(T_1 - T_0 < 0)] \\
= &\iint_{\mathcal{T}_c} \{t_1 - t_0\} F_{T_0,T_1}(dt_0,dt_1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\tau^{Wald_X} \\
&= \frac{\iint_{\mathcal{T}_c} \iint \{G(t_1,x,e) - G(t_0,x,e)\} F_{X,\varepsilon|T_0,T_1}(dx,de|t_0,t_1) F_{T_0,T_1}(dt_0,dt_1)}{\iint_{\mathcal{T}_c} \{t_1 - t_0\} F_{T_0,T_1}(dt_0,dt_1)} \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1} \Big\{ \iint \frac{G(t_1,x,e) - G(t_0,x,e)}{t_1 - t_0} F_{X,\varepsilon|T_0,T_1}(dx,de|t_0,t_1) \Big\} F_{T_0,T_1}(dt_0,dt_1) \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1} \mathbb{E}\left[\frac{Y_{t_1} - Y_{t_0}}{t_1 - t_0} | T_1 = t_1, T_0 = t_0 \right] F_{T_0,T_1}(dt_0,dt_1) \\
&= \iint_{\mathcal{T}_c} w_{t_0,t_1} LATE(t_0,t_1) F_{T_0,T_1}(dt_0,dt_1).
\end{aligned}$$

Proof of Proposition 2: When Assumption 4 monotonicity holds,

$$\pi^{DR} = \frac{\iint \pi(x,v)\Delta q(x,v) f_X(x) dv dx}{\iint \Delta q(x,v) f_X(x) dv dx}.$$

Plug in the expression of $\pi(x,v)$ and $\Delta q(x,v)$, and notice $V = V_z$ when $Z = z$, for $z = 0, 1$. The numerator of π^{DR} is $\int_{\mathcal{X}}\{\int_0^1\{\mathbb{E}[Y|Z=1,X=x,V_1=v] - \mathbb{E}[Y|Z=0,X=x,V_0=v]\}dv\}f_X(x)dx$. Consider the two terms involved in the difference. For $z = 0, 1$, we have

$$\begin{aligned}
&\int_{\mathcal{X}} \left\{ \int_0^1 \mathbb{E}\left[Y|Z=z,X=x,V_1=v\right] dv \right\} f_X(x)dx \\
&= \int_{\mathcal{X}} \left\{ \int_0^1 \mathbb{E}\left[G(T_z(x,v),x,\epsilon)|X=x,V_1=v\right] dv \right\} f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[G\left(T_z\left(x,V_1\right),x,\epsilon\right)|X=x\right] f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[G\left(T_z\left(x,V_1\right),x,\epsilon\right)|Z=1,X=x\right] f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[Y|Z=z,X=x\right] f_X(x)dx,
\end{aligned}$$

where the first equality follows from the models of Y given by (9) and Assumption C2, which implies $Z \perp \epsilon|\left(V_z, X\right)$, the second equality follows from averaging over the conditional distribution of V_z given X, which is $Unif\left(0,1\right)$ by construction, the third equality follows from Assumption C2, which states $Z \perp \left(V_z, \epsilon\right)|X$, the last equality follows from the models of Y given by (9).

Now consider the numerator of π^{DR}. It is given by $\int_{\mathcal{X}}\{\int_0^1\{\mathbb{E}[Y|Z=1,X=x,V_1=v] - \mathbb{E}[Y|Z=0,X=x,V_0=v]\}dv\}f_X(x)dx$. Consider the two terms

involved in the difference. For $z = 0, 1$, we have

$$\begin{aligned}
&\int_{\mathcal{X}} \left\{ \int_0^1 \{\mathbb{E}\left[T|Z=z, X=x, V_z=v\right] dv \right\} f_X(x)dx \\
&= \int_{\mathcal{X}} \left\{ \int_0^1 \mathbb{E}\left[T_z(x,v)|X=x, V_z=v\right] dv \right\} f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[T_z\left(x, V_z\right)|X=x\right] f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[T_z\left(x, V_z\right)|Z=z, X=x\right] f_X(x)dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[T|Z=z, X=x\right] f_X(x)dx,
\end{aligned}$$

where the first equality follows from the models of T given by (10) and Assumption C2, which implies $Z \perp \epsilon | \left(V_z, X\right)$, the second equality follows from averaging over the conditional distribution of V_z given X, which is $Unif\left(0,1\right)$ by construction, the third equality follows from Assumption C2, which states $Z \perp \left(V_z, \epsilon\right) | X$ and further implies $Z \perp V_z | X$, the last equality follows from the model of T given by eq. (10).

Together we have

$$\begin{aligned}
\pi^{DR} &= \frac{\int_{\mathcal{X}} \left\{\mathbb{E}\left[Y|Z=1, X=x\right] - \mathbb{E}\left[Y|Z=0, X=x\right]\right\} f_X(x)dx}{\int_{\mathcal{X}} \left\{\mathbb{E}\left[T|Z=1, X=x\right] - \mathbb{E}\left[T|Z=0, X=x\right]\right\} f_X(x)dx} \\
&= \tau^{Wald_X}.
\end{aligned}$$

Then by Lemma 4, π^{DR} identifies a weighted average of $LATE\left(t_0, t_1\right)$ for $(t_0, t_1) \in \mathcal{T}_c$ under Assumption 4 monotonicity.

Otherwise, when Assumption 4 monotonicity does not hold, but Assumption C5 conditional treatment rank similarity holds,

$$\pi^{DR} := \iint \pi(x,v) w(x,v) dv dx,$$

where $w(x,v) \geq 0$ and $\iint w(x,v)dvdx = 1$. So π^{DR} is a weighted average of $\pi(x,v)$, the conditional average treatment effect given $X = x$ and $V = v$, by Theorem 2.

S.2 Inference for $\pi(x, v)$

A $100(1-\alpha)\%$ confidence interval for $\pi(x,v)$ is constructed as $\big[\hat{\pi}(x,v) - z^*_{1-\alpha}\hat{\sigma}(x,v)/\sqrt{n}, \hat{\pi}(x,v) + z^*_{1-\alpha}\hat{\sigma}(x,v)/\sqrt{n}\big]$, where the critical value $z^*_{1-\alpha}$ can be $\Phi^{-1}(1-\alpha/2)$ by the asymptotically normal approximation. The sieve variance estimator for $\hat{\pi}(x,v)$ is $\hat{\sigma}^2(x,v) = \Delta\hat{\psi}(x,v)'\hat{\mho}\Delta\hat{\psi}(x,v)/\Delta\hat{T}(x,v)^2$, where $\Delta\hat{\psi}(x,v) = \psi^J(x,\hat{T}_1(x,v),1) - \psi^J(x,\hat{T}_0(x,v),0)$.

Theorem 6. *Let Assumptions A1-A3 hold. Then* $\sqrt{n}(\hat{\pi}(x,v)-\pi(x,v))/\hat{\sigma}(x,v) \xrightarrow{d} \mathcal{N}(0,1)$ *uniformly for* $(x,v) \in \Pi_\varrho = \{(x,v) \in \mathcal{X}\times\mathcal{V} : |\Delta T(x,v)| \geq \varrho\}$.

For the uniform confidence interval over $(x,v) \in \Pi_\varrho$, the critical value $z^*_{1-\alpha}$ is simulated from the bootstrap sieve t-statistic $\mathbb{Z}^*_n(x,v)$ for $(x,v) \in \Pi_\varrho$: Let $\varpi_1, ..., \varpi_n$ be i.i.d. random variables independent of the data with mean zero, unit variance, and finite third moment, e.g., $\mathcal{N}(0,1)$. Let

$$\mathbb{Z}^*_n(x,v) = \frac{\Delta\hat{\psi}(x,v)'\hat{G}^{-1}}{\Delta\hat{T}(x,v)\hat{\sigma}(x,v)\sqrt{n}} \sum_{i=1}^n \psi^J(x,T,Z_i)\hat{e}_i\varpi_i.$$

Calculate $\mathbb{Z}^*_n(x,v)$ for a large number of independent draws of $\varpi_1, ..., \varpi_n$. Then the critical value $z^*_{1-\alpha}$ is the $(1-\alpha)$ quantile of $\sup_{(x,v)\in\Pi_\varrho}|\mathbb{Z}^*_n(x,v)|$ over the draws. Theorem 4.1 in Chen and Christensen (2018) implies the result on the consistency of the sieve score bootstrap. $\sup_{s\in\mathcal{R}}\Big|\mathbb{P}\Big(\sup_{(x,v)\in\Pi_\varrho}|\sqrt{n}\big(\hat{\pi}(x,v) - \pi(x,v)\big)/\hat{\sigma}(x,v)| \leq s\Big) - \mathbb{P}^*\Big(\sup_{(x,v)\in\Pi_\varrho}|\mathbb{Z}^*_n(x,v)| \leq s\Big)\Big| = o_p(1)$, where $\mathbb{P}^*$ denotes a probability measure conditional on the data $\{Y_i, T_i, X_i, Z_i\}_{i=1}^n$.

S.3 Proofs: Estimation and Inference

The proofs use the results in Angrist, Chernozhukov, and Fernańdez-Val (2006) (ACF, henceforth) and Chen and Christensen (2018) (CC, henceforth). To simplify exposition, we collect notations used in the proofs below. We suppress the subscripts i, z and dependence on v, when there is no confusion.

Notation:

$$\begin{aligned}
\phi_i(v) &= \vartheta(v)^{-1}\big(1(T_i \le S_i'a(v)) - v\big)S_i \\
S_{1i} &= (1, X_i', 1, X_i')', S_{0i} = (1, X_i', 0, \mathbf{0}_{(d_x\times 1)}')', \Delta S_i = S_{1i} - S_{0i} \\
\partial_t m_z(X, q_z(X,v)) &= \frac{\partial}{\partial t} m_z(X,t)|_{t=q_z(X,v)} \\
q_{zi} &= q_z(X_i, v), \hat{q}_{zi} = \hat{q}_z(X_i, v) \\
\Delta q_i &= \Delta q(X_i, v) = q_{1i} - q_{0i} = (S_{1i} - S_{0i})'a(v) = \Delta S_i'a(v) \\
\Delta \hat{q}_i &= \Delta \hat{q}(X_i, v) = \hat{q}_{1i} - \hat{q}_{0i} = (S_{1i} - S_{0i})'\hat{a}(v) = \Delta S_i'\hat{a}(v) \\
\Delta \psi_i &= \Delta\psi(X_i, v) = \psi^J(X_i, q_1(X_i,v), 1) - \psi^J(x, q_0(X_i,v), 0) \\
\Delta \hat{\psi}_i &= \Delta\hat{\psi}(X_i, v) = \psi^J(X_i, \hat{q}_1(X_i,v), 1) - \psi^J(X_i, \hat{q}_0(X_i,v), 0) \\
\Delta m_i &= \Delta m(X_i, v) = m_1(X_i, q_1(X_i,v)) - m_0(X_i, q_0(X_i,v)) \\
\Delta \hat{m}_i &= \Delta \hat{m}(X_i, v) = \hat{m}_1(X_i, \hat{q}_1(X_i,v)) - \hat{m}_0(X_i, \hat{q}_0(X_i,v)) = \Delta\hat{\psi}_i'\hat{c} \\
\Delta \check{m}_i &= \Delta \check{m}(X_i, v) = \hat{m}_1(X_i, q_1(X_i,v)) - \hat{m}_0(X_i, q_0(X_i,v)) = \Delta\psi_i'\hat{c} \\
\chi_i &= \chi(X_i, v) = 1(|\Delta q(X_i,v)| \ge \varrho|) \\
\chi_i^{\pm} &= \chi^{\pm}(X_i, v) = 1(\pm\Delta q(X_i,v) \ge \varrho|)
\end{aligned}$$

Lemma 5 is for estimating the trimming function.

Lemma 5. *Let Assumption A1 hold. Let* $\sqrt{n}(\varrho_n - \varrho) = o(1)$ *and* $\sqrt{n}l^{-1} = o(1)$. *Then*

1.

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{l}\sum_{v\in V^{(l)}} \Delta m(X_i,v)\left(\hat{\chi}^+(X_i,v) - \chi^+(X_i,v)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^1 \frac{\partial}{\partial\alpha}\mathbb{E}\left[\Delta m(X,v)1(\Delta S'\alpha \ge \varrho)\right]'\big|_{\alpha=a(v)}\phi_i(v)dv + o_p(1).
\end{aligned}$$

2.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{l}\sum_{v\in V^{(l)}}\Delta q(X_i,v)\left(\hat{\chi}^{+}(X_i,v)-\chi^{+}(X_i,v)\right)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1\frac{\partial}{\partial\alpha}\mathbb{E}\left[\Delta q(X,v)1(\Delta S'\alpha\geq\varrho)\right]'\Big|_{\alpha=a(v)}\phi_i(v)dv+o_p(1).$$

Step 1 is $O_p(n^{-1/2})$, so the estimation error of χ is of first order asymptotically by Lemma 5. The rate condition on $\sqrt{n}(\varrho_n-\varrho)=o(1)$ means that using ϱ_n rather than ϱ is first-order asymptotically ignorable.

Lemma 6 is for the approximation error from the numerical integration.

Lemma 6. *Let a function $f(x,v)$ be of bounded variation in $v\in\mathcal{V}$, uniformly in $x\in\mathcal{X}$. Then*

$$\sup_{x\in\mathcal{X}}\Big|l^{-1}\sum_{v\in V^{(l)}}f(x,v)1(\Delta q(x,v)>\varrho)-\int_0^1 f(x,v)1(\Delta q(x,v)>\varrho)dv\Big|=O(l^{-1}).$$

The inference theory for $\pi(v)$ follows analogously to that of π^{DR}, but without integrating over v. Therefore we first present the proof of Theorem 4 for π^{DR}.

Proof of Theorem 4: Define A_+ and A_- as $A_\pm=\int_0^1\int_{\mathcal{X}}\Delta m(x,v)\chi^{\pm}(x,v)f_X(x)dxdv$. So $A=A_+-A_-=\int_0^1\int_{\mathcal{X}}\Delta m(x,v)/\Delta q(x,v)\big(\Delta q(x,v)1(\Delta q(x,v)\geq\varrho)-\Delta q(x,v)1(\Delta q(x,v)\leq-\varrho)\big)f_X(x)dxdv=\int_0^1\int_{\mathcal{X}}\pi(x,v)|\Delta q(x,v)|1(|\Delta q(x,v)|\geq\varrho)f_X(x)dxdv$.

Define B_+ and B_- as $B_\pm=\int_0^1\int_{\mathcal{X}}\Delta q(x,v)\chi^{\pm}(x,v)f_X(x)dxdv$. By a similar argument as A, we can show that $B=B_+-B_-$. Therefore, $\pi^{DR}=A/B$ and $\pi^{DR}_{\pm}=A_\pm/B_\pm$. Linearize $\hat{\pi}^{DR}-\pi^{DR}=(\hat{A}-A)/B-(\hat{B}-B)\pi/B+O_p\left(|\hat{A}-A||\hat{B}-B|/B^2+|\hat{B}-B|^2/B^2\right)$.

The proof is focused on $\hat{A}_+$, the estimator of A_+. The same arguments apply to $\hat{B}_+$, the estimator of B_+. The same arguments apply to $\hat{\pi}^{DR}_-$ and hence $\hat{\pi}^{DR}$.

Write $\hat{\pi}_+^{DR} = \hat{A}_+/\hat{B}_+$, where

$$\hat{A}_+ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{v\in V^{(l)}}\Delta\hat{m}(X_i,v)\hat{\chi}^+(X_i,v),$$
$$\hat{B}_+ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{v\in V^{(l)}}\Delta\hat{q}(X_i,v)\hat{\chi}^+(X_i,v).$$

In the following, we suppress the subscripts of + and superscripts of DR for expositional simplicity. Linearize $\hat{\pi}-\pi = (\hat{A}-A)/B - (\hat{B}-B)\pi/B + O_p\left(|\hat{A}-A||\hat{B}-B|/B^2 + |\hat{B}-B|^2/B^2\right)$.

Let $\tilde{A} = n^{-1}\sum_{i=1}^{n} l^{-1}\sum_{v\in V^{(l)}}\Delta\hat{m}(X_i,v)\chi(X_i,v)$ for a known trimming function. Decompose $\hat{A}-A = \hat{A}-\tilde{A}+\tilde{A}-A$. The estimation error in $\Delta\hat{m}$.

$$\tilde{A}-A = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{v\in V^{(l)}}\left(\Delta\hat{m}(X_i,v)-\Delta m(X_i,v)\right)\chi(X_i,v) \tag{S.1}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{v\in V^{(l)}}\Delta m(X_i,v)\chi(X_i,v) - A. \tag{S.2}$$

By Lemma 6 and assuming $\sqrt{n}l^{-1} = o(1)$, (S.2) is $n^{-1}\sum_{i=1}^{n} R_{A3i} + o_p(n^{-1/2})$, where $R_{A3i} = \int_0^1 \Delta m(X_i,v)\chi^+(X_i,v)dv - A_+$.

We focus on (S.1) next. Decompose $\Delta\hat{m}_i - \Delta m_i = \left(\Delta\hat{m}_i - \Delta\check{m}_i\right) + \left(\Delta\check{m}_i - \Delta m_i\right)$. The first part is for Step 1 estimation error, and the second part is for Step 2 estimation error.

Step 1 Theorem 3 in ACF shows that $\hat{a}(v) - a(v) = n^{-1}\sum_{i=1}^{n}\phi_i(v) + o_p(n^{-1/2})$ uniformly over $v\in\mathcal{V}$ and converges in distribution to a zero mean Gaussian process indexed by v. Decompose

$$\begin{aligned}
&\Delta\hat{m}_i - \Delta\check{m}_i \\
&= m_1(X_i,\hat{q}_{1i}) - m_1(X_i,q_{1i}) - (m_0(X_i,\hat{q}_{0i}) - m_0(X_i,q_{0i})) + so1 \\
&= \partial_t m_1(X_i,q_{1i})(\hat{q}_{1i}-q_{1i}) - \partial_t m_0(X_i,q_{0i})(\hat{q}_{0i}-q_{0i}) + so1 + so2 \\
&= \partial_t m_1(X_i,q_{1i})S_{1i}(\hat{a}(v)-a(v)) - \partial_t m_0(X_i,q_{0i})S_{0i}(\hat{a}(v)-a(v)) + so1 + so2,
\end{aligned}$$

where (We suppress the subscript i for simplicity.)

$$\begin{aligned} so1 &= \hat{m}_1(\hat{q}_1) - m_1(\hat{q}_1) - \big(\hat{m}_0(\hat{q}_0) - m_0(\hat{q}_0)\big) - \big(\hat{m}_1(q_1) - m_1(q_1)\big) \\ &\quad + \big(\hat{m}_0(q_0) - m_0(q_0)\big) \\ &= O_p\left(\|\partial_t \hat{m}_z - \partial_t m_z\|_\infty \|\hat{q}_z - q_z\|_\infty\right), \\ so2 &= O_p\left(\partial_t^2 m_1\big(\hat{q}_1 - q_1\big)^2 + \partial_t^2 m_0\big(\hat{q}_0 - q_0\big)^2\right) = O_p(\|\hat{q}_z - q_z\|_\infty^2), \end{aligned}$$

as $\partial_t^2 m_z$ is uniformly bounded by Assumption A3. ACF and Corollary 3.1(ii) in CC implies that $so1 + so2 = O_p(\|\hat{q}_z - q_z\|_\infty \|\partial_t \hat{m}_z - \partial_t m_z\|_\infty + \|\hat{q}_z - q_z\|_\infty^2) = O_p(n^{-1/2}(J^{-(p-1)} + J\sqrt{(J\log J)/n}) + n^{-1}) = o_p(n^{-1/2})$ uniformly over $v \in \mathcal{V}$, by assuming $J\sqrt{(J\log J)/n} = o(1)$ and $p > 1$.

Then

$$\begin{aligned} &\sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{1}{l}\sum_{v\in V^{(l)}} (\Delta\hat{m}_i - \Delta\check{m}_i)\chi_i \\ &= \frac{1}{n}\sum_{i=1}^n \frac{1}{l}\sum_{v\in V^{(l)}} (\partial_t m_1(X_i, q_{1i}) S_{1i}' \chi_i \sqrt{n}(\hat{a}(v) - a(v)) \\ &\quad - \partial_t m_0(X_i, q_{0i}) S_{0i}' \chi_i \sqrt{n}(\hat{a}(v) - a(v)) + o_p(1) \\ &= \frac{1}{l}\sum_{v\in V^{(l)}} \mathbb{E}\left[(\partial_t m_1(X_i, q_{1i}) S_{1i} - \partial_t m_0(X_i, q_{0i}) S_{0i})\,\chi_i\right]' \sqrt{n}(\hat{a}(v) - a(v)) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^n \frac{1}{l}\sum_{v\in V^{(l)}} \mathbb{E}\left[(\partial_t m_1(X_i, q_{1i}) S_{1i} - \partial_t m_0(X_i, q_{0i}) S_{0i})\,\chi_i\right]' \phi_j(v) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^n \int_0^1 \mathbb{E}\left[(\partial_t m_1(X_i, q_{1i}) S_{1i} - \partial_t m_0(X_i, q_{0i}) S_{0i})\,\chi_i\right]' \phi_j(v) dv + o_p(1), \end{aligned}$$

where the third equality is by ACF, and the last equality is by Lemma 6 and $\sqrt{n}l^{-1} = o(1)$.

For the second equality, let $\mathcal{F} = \left\{1(\Delta S_i' a > \varrho), a \in \mathcal{B}\right\}$ that is a VC subgraph class and hence a bounded Donsker class. Then $\mathcal{F}(\partial_t m_1(X_i, S_{1i}' a) S_{1i} - \partial_t m_0(X_i, S_{0i}' a) S_{0i})$ is also bounded Donsker with a square-integrable envelop $2\sup_{z,x,t} |\partial_t m_z(x,t)| \max_{j\in\{1,2,\ldots,d_x\}} |X_j|$ by Theorem 2.10.6 in Van der Vaart

and Wellner (1996). So $n^{-1}\sum_{i=1}^{n}(\partial_t m_1(X_i,q_{1i})S_{1i} - \partial_t m_0(X_i,q_{0i})S_{0i})\chi_i = \mathbb{E}[(\partial_t m_1(X_i,q_{1i})S_{1i} - \partial_t m_0(X_i,q_{0i})S_{0i})\chi_i] + o_p(n^{-1/2})$ uniformly in $v \in \mathcal{V}$.

Step 2 We show the stochastic equicontinuity, $n^{-1}\sum_{i=1}^{n}(\Delta\check{m}_i - \Delta m_i)\chi_i = \mathbb{E}\left[(\Delta\check{m}_i - \Delta m_i)\chi_i\right] + so3$, where $so3 = o_p(n^{-1/2})$ uniformly in $v \in \mathcal{V}$.

Let $\Delta\tilde{m}_i = \Delta\psi_i'\tilde{c}$ and $\Delta\check{m}_i = \Delta\psi_i'\hat{c}$. Then decompose $so3 = so31 + so32$ to the "standard deviation" term $so31$ and the "bias" term $so32$,

$$so3 = \frac{1}{n}\sum_{i=1}^{n}\chi_i\left(\Delta\check{m}_i - \Delta\tilde{m}_i\right) - \int_{\mathcal{X}}\chi_i\left(\Delta\check{m}_i - \Delta\tilde{m}_i\right)F_X(dX_i) \tag{so31}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\chi_i\left(\Delta\tilde{m}_i - \Delta m_i\right) - \int_{\mathcal{X}}\chi_i\left(\Delta\tilde{m}_i - \Delta m_i\right)F_X(dX_i). \tag{so32}$$

Let $\sqrt{n}so31 = Q_J'(\hat{c}-\tilde{c})$, where $Q_J = \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}\chi_i\Delta\psi_i - \int_{\mathcal{X}}\chi_i\Delta\psi_i F_X(dX_i))$. By $var(Q_J) = \mathbb{E}[\chi_i\Delta\psi_i\Delta\psi_i']$ and the Jensen's inequality, $\mathbb{E}[\|Q_J\|] \leq O(\sqrt{\mathbb{E}[\|\Delta\psi_i\|^2]}) = O(\zeta)$. As given in the proof of Lemma 3.1 in CC, $\|\hat{c} - \tilde{c}\|_{\ell^\infty} = O_p(\sqrt{\log J/(n\lambda_{min}(G))})$, where the minimum eigenvalue $\lambda_{min}(G) > 0$.[16] Then $\mathbb{E}[|so31|] = O(n^{-1/2}\zeta\sqrt{\log J/(n\lambda_{min}(G))})$ by the Cauchy-Schwartz inequality. The Markov's inequality implies $so31 = O_p(n^{-1}\zeta\sqrt{\log J/\lambda_{min}(G)}) = o_p(n^{-1/2})$ implied by Assumption A2.5.

$var(\sqrt{n}so32) = O(\mathbb{E}\left[\chi_i\left(\Delta\tilde{m}_i - \Delta m_i\right)^2\right]) = O(\|m - \Pi_J m\|_\infty^2)$, where $\Pi_J m = \arg\min_{h\in\Psi_J}\|m-h\|_{L^2(X,T,Z)}$, by Theorem 3.1 (i) in CC. The Markov's inequality yields $so32 = O_p(n^{-1/2}\|m - \Pi_J m\|_\infty) = O_p(n^{-1/2}J^{-p}) = o_p(1)$ by the results in the proof of Corollary 3.1 in CC.

By Lemma 6 and assuming $\sqrt{n}l^{-1} = o(1)$, $l^{-1}\sum_{v\in V^{(l)}}\mathbb{E}\left[(\Delta\check{m}_i - \Delta m_i)\chi_i\right] = \int_0^1 \mathbb{E}\left[\Delta\check{m}_i\chi_i\right]dv - A + o_p(n^{-1/2})$.

Note that A is based on a linear functional of m, $L(m) = \int_0^1\int_{\mathcal{X}} m_z(x, q_z(x,v))$ $1(\Delta q(x,v) > \varrho)F_X(dx)dv$. So we use the results on linear functionals of a sieve estimator in CC. Let $\sigma_{A2n}^2 = \mathbb{E}\left[R_{A2i}^2\right]$, where $R_{A2i} = \mathcal{D}^{+\prime}G^{-1}\psi^J(X_i,T_i,Z_i)e_i$ and $\mathcal{D}^+ = \int_0^1 \mathbb{E}\left[\Delta\psi^J(X,v)\chi^+(X,v)\right]dv$, with a consistent estimator $\hat{\sigma}_{A2}^2$.

[16] By Lemma A.1 in CC, $s_{JK}^{-1} \asymp \pi_J = 1$ for the exogenous case.

Lemma 4.1 in CC provides

$$\left| \frac{\sqrt{n}}{\hat{\sigma}_{A2}} \left(\int_0^1 \mathbb{E}\left[\Delta \check{m}_i \chi_i\right] dv - A \right) - \frac{1}{\sigma_{A2n}\sqrt{n}} \sum_{i=1}^n R_{A2i} \right| = o_p(1).$$

The estimation error from the trimming function $\hat{A} - \tilde{A} = n^{-1}\sum_{i=1}^n l^{-1} \sum_{v\in V^{(l)}} \Delta m(X_i, v)\big(\hat{\chi}(X_i, v) - \chi(X_i, v)\big) + o_p(1)$ by $n^{-1}\sum_{i=1}^n l^{-1}\sum_{v\in V^{(l)}} \big(\Delta\hat{m}(X_i, v) - \Delta m(X_i, v)\big)\big(\hat{\chi}(X_i, v) - \chi(X_i, v)\big) = O_p\Big(\|\Delta\hat{m} - \Delta m\|_\infty \|\hat{q}_z - q_z\|_\infty\Big) = o_p(n^{-1/2})$. Together with Lemma 5(i), $\left|\sqrt{n}(\hat{A}-A) - n^{-1/2}\sum_{i=1}^n R_{Ai}\right| = o_p(1)$, where $R_{Ai} = R_{A1i} + R_{A2i} + R_{A3i}$ with

$$\begin{aligned} R_{A1i} = \int_0^1 \Bigg(& \mathbb{E}\left[\left(\partial_t m_1(X, q_1) S_1 - \partial_t m_0(X, q_0) S_0\right) \chi^+(X, v)\right] \\ & + \frac{\partial}{\partial \alpha} \mathbb{E}\left[\Delta m(X, v) 1(\Delta S'\alpha \geq \varrho)\right]\big|_{\alpha = a(v)} \Bigg)' \phi_i(v) dv, \end{aligned}$$

By the similar arguments as for A in (S.1) and (S.2),

$$\tilde{B} - B = \frac{1}{n}\sum_{i=1}^n \frac{1}{l} \sum_{v\in V^{(l)}} \left(\Delta\hat{q}(X_i, v) - \Delta q(X_i, v)\right) \chi(X_i, v) \tag{S.3}$$

$$+ \frac{1}{n}\sum_{i=1}^n \frac{1}{l} \sum_{v\in V^{(l)}} \Delta q(X_i, v)\chi(X_i, v) - B. \tag{S.4}$$

By Lemma 6, (S.4) is $n^{-1}\sum_{i=1}^n \int_0^1 \Delta q(X_i, v)\chi(X_i, v) dv - B + o_p(n^{-1/2})$. (S.3) is

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \frac{1}{l} \sum_{v\in V^{(l)}} \Delta S_i' \left(\hat{a}(v) - a(v)\right) \chi_i \\ & = \frac{1}{n}\sum_{j=1}^n \frac{1}{l} \sum_{v\in V^{(l)}} \frac{1}{n}\sum_{i=1}^n \chi_i \Delta S_i' \phi_j(v) + o_p(n^{-1/2}) \\ & = \frac{1}{n}\sum_{j=1}^n \frac{1}{l} \sum_{v\in V^{(l)}} \mathbb{E}\left[\chi_i \Delta S_i\right]' \phi_j(v) + o_p(n^{-1/2}) \end{aligned}$$

$$= \frac{1}{n}\sum_{j=1}^{n}\int_0^1 \mathbb{E}\left[\Delta S_i'\chi_i\right]\phi_j(v)dv + o_p(n^{-1/2}),$$

where the first equality by ACF, and the third equality by Lemma 6. For the second equality, let $\mathcal{F} = \left\{1(\Delta S_i'a > \varrho), a \in \mathcal{B}\right\}$ that is a VC subgraph class and hence a bounded Donsker class. Then $\mathcal{F}\Delta S$ is Donsker with a square-integrable envelop $\max_{j\in\{1,2,\ldots,d_x\}}|X_j|$ by Theorem 2.10.6 in Van der Vaart and Wellner (1996). So $n^{-1}\sum_{i=1}^n \chi_i\Delta S_i - \mathbb{E}\left[\chi_i\Delta S_i\right] = o_p(1)$ uniformly over $v \in \mathcal{V}$

Together with Lemma 5(ii), we obtain $\left|\sqrt{n}(\hat{B} - B) - n^{-1/2}\sum_{i=1}^n R_{Bi}\right| = o_p(1)$, where $R_{Bi} = R_{B1i} + R_{B3i}$ with

$$R_{B1i} = \int_0^1 \left(\mathbb{E}\left[\Delta S'^{+}(X,v)\right] + \frac{\partial}{\partial\alpha}\mathbb{E}\left[\Delta q(X,v)1(\Delta S'\alpha \geq \varrho)\right]'\Big|_{\alpha=a(v)}\right)\phi_i(v)dv$$
$$R_{B3i} = \int_0^1 \Delta q(X_i,v)\chi^+(X_i,v)dv - B.$$

By a linearization for $\hat{\pi}_+^{DR}$, $\hat{\pi}_+^{DR} - \pi_+^{DR} = \frac{\hat{A}_+}{\hat{B}_+} - \frac{A_+}{B_+} = \frac{\hat{A}_+ - A_+}{B_+} - \frac{\pi_+^{DR}}{B_+}(\hat{B}_+ - B_+) + o_p(n^{-1/2})$. Therefore, we define $R_i^+ = R_{Ai} - \pi_+^{DR}R_{Bi} = R_{1i}^+ + R_{2i}^+ + R_{3i}^+$, where $R_{1i}^+ = R_{A1i} - \pi_+^{DR}R_{B1i}$, $R_{2i}^+ = R_{A2i}$, and $R_{3i}^+ = R_{A3i} - \pi_+^{DR}R_{B3i}$. That is,

$$R_{1i}^+ = \int_0^1 \Bigg(\mathbb{E}\left[\left(\partial_t m_1(X,q_1(X,v))S_1 - \partial_t m_0(X,q_0(X,v))S_0 - \pi_+^{DR}\Delta S\right)\chi^+(X,v)\right]$$
$$+ \frac{\partial}{\partial\alpha}\mathbb{E}\left[\left(\Delta m(X,v) - \pi_+^{DR}\Delta q(X,v)\right)1(\Delta S'\alpha \geq \varrho)\right]\Big|_{\alpha=a(v)}\Bigg)'\phi_i(v)dv,$$
$$R_{2i}^+ = \mathcal{D}^{+'}G^{-1}\psi^J(X_i,T_i,Z_i)e_i, \text{with } \mathcal{D}^+ = \int_0^1 \mathbb{E}\left[\Delta\psi^J(X,v)\chi^+(X,v)\right]dv,$$
$$R_{3i}^+ = \int_0^1 \left(\Delta m(X_i,v) - \pi_+^{DR}\Delta q(X_i,v)\right)\chi^+(X_i,v)dv.$$

Then we obtain $\hat{\pi}_+^{DR} - \pi_+^{DR} = n^{-1}\sum_{i=1}^n \left(R_{Ai} - \pi_+^{DR}R_{Bi}\right)/B_+ + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n R_i^+/B_+ + o_p(n^{-1/2})$.

Asymptotic normality We suppress the subscripts of + and superscripts of DR for expositional simplicity. Because R_{2i} depends on (Y_i, T_i, X_i), R_{1i} depends on (T_i, X_i), and R_{3i} depends on X_i, the law of iterated expectations yields $\sigma_n^2 = \big(\mathbb{E}\big[R_{1i}^2\big] + \mathbb{E}\big[R_{2i}^2\big] + \mathbb{E}\big[R_{3i}^2\big]\big)/B^2 = (\sigma_1^2 + \sigma_{2n}^2 + \sigma_3^2)/B^2$.

We will show the Bahadur representation that

$$\begin{aligned}&\left|\frac{\sqrt{n}(\hat{\pi}-\pi)}{\hat{\sigma}} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{B\sigma_n}\right| \\ &\leq \left|\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sigma_n} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{B\sigma_n}\right| + \left|\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sigma_n}\left(\frac{\sigma_n}{\hat{\sigma}}-1\right)\right| = o_p(1) \qquad \text{(S.5)}\end{aligned}$$

by (i) $n^{-1/2}\sum_{i=1}^{n} R_i/(B\sigma_n) \xrightarrow{d} \mathcal{N}(0,1)$, and (ii) $|\sigma_n/\hat{\sigma} - 1| = o_p(1)$, as shown below.

(i) Asymptotic normality will follow from the Lyapunov central limit theorem with the third absolute moment, $n^{-1/2}\mathbb{E}[|R_i|^3]/(B\sigma_n)^3 \to 0$, since $\{R_i\}_{i=1}^n$ are independent across i, with mean zero and variance 1. By the assumed conditions, it is straightforward to show that $n^{-1/2}\mathbb{E}[|R_{1i}|^3]/(B\sigma_1)^3 \to 0$. We show below that $n^{-1/2}\mathbb{E}[|R_{2i}|^3]/(B\sigma_{2n})^3 \to 0$. Then it implies that all the cross-product terms $n^{-1/2}\mathbb{E}[|R_{1i}R_{2i}R_{3i}|]/(B\sigma_n)^3 \to 0$ and $n^{-1/2}\mathbb{E}[|R_{ji}^2 R_{ki}|]/(B\sigma_n)^3 \to 0$ for $j,k = 1,2,3$, $j \neq k$.

Denote as $\psi_i = \psi^J(X_i, T_i, Z_i)$. By Assumption A2.2(ii),

$$\begin{aligned}\sigma_{2n}^2 &= \mathbb{E}\left[R_{2i}^2\right]/B^2 = \mathbb{E}\left[(\mathcal{D}'G^{-1}\psi_i)^2 e_i^2\right]/B^2 \\ &\geq \mathbb{E}\left[(\mathcal{D}'G^{-1}\psi_i)^2\right]\underline{\sigma}^2/B^2 = \mathcal{D}'G^{-1}\mathcal{D}\underline{\sigma}^2/B^2. \qquad \text{(S.6)}\end{aligned}$$

By the Schwarz inequality, (S.6), and Assumption A2.3(ii),

$$\frac{(\mathcal{D}'G^{-1}\psi_i)^2}{\sigma_{2n}^2} \leq \frac{(\mathcal{D}'G^{-1}\mathcal{D}')(\psi_i'G^{-1}\psi_i)}{\sigma_{2n}^2} \leq \frac{\zeta^2}{\underline{\sigma}^2}. \qquad \text{(S.7)}$$

Then by (S.6), (S.7), and Assumption A2.2(iii),

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{|R_{2i}|^3}{B^3\sigma_{2n}^3}\right] &= \frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{|\mathcal{D}'G^{-1}\psi_i e_i|^3}{B^3\sigma_{2n}^3}\right] \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{(\mathcal{D}'G^{-1}\psi_i)^2}{B^3\sigma_{2n}^2}\frac{|\mathcal{D}'G^{-1}\psi_i|}{\sigma_{2n}}\mathbb{E}\left[|e_i|^3|X_i,T_i,Z_i\right]\right] \\ &\leq \frac{\zeta}{\sqrt{n}B^3\underline{\sigma}^3}\sup_{x,t,z}\mathbb{E}\left[|e_i|^3|X_i=x,T_i=t,Z_i=z\right] = O\left(\frac{\zeta}{\sqrt{n}}\right) = o(1).\end{aligned}$$

(ii) It is straightforward that $\hat{\sigma}_1^2 = n^{-1}\sum_{i=1}^n \hat{R}_{1i}^2/\hat{B}^2 \xrightarrow{p} \sigma_1^2 = \mathbb{E}\left[R_{1i}^2\right]/B^2$ and $\hat{\sigma}_3^2 \xrightarrow{p} \sigma_3^2$. The same arguments in Lemma G.4 in CC give $|\sigma_{2n}/\hat{\sigma}_2 - 1| = O_p(\delta_{V,n}) = o_p(1)$. So $|\sigma_n/\hat{\sigma} - 1| = o_p(1)$.

By (i) that $n^{-1/2}\sum_{i=1}^n R_i/(B\sigma_n) = O_p(1)$ and (ii), the second term $\left|\frac{\sqrt{n}(\hat{\pi}-\pi)}{\hat{\sigma}}\left(\frac{\hat{\sigma}}{\sigma_n}-1\right)\right| = O_p(1)o_p(1) = o_p(1)$. We then obtain the Bahadur representation. The asymptotic normality follows from the result (i).

Therefore, we obtain that when $B_+ > 0$, $\sqrt{n}\big(\hat{\pi}_+^{DR} - \pi_+^{DR}\big)/\hat{\sigma}_{n+} = n^{-1/2}\sum_{i=1}^n R_i^+ /(B_+\sigma_{n+}) + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)$, where $\hat{\sigma}_+^2$ is a consistent estimator of $\sigma_{n+}^2 = \mathbb{E}\left[R_i^{+2}\right]/B_+^2$.

For π_-^{DR}, define

$$\begin{aligned}R_{1i}^- = \int_0^1 \Bigg(&\mathbb{E}\left[\big(\partial_t m_1(X,q_1(X,v))S_1 - \partial_t m_0(X,q_0(X,v))S_0 - \pi_-^{DR}\Delta S\big)\,\chi^-(X,v)\right] \\ &+ \frac{\partial}{\partial\alpha}\mathbb{E}\left[\big(\Delta m(X,v) - \pi_+^{DR}\Delta q(X,v)\big)\,1(\Delta S'\alpha \leq -\varrho)\right]\big|_{\alpha=a(v)}\Bigg)'\phi_i(v)dv.\end{aligned}$$

Define R_i^- as R_i^+ by replacing $+$ with $-$ in all the components in R_i^+. By the same arguments for π_+^{DR}, we obtain that when $B_- > 0$, $\sqrt{n}\big(\hat{\pi}_-^{DR} - \pi_-^{DR}\big)/\hat{\sigma}_- = n^{-1/2}\sum_{i=1}^n R_i^-/(B_-\sigma_{n-}) + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)$, where $\hat{\sigma}_-^2$ is a consistent estimator of $\sigma_{n-}^2 = \mathbb{E}\left[R_i^{-2}\right]/B_-^2$, such that $|\sigma_{n-}/\hat{\sigma}_- - 1| = o_p(1)$.

For π^{DR}, the same linearization yields $\hat{\pi}^{DR} - \pi^{DR} = (\hat{A} - A)/B - (\hat{B} - B)\pi^{DR}/B + O_p\big(|\hat{A} - A||\hat{B} - B|/B^2 + |\hat{B} - B|^2/B^2\big)$. Let $R_i = R_i^+ - R_i^- = R_{1i} + R_{2i} + R_{3i}$, where $R_{li} = R_{li}^+ - R_{li}^-$ for $l = 1,2,3$ by replacing π_+^{DR} and π_-^{DR}

with π^{DR}. Specifically, let $\text{sgn}(x,v) = 1(\Delta q(x,v) \geq \varrho) - 1(\Delta q(x,v) \leq -\varrho)$,

$$
\begin{aligned}
R_{1i} = \int_0^1 \Big(& \mathbb{E}\big[(\partial_t m_1(X, q_1(X,v))S_1 - \partial_t m_0(X, q_0(X,v))S_0 - \pi^{DR}\Delta S)\text{sgn}(X,v)\big] \\
& + \frac{\partial}{\partial \alpha}\mathbb{E}[\big(\Delta m(X,v) - \pi^{DR}\Delta q(X,v)\big)\,(1(\Delta S'\alpha \geq \varrho) \\
& - 1(\Delta S'\alpha \leq -\varrho))]\big|_{\alpha = a(v)}\Big)' \phi_i(v)dv, \\
\text{with } \phi_i(v) &= \vartheta(v)^{-1}\big(1(T_i \leq S_i'a(v)) - v\big)S_i, \\
S_{1i} &= (1, X_i', 1, X_i')', S_{0i} = (1, X_i', 0, \mathbf{0}_{(d_x \times 1)}')', \Delta S_i = S_{1i} - S_{0i}, \\
R_{2i} &= \mathcal{D}' G^{-1}\psi^J(X_i, T_i, Z_i)e_i, \\
\text{with } \mathcal{D} &= \int_0^1 \mathbb{E}\left[(\psi^J(X, q_1(X,v), 1) - \psi^J(X, q_0(X,v), 0))\text{sgn}(X,v)\right] dv. \\
R_{3i} &= \int_0^1 \big(\Delta m(X_i,v) - \pi^{DR}\Delta q(X_i,v)\big)\,\text{sgn}(X_i,v)dv, \\
B &= \int_0^1 \int_{\mathcal{X}} |\Delta q(x,v)| 1(|\Delta q(x,v)| \geq \varrho) f(x)dxdv. \qquad \text{(S.8)}
\end{aligned}
$$

Proof of Theorem 3: The proof follows exactly the same arguments in the proof of Theorem 4 and Lemma 5 by removing all "$\int_0^1 \cdots dv$" and "$l^{-1}\sum_{v \in V^{(l)}}$". We can derive the influence function of $\hat{\pi}(v)$ to be $R_i(v)/B(v)$ defined as the influence function of $\hat{\pi}^{DR}$ given in (S.8) by removing all $\int_0^1 \cdots dv$. Specifically, as π^{DR}, define $\pi_+(v)$ over units experiencing positive changes for $v \in \mathcal{V}_{+\varrho} = \{v \in \mathcal{V} : P(\Delta q(X,v) \geq \varrho) > 0\}$. Define $B_+(v) = \int_{\mathcal{X}} \Delta q(x,v)\chi^+(x,v)f(x)dx$, so $B_+ = \int_0^1 B_+(v)dv$. The influence function of $\hat{\pi}_+(v)$ is $R_i^+(v)/B_+(v) = (R_{1i}^+(v) + R_{2i}^+(v) + R_{3i}^+(v))/B_+(v)$, where

$$
\begin{aligned}
R_{1i}^+(v) = \Bigg(& \frac{\partial}{\partial \alpha}\mathbb{E}\left[(\Delta m(X,v) - \pi_+(v)\Delta q(X,v))\,1(\Delta S'\alpha \geq \varrho)\right]\big|_{\alpha = a(v)} \\
& + \mathbb{E}\big[(\partial_t m_1(X, q_1(X,v))S_1 - \partial_t m_0(X, q_0(X,v))S_0 \\
& - \pi_+(v)\Delta S)\chi^+(X,v)\big]\Bigg)' \phi_i(v), \\
R_{2i}^+(v) &= \mathcal{D}^{+\prime}(v)G^{-1}\psi^J(X_i, T_i, Z_i)e_i, \text{with } \mathcal{D}^+(v) = \mathbb{E}\left[\Delta\psi^J(X,v)\chi^+(X,v)\right], \\
R_{3i}^+(v) &= (\Delta m(X_i,v) - \pi_+(v)\Delta q(X_i,v))\,\chi^+(X_i,v). \qquad \text{(S.9)}
\end{aligned}
$$

Similarly consider $\pi_-(v)$ over units experiencing negative changes for $v \in \mathcal{V}_{-\varrho} = \{v \in \mathcal{V} : P(-\Delta q(X,v) \geq \varrho) > 0\}$. Let $B(v) = B_+(v) - B_-(v)$, where $B_-(v) = \int_{\mathcal{X}} \Delta q(x,v)\chi^-(x,v)f(x)dx$. Let $R_i(v) = R_i^+(v) - R_i^-(v)$, and the influence function of $\hat{\pi}(v)$ is $R_i(v)/B(v)$.

Define $\sigma^2(v) = \mathbb{E}\left[R_i(v)^2\right]/B(v)^2$. The unknown elements are estimated following the same procedure as $\hat{\pi}^{DR}$ by removing "$l^{-1}\sum_{v\in V^{(l)}}$." For example, $\hat{\mathcal{D}}^+(v) = n^{-1}\sum_{i=1}^n \Delta\hat{\psi}_i\hat{\chi}^+(X_i,v)$.

Proof of Theorem 6: We first show that the estimation error of $\hat{q}_z(x,v)$ in Step 1 is of smaller order than the estimation error in Step 2, i.e., the first-order asymptotic distribution of $\hat{\pi}(x,v)$ is as if $q_z(x,v)$ was known. Under Assumption A1, Theorem 3 in ACF implies that $\sup_{(x,v)\in\mathcal{X}\times\mathcal{V}}|\hat{q}_z(x,v) - q_z(x,v)| = O_p(n^{-1/2})$. The Step 2 series least squares estimator converges at a nonparametric rate shower than $\sqrt{n}$. Therefore the first-order asymptotic distribution of $\hat{\pi}(x,v)$ is dominated by Step 2 $\Delta\check{m}(x,v)$.

Step 1 When T_{zi} is observed, i.e., there is no Step 1 estimation error, define $\check{\pi}(x,v) = \Delta\check{m}(x,v)/\Delta q(x,v)$. Decompose $\hat{\pi}(x,v) - \check{\pi}(x,v) = \frac{\Delta\hat{m}}{\Delta\hat{q}} - \frac{\Delta\check{m}}{\Delta q} = \left(\frac{\Delta\hat{m}}{\Delta\hat{q}} - \frac{\Delta\check{m}}{\Delta\hat{q}}\right) + \left(\frac{\Delta\check{m}}{\Delta\hat{q}} - \frac{\Delta\check{m}}{\Delta q}\right)$. The second part is for Step 1 in the denominator: $\frac{\Delta\check{m}}{\Delta\hat{q}} - \frac{\Delta\check{m}}{\Delta q} = \frac{\Delta m}{\Delta q^2}(\Delta q - \Delta\hat{q}) + so1$. The first part is for Step 1 in the argument in the numerator,

$$
\begin{aligned}
&\frac{\Delta\hat{m}}{\Delta\hat{q}} - \frac{\Delta\check{m}}{\Delta\hat{q}} \\
&= \frac{1}{\Delta q}(\Delta\hat{m} - \Delta\check{m}) + so2 \\
&= \frac{1}{\Delta q}(m_1(x,\hat{q}_1) - m_1(x,q_1) - (m_0(x,\hat{q}_0) - m_0(x,q_0))) + so2 + so3 \\
&= \frac{1}{\Delta q}(\partial_t m_1(x,q_1)(\hat{q}_1 - q_1) - \partial_t m_0(x,q_0)(\hat{q}_0 - q_0)) + so2 + so3 + so4,
\end{aligned}
$$

where

$$
\begin{aligned}
so1 &= \frac{\Delta\check{m}}{\Delta\hat{q}\Delta q}(\Delta q - \Delta\hat{q}) - \frac{\Delta m}{\Delta q^2}(\Delta q - \Delta\hat{q}) = (\Delta q - \Delta\hat{q})\frac{1}{\Delta q}\left(\frac{\Delta\check{m}}{\Delta\hat{q}} - \frac{\Delta m}{\Delta q}\right), \\
so2 &= \Delta\hat{m}\left(\frac{1}{\Delta\hat{q}} - \frac{1}{\Delta q}\right) + \Delta\check{m}\left(\frac{1}{\Delta q} - \frac{1}{\Delta\hat{q}}\right) = (\Delta\hat{m} - \Delta\check{m})\left(\frac{1}{\Delta\hat{q}} - \frac{1}{\Delta q}\right), \\
so3 &= \frac{1}{\Delta q}\Big\{\hat{m}_1(x,\hat{q}_1) - m_1(x,\hat{q}_1) - \big(\hat{m}_0(x,\hat{q}_0) - m_0(x,\hat{q}_0)\big) - \big(\hat{m}_1(x,q_1) \\
&\quad - m_1(x,q_1)\big) + \big(\hat{m}_0(x,q_0) - m_0(x,q_0)\big)\Big\} \\
&= O_p\left((\partial_t\hat{m}_1(x,q_1) - \partial_t m_1(x,q_1))(\hat{q}_1 - q_1)\right), \\
so4 &= O_p\left(\partial_t^2 m_1\big(\hat{q}_1 - q_1\big)^2 + \partial_t^2 m_0\big(\hat{q}_0 - q_0\big)^2\right) = O_p(\|\hat{q}_z - q_z\|_\infty^2).
\end{aligned}
$$

Thus $so1 + so2 + so3 + so4 = O_p(\|\hat{T} - T\|_\infty^2 + \|\hat{T} - T\|_\infty\|\partial_t\check{m} - \partial_t m\|_\infty) = O_p(n^{-1} + n^{-1/2}(J^{-(p-1)} + J\sqrt{(J4\log J)/n})) = o_p(n^{-1/2})$ uniformly over $(x,v) \in \Pi_\varrho$, by Corollary 3.1(ii) in CC and assuming $J\sqrt{(J\log J)/n} = o(1)$ and $p > 1$. Therefore,

$$
\begin{aligned}
&\sqrt{n}\big(\hat{\pi}(x,v) - \check{\pi}(x,v)\big) \\
&= \sqrt{n}\left\{\frac{\Delta m}{\Delta q^2}(\Delta q - \Delta\hat{q}) + \frac{1}{\Delta q}(\partial_t m_1(x,q_1)(\hat{q}_1 - q_1) - \partial_t m_0(x,q_0)(\hat{q}_0 - q_0))\right\} \\
&\quad + o_p(1) \\
&= \left\{-\frac{\pi(x,v)}{\Delta q}(S_1 - S_0) + \frac{1}{\Delta q}(\partial_t m_1(x,q_1)S_1 - \partial_t m_0(x,q_0)S_0)\right\}' \sqrt{n}(\hat{a}(v) - a(v)) \\
&\quad + o_p(1) \\
&= \left\{-\frac{\pi(x,v)}{\Delta q}\Delta S + \frac{1}{\Delta q}(\partial_t m_1(x,q_1)S_1 - \partial_t m_0(x,q_0)S_0)\right\}' \frac{1}{\sqrt{n}}\sum_{j=1}^n \phi_j(v) + o_p(1)
\end{aligned}
\tag{S.10}
$$

by Theorem 3 in ACF and $\|\hat{\pi} - \check{\pi}\|_\infty = O_p(n^{-1/2})$.

Step 2 Define $\mathcal{Z}_n \sim \mathcal{N}(0, \mho)$, $\sigma_n^2(x,v) = \Delta\psi(x,v)'\mho\Delta\psi(x,v)/\Delta q(x,v)^2$, and

$$\mathbb{Z}_n^\pi(x,v) = \frac{\Delta\psi(x,v)'}{\Delta q(x,v)\sigma_n(x,v)}\mathcal{Z}_n.$$

Lemma 4.1 in CC provides uniform Bahadur representation and uniform Gaussian process strong approximation

$$\sup_{(x,v)\in\Pi_\varrho}\left|\frac{\sqrt{n}\left(\hat{\pi}(x,v)-\pi(x,v)\right)}{\hat{\sigma}(x,v)} - \mathbb{Z}_n^\pi(x,v)\right| = o_p(1).$$

Proof of Lemma 5: Since $\Delta S_i' = (0, \mathbf{0}'_{(d_x\times 1)}, 1, X_i')'$, let $\Delta S_i'a - \varrho = \Delta S_i'\beta$, where $\beta = (a_0(v), a_1'(v), a_2(v) - \varrho, a_3'(v))'$. Let $\hat{\beta} = (\hat{a}_0(v), \hat{a}_1'(v), \hat{a}_2(v) - \varrho_n, \hat{a}_3'(v))'$.

We show that $(v,\beta) \mapsto \mathbb{G}_n[\Delta m_i\chi_i] = \sqrt{n}\sum_{i=1}^n \left(\Delta m_i\chi_i - \mathbb{E}\left[\Delta m_i\chi_i\right]\right)$ is stochastic equicontinuous over $\mathcal{V}\times\mathcal{B}$, with respect to the $L_2(P)$ pseudometric $\rho((v_1,\beta_1),(v_2,\beta_2))^2 = \mathbb{E}\big[\big(\Delta m(X_i,v_1)(1(\Delta S_i'\beta_1 \geq 0) - \Delta m(X_i,v_2)1(\Delta S_i'\beta_2 \geq 0)\big)^2\big]$.

Following the proof of Theorem 3 in Section A.1.2 in the appendix of ACF, let $\mathcal{F} = \left\{1(\Delta S_i'\beta > 0), \beta\in\mathcal{B}\right\}$ that is a VC subgraph class and hence a bounded Donsker class. $\mathcal{F}\Delta m(X,v)$ is Donsker with a square-integrable envelop $|\Delta m(X,v)|$ by Theorem 2.10.6 in Van der Vaart and Wellner (1996).

By stochastic equicontinuity of $(v,\beta)\mapsto\mathbb{G}_n[\Delta m_i\chi_i]$, $n^{-1/2}\sum_{i=1}^n \Delta m_i\big(\hat{\chi}_i - \chi_i\big) = \sqrt{n}\mathbb{E}\big[\Delta m_i\big(\hat{\chi}_i-\chi_i\big)\big] + o_{p^*}(1) = \frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X_i,v)1(\Delta S_i'\alpha\geq 0)\big]'\big|_{\alpha=\beta(v)} \times \sqrt{n}(\hat{\beta}(v)-\beta(v)) + o_{p^*}(1)$ uniformly over $v\in\mathcal{V}$, which follows from $\|\hat{\beta}(v)-\beta(v)\| = o_{p^*}(1)$, and resulting convergence with respect to the pseudometric $\sup_{v\in\mathcal{V}}\rho((v,\hat{\beta}(v)),(v,\beta(v)))^2 = o_p(1)$. The latter is from $\rho((v,\beta),(v,B))^2 = \mathbb{E}\big[\Delta m(X_i,v)^2(1(\Delta S_i'\beta\geq 0) - 1(\Delta S_i'^2\big] = O\big(\frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X_i,v)^2 1(\Delta S_i'\alpha\geq 0)\big]\big|'_{\alpha=\beta}(B-\beta)\big)$ for $\beta, B\in\mathcal{B}$, which we show below.

We can rewrite $1(\Delta S_i'\beta\geq 0) - 1(\Delta S_i'B\geq 0) = 1(\Delta S_i'\beta\geq 0, \Delta S_i'B < 0) - 1(\Delta S_i'\beta < 0, \Delta S_i'B\geq 0)$, and hence $\big(1(\Delta S_i'\beta\geq 0) - 1(\Delta S_i'B\geq 0)\big)^2 = 1(\Delta S_i'\beta\geq 0, \Delta S_i'B<0) + 1(\Delta S_i'\beta<0, \Delta S_i'B\geq 0)$. By symmetry, we focus on the second term. We can write $1(\Delta S_i'\beta<0, \Delta S_i'B\geq 0) = (1(\Delta S_i'B\geq 0) - 1(\Delta S_i'\beta\geq 0))1(\Delta S_i'(B-\beta)\geq 0)$. Then $\mathbb{E}\big[\Delta m(X_i,v)^2(1(\Delta S_i'B\geq 0) -$

$1(\Delta S_i'\beta \geq 0))1(\Delta S_i'(B-\beta) \geq 0)\big] \leq \mathbb{E}\big[\Delta m(X_i,v)^2(1(\Delta S_i'B \geq 0) - 1(\Delta S_i'\beta \geq 0))\big] = \frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X_i,v)^2 1(\Delta S_i'\alpha \geq 0)\big]\big|'_{\alpha=\bar{\beta}}(B-\beta)$, where $\bar{\beta}$ is between β and B by the mean value theorem.

$n^{-1/2}\sum_{i=1}^n l^{-1}\sum_{v\in V^{(l)}}\Delta m_i\big(\hat{\chi}_i - \chi_i\big) = l^{-1}\sum_{v\in V^{(l)}}\frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X_i,v)1(\Delta S_i'\alpha \geq 0)\big]'\big|_{\alpha=\beta(v)}\sqrt{n}(\hat{\beta}(v)-\beta(v)) + o_{p^*}(1) = n^{-1/2}\sum_{j=1}^n\int_0^1\frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X_i,v)1(\Delta S_i'\alpha \geq \varrho)\big]'\big|_{\alpha=a(v)}\phi_j(v)dv + \int_0^1\frac{\partial}{\partial a_2(v)}\mathbb{E}\big[\Delta m(X_i,v)1(\Delta S_i'a(v) \geq \varrho)\big]dv\sqrt{n}(\varrho_n - \varrho) + o_{p^*}(1)$ by Lemma 6.

The same arguments yield the result in 2. by replacing Δm with Δq.

Proof of Lemma 6: Let $\mathcal{V}(x) = \{v \in \mathcal{V} : \Delta q(x,v) > \varrho\}$. The approximation error of Riemann sum is $\sup_{x\in\mathcal{X}}\big|l^{-1}\sum_{v\in V^{(l)}\cap\mathcal{V}(x)} f(x,v) - \int_{\mathcal{V}(x)} f(x,v)dv\big| = O\big(\sup_{x\in\mathcal{X}} l^{-1}\sum_{v_j\in V^{(l)}}\big(\sup_{v\in(v_{j-1},v_j)} f(x,v) - \inf_{v\in(v_{j-1},v_j)} f(x,v)\big)\big)$
$= O\big(\sup_{x\in\mathcal{X}} l^{-1}\sup_{P\in\mathcal{P}}\sum_{j=0}^{n_P}\big|f(x,v_j) - f(x,v_{j-1})\big|\big) = O(l^{-1})$, where the set of all partitions $\mathcal{P} = \{P = \{v_0,\ldots,v_{n_P}\} \subset \mathcal{V}\}$.

Proof of Theorem 5: Decompose $\hat{\pi}^{DR,K} - \pi^{DR,K} = \sum_{k=1}^K \hat{\lambda}_k\hat{\pi}_k - \lambda_k\pi_k = \sum_{k=1}^K(\hat{\lambda}_k - \lambda_k)\pi_k + \lambda_k(\hat{\pi}_k - \pi_k) + O_p((\hat{\lambda}_k - \lambda_k)(\hat{\pi}_k - \pi_k))$.

Let $n_k = \sum_{i=1}^n D_i^k$. By the proof of Theorem 4, $\sum_{k=1}^K \lambda_k(\hat{\pi}_k - \pi_k) = \sum_{k=1}^K \lambda_k(n_k + n_{k-1})^{-1}\sum_{i=1}^n(D_i^k + D_i^{k-1})R_i^k/B^k + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n\sum_{k=1}^K \lambda_k\frac{(D_i^k + D_i^{k-1})R_i^k}{(p_k+p_{k-1})B^k} + o_p(n^{-1/2})$, where $R_i^k = R_{1i}^k + R_{2i}^k + R_{3i}$,

$$
\begin{aligned}
R_{1i}^k = \int_0^1 \Bigg(& \mathbb{E}[(\partial_t m_1(X, q_1(X,v))S_1 - \partial_t m_0(X, q_0(X,v))S_0 - \pi^{DR}\Delta S) \\
& \operatorname{sgn}(X,v)(D^k + D^{k-1})] + \frac{\partial}{\partial\alpha}\mathbb{E}\big[\big(\Delta m(X,v) - \pi^{DR}\Delta q(X,v)\big)\big(1(\Delta S'\alpha \geq \varrho) \\
& - 1(\Delta S'\alpha \leq -\varrho)\big)(D^k + D^{k-1})\big]\big|_{\alpha=a(v)}\Bigg)' \phi_i^k(v)dv/(p_k + p_{k-1}), \\
& \text{with } \phi_i^k(v) = \vartheta_k(v)^{-1}\big(1(T_i \leq S_i'a_k(v)) - v\big)S_i, \\
& \vartheta_k(v) = \mathbb{E}\left[f_{T|X,Z}(S'a_k(v)|X,Z)SS'(D^k + D^{k-1})\right]/(p_k + p_{k-1}), \\
& S_{1i} = (1, X_i', 1, X_i')', S_{0i} = (1, X_i', 0, \mathbf{0}_{(d_x\times 1)}')', \Delta S_i = S_{1i} - S_{0i},
\end{aligned}
$$

$$
\begin{aligned}
R_{2i}^k &= \mathcal{D}_k' G_k^{-1} \psi^J(X_i, T_i, Z_i) e_i, \\
&\quad \text{with } G_k = \mathbb{E}\left[e^2 \psi^J(X,T,Z)\psi^J(X,T,Z)'(D^k + D^{k-1})\right]/(p_k + p_{k-1}), \\
&\quad \mathcal{D}_k = \int_0^1 \mathbb{E}\big[(\psi^J(X, q_1(X,v), 1) - \psi^J(X, q_0(X,v), 0))\text{sgn}(X,v) \\
&\qquad \times (D^k + D^{k-1})\big] dv/(p_k + p_{k-1}), \\
R_{3i} &= \int_0^1 \big(\Delta m(X_i, v) - \pi^{DR} \Delta q(X_i, v)\big)\, \text{sgn}(X_i, v) dv, \\
B^k &= \int_0^1 \mathbb{E}\left[|\Delta q(X,v)| 1(|\Delta q(X,v)| \geq \varrho)(D^k + D^{k-1})\right] dv/(p_k + p_{k-1}).
\end{aligned} \tag{S.11}
$$

Next we analyze $\sum_{k=1}^K (\hat{\lambda}_k - \lambda_k)\pi_k$. Let $\mathsf{A}_k = Q_k P_k$, where $Q_k = q_k - q_{k-1}$ and $P_k = \sum_{l=k}^K p_l(q_l - \mathbb{E}[T])$. Let $\lambda_k = \mathsf{A}_k/\mathsf{B}$, where $\mathsf{B} = \sum_{k=1}^K \mathsf{A}_k$. So $\pi^{DR,K} = \sum_{k=1}^K \pi_k \mathsf{A}_k/\mathsf{B}$. Then $\sum_{k=1}^K (\hat{\lambda}_k - \lambda_k)\pi_k = \sum_{k=1}^K \big\{(\hat{\mathsf{A}}_k - \mathsf{A}_k)/\mathsf{B} - (\hat{\mathsf{B}} - \mathsf{B})\mathsf{A}_k/\mathsf{B}^2\big\}\pi_k + o_p(n^{-1/2}) = \sum_{k=1}^K (\hat{\mathsf{A}}_k - \mathsf{A}_k)(\pi_k - \pi^{DR,K})/\mathsf{B} + o_p(n^{-1/2})$.

Decompose $\hat{\mathsf{A}}_k - \mathsf{A}_k = (\hat{Q}_k - Q_k)P_k + (\hat{P}_k - P_k)Q_k + o_p(n^{-1/2})$. It is straightforward to show that $\hat{q}_k - q_k = n^{-1}\sum_{i=1}^n \big\{(T_i D_i^k - \mathbb{E}\left[T_i D_i^k\right])/p_k - (D_i^k - p_k)q_k/p_k\big\} + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n (T_i - q_k)D_i^k/p_k + o_p(n^{-1/2})$.

$\hat{P}_k - P_k = \sum_{l=k}^K \Big\{(\hat{p}_l - p_l)(q_l - \mathbb{E}[T]) + p_l\big(\hat{q}_l - q_l - \bar{T} + \mathbb{E}[T]\big)\Big\} + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n \sum_{l=k}^K \Big\{(D_{li} - p_l)(q_l - \mathbb{E}[T]) + p_l\big((T_i - q_l)D_{li}/p_l - T_i + \mathbb{E}[T]\big)\Big\} + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n \sum_{l=k}^K \big(D_{li} - p_l\big)\big(T_i - \mathbb{E}[T]\big) - P_k + o_p(n^{-1/2})$.

Therefore $\sum_{k=1}^K (\hat{\lambda}_k - \lambda_k)\pi_k = \sum_{k=1}^K (\hat{\mathsf{A}}_k - \mathsf{A}_k)(\pi_k - \pi^{DR,K})/\mathsf{B} + o_p(n^{-1/2}) = n^{-1}\sum_{i=1}^n \sum_{k=1}^K R_{4ki} + o_p(n^{-1/2})$, where

$$
\begin{aligned}
R_{4i}^k = \Bigg\{ &\left((T_i - q_k)\frac{D_i^k}{p_k} - (T_i - q_{k-1})\frac{D_i^{k-1}}{p_{k-1}}\right)\sum_{l=k}^K p_l(q_l - \mathbb{E}[T]) + (q_k - q_{k-1}) \\
&(T_i - \mathbb{E}[T])\sum_{l=k}^K \big(D_{li} - p_l\big)\Bigg\} \frac{\pi_k - \pi^{DR,K}}{\sum_{k=1}^K (q_k - q_{k-1})\sum_{l=k}^K p_l(q_l - \mathbb{E}[T])}.
\end{aligned} \tag{S.12}
$$

By R_i^k given in (S.11) and R_{4i}^k given in (S.12), we obtain the influence function

$$R_{Ki} = \sum_{k=1}^{K} \lambda_k \frac{(D_i^k + D_i^{k-1})R_i^k}{(p_k + p_{k-1})B^k} + R_{4i}^k. \tag{S.13}$$

Asymptotic normality follows the same arguments in the proof of Theorem 4 with the following modifications. The law of iterated expectations yields $\sigma_{Kn}^2 = \sigma_{K1}^2 + \sigma_{K2n}^2 + \sigma_{K3}^2$, where $\sigma_{K1}^2 = \mathbb{E}\left[\left(\sum_{k=1}^K \lambda_k \frac{(D_i^k+D_i^{k-1})R_{1i}^k}{(p_k+p_{k-1})B^k} + R_{4i}^k\right)^2\right]$, $\sigma_{K2n}^2 = \mathbb{E}\left[\left(\sum_{k=1}^K \lambda_k \frac{(D_i^k+D_i^{k-1})R_{2i}^k}{(p_k+p_{k-1})B^k}\right)^2\right]$, and $\sigma_{K3}^2 = \mathbb{E}\left[\left(\sum_{k=1}^K \lambda_k \frac{(D_i^k+D_i^{k-1})R_{3i}}{(p_k+p_{k-1})B^k}\right)^2\right]$.

S.4 Variance Estimation

For convenience, we first collect the relevant notations and then discuss implementation details.

S.4.1 Notation:

Let $\phi_i(v) = \vartheta(v)^{-1}\big(1(T_i \le S_i'a(v)) - v\big)S_i$. Let the trimming function $\chi^+(x,v) = 1(\Delta q(x,v) \ge \varrho)$. Let $S_{1i} = (1, X_i', 1, X_i')'$, $S_{0i} = (1, X_i', 0, \mathbf{0}_{(d_x\times 1)}')'$, $\Delta S_i = S_{1i} - S_{0i}$, $\partial_t m_z(X, q_z) = \frac{\partial}{\partial t} m_z(X,t)|_{t=q_z(X,v)}$.

$$\begin{aligned}
R_{1i}^+ &= \int_0^1 \Bigg(\mathbb{E}\left[\left(\partial_t m_1(X, q_1(X,v))S_1 - \partial_t m_0(X, q_0(X,v))S_0 - \pi_+^{DR}\Delta S\right)\chi^+(X,v)\right] \\
&\quad + \frac{\partial}{\partial \alpha}\mathbb{E}\left[\left(\Delta m(X,v) - \pi_+^{DR}\Delta q(X,v)\right)1(\Delta S'\alpha \ge \varrho)\right]\big|_{\alpha=a(v)}\Bigg)' \phi_i(v)dv, \\
R_{2i}^+ &= \mathcal{D}^{+'}G^{-1}\psi^J(X_i, T_i, Z_i)e_i, \text{with } G \equiv \mathbb{E}\left[\psi^J(X,T,Z)\psi^J(X,T,Z)'\right] = \mathbb{E}\left[\Psi'\Psi/n\right], \\
\mathcal{D}^+ &= \mathcal{D}_1^+ - \mathcal{D}_0^+, \mathcal{D}_z^+ = \int_0^1 \mathbb{E}\left[\psi^J(X, q_z(X,v), z)\chi^+(X,v)\right]dv, \\
R_{3i}^+ &= \int_0^1 \left(\Delta m(X_i,v) - \pi_+^{DR}\Delta q(X_i,v)\right)\chi^+(X_i,v)dv, \\
B_+ &= \int_0^1 \int_{\mathcal{X}} \Delta q(x,v)\chi^+(x,v)f(x)dxdv.
\end{aligned}$$

Let $\chi^-(x,v) = 1(\Delta q(x,v) < -\varrho)$ and $B_- = \int_0^1 \int_{\mathcal{X}} \Delta q(x,v)\chi^-(x,v)f(x)dxdv$. Let $B = B_+ - B_-$.

For π_-^{DR}, define R_i^- as R_i^+ by replacing + with − in all the components in R_i^+.

For π^{DR}, define $R_i = R_{1i} + R_{2i} + R_{3i}$, where $R_{ki} = R_{ki}^+ - R_{ki}^-$ for $k = 1, 2, 3$ except that one needs to replace π_+^{DR} and π_-^{DR} with π^{DR} in R_{ki}^+ and R_{ki}^-, $k = 1, 3$.

S.4.2 Implementation

We estimate σ^2 by the sample analogue plug-in estimator, i.e., $\hat{\sigma}^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2$, where $\hat{\sigma}_k^2 = n^{-1}\sum_{i=1}^n \hat{R}_{ki}^2/\hat{B}^2$, $\hat{B}$ and $\hat{R}_{ki}$ are consistent estimators for B and R_{ki} for $k = 1, 2, 3$, respectively, given in (S.8):

For $\hat{R}_{1i}$, $\partial_t \hat{m}_z$ is directly computed from Step 2. From the linear quantile regression literature, it is standard $\hat{\vartheta}(v) = n^{-1}\sum_{i=1}^n \hat{f}_{T|X,Z}(S_i'\hat{a}(v)|X_i, Z_i)S_iS_i'$. The derivative $\frac{\partial}{\partial\alpha}\mathbb{E}\big[\Delta m(X,v)1(\Delta S_i'\alpha \geq \varrho)\big]\big|_{\alpha=a(v)}$ may be estimated by a numerical differentiation, i.e., $n^{-1}\sum_{i=1}^n \Delta\hat{m}(X_i,v)\big(1(\Delta S_i'(\hat{a}(v)+\iota/2) \geq \varrho_n) - 1(\Delta S_i'(\hat{a}(v)-\iota/2) \geq \varrho_n)\big)/\iota$ for some small $\iota > 0$.

For $\hat{R}_{2i}$, let $\hat{e}_i = Y_i - \psi^J(X_i,T_i,Z_i)'\hat{c}$, $\hat{\Omega} = n^{-1}\sum_{i=1}^n \hat{e}_i^2\psi^J(X_i,T_i,Z_i)\psi^J(X_i, T_i, Z_i)'$, $\hat{G} = \Psi'\Psi/n$, and $\hat{\mho} = \hat{G}^{-1}\hat{\Omega}\hat{G}^{-1}$. Let $\hat{\mathcal{D}}^+ = n^{-1}\sum_{i=1}^n l^{-1}\sum_{v\in V^{(l)}} \hat{\psi}_i^J$ $\hat{\chi}^+(X_i,v)$. Let $\hat{\mathcal{D}} = \hat{\mathcal{D}}^+ - \hat{\mathcal{D}}^-$. Then $\hat{\sigma}_{2n}^2 = \hat{\mathcal{D}}'\hat{\mho}\hat{\mathcal{D}}$, $\hat{\sigma}_{+2}^2 = \hat{\mathcal{D}^+}'\hat{\mho}\hat{\mathcal{D}^+}$, and $\hat{\sigma}_{-2}^2 = \hat{\mathcal{D}^-}'\hat{\mho}\hat{\mathcal{D}^-}$.

$\hat{R}_{3i}^+ = l^{-1}\sum_{v\in V^{(l)}} \big(\Delta\hat{m}(X_i,v) - \hat{\pi}_+^{DR}\Delta\hat{q}(X_i,v)\big)\hat{\chi}^+(X_i,v)$, and $\hat{B}_+$ is analogous.

Consider $\varrho = 0$. In practice, one may choose $\varrho_n = 1.96 \times \min_{v\in V^{(l)},\{X_i\}_{i=1,\ldots,n}} se(\Delta\hat{T}(X_i,v))/\log(n)$. This procedure includes insignificant estimates of $\Delta\hat{T}(X_i,v)$ (at the 5% significance level).[17]

[17]Step 1 is $O_p(n^{-1/2})$, so the estimation error of χ is of first order asymptotically by Lemma 5. The rate condition on $\sqrt{n}(\varrho_n - \varrho) = o(1)$ means that using ϱ_n rather than ϱ is first-order asymptotically ignorable.

References

[1] Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the U.S. wage structure," Econometrica, 74(2), 539-63.

[2] Chen, X. and T. Christensen (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV Regression," Quantitative Economics, 9(1), 39-85.

[3] Van Der Vaart, A. W. and J. A. Wellner (1996): Weak convergence and empirical processes. with applications to statistics. New York: Springer-Verlag.