On Quantile Treatment Effects, Rank Similarity, and Variation of 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

April 23, 2023

Abstract

This paper investigates how certain relationship between observed and counterfactual distributions serves as an identifying condition for distributional treatment effects under endogeneity, and shows that this condition holds in a range of nonparametric models for treatment effects. To this end, we first provide a novel characterization of prevalent assumptions restricting treatment heterogeneity in the literature, namely rank similarity. Our characterization demonstrates the stringency of this assumption and allows us to relax it in a economically meaningful way, resulting in our identifying condition. It also justifies the quest of richer exogenous variations in the data (e.g., multi-valued or multiple instrumental variables) in exchange for the weaker identifying condition. The primary goal of this investigation is to provide empirical researchers with tools that are robust and easy to implement but still yield tight policy evaluations.

*For insightful discussions, the authors are grateful to Victor Chernozhukov and participants at the 2023 North American Winter Meeting of the Econometric Society and the CeMMAP/SNU conference and in the seminars at UCL and Oxford.

JEL Numbers: C14, C31, 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 serves as an identifying condition for distributional treatment effects under endogeneity, and shows that this condition holds in a range of nonparametric models for treatment effects. To this end, we first provide a novel characterization of prevalent assumptions restricting treatment heterogeneity in the literature, namely rank similarity. Our characterization demonstrates the stringency of this assumption and allows us to relax it in a economically meaningful way, resulting in our identifying condition. It also justifies the quest of richer exogenous variations in the data (e.g., multi-valued or multiple instrumental variables) in exchange for the weaker identifying condition.

The primary goal of this investigation is to provide empirical researchers with (i) a framework where validity of identifying conditions 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.

Our analysis centers on the relationship between observed and counterfactual distributions, specifically on the preservation of first-order stochastic dominance (FOSD) of one distribution over the other to their corresponding counterfactual distributions: for arbitrary compliance types $t, t' \in \mathcal{T}$ induced by induced by individuals' potential treatment responses to 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 counterfactual outcome given treatment $D = d$.[1] This condition relates a partial order between observed distributions to that of the counterfactual distributions. For the sake of illustration, consider binary instrument $Z \in \{0, 1\}$ that affects the treatment participation in a monotone way (Imbens and Angrist (1994)) 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 $Y_1|AT \prec_{FOSD} Y_1|C$ is equivalent to the following expression of the distribution of observables

$$P[Y \leq \cdot|D = 1, Z = 0] \leq P[Y \leq \cdot|D = 1, Z = 1] \tag{1.3}$$

and $Y_0|AT \prec_{FOSD} Y_0|C$ is equivalent to the following expression of the distribution of counterfactual

$$P[Y_0 \leq \cdot|D = 1] \leq \frac{P[Y \leq \cdot, D = 0|Z = 1] - P[Y \leq \cdot, D = 0|Z = 0]}{P[D = 1|Z = 0] - P[D = 1|Z = 1]}. \tag{1.4}$$

Therefore, (1.4) provides an informative upper bound for $P[Y_0 \leq y|D = 1]$ (and a symmetric analysis provides a lower bound). Note $P[Y_0 \leq y|D = 1]$ is a necessary component in calculating the distributional treatment effect, such as the quantile treatment effect on the treated (QTT). We also provide an analogous analysis to bound other treatment parameters (e.g., the average effects). The proposed bounds can be informative for practitioners provided (1.3) holds. Although (1.3) may seem restrictive, this is *not* generally the case when Z departs from a scalar binary variable. In this sense, our approach underscores the significance of searching for richer exogenous variations of IVs, such as multi-valued or multiple instrumental

[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.

variables, as a means of trading for less restrictive identifying conditions. Still, the benefit of our approach can be manifested without requiring continuous or large support. We also show that, when IVs take more than four values, the preservation of FOSD (i.e., if (1.1) then (1.2)) yields a testable restriction.

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 literature, Manski (1990), Manski (1997), and Manski and Pepper (2000), among many others, construct sharp bounds on the ATE under a set of assumptions on 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 alternative when researchers are equipped with discrete IVs and impose a monotonicity assumption on the selection to treatment. However, the local group for which the treatment effect is identified may not be the group of policy interest. Therefore, 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 (see e.g., Mogstad et al. (2018), Han and Yang (2020)).

Another prevalent approach in the literature is to restrict the degree of treatment heterogeneity via rank similarity (or rank invariance). This assumption is shown to have substantial identifying power for distributional treatment effects and the ATE and used in various nonparametric contexts implicitly or explicitly (Heckman et al. (1997), Chesher (2003, 2005), Chernozhukov and Hansen (2005), Vytlacil and Yildiz (2007), Jun et al. (2011), Shaikh and Vytlacil (2011), D'Haultfœuille and Février (2015), Torgovitsky (2015), 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 proposed as one reaction to the skepticism (Frandsen and Lefgren (2018), Dong and Shen (2018), Kim and Park (2022)).

In this paper, we clarify the stringency of the rank similarity assumption by characterizing its restrictions on the relationship between observed and counterfactual distributions. In particular, we show that the *strong* preservation of FOSD (i.e., (1.1) holds *if and only if* (1.2) holds) is equivalent to *rank linearity*, a slight relaxation of rank similarity that allows for a linear transformation of an individual's rank to the counterfactual rank. By doing so, we establish the connection between the rank similarity condition in Chernozhukov and Hansen (2005)'s structural IV model and its corresponding assumptions in terms of counterfactual outcomes within Rubin (1974)'s causal framework. Furthermore, we propose less stringent FOSD preservation conditions that allow us to identify certain treatment parameters. We provide economic justifications for weak preservation of FOSD by proposing a variety of non-separable structural IV models that imply the FOSD preservation condition, but that do not satisfy rank similarity.

Based on our identification strategy, we develop a statistical linear programming (LP) approach to estimate optimal bounds on the treatment parameters. These bounds are defined as optimal values of LP (with a discrete outcome) or semi-infinite LP (SILP) (with a continuous outcome). To address the infeasibility of the SILP problem, we transform the optimization problem by (i) randomizing the constraints or (ii) invoking duality and approximates the Lagrangian measure using sieves.

The next section formally introduces the main identifying conditions (i.e., the preservation of stochastic ordering) and shows how to construct bounds on treatment effects. Section 3 introduces structural models as sufficient conditions for the identifying conditions. Section 4 shows that point identification can be achieved with sufficient (but not infinite) variation of IVs. Section 5 discusses how to systematically calculate bounds using linear programming and, finally, Section 6 presents numerical studies.

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, which 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 $Y \in \mathcal{Y} \subseteq \mathbb{R}$ satisfies $Y = DY_1 + (1-D)Y_0$. Finally, $X \in \mathcal{X} \subseteq \mathbb{R}^k$ 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 and satisfy the exclusion restriction. Let $Y_{d,z}$ be the counterfactual outcome given $(D,Z) = (d,z)$.

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

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

2.1 Introducing Key Conditions

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

Condition $\mathbf{S}_1$. *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*

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

then

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

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), whose insights we share. Suppose that $Z \perp (Y_d, D_z)|X$ additionally holds, where D_z is the counterfactual treatment given $Z = z$. Under this assumption, Condition S_1 is equivalent to assuming if

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

then

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

Note that the event $\{D_{z_\ell} = 1\}$ (for $\ell = 1, ..., L$) captures a type of compliance to a given $Z = z_\ell$. Then, $\sum_{\ell=1}^L w_\ell P[Y_d \le y | D_{z_\ell} = 1, X = x]$ 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 S_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., by Imbens and Angrist (1994)'s monotonicity assumption for the LATE), then Condition S_1 simply describes the stochastic ordering between always-takers and compliers. When $L \geq 3$, however, the composition becomes more complex as illustrated in Section 2.3. We provide a condition sufficient to Condition S_1 that may be easier to interpret. To state this sufficient condition, we introduce a general model for treatment selection:

Assumption D. *Assume that*

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

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

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). Although Assumption D is not necessary for our main procedure, it is useful in defining the types of compliance behavior via the unobservable η. Under this assumption, the following condition is sufficient but not necessary for Condition S_1. Let $F_{Y_d|\eta,X}(y|t,x) \equiv P[Y_d \leq y|\eta = t, X = x]$.

Condition S_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 $\int w(t,x)dt = \int \tilde{w}(t,x)dt = 1$, 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 $w(\cdot, x)$ is non-negative and $\int w(t,x)dt = 1$, note that $\int w(t,x) F_{Y_d|\eta,X}(\cdot|t,x)dt$

is a mixture of conditional CDFs (with $w(\cdot, x)$ being the mixture weight) and thus itself a CDF. In other words, 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 S_1^* assumes that the FOSD ordering of Y_1 distributions conditional on η conforming to two different type distributions ($W_x(\cdot)$ and $\tilde{W}_x(\cdot)$) is preserved in the ordering of Y_0 distributions conditional on the corresponding type distributions. Note that Condition S_1^* (or Condition S_1) is *not* an "if and only if" statement. It would be stringent to impose the preservation of ordering to hold in both directions. In fact, such a condition is closely related to the rank similarity condition (Chernozhukov and Hansen (2005)); see Section 3 for full details. The following lemma establishes the sufficiency of Condition S_1^* for Condition S_1.

Lemma 2.1. *Under Assumption D, Condition S_1^* implies Condition S_1.*

The proof of this lemma and most of other proofs are contained in the appendix.

2.2 Bounds on Treatment Effects

Now, we show that Condition S_1 is useful in constructing bounds on $F_{Y_0|D,X}(\cdot|1,x)$ and subsequently on $QTE_\tau(1,x)$. Let

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

Theorem 2.1. *Suppose that Assumption Z and Condition S_1 hold. Fix $x \in \mathcal{X}$. For $\gamma \equiv (\gamma_1, ..., \gamma_L)$ and $\tilde{\gamma} \equiv (\tilde{\gamma}_1, ..., \tilde{\gamma}_L)$ in $\Gamma_p(x)$, suppose*

$$P[Y \leq \cdot|D = 1, X = x] \leq \sum_{\ell=1}^{L} \gamma_\ell P[Y \leq \cdot, D = 1|Z = z_\ell, X = x], \tag{2.6}$$

$$\sum_{\ell=1}^{L} \tilde{\gamma}_\ell P[Y \leq \cdot, D = 1|Z = z_\ell, X = x] \leq P[Y \leq \cdot|D = 1, X = x]. \tag{2.7}$$

[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

$$
\begin{aligned}
&-\sum_{\ell=1}^{L}\tilde{\gamma}_\ell P[Y\le\cdot,D=0|Z=z_\ell,X=x] && (2.8)\\
&\le P[Y_0\le\cdot|D=1,X=x]\\
&\le -\sum_{\ell=1}^{L}\gamma_\ell P[Y\le\cdot,D=0|Z=z_\ell,X=x] && (2.9)
\end{aligned}
$$

Note that there can be multiple γ and $\tilde{\gamma}$ 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 S_1 hold. Fix $x\in\mathcal{X}$. Then, $F_{Y_0|D,X}(\cdot|1,x)$ is upper and lower bounded by*

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

Theorem 2.1 and Corollary 2.1 highlight the identifying power of multi-valued IVs. The key step in Theorem 2.1 to calculate the bounds is to find γ (resp. $\tilde{\gamma}$) in $\Gamma_p(x)$ that satisfies (2.6) (resp. (2.7)), which serves as a rank condition. Note that this condition is verifiable with the data. Corollary 2.1 additionally implies that the bounds can be further tightened if one increases the degree of freedom in the feasible set $\Gamma_p(x)$ by increasing L, in which case (2.6)–(2.7) are more likely to hold. See below and Section 6 for related discussions.

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)):

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

where $Q^{LB}_{Y_0|D,X}(\tau|1,x)$ and $Q^{UB}_{Y_0|D,X}(\tau|1,x)$ are the τ-th quantiles of $F^{LB}_{Y_0|D,X}(\cdot|1,x)$ and $F^{UB}_{Y_0|D,X}(\cdot|1,x)$, respectively. 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 S_1.

Remark 2.1 (Constraints on γ)**.** *In Theorem 2.1, $\Gamma_p(x)$ imposes two restrictions on γ: (i) $\sum_{\ell=1}^L \gamma_\ell = 0$ and (ii) $\sum_{\ell=1}^L \gamma_\ell p(z_\ell,x) = 1$. 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, (ii) implicitly introduces a scale normalization. That is, for any γ satisfying $\sum_{\ell=1}^L \gamma_\ell p(z_\ell,x) \neq 0$, we can always rescale it as $\gamma^* = \frac{\gamma}{\sum_{\ell=1}^L \gamma_\ell p(z_\ell,x)}$ so that $\sum_{\ell=1}^L \gamma_\ell p(z_\ell,x) = 1$. It can be shown that this normalization does not affect the bounds obtained in* (2.8) *and* (2.9).

If we assume the converse of Condition S_1, we can calculate bounds on the $QTE(0,x)$.

Condition S_0. *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 S_0 hold. Fix $x \in \mathcal{X}$. For $\gamma \equiv$*

$(\gamma_1, ..., \gamma_L)$ *and* $\tilde{\gamma} \equiv (\tilde{\gamma}_1, ..., \tilde{\gamma}_L)$ *in* $\Gamma_p(x)$, *suppose*

$$P[Y \leq \cdot | D = 0, X = x] \leq \sum_{\ell=1}^{L} \gamma_\ell P[Y \leq \cdot, D = 0 | Z = z_\ell, X = x], \tag{2.12}$$

$$\sum_{\ell=1}^{L} \tilde{\gamma}_\ell P[Y \leq \cdot, D = 0 | Z = z_\ell, X = x] \leq P[Y \leq \cdot | D = 0, X = x]. \tag{2.13}$$

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

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

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

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

The proof of this theorem is analogous to that of Theorem 2.1. The bounds on $QTE_\tau(0, x)$ can be derived symmetrically as in the case of $QTE_\tau(1, x)$ and thus are omitted. Notably, which treatment parameter we can obtain bounds for is determined by which identifying condition we impose (i.e., Condition S_1 or S_0). In Section 3, we investigate this aspect within economic structural models. Finally, in the Appendix, we introduce weaker conditions to bound average treatment effects.

2.3 Understanding Key Conditions

We further explore Conditions S_1 and S_0 to give additional interpretation and discuss testability. Suppress $X = x$ to simplify our discussions. Under $Z \perp (Y_d, D_z)$, the inequalities for FOSD in Conditions S_1 and S_0 can be rewritten as

$$\sum_{\ell=1}^{L} w_\ell P[Y_d \leq y | D_{z_\ell} = 1] \leq \sum_{\ell=1}^{L} \tilde{w}_\ell P[Y_d \leq y | D_{z_\ell} = 1]. \tag{2.16}$$

Recall, 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, $P[Y \le y|D = 1] \le \sum_{\ell=1}^{L} \gamma_\ell P[Y \le y, D = 1|Z = z_\ell]$ for all y. Note that (2.6) is a special case of (2.16) with $d = 1$, which is the "if" part of Condition S_1. Let $p(z) \equiv (D = 1|Z = z)$. Assume weak separability in the selection equation (which is equivalent to the LATE monotonicity assumption by Vytlacil (2002)) for the sake of discussion:

Assumption D*. *Assume that $D = h(Z, \eta) = 1\{\eta \le p(Z)\}$ with $\eta \sim U[0, 1]$.*

Then, $\{D_{z_\ell} = 1\}$ in (2.16) are a mix of individuals who are compliers (C) and always-takers (AT). To understand this better, we first consider a simple case of binary IV before considering a general case.

Lemma 2.2. *Suppose $L = 2$ with $\mathcal{Z} = \{0, 1\}$ and Assumption D* holds. 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) in the introduction and is in turn equivalent to $Y_1|AT \prec_{FOSD} Y_1|C$.*

Now, consider the general L. For $Z \in \mathcal{Z} = \{z_1, ..., z_L\}$, let $(z_{\ell-1}, z_\ell)$-compliers be compliers induced by the change of Z from $z_{\ell-1}$ to z_ℓ. When $L = 3$ and $(z_1, z_2, z_3) = (0, 1, 2)$, for example, $\{(0, 1)\text{-C}\} = \{i : D_{0,i} = 0, D_{1,i} = D_{2,i} = 1\}$ is the set of eager compliers (E-C) and $\{(1, 2)\text{-C}\} = \{i : D_{0,i} = D_{1,i} = 0, D_{2,i} = 1\}$ is the set of reluctant compliers (R-C), following the language of Mogstad et al. (2021). Also, $\{\text{AT}\} = \{i : D_{0,i} = D_{1,i} = D_{2,i} = 1\}$ is the set of always-takers. Let p_ℓ for $\ell = \{2, ..., L\}$ is the proportion of $(z_{\ell-1}, z_\ell)$-compliers and let $p_1 \equiv P[\text{AT}]$. We show that (2.16) establishes the FOSD relationship between the mixtures of observed distributions of Y conditional on various always-takers and compliers groups:

Lemma 2.3. *Suppose Assumption D* holds and $Z \perp (Y_d, D_z)$ and $0 < p(z_\ell) < 1$ for all ℓ.*

(i) Then, (2.16) *is equivalent to*

$$\omega_1 P[Y_d \leq y|AT] + \sum_{\ell=2}^{L} \omega_\ell P[Y_d \leq y|(z_{\ell-1}, z_\ell)\text{-}C]$$
$$\leq \tilde{\omega}_1 P[Y_d \leq y|AT] + \sum_{\ell=2}^{L} \tilde{\omega}_\ell P[Y_d \leq y|(z_{\ell-1}, z_\ell)\text{-}C]$$

for some non-negative ω_ℓ *and* $\tilde{\omega}_\ell$ *for* $\ell = 1, ..., L$. *(ii) Moreover, suppose* $L \geq 2$ *and* $w_1 + p_1 \sum_{\ell=2}^{L} \frac{w_\ell}{p(z_\ell)} = \tilde{w}_1 + p_1 \sum_{\ell=2}^{L} \frac{\tilde{w}_\ell}{p(z_\ell)}$. *Then* (2.16) *with* $w \neq \tilde{w}$ *can be expressed as*

$$\sum_{\ell=2}^{L} \omega_\ell P[Y_d \leq y|(z_{\ell-1}, z_\ell)\text{-}C] \leq \sum_{\ell=2}^{L} \tilde{\omega}_\ell P[Y_d \leq y|(z_{\ell-1}, z_\ell)\text{-}C] \tag{2.17}$$

for some non-negative ω_ℓ *and* $\tilde{\omega}_\ell$ *for* $\ell = 2, ..., L$.

To illustrate the intuition of Lemma 2.3(i), consider $L = 3$ and $(z_1, z_2, z_3) = (0, 1, 2)$. Then,

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

because $\{D_0 = 1, D_1 = 1, D_2 = 0\} = \emptyset$ and $\{D_0 = 0, D_1 = 1, D_2 = 0\} = \emptyset$. Also, $\{D_2 = 1\} = \{\text{AT}\} \cup \{\text{E-C}\} \cup \{\text{R-C}\}$ and $\{D_0 = 1\} = \{\text{AT}\}$.

Lemma 2.3(ii) can be used as the basis to test (2.16) and thus Condition S_1. The intuition is as follows. With a binary IV, the marginal distributions of Y_1 and Y_0 are identified for compliers (Abadie et al. (2002)). This result holds for any complier group defined by a pair of instrument values, such as $\{(z_{\ell-1}, z_\ell)\text{-C}\}$ in the lemma. Then, when $L \geq 2$, we can find vectors w and $\tilde{w}$ in $\mathbb{R}_+^L$ that assign zero weights to the distributions for AT and still make (2.16) a non-trivial inequality where all the associated distributions for compliers are identified for all $d = 1, 0$.

Remark 2.2 (Conditions w.r.t. Compliance Types)**.** *Motivated from the discussion of this section, we can rewrite Condition* S_1 *(and all the relevant conditions) without even invoking*

Assumption D. 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 S_1 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.

3 Structural Models as Sufficient Conditions

We show that Conditions S_1 and S_0 can be justified in a range of nonparametric structural models for the counterfactual outcomes. By relating the conditions with the structural models, we provide additional intuitions for the conditions. We present a leading model here and the rest in the Appendix. 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

$$Y = q(D, X, U_D), \tag{3.1}$$

$$D = h(Z, X, \eta), \tag{3.2}$$

where $q(d, x, \cdot)$ is continuous and monotone increasing and $U_D = DU_1 + (1 - D)U_0$, (ii) conditional on (η, X, Z), $U_d \overset{d}{=} U + \xi_d$ where $\xi_d \perp (\eta, U)$, (iii) 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(ii)–(iii) implies that $U_0 \overset{d}{=} U_1 + V$ conditional on (η, X, Z). 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). 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(ii), we have $U_0 \overset{d}{=} U_1 \overset{d}{=} U$ conditional on (η, X, Z), which proves the claim.

Model 1(iii) 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 S_1^* (and thus Condition S_1 by Lemma 2.1).

Theorem 3.1. *Under Assumptions Z, Model 1 (with ξ_0 being weakly more noisy than ξ_1) implies Condition S_1^*.*

Analogous to Theorem 3.1, one can readily show that Model 1 with ξ_0 being weakly less noisy than ξ_1 implies Condition S_0.

Now we provide examples that are consistent with Model 1.

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

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 $U_d \overset{d}{=} U + \xi_d$ 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 $U_d \overset{d}{=} U + \xi_d$ 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, $U_d \overset{d}{=} U + \xi_d$ 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 S_1 via Theorem 3.1. Then, under Condition S_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 S_0, 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 S_1 has to be strengthened to be equivalent to rank similarity? To answer this question, recall that Condition S_1^* is stronger than Condition S_1 (by Lemma 2.1). We strengthen Condition S_1^* further by making it an "if and only if" condition:

Condition S^*. *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 3.2. *Model 1(i) with $F_{U_0|\eta,X,Z} = F_{U_1|\eta,X,Z}$ (i.e., rank similarity) implies Condition S^*.*

This theorem highlights the stringency of rank similarity relative to Condition S_1. The proof is trivial so omitted. It is worth noting that the converse of Theorem 3.2 is *not* true. Here is a counter-example for the converse statement.

Definition 3.1 (Rank Linearity)**.** *Assume Model 1(i) and*

$$F_{Y_0|\eta,X,Z}(\cdot|t,x,z) = \lambda(\cdot,x)F_{Y_1|\eta,X,Z}(\psi(\cdot,x)|t,x,z) \tag{3.3}$$

for every $t \in \mathcal{T}$ 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 $\lambda(\cdot,x) : \mathcal{Y} \to \mathbb{R}_+$ is consistent with $F_{Y_d|\eta,X,Z}$ being a proper CDF.

This *rank linearity* implies Condition S^*, 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 $Z = z$ (and suppressing X), (3.3) with Model 1(i) yields $F_{U_0|\eta}(q^{-1}(0,y)|t) = \lambda(y)F_{U_1|\eta}(q^{-1}(1,\psi(y))|t)$. Then, by choosing $\lambda(y) = 1$ and $\psi(y) = q(1,q^{-1}(0,y))$, we have $F_{U_0|\eta}(\cdot|t) = F_{U_1|\eta}(\cdot|t)$. In general, while the ranks between Y_0 and Y_1 should stay the same under rank similarity, they can be arbitrarily different (e.g., the rank can be reversed) under rank linearity because of the multiplying term $\lambda(\cdot)$ in $F_{U_0|\eta}(u|t) = \lambda(q(0,u))F_{U_1|\eta}(u|t)$.

Interestingly, rank linearity is equivalent to Condition S^*. The following theorem is one of the main contributions of this paper. Suppress (Z,X) for simplicity.

Theorem 3.3. *Suppose for any CDF $F_1(\cdot)$ supported on $\mathbb{R}$, there always exists a function $c : \mathcal{T} \to \mathbb{R}$ such that*

$$F_d(\cdot) = \int c(t)F_{Y_d|\eta}(\cdot|t)dt. \tag{3.4}$$

Then Condition S^* holds if and only if there exits some $\psi(\cdot)$ that is strictly increasing and $\lambda(\cdot) > 0$ such that

$$F_{Y_0|\eta}(\cdot|t) = \lambda(\cdot)F_{Y_1|\eta}(\psi(\cdot)|t) \qquad \text{for } t \in \mathcal{T}. \tag{3.5}$$

We prove this equivalence in the Appendix. The proof with continuous Y_d is more involved than that with discrete Y_d; we recommend that the interested reader reads the latter first. The condition (3.4) 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 bound analysis. This condition would be violated when there is no endogeneity (i.e., $Y_d \perp \eta$), which is not our focus.

Condition S* 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 S$_1$) 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 S$_0$) 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 S* is weaker than rank similarity illustrates the importance of rank similarity in the identification of the QTE and ATE.

Remark 3.1 (Testability of the Conditions). *It is immediate from Lemma 2.3(ii) that Condition S* can be tested from the data when $L \geq 2$ and under the LATE monotonicity assumption. Given the established connection between Condition S* and rank similarity (Theorem 3.2), when Condition S* is refuted from the data, rank similarity can be refuted. This result relates to the testability of rank similarity (Frandsen and Lefgren (2018), Dong and Shen (2018), Kim and Park (2022)).*

Remark 3.2 (Conditions w.r.t. Compliance Types, continued). *In the framework proposed in Remark 2.2, if we additionally impose Assumption D, 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)–(ii), $F_{U_1|\eta} = F_{U_0|\eta}$ implies $F_{U_1|T} = F_{U_0|T}$. Moreover, Condition C.1 can be motivated by this framework as the discrete η can be viewed as T (with $k_\eta = 2^L$).*

4 Point Identification

Point identification of $QTE_\tau(d)$ and $ATE(d)$ can be achieved as long as the stochastic dominance ordering is preserved (i.e., Condition S_1 or S_0) and instruments have sufficient variation in a specific sense. As is clear below, however, we do *not* require $p(z) \to 1$ or 0 (i.e., instruments with large support). In this sense, our approach to point identification complements the approach of identification at infinity. To see this, consider the following theorem.

Theorem 4.1. *Suppose that Assumption Z and Condition S_1 hold. Fix $x \in \mathcal{X}$. For $\gamma \equiv (\gamma_1, ..., \gamma_L)$ in $\Gamma_p(x)$, suppose*

$$P[Y \leq \cdot | D = 1, X = x] = \sum_{\ell=1}^{L} \gamma_\ell P[Y \leq \cdot, D = 1 | Z = z_\ell, X = x]. \tag{4.1}$$

Then $F_{Y_0|D,X}(\cdot|1, x)$ is identified as

$$P[Y_0 \leq \cdot | D = 1, X = x] = -\sum_{\ell=1}^{L} \gamma_\ell P[Y \leq \cdot, D = 0 | Z = z_\ell, X = x] \tag{4.2}$$

The key for this point identification result is that there exists γ such that (4.1) holds, which is a stronger requirement than the inequality version (2.6). The equation (4.1) is more likely to hold when L is large, that is, when instruments take more values. In particular, when $L \to \infty$ (e.g., continuous Z), we may invoke the notion similar to sieve approximation for $P[Y \leq y | D = 1, X = x]$ with $\{P[Y \leq y | D = 1, Z = z_\ell, X = x]\}_{\ell=1}^{L}$ being the "basis functions" and assume

$$P[Y \leq y | D = 1, X = x] = \lim_{L \to \infty} \sum_{\ell=1}^{L} \gamma_\ell p(z_\ell, x) P[Y \leq y | D = 1, Z = z_\ell, X = x],$$

where $\sum_{\ell=1}^{L} \gamma_\ell p(z_\ell, x) = 1$ and $\sum_{\ell=1}^{L} \gamma_\ell = 0$. Note that, although this does *not* demand an infinite support for Z, it implicitly assumes that Z sufficiently influences the distribution of Y conditional on $(D, X) = (1, x)$ in a way that the resulting basis functions generate

$P[Y \leq y|D=1, X=x]$. Importantly, whether this is possible or not can be confirmed from the data.

Given Theorem 4.1, we identify $QTE_\tau(1,x) = Q_{Y|D,X}(\tau|1,x) - Q_{Y_0|D,X}(\tau|1,x)$ where $Q_{Y_0|D,X}(\tau|1,x)$ is a solution to $\tau = -\sum_{\ell=1}^{L} \gamma_\ell P[Y \leq \cdot, D=0|Z=z_\ell, X=x]$. Similarly, under Condition S_0, we can identify $F_{Y_0|D,X}(\cdot|1,x)$ and thus $QTE_\tau(0,x)$. We omit this result for succinctness.

It is worth comparing the point identification result with that in Chernozhukov and Hansen (2005). While they point identify $QTE_\tau(x)$ with a binary instrument by assuming rank similarity, the result of this section suggests that we may achieve the identification of $QTE_\tau(x)$ under rank linearity if we have a continuous instrument in the sense of (4.1). This is because rank linearity is equivalent to Condition S^* (by Theorem 3.3), but the latter implies Conditions S_1 and S_0 that identify $QTE_\tau(1,x)$ and $QTE_\tau(0,x)$, respectively.

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

$$\Gamma_p = \{\gamma : \gamma'[\begin{array}{cc} \mathbf{1} & \boldsymbol{p} \end{array}] = [\begin{array}{cc} 0 & 1 \end{array}]\} \subset \mathbb{R}^L.$$

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 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

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

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

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 S_1^*. 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 $Y_0 = 1 + U_0$
- $(U, \eta) \sim BVN((0,0)', \Sigma)$
- $V \sim N(0, \sigma_V^2)$ and $\xi_1 \sim N(0, \sigma_\xi^2)$
- $\xi_0 = \xi_1 + V$
- $U_d = U + \xi_d$
- $Z \sim Bin(L-1, p)/(L-1) \in [0,1]$ 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 1–3 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.

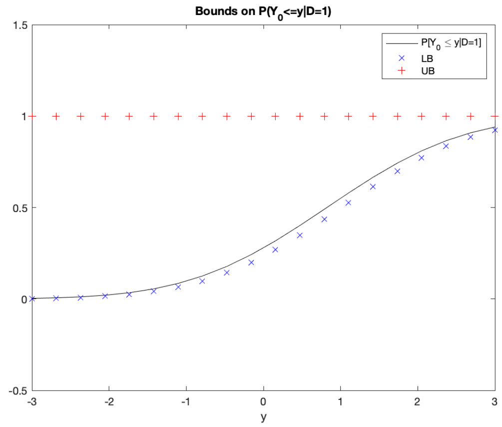

Figure 1: Bounds on $\Pr[Y_0 \leq y|D=1]$ When $L=2$

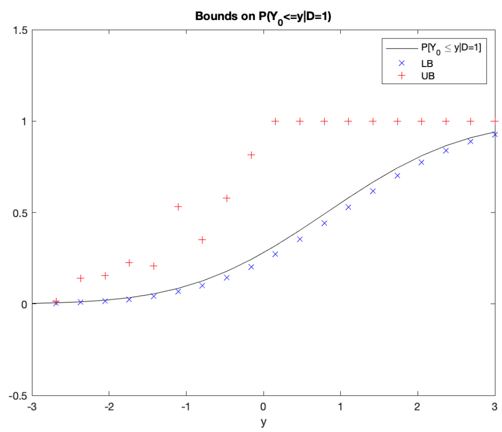

Figure 2: Bounds on $\Pr[Y_0 \leq y|D=1]$ When $L=5$

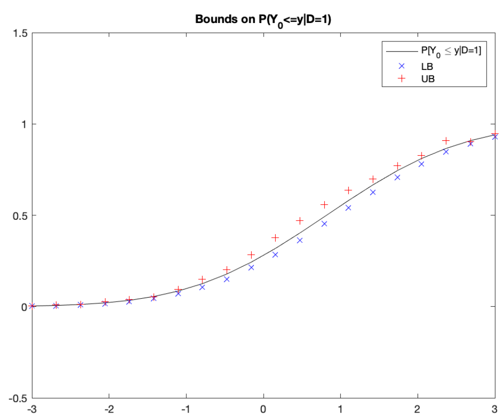

Figure 3: Bounds on $\Pr[Y_0 \leq y|D=1]$ When $L=6$

A Conditions for Average Treatment Effects

To calculate bounds on $ATE(1,x)$ and $ATE(0,x)$, we introduce conditions that are weaker that Conditions S_1 and S_0.

Condition S_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*

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

then

$$\sum_{\ell=1}^L w_\ell E[Y_0 | D=1, Z=z_\ell, X=x] \leq \sum_{\ell=1}^L \tilde{w}_\ell E[Y_0 | D=1, Z=z_\ell, X=x]. \tag{A.2}$$

Condition S_1' can be used to bound the $ATE(1,x)$. An analogous condition can be imposed to bound $ATE(0,x)$.

Condition S_0'. *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*

$$\sum_{\ell=1}^L w_\ell P[Y_0 \leq \cdot | D=1, Z=z_\ell, X=x] \leq \sum_{\ell=1}^L \tilde{w}_\ell P[Y_0 \leq \cdot | D=1, Z=z_\ell, X=x], \tag{A.3}$$

then

$$\sum_{\ell=1}^L w_\ell E[Y_1 | D=1, Z=z_\ell, X=x] \leq \sum_{\ell=1}^L \tilde{w}_\ell E[Y_1 | D=1, Z=z_\ell, X=x]. \tag{A.4}$$

B Other Structural Models as Sufficient Conditions

We present two more structural models that are not nested to Model 1 in the text. Model1(i) are maintained in these models, that is, $Y = q(D, X, U_D)$ where $q(d, x, \cdot)$ is continuous and monotone increasing and $D = h(Z, X, \eta)$.

Model 2. (ii) $U_0 \overset{d}{=} \phi(U_1, V)$ conditional on (η, X) where $V \perp (U_1, \eta)|X$ and $\phi(\cdot, v)$ is strictly increasing for all v.

Model 2(ii) defines that U_0 is "noisier" than U_1. Therefore, Model 2 is weaker than the model in Chernozhukov and Hansen (2005). Model 2 and Model 1 are not nested because, in $U_0 = U_1 + V$ of Model 1, V is not independent of U_1. We show below that Model 2 implies Condition S_1. Interestingly, Model 2(ii) with $U_0 = \phi(U_1, V)$ (instead of "$\overset{d}{=}$") is a generalization of the definition that U_0 is "noisier" than U_1 if $U_0 = U_1 + V$ with $U_1 \perp V$ in Pomatto et al. (2020, p. 1880).

Model 3. (ii) $U_0 \overset{d}{=} \max\{\phi(U_1), V\}$ conditional on (η, X) where $V \perp (U_1, \eta)|X$ and $\phi(\cdot)$ is strictly increasing.

We show below that Model 3 implies rank linearity. Model 3 can alternatively be defined as follows: $Y_0 \overset{d}{=} \max\{\phi(Y_1), V\}$ conditional on (η, X) where $V \perp (Y_1, \eta)|X$ and $\phi(\cdot)$ is strictly increasing. Then, this model also implies rank linearity with $\psi(\cdot) = \phi^{-1}(\cdot)$ because

$$\Pr[Y_0 \leq y|\eta, X] = \Pr[\phi(Y_1) \leq y, V \leq y|\eta, X] = \Pr[Y_1 \leq \phi^{-1}(y)|\eta, X] \Pr[V \leq y|X].$$

This model provides another interpretation of an insurance policy $(D = 1)$ as $Y_1 = \max\{Y_0, V\}$ guarantees at least Y_0. Models 2 and 3 are not nested.

Lemma B.1. *(i) Model 2 implies Condition S_1^*; (ii) Model 3 implies rank linearity.*

The proof of this lemma is contained in the next section.

C Proofs

C.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,

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

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. $\square$

C.2 Proof of Theorem 2.1

We suppress X for simplicity and prove the upper bound; the lower bound can be analogously derived. Without loss of generality, for some $\ell^* \leq L$, let $\gamma_\ell \leq 0$ for $\ell \leq \ell^*$ 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

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

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

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

where $\sum_{\ell=1}^{\ell^*} \frac{q(z_\ell)-\gamma_\ell p(z_\ell)}{a} + \sum_{\ell=\ell^*+1}^{L} \frac{q(z_\ell)}{a} = 1$ and $\sum_{\ell=\ell^*+1}^{L} \frac{\gamma_\ell p(z_\ell)}{a} = 1$. Therefore, by Condition S_1, we have

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

Equivalently, we have

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

where the last equality is by $\sum_{\ell=1}^{L} \gamma_\ell = 0$. $\square$

C.3 Proof of Lemma 2.2

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

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

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

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

Hence,

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

that is,

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

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

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

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

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

□

C.4 Proof of Lemma 2.3

Note that $P[Y_d \leq y|D_{z_1} = 1] = P[Y_d \leq y|\text{AT}]$ and, for $\ell = \{2, ..., L\}$,

$$P[Y_d \leq y|D_{z_\ell} = 1] = \frac{1}{p(z_\ell)} P\left[Y_d \leq y, \{\text{AT}\} \cup \bigcup_{\ell'=2}^{\ell} \{(z_{\ell'-1}, z_{\ell'})\text{-C}\}\right]$$

by Assumption D* and $p(z_\ell) = P[D_{z_\ell} = 1]$. Then, in (2.16),

$$
\begin{aligned}
&\sum_{\ell=1}^{L} w_\ell P[Y_d \le y | D_{z_\ell} = 1] \\
&= w_1 P[Y_d \le y | \text{AT}] + \sum_{\ell=2}^{L} \frac{w_\ell}{p(z_\ell)} \left(p_1 P[Y_d \le y | \text{AT}] + \sum_{\ell'=2}^{\ell} p_{\ell'} P[Y_d \le y | (z_{\ell'-1}, z_{\ell'})\text{-C}] \right)
\end{aligned}
$$

and similarly for the right-hand side of (2.16). This proves (i). To remove the distributions for AT in the expressions, we set

$$
w_1 + p_1 \sum_{\ell=2}^{L} \frac{w_\ell}{p(z_\ell)} = \tilde{w}_1 + p_1 \sum_{\ell=2}^{L} \frac{\tilde{w}_\ell}{p(z_\ell)}. \tag{C.2}
$$

Then, note that when $L \ge 2$, $w \ne \tilde{w}$ even if w and $\tilde{w}$ satisfy (C.2). Therefore, the resulting (2.16) is the dominance between the two distinct weight sums of $P[Y_d \le y | j\text{-C}]$'s:

$$
\sum_{\ell=2}^{L} \frac{w_\ell}{\sum_{\ell'=1}^{\ell} p_{\ell'}} \sum_{\ell'=2}^{\ell} p_{\ell'} P[Y_d \le y | (z_{\ell'-1}, z_{\ell'})\text{-C}] \le \sum_{\ell=2}^{L} \frac{\tilde{w}_\ell}{\sum_{\ell'=1}^{\ell} p_{\ell'}} \sum_{\ell'=2}^{\ell} p_{\ell'} P[Y_d \le y | (z_{\ell'-1}, z_{\ell'})\text{-C}],
$$

which can be simplified as (2.17) in (ii). $\square$

C.5 Proof of Theorem 3.1

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 (3.1) in Model 1(i), $F_{Y_d}^w \le F_{Y_d}^{\tilde{w}}$ if and only if $F_{U_d}^w \le F_{U_d}^{\tilde{w}}$. So it suffices to show that, if $F_{U_1}^w \le F_{U_1}^{\tilde{w}}$, then $F_{U_0}^w \le F_{U_0}^{\tilde{w}}$.

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

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

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(ii), and the last eq. is by change of variables. By Model 1(iii), $f_{\xi_0}(s) = \int f_{\xi_1}(s-v)f_V(v)dv = \int f_{\xi_1}(v)f_V(s-v)dv$ where $f_A(\cdot)$ is the PDF of an arbitrary r.v. A. Therefore,

$$\begin{aligned}
&\int GdF^w_{U_0} - \int GdF^{\tilde{w}}_{U_0} \\
&= \int\int\int [\tilde{w}(t) - w(t)]F_{U|\eta}(u|t) \int f_{\xi_1}(v)f_V(s-v)g(u+s)dvdudsdt \\
&= \int\int [\tilde{w}(t) - w(t)]F_{U|\eta}(u|t) \int f_{\xi_1}(v)[\int f_V(s)g(u+s+v)ds]dvdudt.
\end{aligned}$$

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

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

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 $F^{\tilde{w}}_{U_0}$. $\square$

C.6 Proof of Theorem 3.3: Equivalence Between Rank Linearity and Condition S*

The "if" part is trivial. We prove "only if" part. Suppose Condition S* holds. Let $\mathcal{Y}_\infty \equiv \{y_k \in \mathbb{R} : k = 1, \cdots, \infty\}$ be a sequence that is dense on $\mathbb{R}$. Denote $\mathcal{Y}_n \equiv \{y_k \in \mathbb{R} : k = 1, \cdots, n\}$. Because $\mathcal{Y}_\infty$ is dense in $\mathbb{R}$ and CDFs are right-continuous, it suffices to show the existence of $\lambda(\cdot)$ and $\psi(\cdot)$ on $\mathcal{Y}_\infty$ such that

$$F_{Y_0|\eta}(\psi(y)|t) = \lambda(y) F_{Y_1|\eta}(y|t) \tag{C.3}$$

holds for all $t \in \mathcal{T}$ and $y \in \mathcal{Y}_\infty$.

Fix $n \in \mathbb{N}$. Let $G_{1,k} : \mathbb{R} \to \{0,1\}$ be a simple function defined as $G_{1,k}(\cdot) \equiv 1\{y_k \leq \cdot\}$ for $k = 1, \cdots, n$. By the full rank condition (3.4), for each $1 \leq k \leq n$, there exists a function $c_k : \mathcal{T} \to \mathbb{R}$ such that

$$G_{1,k}(\cdot) = \int c_k(t) F_{Y_1|\eta}(\cdot|t)dt.$$

Define $G_{0,k} : \mathbb{R} \to [0,1]$ as

$$G_{0,k}(\cdot) \equiv \int c_k(t) F_{Y_0|\eta}(\cdot|t)dt.$$

Note that $G_{0,k}$ is a proper CDF. Now, for any vectors $\pi \equiv (\pi_1, \cdots, \pi_n)$ and $\tilde{\pi} \equiv (\tilde{\pi}_1, \cdots, \tilde{\pi}_n)$ such that $\sum_{k=1}^n \pi_k = \sum_{k=1}^n \tilde{\pi}_k = 1$, suppose

$$\sum_{k=1}^n \pi_k G_{1,k}(\cdot) \leq \sum_{k=1}^n \tilde{\pi}_k G_{1,k}(\cdot).$$

It follows that

$$\int b_n(t) F_{Y_1|\eta}(\cdot|t)dt \leq \int \tilde{b}_n(t) F_{Y_1|\eta}(\cdot|t)dt,$$

where $b_n(t) \equiv \sum_{k=1}^n \pi_k c_k(t)$ and $\tilde{b}_n(t) \equiv \sum_{k=1}^n \tilde{\pi}_k c_k(t)$. Let $b_n^+(t) = \max\{b_n(t), 0\}$ and $b_n^-(t) = \min\{b_n(t), 0\}$ and similarly define $\tilde{b}_n^+(t)$ and $\tilde{b}_n^-(t)$. Then, the above inequality can be written as

$$\int \{b_n^+(t) - \tilde{b}_n^-(t)\} F_{Y_1|\eta}(\cdot|t)dt \leq \int \{\tilde{b}_n^+(t) - b_n^-(t)\} F_{Y_1|\eta}(\cdot|t)dt,$$

where the resulting weight functions on both sides are non-negative. Then, by Condition S*, we have

$$\sum_{k=1}^n \pi_k G_{0,k}(\cdot) \leq \sum_{k=1}^n \tilde{\pi}_k G_{0,k}(\cdot)$$

By a similar argument, the converse is also true and thus we have

$$\sum_{k=1}^{n} \pi_k G_{1,k}(\cdot) \leq \sum_{k=1}^{n} \tilde{\pi}_k G_{1,k}(\cdot).$$

if and only if

$$\sum_{k=1}^{n} \pi_k G_{0,k}(\cdot) \leq \sum_{k=1}^{n} \tilde{\pi}_k G_{0,k}(\cdot)$$

for any non-negative weights π and $\tilde{\pi}$. Therefore, it follows that

$$\sum_{k=1}^{n} \delta_k G_{1,k}(\cdot) \leq 0 \text{ if and only if } \sum_{k=1}^{n} \delta_k G_{0,k}(\cdot) \leq 0 \tag{C.4}$$

for any n-dimensional vector $\delta \equiv (\delta_1, \cdots, \delta_n)$ that satisfies $\sum_{k=1}^{n} \delta_k = 0$.

For $d \in \{0, 1\}$, define

$$\Delta_d^G \equiv \left\{ \delta \in \mathbb{R}^n : \sum_{k=1}^{n} \delta_k G_{d,k}(y) \leq 0 \ \forall y \in \mathbb{R}; \sum_{k=1}^{n} \delta_k = 0 \right\}.$$

Note that $\{(G_{1,1}(y), \cdots, G_{1,n}(y)) : y \in \mathbb{R}\} = \{(G_{1,1}(y), \cdots, G_{1,n}(y)) : y \in \mathcal{Y}_n\}$ by definition. Therefore, Δ_1^G is a *finite* cone and its dimension is $n-1$. Define the polar cone of Δ_d^G as $\Delta_d^{G*} \equiv \{G_d \in \mathbb{R}^n : G_d'\delta \leq 0, \forall \delta \in \Delta_d^G\}$. Note that by definition, $\big(G_{1,1}(y), \cdots, G_{1,n}(y)\big)$ for $y \in \mathcal{Y}_n/\{y_n\}$ are $n-1$ linearly independent vectors and therefore generate extreme rays of Δ_1^{G*}. Also note that any element in Δ_0^{G*} is written as $\big(G_{0,1}(y), \cdots, G_{0,n}(y)\big)$ for some $y \in \mathbb{R}$, and so is a vector that generates its extreme ray. But by (C.4), we have that $\Delta_1^G = \Delta_0^G$ and thus $\Delta_1^{G*} = \Delta_0^{G*}$, and therefore, for each $y_k \in \mathcal{Y}_n/\{y_n\}$, there exists $y_k^* \in \mathbb{R}$ and $\lambda_n(\cdot) > 0$ such that

$$\big(G_{0,1}(y_k^*), \cdots, G_{0,n}(y_k^*)\big) = \lambda_n(y_k) \times \big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k)\big). \tag{C.5}$$

If there exists multiple values of y_k^* satisfying (C.5), we define y_k^* as the infimum of $\{\tilde{y}_k^* : \big(G_{0,1}(\tilde{y}_k^*), \cdots, G_{0,n}(\tilde{y}_k^*)\big) = \lambda_n(y_k) \times \big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k)\big)\}$. Because CDFs are right-continuous function, the infimum should also satisfy (C.5).

For any $j, k = 1, ..., n$, if $G_{1,j}(y_k) = 0$ then $G_{0,j}(y_k^*) = 0$ by (C.5), which further implies that $G_{0,j}(y^*) = 0$ for all $y^* \leq y_k^*$ because $G_{0,j}(\cdot)$ is monotone increasing. Let $\{j_1, \cdots, j_n\}$ be a permutation of $\{1, \cdots, n\}$ such that $y_{j_1} < y_{j_2} < \cdots < y_{j_n}$. Note that $G_{1,j_1}(y_{j_1})$ is the only non-zero component in the set $\{G_{1,k}(y_{j_1}) : k = 1, \cdots, n\}$. Then, by (C.5), $G_{0,j_1}(y_{j_1}^*) \neq 0$ and $G_{0,j_k}(y_{j_1}^*) = 0$ for $k \geq 2$. Similarly, there are two elements of $\{G_{0,k}(y_{j_2}^*) : k = 1, \cdots, n\}$ which are non-zero, namely, $G_{0,j_1}(y_{j_2}^*)$ and $G_{0,j_2}(y_{j_2}^*)$. Therefore, by $G_{0,j_2}(y_{j_1}^*) = 0$ and $G_{0,j_2}(y_{j_2}^*) \neq 0$ and the fact that $G_{0,j_2}(\cdot)$ is monotone increasing, we can conclude $y_{j_1}^* < y_{j_2}^*$. Continuing this argument, we can conclude that

$$y_{j_1}^* < y_{j_2}^* < \cdots < y_{j_n}^*.$$

Define a function $\psi_n : y_k \mapsto y_k^*$ for $k = 1, \cdots, n$. By the above analysis, $\psi_n(\cdot)$ is a monotone increasing function. Note that the support of ψ_n is $\mathcal{Y}_n$, which we extend to $\mathbb{R}$ as follows: for any $y \in \mathbb{R}$,

$$\psi_n(y) = \begin{cases} \max\{\psi_n(y_k) : y_k \leq y,\ k = 1, \cdots, n\} & \text{if } y \geq \min\{y_1, \cdots, y_n\} \\ -\infty & \text{otherwise} \end{cases}$$

Then, $\psi_n : \mathbb{R} \to \mathbb{R}$ is still a monotone increasing function.

We now consider increasing n to $n+1$. By a similar argument, there exists a sequence $\{y_1^\dagger, \cdots, y_n^\dagger, y_{n+1}^\dagger\}$ and $\lambda_{n+1}(\cdot) > 0$ such that for $k = 1, \cdots, n+1$, we have

$$\big(G_{0,1}(y_k^\dagger), \cdots, G_{0,n}(y_k^\dagger), G_{0,n+1}(y_k^\dagger)\big) = \lambda_{n+1}(y_k) \times \big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k), G_{1,n+1}(y_k)\big), \quad \text{(C.6)}$$

If there exists multiple values of $y_k^\dagger$, we define $y_k^\dagger$ as the infimum of them. Note that, by (C.6) and (C.5), $y_k^\dagger$ is one of the candidates $\tilde{y}_k^*$'s that make $\big(G_{0,1}(\tilde{y}_k^*), \cdots, G_{0,n}(\tilde{y}_k^*)\big)$ proportional to $\big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k)\big)$ satisfy (C.5). While y_k^* is the infimum of those candidates, $y_k^\dagger$ cannot reach that infimum because it has to satisfies the additional restriction, $G_{0,n+1}(y_k^\dagger) = \lambda_{n+1}(y_k)G_{1,n+1}(y_k)$. Therefore, we can conclude that $y_k^\dagger \geq y_k^*$ for $k = 1, \cdots, n$. Using

$\{y_1, \cdots, y_n, y_{n+1}\}$ and $\{y_1^\dagger, \cdots, y_n^\dagger, y_{n+1}^\dagger\}$, define $\psi_{n+1}(\cdot)$ analogous to $\psi_n(\cdot)$ above. Then, $\psi_{n+1}(y_k) = y_k^\dagger \geq y_k^* = \psi_n(y_k)$ for $k = 1, \cdots, n$. Furthermore, by definition, $\psi_{n+1}(y_{n+1}) \geq \psi_n(y_{n+1})$ regardless of the rank order of y_{n+1} in $\mathcal{Y}_{n+1}$. Therefore, for any $y \in \mathbb{R}$,

$$\psi_{n+1}(y) \geq \psi_n(y),$$

and thus the limit of the sequence of functions $\psi_n(\cdot)$ exists as $n \to \infty$, which we denote as $\psi_\infty(\cdot)$. Recall each $\psi_n(\cdot)$ is weakly increasing. It is easy to prove by contradiction that $\psi_\infty(\cdot)$ is strictly increasing. Fix $y_k \in \mathcal{Y}_\infty$. For any $n \geq k$, $\big(G_{0,1}(\psi_\infty(y_k)), \cdots, G_{0,n}(\psi_\infty(y_k))\big)$ is proportional to $\big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k)\big)$ and therefore there exists $\lambda_\infty(y_k)$ such that

$$\big(G_{0,1}(\psi_\infty(y_k)), \cdots, G_{0,n}(\psi_\infty(y_k))\big) = \lambda_\infty(y_k) \times \big(G_{1,1}(y_k), \cdots, G_{1,n}(y_k)\big) \tag{C.7}$$

for any $n \in \mathbb{N}$. Moreover, because $\mathcal{Y}_\infty$ is dense in $\mathbb{R}$ and ψ_∞ and $G_{d,k}$ are right-continuous functions, the above condition holds for all $y \in \mathbb{R}$.

Note $\{G_{1,k}(\cdot) : k = 1, \cdots, \infty\}$ is a class of simple functions. Therefore, any $F_{Y_1|\eta}(\cdot|t)$ can be written as

$$F_{Y_1|\eta}(\cdot|t) = \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) G_{1,k}(\cdot)$$

for some triangular array $\{a_{Kk}(t) : 1 \leq k \leq K, K = 1, 2, \cdots, \infty\}$. By the definition of $G_{1,k}(\cdot)$, it follows that

$$\begin{aligned} F_{Y_1|\eta}(\cdot|t) &= \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) \int w_k(s) F_{Y_1|\eta}(\cdot|s) ds = \int \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) w_k(s) F_{Y_1|\eta}(\cdot|s) ds \\ &\equiv \int \kappa(t,s) F_{Y_1|\eta}(\cdot|s) ds, \end{aligned} \tag{C.8}$$

where $\kappa(t,s) \equiv \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) w_k(s)$ serves as a Dirac delta function. Because $F_{Y_1|\eta}(\cdot|t) = \int \kappa(t,s) F_{Y_1|\eta}(\cdot|s) ds$ if and only if $F_{Y_1|\eta}(\cdot|t) \leq \int \kappa(t,s) F_{Y_1|\eta}(\cdot|s) ds$ and $F_{Y_1|\eta}(\cdot|t) \geq \int \kappa(t,s) F_{Y_1|\eta}(\cdot|s) ds$, we have, by Condition S*,

$$F_{Y_0|\eta}(\cdot|t) = \int \kappa(t,s) F_{Y_0|\eta}(\cdot|s) ds = \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) G_{0,k}(\cdot) \tag{C.9}$$

using the definition of $G_{0,k}(\cdot)$. Combining (C.8), (C.9) and (C.7), for any $y \in \mathbb{R}$ and $t \in \mathcal{T}$, we have

$$\begin{aligned} F_{Y_0|\eta}(\psi_\infty(y)|t) &= \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) G_{0,k}(\psi_\infty(y)) = \lim_{K\to\infty} \sum_{k=1}^{K} a_{K,k}(t) \lambda_\infty(y) G_{1,k}(y) \\ &= \lambda_\infty(y) F_{Y_1|\eta}(y|t), \end{aligned}$$

which completes the proof. □

C.7 Equivalence Between Rank Linearity and Condition S*: Discrete Y_d

For $d \in \{0,1\}$, suppose Y_d and η are discretely distributed. Specifically, let $\mathcal{Y}_d \equiv \{y_{d,1}, \cdots, y_{d,K_d}\}$ and $\mathcal{T} \equiv \{t_1, \cdots, t_{K_\eta}\}$ be the support of Y_d and η, respectively. 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"). Suppress (Z, X) for simplicity.

Condition C.1. *For arbitrary non-negative weights $\{w_1, \cdots, w_{K_\eta}\}$ and $\{\tilde{w}_1, \cdots, \tilde{w}_{K_\eta}\}$ such that $\sum_{k=1}^{K_\eta} w_k = 1$ and $\sum_{k=1}^{K_\eta} \tilde{w}_k = 1$, it holds that*

$$\sum_{k=1}^{K_\eta} w_k F_{Y_1|\eta}(\cdot|t_k) \leq \sum_{k=1}^{K_\eta} \tilde{w}_k F_{Y_1|\eta}(\cdot|t_k)$$

if and only if

$$\sum_{k=1}^{K_\eta} w_k F_{Y_0|\eta}(\cdot|t_k) \leq \sum_{k=1}^{K_\eta} \tilde{w}_k F_{Y_0|\eta}(\cdot|t_k).$$

Theorem C.1. *For any probability distribution function $\tilde{F}_d$ supported on $\mathcal{Y}_d \equiv \{y_{d,1}, \cdots, y_{d,K_d}\}$,* *suppose there always exists a sequence* $\{c_{d,1}, \cdots, c_{d,K_\eta}\}$ *such that*

$$\tilde{F}_d(\cdot) = \sum_{k=1}^{k_\eta} c_{d,k} F_{Y_d|\eta}(\cdot|t_k), \tag{C.10}$$

Then, Condition C.1 holds if and only if (i) $K_0 = K_1$ *and (ii) for some strictly increasing mapping* $\psi : \{y_{0,1}, \cdots, y_{0,K_0}\} \to \{y_{1,1}, \cdots, y_{1,K_1}\}$ *and some* $\lambda : \{y_{0,1}, \cdots, y_{0,K_0}\} \to \mathbb{R}_+$,

$$F_{Y_0|\eta}(y|t_k) = \lambda(y) F_{Y_1|\eta}(\psi(y)|t_k), \qquad \text{for } y \in \mathcal{Y}_0, k = 1, \cdots, K_\eta. \tag{C.11}$$

The condition C.10 is a rank condition as the rank of matrix $\{F_{Y_d|\eta}(y_{d,j}|t_{j'}) : j = 1, ..., K_d, \quad j' = 1, \cdots, k_\eta\}$ should be no smaller than K_d. A necessary condition is $K_\eta \geq K_d$, namely, the support of η is no coarser than the support of Y_d. The rank condition would be violated when there is no endogeneity (i.e., $Y_d \perp \eta$), which is not our focus. Again, 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 bound analysis.

Proof. By Condition C.1, we have

$$\sum_{k=1}^{K_\eta} \delta_k F_{Y_1|\eta}(\cdot|t_k) \leq 0 \text{ if and only if } \sum_{k=1}^{K_\eta} \delta_k F_{Y_0|\eta}(\cdot|t_k) \leq 0 \tag{C.12}$$

for any K_η-dimensional vector $\delta \equiv (\delta_1, \cdots, \delta_n)$ that satisfies $\sum_{k=1}^{K_\eta} \delta_k = 0$.

Note that $\big(F_{Y_1|\eta}(y|t_1), \cdots, (F_{Y_1|\eta}(y|t_{K_\eta})\big)$ for each $y \in \mathcal{Y}_1/\{y_{K_1}\}$ generates an extreme ray of the $(K_\eta - 1)$-dimensional polar cone of a cone

$$\left\{\delta \in \mathbb{R}^n : \sum_{k=1}^{K_\eta} \delta_k F_{Y_1|\eta}(\cdot|t_k) \leq 0; \sum_{k=1}^{K_\eta} \delta_k = 0\right\}.$$

A similar argument holds for $\big(F_{Y_0|\eta}(\cdot|t_1), \cdots, (F_{Y_0|\eta}(\cdot|t_{K_\eta})\big)$. By (C.12), these two polar cones are the same. Therefore, for each $y_k \in \mathcal{Y}_1/\{y_{K_1}\}$, there exists a $y_k^* \in \mathcal{Y}_0/\{y_{K_0}\}$ such that

$$\big(F_{Y_0|\eta}(y_k^*|t_1), \cdots, F_{Y_0|\eta}(y_k^*|t_{K_0}))\big) = \lambda(y_k) \times \big(F_{Y_1|\eta}(y_k|t_1), \cdots, F_{Y_1|\eta}(y_k|t_{K_1}))\big).$$

Finally it is easy to show that if $y_k < y_{k'}$ then $y_k^* < y_{k'}^*$ and thus $\psi(y_k) = y_k^*$ is a strictly increasing function.

C.8 Proof of Theorem 4.1

We suppress X for simplicity. The proof is analogous to that of Theorem 2.1. Using the same notation as the earlier proof, (4.1) can be rewritten as

$$\begin{aligned}\sum_{\ell=1}^{\ell^*} \frac{q(z_\ell) - \gamma_\ell p(z_\ell)}{a} \times P[Y_1 \leq y|D=1, Z=z_\ell] + \sum_{\ell=\ell^*+1}^{L} \frac{q(z_\ell)}{a} \times P[Y_1 \leq y|D=1, Z=z_\ell]& \\ = \sum_{\ell=\ell^*+1}^{L} \frac{\gamma_\ell p(z_\ell)}{a} \times P[Y_1 \leq y|D=1, Z=z_\ell]&.\end{aligned}$$

The above equation being satisfied as equality can be viewed as being satisfied as inequalities "$\leq$" and "$\geq$." Therefore, by Condition S_1 applied for both inequalities, we have

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

Equivalently, we have

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

by $\sum_{\ell=1}^{L} \gamma_\ell = 0$. $\square$

C.9 Proof of Lemma B.1

Part (i) can be shown analogous to the proof of Theorem 3.1. Suppose

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

holds for some w and $\tilde{w}$. We want to show that

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

First, because of the strict monotonicity of $q(d,x,\cdot)$, we have

$$\int w(t,x)F_{U_1|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x)F_{U_1|\eta,X}(\cdot|t,x)dt$$

and it suffices to show

$$\int w(t,x)F_{U_0|\eta,X}(\cdot|t,x)dt \leq \int \tilde{w}(t,x)F_{U_0|\eta,X}(\cdot|t,x)dt.$$

Second, for any $v \in \text{Supp}(V|X=x)$, because of the strict monotonicity of $\phi(\cdot,v)$, we have $1(U_1 \leq u_1) \overset{a.s.}{=} 1(\phi(U_1,v) \leq \phi(u_1,v))$. Because $V \perp (U_1,\eta)|X$, we have

$$\int w(t,x)F_{\phi(U_1,V)|\eta,X,V}(\phi(\cdot,v)|t,x,v)dt \leq \int \tilde{w}(t,x)F_{\phi(U_1,V)|\eta,X,V}(\phi(\cdot,v)|t,x,v)dt$$

and thus,

$$\int w(t,x)F_{U_0|\eta,X,V}(\phi(\cdot,v)|t,x,v)dt \leq \int \tilde{w}(t,x)F_{U_0|\eta,X,V}(\phi(\cdot,v)|t,x,v)dt.$$

Conditional on (η, X, V), $\mathrm{Supp}(\phi(U_1, v)) = \mathrm{Supp}(\phi(U_1, V)) = \mathrm{Supp}(U_0)$. Therefore, for u_0 in that support,

$$\int w(t,x) F_{U_0|\eta,X,V}(u_0|t,x,v)dt \leq \int \tilde{w}(t,x) F_{U_0|\eta,X,V}(u_0|t,x,v)dt.$$

It follows that

$$\begin{aligned} &\int \int w(t,x) F_{U_0|\eta,X,V}(u_0|t,x,v) f_{V|X}(v|x)dtdv \\ \leq &\int \int \tilde{w}(t,x) F_{U_0|\eta,X,V}(u_0|t,x,v) f_{V|X}(v|x)dtdv \end{aligned}$$

Note that $f_{V|X} = f_{V|\eta,X}$. Then, by the law of iterated expectation, we have

$$\int w(t,x) F_{U_0|\eta,X}(u_0|t,x)dt \leq \int \tilde{w}(t,x) F_{U_0|\eta,X}(u_0|t,x)dt.$$

Next, we prove part (ii) by first noting that

$$\Pr[U_0 \leq u|\eta, X] = \Pr[\phi(U_1) \leq u, V \leq u|\eta, X] = \Pr[\phi(U_1) \leq u|\eta, X] \Pr[V \leq u|X].$$

Therefore,

$$\begin{aligned} F_{Y_0|\eta,X}(y|t,x) &= \Pr[g(0,x,U_0) \leq y|\eta = t, X = x] = \Pr[U_0 \leq g^{-1}(0,x,y)|\eta = t, X = x] \\ &= \Pr[\phi(U_1) \leq g^{-1}(0,x,y)|\eta = t, X = x] \Pr[V \leq g^{-1}(0,x,y)] \\ &= \Pr[Y_1 \leq g(1,x,\phi^{-1}(g^{-1}(0,x,y)))|\eta = t, X = x] \Pr[V \leq g^{-1}(0,x,y)] \\ &= F_{Y_1|\eta,X}(\psi(y,x)|t,x)\lambda(y,x), \end{aligned}$$

where $\psi(y,x) \equiv g(1,x,\phi^{-1}(g^{-1}(0,x,y)))$ and $\lambda(y,x) \equiv F_V(g^{-1}(0,x,y))$. $\square$

D 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 D.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 D.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 D.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 D.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, 2.3

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. 2.2

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

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261. 1, 2.1, 3, 4, B

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

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441. 1

——— (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550. 1

D'Haultfœuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83, 1199–1210. 1

Dong, Y. and S. Shen (2018): "Testing for rank invariance or similarity in program evaluation," *Review of Economics and Statistics*, 100, 78–85. 1, 3.1

Frandsen, B. R. and L. J. Lefgren (2018): "Testing rank similarity," *Review of Economics and Statistics*, 100, 86–91. 1, 3.1

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, 1, 2.1

Jun, S. J., J. Pinkse, and H. Xu (2011): "Tighter bounds in triangular systems," *Journal of Econometrics*, 161, 122–128. 1

Kim, J. H. and B. G. Park (2022): "Testing rank similarity in the local average treatment effects model," *Econometric Reviews*, 1–22. 1, 3.1

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. 2.2

——— (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, 2.3

Pomatto, L., P. Strack, and O. Tamuz (2020): "Stochastic dominance under independent noise," *Journal of Political Economy*, 128, 1877–1900. B

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688. 1

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

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83, 1185–1197. 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, 2.1

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341. 2.3

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