On Quantile Treatment Effects, Rank Similarity, and Multiple Instrumental Variables

Sukjin Han
School of Economics
University of Bristol
sukjin.han@gmail.com

Haiqing Xu
Department of Economics
University of Texas at Austin
h.xu@austin.utexas.edu

October 17, 2022

Abstract

This paper investigates how certain relationship between observed and counterfactual distributions plays a role in the identification of distributional treatment effects under endogeneity, and shows that this relationship holds in a range of nonparametric models for treatment effects. To motivate the new identifying assumption, we first provide a novel characterization of popular assumptions restricting treatment heterogeneity in the literature, specifically rank similarity. We show the stringency of this type of assumptions and propose to relax them in economically meaningful ways. This relaxation will justify certain parameters (e.g., treatment effects on the treated) against others (e.g., treatment effects for the entire population). It will also justify the quest of richer exogenous variation in the data (e.g., the use of multiple instrumental variables). The prime goal of this investigation is to provide empirical researchers with tools for identifying and estimating treatment effects that are flexible enough to allow for treatment heterogeneity, but still yield tight policy evaluation and are easy to implement.

JEL Numbers: C14, C32, C33, C36

Keywords: quantile treatment effects, rank similarity, average treatment effects, endogeneity, multi-valued instrumental variables, partial identification.

1 Introduction

This paper investigates how certain relationship between observed and counterfactual distributions plays a role in the identification of distributional treatment effects under endogeneity and shows that this relationship holds in a range of nonparametric models for treatment effects. To motivate the new identifying assumption, we first provide a novel characterization of popular assumptions restricting treatment heterogeneity in the literature, specifically the rank similarity assumptions. This characterization allows us to show the stringency of the assumptions and motivates us ways of relaxing them in economically meaningful ways. This relaxation will justify certain parameters (e.g., treatment effects on the treated) against others (e.g., treatment effects for the entire population). It will also justify the quest of richer exogenous variation in the data (e.g., the use of multiple instrumental variables).

The prime goal of this investigation is to provide empirical researchers with (i) a framework where validity of identifying assumptions prescribes the parameters of interest, (ii) tools for identifying and estimating treatment effects that are flexible enough to allow for treatment heterogeneity, but that still yield tight policy evaluation and are easy to implement, and (iii) guidance on data collection that leads to drawing meaningful causal conclusions.

We consider a mapping between observed and counterfactual distributions. A version of this mapping can be informally stated in terms of the following preservation of first order stochastic dominance (FOSD): for arbitrary compliance types $t, t' \in \mathcal{T}$ induced by instrumental variables (IVs), if

$$Y_1|t \prec_{FOSD} Y_1|t' \tag{1.1}$$

then

$$Y_0|t \prec_{FOSD} Y_0|t', \tag{1.2}$$

where Y_d denotes the potential outcome given treatment $D = d$.[1] This condition relates the observed distributions to the counterfactual distributions. For instance, consider instrument $Z \in \{0, 1\}$ and $\mathcal{T} = \{C, AT, NT\}$ where C, AT, and NT stand for compliers, always-takers, and never-takers, respectively. Let Y be the observed outcome given by $Y \equiv DY_1 + (1-D)Y_0$. A simple algebra shows that $Y_1|AT \prec_{FOSD} Y_1|C$ can be expressed as

$$P[Y \leq y|D = 1, Z = 0] \leq P[Y \leq y|D = 1, Z = 1] \quad \text{for all } y \tag{1.3}$$

and $Y_0|AT \prec_{FOSD} Y_0|C$ can be expressed as

$$P[Y_0 \leq y|D = 1] \leq \frac{P[Y \leq y, D = 0|Z = 1] - P[Y \leq y, D = 0|Z = 0]}{P[D = 1|Z = 0] - P[D = 1|Z = 1]} \quad \text{for all } y. \tag{1.4}$$

Then, (1.4) produces an informative upper bound for $P[Y_0 \leq y|D = 1]$ and a symmetric analysis produces a lower bound. Note $P[Y_0 \leq y|D = 1]$ is a key component in calculating the distributional treatment effect, such as the quantile treatment effect on the treated (QTT). We also provide analogous conditions to bound average treatment effects. The proposed bounds can be much tighter than other bounds available in the literature but is not guaranteed without (1.3) (or equivalently, (1.1)) being satisfied. Note that (1.3) is a testable restriction from the data. Although (1.3) may seem too restrictive to hold in the data, this is not generally the case when Z departs from a scalar binary variable. One of the main messages we hope to deliver in this paper is precisely that: (1.1) justifies the need of Z that takes multiple values (or equivalently the need of multiples IVs). In that case, (1.1) is likely to satisfy (and can be tested) and we can eventually obtain informative bounds

[1] For r.v.'s A and B, let $A \prec_{FOSD} B$ denotes $F_B(t) \leq F_A(t)$ where F_A and F_B are CDFs of A and B, respectively.

on the distributional treatment effects under relatively weak assumptions on heterogeneity. Although the procedure may requires multi-valued IVs, the benefit of such variables can be manifested without requiring continuous or large support. This is practically relevant as infinitely supported IVs and continuously varying IVs are typically hard to find in practice, whereas discrete IVs are more common in an experimental and observational settings.

Nonparametric identification of treatment effects using IVs with limited support has been a challenging goal even when the focus is on mean treatment effects, such as the average treatment effect (ATE) and the ATE on the treated (ATT). In an influential line of work (Manski (1990), Manski (1997), Manski and Pepper (2000)), Manski proposes the approach of partial identification and shows that sharp bounds on the ATE can be constructed under a set of assumptions relating to the directions of treatment effects and treatment selection while allowing instruments to be invalid in a specific sense. Even with valid instruments, however, bounds on the ATE are typically wide and uninformative to yield precise policy prediction. The local ATE (LATE) (Imbens and Angrist (1994)) and local QTE (Abadie et al. (2002)) have been a popular (point-identified) alternative when researchers are equipped with discrete IVs and impose a monotonicity assumption in the treatment selection. Unfortunately, it has intrinsic limitation that the local group for which the treatment effect is identified may not be the group of policy interest. Correspondingly, the extrapolation of the local parameters becomes an important issue for policy analysis (e.g., treatment allocation), in which case the identification challenge still remains (Mogstad et al. (2018), Han and Yang (2020)).

In another seminal work, Chernozhukov and Hansen (2005) propose assumptions that restrict the degree of treatment heterogeneity, called rank similarity and rank invariance, and point identify the QTE and the ATE. These assumptions (which are in fact observationally equivalent to each other) are shown to have great identifying power and used in various nonparametric contexts implicitly or explicitly (Heckman et al. (1997), Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), Vuong and Xu (2017), Han (2019) to name a few). However, their plausibility can be questionable in many applications (e.g., Maasoumi and Wang (2019)) and testing methods are also proposed as one reaction to the skepticism (Frandsen

and Lefgren (2018), Dong and Shen (2018), Kim and Park (2022)). In this paper, we provide an alternative interpretation of rank similarity to clarify its stringency. This motivates us to relax rank similarity, which in turn generates our new identifying conditions.

This paper embraces the difficulty of identifying distributional and average effects that are policy relevant under arbitrary heterogeneity without requiring IVs to have infinite or continuous support. Our approach is to exploit the trade-off between flexibility of models and data requirements in a way that is empirically relevant. The condition we introduce (namely, the preservation of FOSD) substantially weakens rank similarity and reveals the usefulness of multiple IVs. We believe this particular trade-off that has not been previously explored. We verify that the preservation of FOSD ordering is satisfied in a range of nonseparable models that includes (but is not restricted to) models under rank similarity. Researchers may justify these models in their particular applications or directly argue that the preservation condition holds (which is weaker to assume). A prevailing feature in these models is that they contain multi-dimensional unobservables, which is crucial to allow for essential treatment heterogeneity. A notable aspect of our approach is that depending on the direction of the FOSD preservation (i.e., from (l.1) to (1.2) or the converse) the procedure calculates bounds on either QTT or the QTE on the untreated (QTUT). In the context of the proposed models, we provide economic interpretation of a specific direction of preservation and justify why a policymaker wants to consider a particular type of treatment parameters depending on the plausibility of this assumption. This way, we argue the usefulness of "assumption-driven" treatment parameters.

We develop a procedure of calculating bounds on the QTT and QTUT, as well as the ATT and ATE on the untreated (ATUT), when the data requirement is met. The bound calculation is straightforward; it only requires to solve a set of linear programs to gather information from the data via (1.1) or (1.2) (or its observable equivalence). The bounds are initially characterized by optimal values of semi-infinite programs, especially with a continuous outcome variable. To tackle the infeasibility of the program, we propose two alternative approaches to transform it into a feasible linear program: (i) randomizing the constraints in the semi-

infinite programs; (ii) invoking duality of the semi-infinite programs and approximating the Lagrangian measure using sieve.

2 Key Conditions and Bounds on Treatment Effects

Let $D \in \{0,1\}$ be the observed treatment indicator, which represents the endogenous decision of an individual responding to IVs Z. We assume Z is either a vector of binary IVs or a multi-valued IV. In either case, we assume that Z takes L distinct values: $Z \in \mathcal{Z} \equiv \{z_1, ..., z_L\}$. Multi-valued or multiple IVs are common in many observational studies (e.g., natural experiments typically provide more than one instrument) and experimental studies (e.g., randomized control trials where multiple treatment arms are implemented either simultaneously or sequentially).[2] One of the main purposes of this paper is to motivate this type of IVs from the perspective of identification analysis. Let Y_1 be the counterfactual outcome of being treated and Y_0 be that of not being treated. They can be either continuously or discretely distributed. The observed outcome $Y \in \mathcal{Y}$ satisfies $Y = DY_1 + (1-D)Y_0$. Finally, $X \in \mathcal{X}$ denotes other covariates that may be endogenous.

Define QTE and ATE for treated and untreated populations. For $d \in \{0,1\}$ and $x \in \mathcal{X}$, define

$$QTE_\tau(d,x) = Q_{Y_1|D,X}(\tau|d,x) - Q_{Y_0|D,X}(\tau|d,x)$$

and

$$ATE(d,x) = E[Y_1 - Y_0|D=d, X=x].$$

These parameters are what researchers and policymakers are potentially interested. The unconditional QTE and ATE (with respect to $D=d$) can be recovered when these parameters are identified for all $d \in \{0,1\}$. Throughout the paper, we maintain that the IVs are valid

[2] See Mogstad et al. (2021) for a recent survey.

and satisfy the exclusion restriction.

Assumption Z. *For $d \in \{0,1\}$ and $z \in \{z_1, ..., z_L\}$, (i) $Y_{d,z} = Y_d$; (ii) $Z \perp Y_d|X$.*

Now we introduce the main identifying condition of this paper that establishes the mapping between observed and counterfactual distributions.

Condition 2.1. *For arbitrary non-negative weight vectors $(w_1, ..., w_L)$ and $(\tilde{w}_1, ..., \tilde{w}_L)$ that satisfy $\sum_{\ell=1}^{L} w_\ell = \sum_{\ell=1}^{L} \tilde{w}_\ell = 1$, if*

$$\sum_{\ell=1}^{L} w_\ell P[Y_1 \leq \cdot | D = 1, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_1 \leq \cdot | D = 1, Z = z_\ell, X = x], \tag{2.1}$$

then

$$\sum_{\ell=1}^{L} w_\ell P[Y_0 \leq \cdot | D = 1, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_0 \leq \cdot | D = 1, Z = z_\ell, X = x]. \tag{2.2}$$

Note that the probabilities in (2.1) are observed as $Y_D = Y$. The mapping between observed and counterfactual distributions has been considered in Vuong and Xu (2017), which insights we share. Suppose $Z \perp (Y_d, D_z)|X$ where D_z is the counterfactual treatment that satisfies $D = \sum_{z \in \mathcal{Z}} 1\{Z = z\} D_z$; this assumption is stronger than Assumption Z that we require. Under this assumption, Condition 2.1 is equivalent to assuming if

$$\sum_{\ell=1}^{L} w_\ell P[Y \leq \cdot | D_{z_\ell} = 1, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y \leq \cdot | D_{z_\ell} = 1, X = x], \tag{2.3}$$

then

$$\sum_{\ell=1}^{L} w_\ell P[Y_0 \leq \cdot | D_{z_\ell} = 1, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_0 \leq \cdot | D_{z_\ell} = 1, X = x]. \tag{2.4}$$

Note that the event $\{D_{z_\ell} = 1\}$ (for $z_1, ..., z_L$) captures a type of compliance to a given $Z = z_\ell$. Then, $\sum_{\ell=1}^{L} w_\ell P[Y_d \leq y | D_{z_\ell} = 1, X = x]$ can be viewed as a distribution of Y_d weighted across different compliance types, and thus resulting in a distribution for a hypothetical population

with a specific composition of compliance types. Therefore, Condition 2.1 posits that the FOSD ordering between the distributions of Y of two compliance compositions is preserved between the distributions of Y_0 of the same pair of compliance compositions. For example, when $L = 2$ and defiers are excluded as a possible compliance type (e.g., the monotonicity assumption for local ATE (LATE) by Imbens and Angrist (1994)), then Condition 2.1 simply concerns the stochastic ordering between always-takers and compliers. When $L \geq 3$, the composition becomes more complicated; see Section 3. We provide a condition sufficient to Condition 2.1 that may be easier to interpret. To state this sufficient condition, we introduce a very general model for the treatment selection:

Assumption SEL. *Assume that*

$$D = h(Z, X, \eta) \tag{2.5}$$

where $\eta \in \mathcal{T}$ can be an arbitrary vector.

Although it is not necessary for our main procedure, this assumption is useful in capturing the types of compliance behavior via the unobservable η. Note that (2.5) permits a more general compliance behavior than what a weakly separable model $D = 1\{\eta \leq h(Z, X)\}$ does (or equivalently, what the LATE monotonicity does). This generalized selection equation is introduced to state the following sufficient condition and is not necessary for the main results of our identification analysis. Let $F_{Y_d|\eta,X}(y|t,x) \equiv P[Y_d \leq y|\eta = t, X = x]$.

Condition 2.2. *Fix $x \in \mathcal{X}$. For arbitrary weight functions $w : \mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ and $\tilde{w} : \mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ such that $\int w(t,x)dt = \int \tilde{w}(t,x)dt = 1$, if*

$$\int w(t,x) F_{Y_1|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x) F_{Y_1|\eta,X}(\cdot|t,x)dt,$$

then

$$\int w(t,x) F_{Y_0|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x) F_{Y_0|\eta,X}(\cdot|t,x)dt.$$

Because $\int w(t,x)dt = 1$, note that $\int w(t,x)F_{Y_d|\eta,X}(\cdot|t,x)dt$ is a mixture of conditional CDFs (with $w(t,x)$ being the mixture weight) and thus itself a CDF. Defining a type distribution $W_x(t) = \int^t w(\eta,x)d\eta$, we can write $\int w(t,x)F_{Y_d|\eta,X}(\cdot|t,x)dt = \int F_{Y_d|\eta,X}(\cdot|t,x)dW_x(t)$.[3] Therefore, Condition 2.2 assumes that the FOSD ordering of Y_1 distributions conditional on η conforming to two different type distributions is preserved in the ordering of Y_0 distributions conditional on the same type distributions. Note that Condition 2.2 (or Condition 2.1) is *not* an "if and only if" statement. If a condition imposes that the preservation of ordering holds in both directions, it would be a very stringent one. In fact, such a condition is closely related to the rank similarity condition (Chernozhukov and Hansen (2005)); see Section 4 for full discussions. The following lemma establishes the sufficiency of Condition 2.2 for Condition 2.1.

Lemma 2.1. *Under Assumption SEL, Condition 2.2 implies Condition 2.1.*

The proof of this lemma and most of other proofs are contained in the appendix. Now, we show that Condition 2.1 is useful in constructing bounds on $F_{Y_0|D,X}(\cdot|1,x)$ and subsequently on $QTE_\tau(1,x)$. Let

$$\Gamma_p(x) \equiv \left\{ (\gamma_1(x),...,\gamma_L(x)) \in \mathbb{R}^L : \sum_{\ell=1}^L \gamma_\ell(x) = 0 \text{ and } \sum_{\ell=1}^L \gamma_\ell(x)p(z_\ell,x) = 1 \right\}.$$

Theorem 2.1. *Suppose that Assumption Z and Condition 2.1 hold. Fix $x \in \mathcal{X}$. For $\gamma(x) \equiv (\gamma_1(x),...,\gamma_L(x))$ and $\tilde{\gamma}(x) \equiv (\tilde{\gamma}_1(x),...,\tilde{\gamma}_L(x))$ in $\Gamma_p(x)$, suppose*

$$P[Y \le y|D=1,X=x] \le \sum_{\ell=1}^L \gamma_\ell(x)P[Y \le y, D=1|Z=z_\ell,X=x], \quad \forall y \in \mathcal{Y} \tag{2.6}$$

$$\sum_{\ell=1}^L \tilde{\gamma}_\ell(x)P[Y \le y, D=1|Z=z_\ell,X=x] \le P[Y \le y|D=1,X=x], \quad \forall y \in \mathcal{Y} \tag{2.7}$$

[3] Since η has arbitrary dimensions, the integral with respect to t is understood to be a multivariate integral.

Then $F_{Y_0|D,X}(\cdot|1,x)$ *is bounded by*

$$-\sum_{\ell=1}^{L}\tilde{\gamma}_\ell(x)P[Y\le\cdot,D=0|Z=z_\ell,X=x] \tag{2.8}$$

$$\le P[Y_0\le\cdot|D=1,X=x]$$

$$\le -\sum_{\ell=1}^{L}\gamma_\ell(x)P[Y\le\cdot,D=0|Z=z_\ell,X=x] \tag{2.9}$$

Proof. We suppress X for simplicity and prove the upper bound; the lower bound can be analogously derived. Without loss of generality, for some $\ell^*\le L$, let $\gamma_\ell\le 0$ for $\ell\le\ell^*$ and $\gamma_\ell>0$ for $\ell>\ell^*$. Let $q(z_\ell)\equiv P[Z=z_\ell|D=1]$. Then, (2.6) can be rewritten as

$$\sum_{\ell=1}^{L}q(z_\ell)\times P[Y\le y|D=1,Z=z_\ell]-\sum_{\ell=1}^{\ell^*}\gamma_\ell q(z_\ell)\times\mathbb{P}(Y\le y|D=1,Z=z_\ell)$$
$$\le\sum_{\ell=\ell^*+1}^{L}\gamma_\ell p(z_\ell)\times\mathbb{P}(Y\le y|D=1,Z=z_\ell).$$

Let $a\equiv 1-\sum_{\ell=1}^{\ell^*}\gamma_\ell p(z_\ell)$. By definition and that $\sum_{\ell=1}^{L}\gamma_\ell p(z_\ell)=1$, we have $a=\sum_{\ell=\ell^*+1}^{L}\gamma_\ell p(z_\ell)$. Therefore, we have

$$\sum_{\ell=1}^{\ell^*}\frac{q(z_\ell)-\gamma_\ell p(z_\ell)}{a}\times\mathbb{P}(Y_1\le y|D=1,Z=z_\ell)+\sum_{\ell=\ell^*+1}^{L}\frac{q(z_\ell)}{a}\times\mathbb{P}(Y_1\le y|D=1,Z=z_\ell)$$
$$\le\sum_{\ell=\ell^*+1}^{L}\frac{\gamma_\ell p(z_\ell)}{a}\times\mathbb{P}(Y_1\le y|D=1,Z=z_\ell).$$

Note that $\sum_{\ell=1}^{\ell^*}\frac{q(z_\ell)-\gamma_\ell p(z_\ell)}{a}+\sum_{\ell=\ell^*+1}^{L}\frac{q(z_\ell)}{a}=1$ and $\sum_{\ell=\ell^*+1}^{L}\frac{\gamma_\ell p(z_\ell)}{a}=1$. Therefore, by Condition 2.1, we have

$$\sum_{\ell=1}^{k}\frac{q(z_\ell)-\gamma_\ell p(z_\ell)}{a}\times P[Y_0\le y|D=1,Z=z_\ell]+\sum_{\ell=\ell^*+1}^{L}\frac{q(z_\ell)}{a}\times P[Y_0\le y|D=1,Z=z_\ell]$$
$$\le\sum_{\ell=\ell^*+1}^{L}\frac{\gamma_\ell p(z_\ell)}{a}\times P[Y_0\le y|D=1,Z=z_\ell].$$

Equivalently, we have

$$
\begin{aligned}
& P[Y_0 \le y | D = 1] \\
\le & \sum_{\ell=1}^{L} \gamma_\ell \times P[Y_0 \le y, D = 1 | Z = z_\ell] \\
= & \sum_{\ell=1}^{L} \gamma_\ell \times [P[Y_0 \le y | Z = z_\ell] - P[Y_0 \le y, D = 0 | Z = z_\ell]] \\
= & \sum_{\ell=1}^{L} \gamma_\ell P[Y_0 \le y | Z = z_\ell] - \sum_{\ell=1}^{L} \gamma_\ell \times P[Y_0 \le y, D = 0 | Z = z_\ell] \\
= & P[Y_0 \le y] \times \sum_{\ell=1}^{L} \gamma_\ell - \sum_{\ell=1}^{L} \gamma_\ell \times P[Y \le y, D = 0 | Z = z_\ell] \\
= & -\sum_{\ell=1}^{L} \gamma_\ell \times P[Y \le y, D = 0 | Z = z_\ell].
\end{aligned}
$$

□

Note that there can be multiple $\gamma(x)$ and $\tilde{\gamma}(x)$ in $\Gamma_p(x)$ that satisfy (2.6) and (2.7), respectively. Therefore, we can further tighten the bounds as follows.

Corollary 2.1. *Suppose that Assumption Z and Condition 2.1 hold. Fix $x \in \mathcal{X}$. Then, $F_{Y_0|D,X}(\cdot|1,x)$ is upper and lower bounded by*

$$
\begin{aligned}
F^{UB}_{Y_0|D,X}(y|1,x) &\equiv \min_{\gamma(x) \in \Gamma_p(x): (2.6) \text{ holds}} -\sum_{\ell=1}^{L} \gamma_\ell(x) P[Y \le y, D = 0 | Z = z_\ell], \\
F^{LB}_{Y_0|D,X}(y|1,x) &\equiv \max_{\tilde{\gamma}(x) \in \Gamma_p(x): (2.7) \text{ holds}} -\sum_{\ell=1}^{L} \tilde{\gamma}_\ell(x) P[Y \le y, D = 0 | Z = z_\ell].
\end{aligned}
$$

Theorem 2.1 and Corollary 2.1 highlight the usefulness of multi-valued IVs. The key to calculating bounds is to find $\gamma(x)$ (and $\tilde{\gamma}(x)$) in $\Gamma_p(x)$ that satisfies (2.6) (and (2.7)). The inequality is more likely to hold when instruments take more values (i.e., when L is large) because, in this case, the degree of freedom in Γ_p increases. The corollary additionally shows how the bounds can be further tightened if the set of $\gamma(x) \in \Gamma_p(x)$ that satisfies (2.6) becomes "large" with large L. See Sections 3 and 6 for related discussions. Note that (2.6)–(2.7) are

testable in the data.

Finally, note that

$$QTE_\tau(1,x) = Q_{Y|D,X}(\tau|1,x) - Q_{Y_0|D,X}(\tau|1,x)$$

and the bounds on the second quantity on the right-hand side can be calculated using the worst case bounds for the conditional quantile (Manski (1994), Blundell et al. (2007)):

$$Q^{LB}_{Y_0|D,X}(\tau|1,x) \leq Q_{Y_0|D,X}(\tau|1,x) \leq Q^{UB}_{Y_0|D,X}(\tau|1,x)$$

with $Q^{LB}_{Y_0|D=1}(\tau)$ and $Q^{UB}_{Y_0|D=1}(\tau)$ being the solutions to

$$\tau = F^{UB}_{Y_0|D,X}(y|1,x),$$
$$\tau = F^{LB}_{Y_0|D,X}(y|1,x),$$

respectively, using Corollary 2.1. Although the bounds on $ATE(1,x) = E[Y|D=1,X=x] - E[Y_0|D=1,X=x]$ can be calculated based on $E[Y_0|D=1,X=x] = \int_0^1 Q_{Y_0|D,X}(\tau|1,x)d\tau$, we present later how the bounds on the $ATE(d,x)$ can be calculated under a weaker condition than Condition 2.1.

Remark 2.1 (Multi-Valued IVs)**.** *In the introduction we claim that, when $L = 2$ and $(\gamma_1,\gamma_2) = \left(\frac{1}{p(z_1)-p(z_2)}, -\frac{1}{p(z_1)-p(z_2)}\right)$, (2.6) can be expressed as $P[Y \leq y|D=1,Z=0] \leq P[Y \leq y|D=1,Z=1]$ for all y, and that the latter is in turn establishes FOSD of Y_1 between always takers and compliers, assuming there is no defiers. In the next section, we formally prove this claim and provide further insights by giving interpretations of (2.6) for $L = 3$.*

Remark 2.2 (More on $\gamma(x)$)**.** *The restrictions on $\gamma(x)$ (i.e., $\sum_{\ell=1}^L \gamma_\ell(x) = 0$ and $\sum_{\ell=1}^L \gamma_\ell p(z_\ell,x) = 1$) are worth discussing. First, note that the existence of such a sequence requires the relevance of the IV: $p(z_\ell,x) \neq p(z_{\ell'},x)$ for some $z_\ell, z_{\ell'}$. Moreover, any sequence $\gamma(x)$ satisfying*

$\sum_{\ell=1}^{L} \gamma_\ell(x) = 0$ *and* $\sum_{\ell=1}^{L} \gamma_\ell(x) p(z_\ell, x) \neq 0$ *is enough. We can always rescale* $\gamma_\ell(x)$*'s as* $\gamma_\ell^*(x) = \frac{\gamma_\ell(x)}{\sum_{\ell=1}^{L} \gamma_\ell(x) p(z_\ell, x)}$*, then* $\gamma^*(x) = (\gamma_1^*(x), ..., \gamma_L^*(x))$ *would satisfy* (2.6) *and* (2.9) *in Theorem 2.1.*

If we assume the converse of Condition 2.1, we can calculate bounds on the $QTE(0, x)$.

Condition 2.3. *For arbitrary non-negative weight vectors* $(w_1, ..., w_L)$ *and* $(\tilde{w}_1, ..., \tilde{w}_L)$ *that satisfy* $\sum_{\ell=1}^{L} w_\ell = \sum_{\ell=1}^{L} \tilde{w}_\ell = 1$*, if*

$$\sum_{\ell=1}^{L} w_\ell P[Y_0 \leq \cdot | D = 0, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_0 \leq \cdot | D = 0, Z = z_\ell, X = x], \tag{2.10}$$

then

$$\sum_{\ell=1}^{L} w_\ell P[Y_1 \leq \cdot | D = 0, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_1 \leq \cdot | D = 0, Z = z_\ell, X = x]. \tag{2.11}$$

Theorem 2.2. *Suppose that Assumption Z and Condition 2.3 hold. Fix* $x \in \mathcal{X}$*. For* $\gamma(x) \equiv (\gamma_1(x), ..., \gamma_L(x))$ *and* $\tilde{\gamma}(x) \equiv (\tilde{\gamma}_1(x), ..., \tilde{\gamma}_L(x))$ *in* $\Gamma_p(x)$*, suppose*

$$P[Y \leq y | D = 0, X = x] \leq \sum_{\ell=1}^{L} \gamma_\ell(x) P[Y \leq y, D = 0 | Z = z_\ell, X = x], \quad \forall y \in \mathcal{Y} \tag{2.12}$$

$$\sum_{\ell=1}^{L} \tilde{\gamma}_\ell(x) P[Y \leq y, D = 0 | Z = z_\ell, X = x] \leq P[Y \leq y | D = 0, X = x], \quad \forall y \in \mathcal{Y} \tag{2.13}$$

Then $F_{Y_1|D,X}(\cdot|0, x)$ *is bounded by*

$$-\sum_{\ell=1}^{L} \tilde{\gamma}_\ell(x) P[Y \leq \cdot, D = 1 | Z = z_\ell, X = x] \tag{2.14}$$

$$\leq P[Y_1 \leq \cdot | D = 0, X = x]$$

$$\leq -\sum_{\ell=1}^{L} \gamma_\ell(x) P[Y \leq \cdot, D = 1 | Z = z_\ell, X = x] \tag{2.15}$$

The proof of this theorem is analogous to that of Theorem 2.1. Also the derivation of the bounds on $QTE(0, x)$ is symmetric to the case of $QTE(1, x)$ so is omitted. Notably, which

treatment parameter we can obtain bounds for is determined by which identifying condition we impose. In Section 4, we investigate this aspect within a structural model. Finally, the next version of the condition can be used to bound the ATT.

Condition 2.4. *For arbitrary non-negative weight vectors* $(w_1, ..., w_L)$ *and* $(\tilde{w}_1, ..., \tilde{w}_L)$ *that satisfy* $\sum_{\ell=1}^{L} w_\ell = \sum_{\ell=1}^{L} \tilde{w}_\ell = 1$, *if*

$$\sum_{\ell=1}^{L} w_\ell P[Y_1 \leq \cdot | D = 1, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_1 \leq \cdot | D = 1, Z = z_\ell, X = x], \tag{2.16}$$

then

$$\sum_{\ell=1}^{L} w_\ell E[Y_0 | D = 1, Z = z_\ell, X = x] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell E[Y_0 | D = 1, Z = z_\ell, X = x]. \tag{2.17}$$

Condition 2.4 is weaker than Condition 2.1, but would be enough to bound the $ATE(1, x)$. An analogous condition can be imposed to bound $ATE(0, x)$.

3 Understanding Key Conditions

Theorem 2.1 relies on the existence of a sequence $\gamma = (\gamma_1, ..., \gamma_L)$ satisfying $\sum_{\ell=1}^{L} \gamma_\ell = 0$, $\sum_{\ell=1}^{L} \gamma_\ell p(z_\ell, x) = 1$, and the inequality (2.6), that is, $P[Y \leq y | D = 1] \leq \sum_{\ell=1}^{L} \gamma_\ell P[Y \leq y, D = 1 | Z = z_\ell]$ for all y. In this section, we suppress $X = x$ to simplify our discussions. To gain the interpretation of (2.6), we consider simple cases of binary and trinary IVs. Note that (2.6) is a special case of (2.1) in Condition 2.1. Under $Z \perp (Y_d, D_z)$, the latter inequality can be rewritten as

$$\sum_{\ell=1}^{L} w_\ell P[Y \leq y | D_{z_\ell} = 1] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y \leq y | D_{z_\ell} = 1]. \tag{3.1}$$

Let $p(z) \equiv (D = 1 | Z = z)$. Assume weak separability $D = h(Z, \eta) = 1\{\eta \leq p(Z)\}$ with $\eta \sim U[0, 1]$ for the sake of discussion. Then, $\{D_{z_\ell} = 1\}$ are a set of individuals who are

compliers (C) and always-takers (AT). This observation is detailed in the two cases of IV.

3.1 The Case of $L = 2$

Lemma 3.1. *Suppose* $L = 2$ *with* $\mathcal{Z} = \{0, 1\}$ *and* $D = 1\{\eta \leq p(Z)\}$. *Also suppose* $Z \perp (Y_d, D_z)$ *and* $0 < p(z) < 1$ *for* $z = 0, 1$. *Then,* (2.6) *uniquely holds with* $(\gamma_1, \gamma_2) = \left(\frac{1}{p(1)-p(0)}, -\frac{1}{p(1)-p(0)}\right)$, *which is equivalent to* (1.3) *and is in turn equivalent to* $Y_1|AT \prec_{FOSD} Y_1|C$.

Remark 3.1. *The case of* $L = 2$ *is illustrative to understand the role of IV support. In general, when the inequality* (2.6) *is changing with* y, *then we want to ensure the resulting segments in the space of* γ *have a nonempty intersection. The inequality is more likely to hold when* Z *has sufficient support and trivially hold with infinite support. To see this when* $L = 2$, *note that with* $z_1 = +\infty$ *and* $z_2 = -\infty$, (2.6) *trivially holds. A similar observation applies to the case of* $L > 2$.

3.2 The Case of $L = 3$

In this case, it is more convenient to focus on (2.1) or (3.1) rather than its special case, (2.6). Suppose for $z, z' \in \mathcal{Z} = \{0, 1, 2\}$ such that $z < z'$, we have $p(z) < p(z')$ in Assumption 2. Borrowing the language from Mogstad et al. (2021), let $\{i : D_{0,i} = D_{1,i} = D_{2,i} = 1\}$ be always-takers (AT), $\{i : D_{0,i} = 0, D_{1,i} = D_{2,i} = 1\}$ be eager compliers (E-C), $\{i : D_{0,i} = D_{1,i} = 0, D_{2,i} = 1\}$ be reluctant compliers (R-C), and $\{i : D_{0,i} = D_{1,i} = D_{2,i} = 0\}$ be never-takers NT). Then, for example

$$\begin{aligned}\{D_1 = 1\} &= \{D_0 = 1, D_1 = 1, D_2 = 1\} \\ &\cup \{D_0 = 0, D_1 = 1, D_2 = 1\} \\ &= \{AT\} \cup \{E\text{-}C\}\end{aligned}$$

because $\{D_0 = 1, D_1 = 1, D_2 = 0\} = \emptyset$ and $\{D_0 = 0, D_1 = 1, D_2 = 0\} = \emptyset$. Also,

$$\begin{aligned}\{D_2 = 1\} &= \{D_0 = 1, D_1 = 1, D_2 = 1\} \\ &\cup \{D_0 = 0, D_1 = 1, D_2 = 1\} \\ &\cup \{D_0 = 0, D_1 = 0, D_2 = 1\} \\ &= \{AT\} \cup \{E\text{-}C\} \cup \{R\text{-}C\}.\end{aligned}$$

Therefore, (3.1) establishes the FOSD relationship between the mixtures of observed distributions of Y conditional on various always-takers and compliers groups.

Remark 3.2 (Conditions w.r.t. Compliance Types)**.** *Motivated from the discussion of this section, we can rewrite Condition 2.2 (and all the relevant conditions) without even invoking Assumption SEL. This modification will provide an interpretation of the condition that solely relies on compliance types. Let $T \equiv \{D(z_1), ..., D(z_L)\}$ be a random vector that indicates a particular compliance type with its realized value in $\{0,1\}^L \equiv \mathcal{T}$. For example, when $L = 2$ (i.e., binary IV), $T = (D(0), D(1)) \in \{(0,0), (1,0), (0,1), (1,1)\} \equiv \mathcal{T}$. Since D and Z are discrete, T is naturally a discrete random vector. Note that this framework do not rely on any selection models, and therefore T captures* all *possible compliance types given D and Z. Then Condition 2 can be modified as follows:*

Fix $x \in \mathcal{X}$. For arbitrary weight functions $w : \mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ and $\tilde{w} : \mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ such that $\sum_{t\in\mathcal{T}} w(t,x) = \sum_{t\in\mathcal{T}} \tilde{w}(t,x) = 1$, if

$$\sum_{t\in\mathcal{T}} w(t,x) F_{Y_1|T,X}(\cdot|t,x) \leq \sum_{t\in\mathcal{T}} \tilde{w}(t,x) F_{Y_1|T,X}(\cdot|t,x),$$

then

$$\sum_{t\in\mathcal{T}} w(t,x) F_{Y_0|T,X}(\cdot|t,x) \leq \sum_{t\in\mathcal{T}} \tilde{w}(t,x) F_{Y_0|T,X}(\cdot|t,x).$$

Then, the weighted sum in each inequality can be interpreted as the distribution of Y_d

weighted across all compliance types.

4 Structural Models and Policymaker's Problems

We show that Conditions 2.1 and 2.3 can be justified in a nonparametric structural model for the counterfactual outcomes. By relating the conditions with the structural model, we provide additional intuitions for the conditions. For arbitrary r.v.'s A and $\tilde{A}$, let $A \overset{d}{=} \tilde{A}$ denote $F_A = F_{\tilde{A}}$.

Model 1. (i) We have

$$Y_d = q(d, X, U_d) \qquad \text{for } d \in \{0, 1\}, \tag{4.1}$$

where (ii) $q(d, x, \cdot)$ is continuous and monotone increasing, (iii) $D = h(Z, X, \eta)$ (i.e., SEL), (iv) conditional on (η, X, Z), $U_d \overset{d}{=} U + \xi_d$ where $\xi_d \perp (\eta, U)$, (v) conditional on (X, Z), ξ_0 is (weakly) more or less noisy than ξ_1, that is, $\xi_0 \overset{d}{=} \xi_1 + V$ for some V independent of ξ_1.

Note that U is the source of endogeneity in that it allowed to be dependent on η. Model 1(iv)–(v) implies that $U_0 \overset{d}{=} U_1 + V$ conditional on (η, X). Importantly, Model 1 nests the model in Chernozhukov and Hansen (2005) as a special case. This can be shown as follows. First, Chernozhukov and Hansen (2005) assume Model 1(i)–(iii). Additionally, they assume, conditional on (X, Z), either rank similarity ($F_{U_0|\eta} = F_{U_1|\eta}$) or rank invariance ($U_0 = U_1$).[4] Then, by taking $\xi_d = 0$ for all d in Model 1(iv), we have $U_0 \overset{d}{=} U_1 \overset{d}{=} U$ conditional on (η, X, Z), which proves the claim.

Model 1(v) assumes that the unobservable under the counterfactual status of being treated are more (or less) dispersed than that under the counterfactually untreated status. Although this may seem stringent, it is substantially weaker than rank similarity (or invariance) and can be plausible in various scenarios. Before providing examples of these scenarios, we first establish the connection between Model 1 and Condition 2.2 (and thus Condition 2.1 by

[4]Note that rank similarity and rank invariance are observationally equivalent under Model 1(i)–(iii) in that they produce the same distribution of observables (Chernozhukov and Hansen (2013)).

Lemma 2.1).

Theorem 4.1. *Under Assumptions Z, Model 1 (with ξ_0 being weakly more noisy than ξ_1) implies Condition 2.2.*

Analogous to Theorem 4.1, one can readily show that Model 1 with ξ_0 being weakly less noisy than ξ_1 implies Condition 2.3.

Now we provide examples that are consistent with Model 1.

Example 1 (Auction)**.** *Consider online and offline auctions. Let Y be the bid (which subsequently forms revenue) and D be participating in an auction with different format ($D = 1$ if online and $= 0$ if offline). Let $U_d \stackrel{d}{=} U + \xi_d$ be the valuation of the item where U is the common valuation (correlated with D) and ξ_d is format specific random shocks satisfying $\xi_d \perp (\eta, U)$. We assume that bidders have limited information on certain features of the auction that affect valuation (e.g., they know the distribution of ξ_d but not its realization). In this example, what would justify $var(\xi_0) > var(\xi_1)$? It may be the case that, in the offline auction, bidders are more emotionally affected by other bidders, which makes their bids more variable.*

Example 2 (Insurance)**.** *We are interested in the effect of insurance on health outcomes. Let Y be the health outcome and D be the decision of getting insurance ($D = 1$ being insured). Let $U_d \stackrel{d}{=} U + \xi_d$ be underlying health conditions where U captures health conditions known to participant (and thus correlated with D) while ξ_d is health conditions not fully known a priori and thus random. In this example, $var(\xi_0) > var(\xi_1)$ may hold because insurance by definition ensures a certain level of health conditions.*

Example 3 (Vaccination)**.** *Similar to Example 2, suppose that D is instead getting vaccination (of an established vaccine). Again, $U_d \stackrel{d}{=} U + \xi_d$ is health conditions where U captures conditions known to participant (and correlated with D) and ξ_d is vaccination-status-specific health conditions, which are not fully known a priori. Then, similarly as before $var(\xi_0) > var(\xi_1)$ may hold because, when not vaccinated, one is exposed to the risk of a serious illness, while vaccination ensures a certain level of immunity.*

The scenarios in Examples 1–3 justify Condition 2.1 via Theorem 4.1. Then, under Condition 2.1, Theorem 2.1 and Corollary 2.1 yield bounds on $QTE_\tau(1,x)$, the effects of treatment for those who take the treatment. The final example illustrates the converse case.

Example 4 (Medical Trial)**.** *In contrast to Example 3, suppose the treatment itself is risky. That is, let D be participating in a frontier medical trial ($D=1$ being participation). In this case, $var(\xi_0) < var(\xi_1)$ is more plausible because, with a newly developed medicine, there is the high risk of unknown side effects.*

The scenario in Example 4 justifies Condition 2.3, under which bounds on $QTE_\tau(0,x)$, the effects of treatment for those who abstain from it, can be obtained.

Model 1 and these examples show how a certain treatment parameter may be more relevant for policy than others depending on the plausibility of assumptions. Consider the problem of a policymaker. Assume that the policymaker concerns risk-averse individuals, *which are typically the majority.* For this policymaker, a candidate policy would aim at providing "insurance," which can be either literally insurance or policies that serve as insurance (e.g., vaccination, subsidies). Therefore, she would be interested in understanding the treatment effects for the target individuals that are risk-averse. Our procedure provides a statistical tool for such a policymaker. That is, under Model 1, our procedure has the ability to bound the treatment effects for individuals with $D=d$ such that $var(\xi_d) < var(\xi_{1-d})$. This is a unique feature of our setting: the plausibility of assumptions dictates the parameters of interest, which then can be terms as *assumption-driven* treatment parameters.

A remaining question one might have is as follows. How much Condition 2.1 has to be strengthened to be equivalent to rank similarity? To answer this question, recall that Condition 2.2 is stronger than Condition 2.1 (by Lemma 2.1). We strengthen Condition 2.2 further by making it an "if and only if" condition:

Condition 4.1. *Fix $x \in \mathcal{X}$. For arbitrary weight functions $w : \mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ and $\tilde{w}$:*

$\mathcal{T} \times \mathcal{X} \to \mathbb{R}_+$ *such that* $\int w(t,x)dt = \int \tilde{w}(t,x) = 1$*, it holds that*

$$\int w(t,x) F_{Y_1|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x) F_{Y_1|\eta,X}(\cdot|t,x)dt$$

if and only if

$$\int w(t,x) F_{Y_0|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x) F_{Y_0|\eta,X}(\cdot|t,x)dt.$$

It turns out that we can establish the following result.

Theorem 4.2. *Model 1(i)–(iii) with* $F_{U_0|\eta,X,Z} = F_{U_1|\eta,X,Z}$ *(i.e., rank similarity) implies Condition 4.1.*

This theorem highlights the stringency of rank similarity relative to Condition 2.1. The proof is trivial so omitted. It is worth noting that the converse of Theorem 4.2 is *not* true.

Remark 4.1 (Rank Linearity)**.** *Here is a counter-example for the converse statement of Theorem 4.2. Assume Model 1(i)–(iii) and*

$$F_{Y_0|\eta,X,Z}(\cdot|t,x,z) = \lambda(\cdot) F_{Y_1|\eta,X,Z}(\psi(\cdot,x)|t,x,z) \tag{4.2}$$

for every $t \in \mathbb{R}^{d_\eta}$ *and* $x \in \mathcal{X}$*, where* $\psi(\cdot,x) : \mathcal{Y} \to \mathcal{Y}$*, a one-to-one and onto mapping, is strictly increasing, and* $\lambda : \mathcal{Y} \to \mathbb{R}$ *is consistent with* $F_{Y_d|\eta,X,Z}$ *being a proper CDF. This* rank linearity *implies Condition 4.1, which is trivial to show. However, rank linearity is weaker than rank similarity as the latter is a special case of the former. To see this, conditional on* $Z = z$*, (4.2) with Model 1(i)–(ii) yields* $F_{U_0|\eta,X}(q^{-1}(0,x,y)|t,x) = \lambda(y) F_{U_1|\eta,X}(q^{-1}(1,x,\psi(y))|t,x)$*. Then, by choosing* $\lambda(y) = 1$ *and* $\psi(y,x) = q(1,x,q^{-1}(0,x,y))$*, we have* $F_{U_0|\eta,X}(\cdot|t,x) = F_{U_1|\eta,X}(\cdot|t,x)$*.*

Interestingly, rank linearity appears to be equivalent to Condition 4.1. We prove this equivalence in the discrete case. We conjecture it would hold in the continuous case as well. For $d \in \{0,1\}$, suppose Y_d and η are discrete random variables, supported on $\{y_{d,1}, \cdots, y_{d,k_d}\}$ and $\{t_1, \cdots, t_{k_\eta}\}$, respectively. Suppress (X,Z) for simplicity.

Condition 4.2. *For arbitrary non-negative weights* $\{w_1, \cdots, w_{k_\eta}\}$ *and* $\{\tilde{w}_1, \cdots, \tilde{w}_{k_\eta}\}$ *such that* $\sum_{j=1}^{k_\eta} w_j = 1$ *and* $\sum_{j=1}^{k_\eta} \tilde{w}_j = 1$*, it holds that*

$$\sum_{j=1}^{k_\eta} w_j F_{Y_1|\eta}(\cdot|t_j) \leq \sum_{j=1}^{k_\eta} \tilde{w}_j F_{Y_1|\eta}(\cdot|t_j)$$

if and only if

$$\sum_{j=1}^{k_\eta} w_j F_{Y_0|\eta}(\cdot|t_j) \leq \sum_{j=1}^{k_\eta} \tilde{w}_j F_{Y_0|\eta}(\cdot|t_j).$$

Theorem 4.3. *For any probability distribution function* $\tilde{F}_d$ *supported on* $\{y_{d,1}, \cdots, y_{d,k_d}\}$*, suppose there always exists a non-negative sequence* $\{c_{d,1}, \cdots, c_{d,k_\eta}\}$ *such that*

$$\tilde{F}_d(\cdot) = \sum_{j=1}^{k_\eta} c_{d,j} F_{Y_d|\eta}(\cdot|t_j). \tag{4.3}$$

Then, Condition 4.2 holds if and only if (i) $k_0 = k_1$ *and (ii) for some one-to-one and onto mapping* $\psi : \{y_{0,1}, \cdots, y_{0,k_0}\} \to \{y_{1,1}, \cdots, y_{1,k_1}\}$ *and* $\lambda(\cdot) > 0$*,*

$$F_{Y_0|\eta}(\cdot|t_j) = \lambda(\cdot) F_{Y_1|\eta}(\psi(\cdot)|t_j), \qquad \text{for } j = 1, \cdots, k_\eta.$$

Note that even with $k_0 = k_1$, we allow that Y_0 and Y_1 have different supports (i.e., there can be a "drift"). The rank condition 4.3 would be violated when there is no endogeneity (i.e., $Y_d \perp \eta$), which is not our focus. A necessary condition is $k_\eta \geq k_d$, namely, the support of η is no coarser than the support of Y_d. Obviously, the rank condition is only introduced in this theorem to establish the relationship between rank linearity (and hence rank similarity) and the range of identifying conditions of this paper, and it is not necessary for our procedure.

Proof. The "if" part is straightforward. We show the "only if" part. Suppose Condition 4.2 holds. For $d = 0, 1$, let Δ_d be a collection of all sequences $\delta \equiv \{\delta_1, \cdots, \delta_{k_\eta}\}$ which satisfies (i) $\sum_{j=1}^{k_\eta} \delta_j \times F_{Y_d|\eta}(\cdot|t_j) \leq 0$; and (ii) $\sum_{j=1}^{k_\eta} \delta_j = 0$. By definition, Δ_d is a linear cone, i.e., if

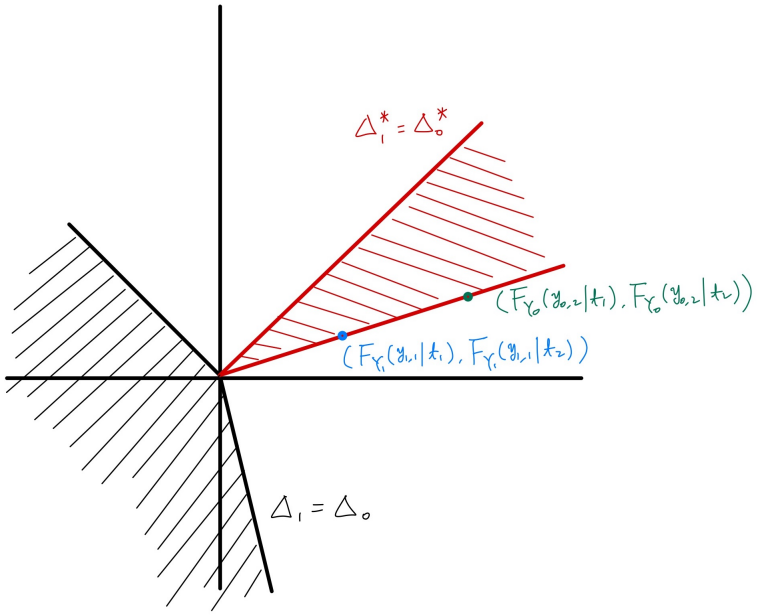

Figure 1: Illustration of the Proof of Theorem 4.3

$\delta \in \Delta_d$, then $\lambda \times \delta \in \Delta_d$ for all strictly positive scaler λ. Consider the polar cone of Δ_d, i.e.,

$$\Delta_d^* \equiv \{F_d \in \mathbb{R}^{k_\eta} | F_d' \delta \leq 0, \ \forall \delta \in \Delta_d\}.$$

By definition, Δ_d^* is a closed convex cone and its extreme ray is generated by the following k_d vectors:

$$\left\{ \left(F_{Y_d|\eta}(y|t_1), \cdots, F_{Y_d|\eta}(y|t_{k_\eta}) \right)' : y = y_{d,1}, \cdots, y_{d,k_d} \right\}.$$

Note that any vector $\left(F_{Y_d|\eta}(y|t_1), \cdots, F_{Y_d|\eta}(y|t_{k_\eta}) \right)'$ in the above set cannot be written as a linear combination of the others because the rank of matrix $\{F_{Y_d|\eta}(y_{d,j}|t_{j'}) : j = 1, ..., k_d, \quad j' = 1, \cdots, k_\eta\}$ is no smaller than k_d by the condition (4.3).

Because Condition 4.2 implies $\Delta_0 = \Delta_1$, we have $\Delta_0^* = \Delta_1^*$ and thus $k_0 = k_1$. Moreover, since every polyhedral cone has a unique representation as a conical hull of its extreme generators, any extreme ray in Δ_0^* must be identical some extreme ray in Δ_1^*. In terms of

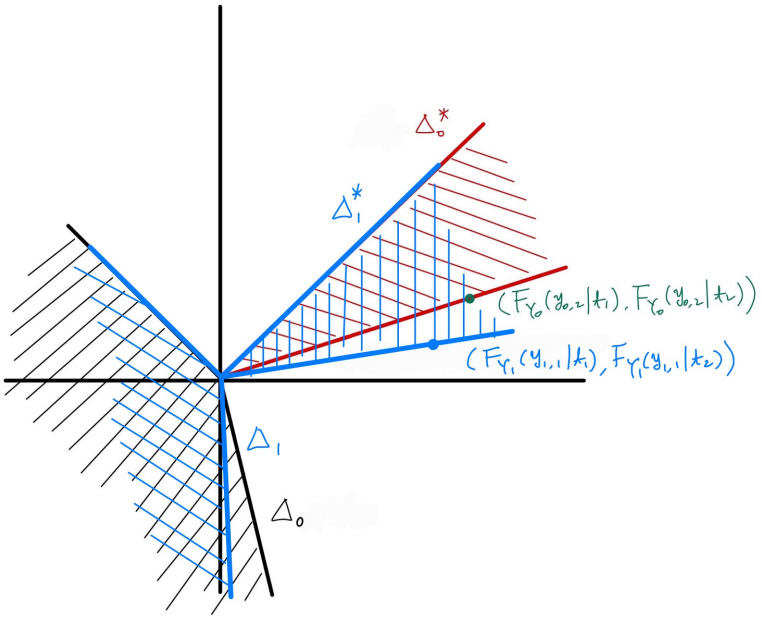

Figure 2: Illustration of the Proof of Theorem 4.3

the vectors that generate extreme rays, this means that, for any $y \in \{y_{0,1}, \cdots, y_{0,k_0}\}$,

$$\left(F_{Y_0|\eta}(y|t_1), \cdots, F_{Y_0|\eta}(y|t_{k_\eta})\right) = \lambda \times \left(F_{Y_1|\eta}(y'|t_1), \cdots, F_{Y_1|\eta}(y'|t_{k_\eta})\right)$$

for some scalar $\lambda > 0$ (that conforms with the definition of CDF) and $y' \in \{y_{1,1}, \cdots, y_{1,k_1}\}$. This last part of the proof is illustrated in Figure 1 when $k_d = k_\eta = 2$. □

Condition 4.1 is crucial in bounding $QTE_\tau(x) = Q_{Y_1|X}(\tau|x) - Q_{Y_0|X}(\tau|x)$ unconditional with respect to $D = d$. The "only if" part (i.e., Condition 2.1) will bound $Q_{Y_0|D=1}(\tau)$ and thus $Q_{Y_0}(\tau)$ by Theorem 2.1, while the "if" part (i.e., Condition 2.3) will bound $Q_{Y_1|D=0}(\tau)$ and thus $Q_{Y_1}(\tau)$ by the symmetric version of Theorem 2.1. The fact that Condition 4.1 is weaker than rank similarity illustrates the importance of rank similarity in the identification of the QTE and ATE.

Remark 4.2. *Analogous to the analysis in Theorem 4.3, Condition 2.2 is equivalent to* $\Delta_1 \subseteq \Delta_0$, *which implies* $\Delta_1^* \supseteq \Delta_0^*$. *Interestingly, the sufficiency of Model 1 for Condition 2.2 can be shown using the polar cones. Suppose* $k_1 = k_0 = k$. *Under Model 1,*

$var(Y_0|\eta) > var(Y_1|\eta)$. Therefore, $F_{Y_1|\eta}(\cdot)$ tend to be closer to 0 or 1 compared to $F_{Y_0|\eta}(\cdot)$, in which case $\left(F_{Y_1|\eta}(y|t_1), \cdots, F_{Y_1|\eta}(y|t_{k_\eta})\right)'$ would be closer than $\left(F_{Y_0|\eta}(y|t_1), \cdots, F_{Y_0|\eta}(y|t_{k_\eta})\right)'$ to e_j (for some $j = 1, ..., k_\eta$), a vector where the j-th element is one and the other elements are zero. Then, the resulting polar cone Δ_1^* generated by the extreme rays $\lambda \times \left(F_{Y_1|\eta}(y|t_1), \cdots, F_{Y_1|\eta}(y|t_{k_\eta})\right)'$ (for $\lambda > 0, y = y_{d,1}, \cdots, y_{d,k}$) would contain the polar cone Δ_0^* generated by $\lambda \times \left(F_{Y_0|\eta}(y|t_1), \cdots, F_{Y_0|\eta}(y|t_{k_\eta})\right)'$ (for $\lambda > 0, y = y_{d,1}, \cdots, y_{d,k}$). This implies that $\Delta_1 \subseteq \Delta_0$, which is an alternative statement of Condition 2.2. For example, consider the simple case of binary $Y_d \in \{0, 1\}$ and $\eta \in \{t_1, t_2\}$ as in Figure 1. Since $var(Y_d|\eta) = P[Y_d = 0|\eta](1 - P[Y_d = 0|\eta])$, smaller variance implies $P[Y_d = 0|\eta = t] = F_{Y_d|\eta}(0|t)$ being closer to 1 or 0. Therefore, the resulting cone generated by the extreme rays $\lambda \times (F_{Y_1|\eta}(0|t_1), F_{Y_1|\eta}(0|t_2))$ and $\tilde{\lambda} \times (F_{Y_1|\eta}(1|t_1), F_{Y_1|\eta}(1|t_2)) = \tilde{\lambda} \times (1, 1)$ (for any $\lambda, \tilde{\lambda} > 0$) would contain the cone generated by $\lambda \times (F_{Y_0|\eta}(0|t_1), F_{Y_0|\eta}(0|t_2))$ and $\tilde{\lambda} \times (1, 1)$ (for any $\lambda, \tilde{\lambda} > 0$). This is illustrated in Figure 2.

Remark 4.3 (Conditions w.r.t. Compliance Types, continued)**.** *In the framework proposed in Remark 3.2, if we additionally impose Assumption SEL, we have $\sigma(T) \subset \sigma(\eta)$ where $\sigma(A)$ is a σ-field generated by a random vector A. Therefore, given Model 1(i)–(iii), $F_{U_1|T} = F_{U_0|T}$ is weaker than $F_{U_1|\eta} = F_{U_0|\eta}$. This way by using T, we can define a weaker version of rank similarity. Moreover, Condition 4.2 can be motivated by this framework as the discrete η can be viewed as T (with $k_\eta = 2^L$).*

5 Systematic Calculation of Bounds

In Theorem 2.1, there can be many γ's that satisfy the condition (2.6) (and $\tilde{\gamma}$ for (2.7)), especially with $L \geq 3$. This motivates the use of optimization in calculating the bounds via Corollary 2.1. We only focus on the upper bound and suppress X henceforth for brevity.

5.1 Semi-Infinite Programming

To simplify notation, let $\boldsymbol{p}(y,d) \equiv (p(y,d|z_1),...,p(y,d|z_L))'$ where $p(y,d|z_\ell) \equiv P[Y \leq y, D = d|Z = z_\ell]$ and $p(y|d) \equiv P[Y \leq y|D = d]$. Also, let $\mathbf{1} \equiv (1,...,1)'$ and $\boldsymbol{p} \equiv (p(z_1),...,p(z_L))'$ with $p(z) \equiv P[D = 1|Z = z]$ so that

$$\Gamma_p = \{\gamma : \gamma'[\ \mathbf{1} \quad \boldsymbol{p}\ ] = [\ 0 \quad 1\ ]\} \subset \mathbb{R}^L.$$

Consider the following linear semi-infinite programming for the upper bound on $P[Y_0 \leq \bar{y}|D = 1]$:

$$UB(\bar{y}) = \min_{\gamma \in \Gamma_p} -\boldsymbol{p}(\bar{y},0)'\gamma \tag{5.1}$$

$$s.t. \quad \boldsymbol{p}(y,1)'\gamma \geq p(y|1), \quad \forall y \in \mathcal{Y} \tag{5.2}$$

Note that the condition (2.6) guarantees the feasible set is non-empty. Also note that this condition is allowed to satisfy only almost everywhere (a.e.), which we suppress for simplicity. This program is infeasible to solve in practice as there are infinitely many constraints. We propose two approaches to approximate it with a linear program (LP).

5.2 Linear Program with Randomized Constraints

One approach to the semi-infinite program (5.1)–(5.2) is to approximate (5.2) using an i.i.d. simulated sample $\{Y_i\}_{i=1}^{s_n}$ as is done in the literature (e.g., Calafiore and Campi (2005)). This approach is also reminiscent of Han and Yang (2020). An obvious candidate of this sample would be $\{Y_i\}_{i=1}^{n}$ with $s_n = n$. Consider a sampled LP of the following:

$$\overline{UB}_n(\bar{y}) = \min_{\gamma \in \Gamma_p} -\boldsymbol{p}(\bar{y},0)'\gamma \tag{5.3}$$

$$s.t. \quad \boldsymbol{p}(Y_i,1)'\gamma \geq p(Y_i|1). \quad \forall i = 1,...,n \tag{5.4}$$

In the appendix, we show that the probability of violating the original constraints (5.2) by using (5.4) can be bounded by $O(1/n)$.

5.3 Dual Program and Sieve Approximation

Another approach to the semi-infinite program (5.1)–(5.2) is to invoke its dual and approximate the Lagrangian measure using sieve. With the constraint $p(y|1) - \boldsymbol{p}(y,1)'\gamma \leq 0$, the Lagrangian for (5.1)–(5.2) is

$$\begin{aligned}\mathcal{L}(\gamma, \Lambda, \lambda) &= -\boldsymbol{p}(\bar{y}, 0)'\gamma + \int_{\mathcal{Y}} [p(y|1) - \boldsymbol{p}(y,1)'\gamma] d\Lambda(y) + \lambda'([\ \mathbf{1} \quad \boldsymbol{p}\ ]'\gamma - [\ 0 \quad 1\ ]') \\ &= \int_{\mathcal{Y}} p(y|1) d\Lambda(y) - [\ 0 \quad 1\ ]\lambda + (\lambda'[\ \mathbf{1} \quad \boldsymbol{p}\ ]' - \int_{\mathcal{Y}} \boldsymbol{p}(y,1)' d\Lambda(y) - \boldsymbol{p}(\bar{y}, 0)')\gamma\end{aligned}$$

and

$$UB(\bar{y}) = \min_{\gamma \in \mathbb{R}^L} \sup_{\Lambda \succeq 0, \lambda \in \mathbb{R}^2} \mathcal{L}(\gamma, \Lambda, \lambda),$$

where Λ is a non-negative (not necessarily probability) measure (i.e., $\Lambda \succeq 0$) that assigns weights to binding constraints. Therefore, the dual problem to (5.1)–(5.2) is

$$\widetilde{UB}(\bar{y}) = \sup_{\Lambda \succeq 0, \lambda \in \mathbb{R}^2} \int_{\mathcal{Y}} p(y|1) d\Lambda(y) - [\ 0 \quad 1\ ]\lambda \tag{5.5}$$

$$s.t. \qquad [\ \mathbf{1} \quad \boldsymbol{p}\ ]\lambda - \int_{\mathcal{Y}} \boldsymbol{p}(y,1) d\Lambda(y) - \boldsymbol{p}(\bar{y}, 0) = \mathbf{0}, \tag{5.6}$$

which now has a finite number of constraints (i.e., L constraints). It is trivial to show weak duality, $\widetilde{UB}(\bar{y}) \leq UB(\bar{y})$.[5] It is conjectured that strong duality may also hold because of

[5] This is because, by (5.1)–(5.2) and (5.5)–(5.6), we have

$$\begin{aligned}-\boldsymbol{p}(\bar{y}, 0)'\gamma &= \left\{ \int_{\mathcal{Y}} \boldsymbol{p}(y,1) d\Lambda(y) - [\ \mathbf{1} \quad \boldsymbol{p}\ ]\lambda \right\}' \gamma = \int_{\mathcal{Y}} \boldsymbol{p}(y,1)'\gamma d\Lambda(y) - \lambda'[\ \mathbf{1} \quad \boldsymbol{p}\ ]'\gamma \\ &\geq \int_{\mathcal{Y}} p(y|1) d\Lambda(y) - \lambda'[\ 0 \quad 1\ ]'.\end{aligned}$$

the structure of the problem (e.g., linearity, continuity of $\boldsymbol{p}(\cdot, d)$ and $p(\cdot|d)$, and etc.), which is yet to be explored. Note that $\Lambda(y)$ is smooth as the feasible set of the primal problem is smooth due to the smoothness of $p(y|d)$ and $\boldsymbol{p}(y, d)$, which are CDFs. This motivates us to use sieve approximation for $\Lambda(y)$ to turn the dual into a linear programming problem. The smoothness class for $\Lambda(y)$ will be determined by the smoothness class of CDFs. Let $\mathcal{Y}$ is normalized to be $[0, 1]$ and $\lambda(y) \equiv d\Lambda(y)/dy$. Consider the following sieve approximation:

$$\lambda(y) \approx \sum_{j=1}^{J} \theta_j b_j(y),$$

where $b_j(y) \equiv b_{j,J}(y)$ is a Bernstein basis function. Then, the LP can be written as

$$\begin{aligned} \widetilde{UB}_J(\bar{y}) = \max_{\theta \in \mathbb{R}^J_+, \lambda \in \mathbb{R}^2} & \sum_{j=1}^{J} \theta_j b_j^1 - [\ 0 \quad 1\ ]\lambda \\ s.t. \qquad & [\ \mathbf{1} \quad \boldsymbol{p}\ ]\lambda - \sum_{j=1}^{J} \theta_j \boldsymbol{b}_{1,j} - \boldsymbol{p}(\bar{y}, 0) = \mathbf{0}, \end{aligned}$$

or equivalently,

$$\widetilde{UB}_J(\bar{y}) = \max_{\theta \in \mathbb{R}^J_+, \lambda \in \mathbb{R}^2} \theta' b^1 - [\ 0 \quad 1\ ]\lambda \tag{5.7}$$

$$s.t. \qquad [\ \mathbf{1} \quad \boldsymbol{p}\ ]\lambda - B_1\theta - \boldsymbol{p}(\bar{y}, 0) = \mathbf{0}, \tag{5.8}$$

where $\theta \equiv (\theta_1, ..., \theta_J)'$, $b^d \equiv (b_1^d, ..., b_J^d)'$ with $b_j^d \equiv \int_{\mathcal{Y}} b_j(y)p(y|d)dy$, $\boldsymbol{b}_{d,j} \equiv (b_{d,j,1}, ..., b_{d,j,L})'$ with $b_{d,j,\ell} \equiv \int_{\mathcal{Y}} b_j(y)p(y, d|z_\ell)dy$, and $B_d \equiv [\ \boldsymbol{b}_{d,1} \ \cdots \ \boldsymbol{b}_{d,J}\ ]$ is an $L \times J$ matrix. Using Bernstein polynomials to approximate infinite-dimensional decision variables is also used in Han and Yang (2020).

Remark 5.1 (Local Approximation)**.** *The LP* (5.7)–(5.8) *may be more stable than the LP* (5.3)–(5.4)*. In terms of dual, the latter approach is equivalent to having* $\sum_{i=1}^{n} p(Y_i|1)\lambda_i$ *as an approximation for* $\int_{\mathcal{Y}} p(y|1)\lambda(y)dy$*. This can be viewed as a crude local approximation that involves a uniform kernel.*

6 Numerical Studies

To illustrate the importance of multiple IVs and the informativeness of resulting bounds, we conduct numerical exercises. We generate the data so that they are consistent with Model 1 and hence satisfy Condition 2.2. The variables (Y, D, Z) are generated in the following fashion:

- $Y_d = q(d, U_d) = 1 - d + (d+1)U_d$ for $\mathcal{Y} = \mathbb{R}$, that is, $Y_1 = 2U_1$ and $Y_0 = 1 + U_0$
- $Y_d = q(d, U_d) = \Phi(1 - d + (d+1)U_d)$ for $\mathcal{Y} = [0, 1]$, that is, $Y_1 = \Phi(2U_1)$ and $Y_0 = \Phi(1 + U_0)$
- $(U, \eta) \sim BVN((0, 0)', \Sigma)$
- $V \sim N(0, \sigma_V^2)$ and $\xi_1 \sim N(0, \sigma_V^2)$
- $\xi_0 = \xi_1 + V$
- $U_d = U + \xi_d$
- $Z \sim Bin(L-1, p)/(L-1) \in [0, 1]$ with $L \in \{2, 3, 4, 5, 6, 7, 8\}$
- $D = 1\{\pi_0 + \pi_1 Z \geq \eta\}$
- $Y = DY_1 + (1 - D)Y_0$

Here, Z is normalized so that the endpoints of the support are invariant regardless of the value of L. This is intended to understand the role of the number of values Z takes while fixing the role of instrument strength. Figures 3–5 presents the bounds on $\Pr[Y_0 \leq y | D = 1]$ while varying L. The bounds are calculated using the approach proposed in Section 5.2. We only report $L \in \{2, 5, 6\}$ for succinctness. In these figures, the black solid line indicates the true value of $\Pr[Y_0 \leq y | D = 1]$ and the red and blue crosses depict the upper and lower bounds. Although the upper bound is a trivial upper bound for the CDF when $L = 2$, it quickly becomes informative as L increases beyond 5. To put this in a context, this

corresponds to the number of instrument values that three binary IVs can easily surpass or a single continuous IV.

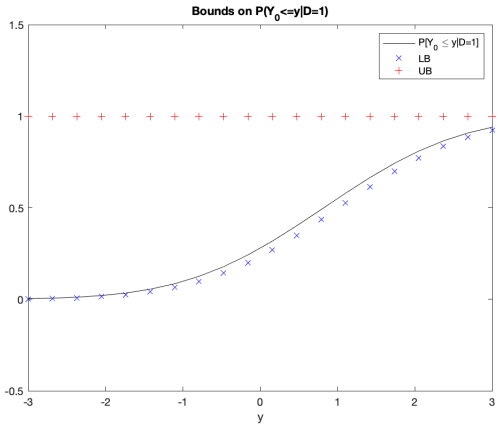

Figure 3: Bounds on $\Pr[Y_0 \leq y|D = 1]$ When $L = 2$

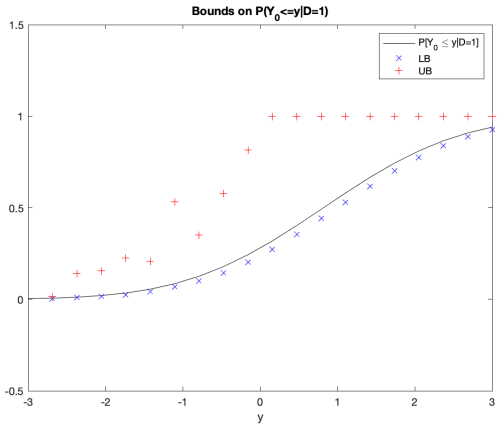

Figure 4: Bounds on $\Pr[Y_0 \leq y|D = 1]$ When $L = 5$

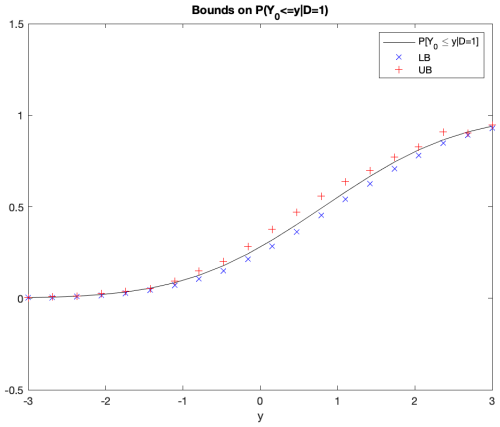

Figure 5: Bounds on $\Pr[Y_0 \leq y|D = 1]$ When $L = 6$

A Proofs

A.1 Proof of Lemma 2.1

Let $p(z,x) \equiv P[D=1|Z=z,X=x]$ and let $H(z,x) \equiv \{\eta : h(z,x,\eta)=1\}$ be a level set. Then,

$$\begin{aligned}\sum_\ell w_\ell P[Y_1 \le y|D=1,Z=z_\ell,X=x] &= \sum_\ell w_\ell P[Y_1 \le y|\eta \in H(z_\ell,x),X=x] \\ &= \int \frac{\sum_\ell w_\ell 1[t \in H(z_\ell,x)]}{p(z_\ell,x)} P[Y_1 \le y|\eta=t,X=x]dt.\end{aligned}$$

Take $w(t,x) = \frac{\sum_\ell w_\ell 1[t\in H(z_\ell,x)]}{p(z_\ell,x)}$. Then, $w(t,x)$ satisfies

$$\int \frac{\sum_\ell w_\ell 1[t \in H(z_\ell,x)]}{p(z_\ell,x)} dt = 1.$$

The same argument applies to $\tilde{w}$ and $\tilde{w}(t,x)$, and also for the distribution of Y_0. □

A.2 Proof of Lemma 3.1

Let $p(1) > p(0)$ without loss of generality. Because

$$\begin{aligned}&P[Y \le y|D=1] \\ &\quad = P[Y \le y|D=1,Z=1]P[Z=1|D=1] + P[Y \le y|D=1,Z=0]P[Z=0|D=1] \\ &\qquad = P[Y \le y|D=1,Z=1]\frac{p(1)P[Z=1]}{P[D=1]} + P[Y \le y|D=1,Z=0]\frac{p(0)P[Z=0]}{P[D=1]},\end{aligned}$$

the inequality (2.6) with $(\gamma_1, \gamma_2) = \left(\frac{1}{p(1)-p(0)}, -\frac{1}{p(1)-p(0)}\right)$ is equivalent to

$$\begin{aligned} P[Y \le y|D=1, Z=1]\frac{[p(1)-p(0)]p(1)P[Z=1]}{P[D=1]} \\ + P[Y \le y|D=1, Z=0]\frac{[p(1)-p(0)]p(0)P[Z=0]}{P[D=1]} \\ \le P[Y \le y|D=1, Z=1]p(1) - P[Y \le y|D=1, Z=0]p(0). \end{aligned}$$

Hence,

$$\begin{aligned} P[Y \le y|D=1, Z=0]p(0) + P[Y \le y|D=1, Z=0]\frac{[p(1)-p(0)]p(0)P[Z=0]}{P[D=1]} \\ \le P[Y \le y|D=1, Z=1]p(1) - P[Y \le y|D=1, Z=1]\frac{[p(1)-p(0)]p(1)P[Z=1]}{P[D=1]}, \end{aligned}$$

that is,

$$\begin{aligned} P[Y \le y|D=1, Z=0]p(0)\left[1 + \frac{[p(1)-p(0)]P[Z=0]}{P[D=1]}\right] \\ \le P[Y \le y|D=1, Z=1]p(1)\left[1 - \frac{[p(1)-p(0)]P[Z=1]}{P[D=1]}\right]. \end{aligned}$$

Note that $P[D=1] = p(1)P[Z=1] + p(0)P[Z=0]$. Therefore,

$$P[Y \le y|D=1, Z=0]p(0)p(1) \le P[Y \le y|D=1, Z=1]p(1)p(0).$$

Note that $p(z) \neq 0$ for $z = 0, 1$, and therefore (γ_1, γ_2) are well-defined. Then we have

$$P[Y \le y|D=1, Z=0] \le P[Y \le y|D=1, Z=1],$$

which proves the first claim of the lemma.

Next, the inequality (1.3) can equivalently be written as

$$P[Y \leq y, D = 1|Z = 1]/p(1) \leq P[Y \leq y, D = 1|Z = 0]/p(0).$$

Then, by the assumed selection equation which excludes defiers, note that $P[Y \leq y, D = 1|Z = z] = P[Y_1 \leq y, D_z = 1]$ by the joint independence assumption and that $P[Y_1 \leq y, D_1 = 1] = P[Y_1 \leq y, AT] + P[Y_1 \leq y, C]$ and $P[Y_1 \leq y, D_0 = 1] = P[Y_1 \leq y, AT]$. Also, note that $p(1) = P(C) + P(AT)$ and $p(0) = P(AT)$ by the selection equation. Then, by a simple algebra, we can show that (1.3) is equivalent to

$$P[Y_1 \leq y|C] \leq P[Y_1 \leq y|AT], \quad \forall y. \tag{A.1}$$

□

A.3 Proof of Theorem 4.1

We suppress X for simplicity. For an arbitrary r.v. A, let $F_A^w(\cdot) \equiv \int w(t)F_{A|\eta}(\cdot|t)dt$, which itself is a CDF. By Model 1(i)–(ii), $F_{Y_d}^w \leq F_{Y_d}^{\tilde{w}}$ if and only if $F_{U_d}^w \leq F_{U_d}^{\tilde{w}}$. So it suffices to show that, if $F_{U_1}^w \leq F_{U_1}^{\tilde{w}}$, then $F_{U_0}^w \leq F_{U_0}^{\tilde{w}}$.

Let $G(\cdot)$ be an arbitrary monotone increasing function and $g(\cdot) \equiv G'(\cdot)$. Note that

$$\begin{aligned}
&\int GdF_{U_0}^w - \int GdF_{U_0}^{\tilde{w}} = \int [\int \tilde{w}(t)F_{U_0|\eta}(u|t)dt - \int w(t)F_{U_0|\eta}(u|t)dt]g(u)du \\
&= \int [\int \tilde{w}(t) \int F_{U|\eta}(u-s|t)f_{\xi_0}(s)dsdt - \int w(t) \int F_{U|\eta}(u-s|t)f_{\xi_0}(s)dsdt]g(u)du \\
&= \int \int \int [\tilde{w}(t) - w(t)]F_{U|\eta}(u|t)f_{\xi_0}(s)g(u+s)dudsdt,
\end{aligned}$$

where the first eq. is due to the integration by part, the second eq. is by $F_{U_d|\eta}(u|t) = \int F_{U|\eta}(u-s|t)f_{\xi_0|\eta}(s|t)ds = \int F_{U|\eta}(u-s|t)f_{\xi_0}(s)ds$ under Model 1(iv), and the last eq. is by change of variables. By Model 1(v), $f_{\xi_0}(s) = \int f_{\xi_1}(s-v)f_V(v)dv = \int f_{\xi_1}(v)f_V(s-v)dv$

where $f_A(\cdot)$ is the PDF of an arbitrary r.v. A. Therefore,

$$\begin{aligned}
&\int G dF^{w}_{U_0} - \int G dF^{\tilde{w}}_{U_0} \\
&= \int\int\int [\tilde{w}(t) - w(t)] F_{U|\eta}(u|t) \int f_{\xi_1}(v) f_V(s-v) g(u+s) dv du ds dt \\
&= \int\int [\tilde{w}(t) - w(t)] F_{U|\eta}(u|t) \int f_{\xi_1}(v) [\int f_V(s) g(u+s+v) ds] dv du dt.
\end{aligned}$$

Let $\psi(s) \equiv \int f_V(t) g(t+s) dt$. By definition, $\psi \geq 0$ since $g \geq 0$. Therefore,

$$\begin{aligned}
&\int G dF^{w}_{U_0} - \int G dF^{\tilde{w}}_{U_0} \\
&= \int\int [\tilde{w}(t) - w(t)] F_{U|\eta}(u|t) \int f_{\xi_1}(v) \psi(u+v) dv du dt \\
&= \int\int [\tilde{w}(t) - w(t)] \int F_{U|\eta}(u-v|t) f_{\xi_1}(v) dv \psi(u) du dt \\
&= \int\int [\tilde{w}(t) - w(t)] \int F_{U_1|\eta}(u|t) \psi(u) du dt \\
&= \int [\int \tilde{w}(t) F_{U_1|\eta}(u|t) dt - \int w(t) F_{U_1|\eta}(u|t) dt] \psi(u) du \geq 0,
\end{aligned}$$

where the last ineq. is by $F^{w}_{U_1} \leq F^{\tilde{w}}_{U_1}$. Because $G(\cdot)$ is arbitrary, then $F^{w}_{U_0}$ is first order stochastic dominant over $F^{\tilde{w}}_{U_0}$. □

B Bounding Violation Probability in Linear Program with Randomized Constraints

Let $h(\gamma, y) \equiv p(y|1) - \boldsymbol{p}(y,1)'\gamma$. Following Calafiore and Campi (2005), define a violation probability and a robustly feasible solution.

Definition B.1 (Violation probability)**.** *Let $\gamma \in \Gamma$ be a candidate solution for (5.1)–(5.4). The probability of violation of γ is defined as*

$$V(\gamma) = \mathbb{P}\{Y \in \mathcal{Y} : h(\gamma, Y) > 0\},$$

where $\{Y \in \mathcal{Y} : h(\gamma, Y) > 0\}$ *is assumed to be measurable.*

Note that $V(\gamma^*) = 0$ where γ^* is the solution to (5.1)–(5.4).

Definition B.2 (ϵ-level solution)**.** *For* $\epsilon \in [0, 1]$, $\gamma \in \Gamma$ *is an* ϵ*-level robustly feasible solution if* $V(\gamma) \leq \epsilon$.

Then, we can show that the violation probability at the solution, denoted as $\bar{\gamma}_n$, to (5.3)–(5.4) is on average bounded by $1/n$.

Proposition B.1. *Let* $\bar{\gamma}_n$ *be the solution to* (5.3)–(5.4)*. Then,*

$$\mathbb{E}_{P^n}[V(\bar{\gamma}_n)] \leq \frac{1}{n+1},$$

where P^n *is the probability measure in the space* $\mathcal{Y}^n$ *of the multi-sample extraction* $Y_1, ..., Y_n$.

Corollary B.1. *Fix* $\epsilon \in [0, 1]$ *and* $\beta \in [0, 1]$ *and let*

$$n \geq \frac{1}{\epsilon\beta} - 1.$$

Then, with probability no smaller than $1 - \beta$*, the sampled LP* (5.3)–(5.4) *returns an optimal solution* $\hat{\gamma}_n$ *which is* ϵ*-level robustly feasible.*

The above results implicitly assume a particular rule of tie-breaking when there are multiple solutions in the sampled LP (see Theorem 3 in Calafiore and Campi (2005)). There is also discussions on no solution in the paper.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117. 1

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the distribution of male and female wages accounting for employment composition using bounds," *Econometrica*, 75, 323–363. 2

Calafiore, G. and M. C. Campi (2005): "Uncertain convex programs: randomized solutions and confidence levels," *Mathematical Programming*, 102, 25–46. 5.2, B, B

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261. 1, 2, 4

——— (2013): "Quantile models with endogeneity," *Annu. Rev. Econ.*, 5, 57–81. 4

Dong, Y. and S. Shen (2018): "Testing for rank invariance or similarity in program evaluation," *Review of Economics and Statistics*, 100, 78–85. 1

Frandsen, B. R. and L. J. Lefgren (2018): "Testing rank similarity," *Review of Economics and Statistics*, 100, 86–91. 1

Han, S. (2019): "Identification in Nonparametric Models for Dynamic Treatment Effects," *UT Austin*. 1

Han, S. and S. Yang (2020): "Sharp Bounds on Treatment Effects for Policy Evaluation," *arXiv preprint arXiv:2009.13861*. 1, 5.2, 5.3

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, 64, 487–535. 1

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 2

Kim, J. H. and B. G. Park (2022): "Testing rank similarity in the local average treatment effects model," *Econometric Reviews*, 1–22. 1

Maasoumi, E. and L. Wang (2019): "The gender gap between earnings distributions," *Journal of Political Economy*, 127, 2438–2504. 1

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *The American Economic Review*, 80, 319–323. 1

——— (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress, ed. by C. Sims*, vol. 1, 143–70. 2

——— (1997): "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1311–1334. 1

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. 1

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. 1

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, 3663–98. 2, 3.2

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79, 949–955. 1

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. 1, 2

Vytlacil, E. and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75, 757–779. 1