cemmap

centre for microdata methods and practice

Layered policy analysis program evaluation using the marginal treatment effect

Ismael Mourifié
Yuanyuan Wan

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP21/21

Economic and Social Research Council

LAYERED POLICY ANALYSIS IN PROGRAM EVALUATION USING THE MARGINAL TREATMENT EFFECT

ISMAEL MOURIFIÉ[†] AND YUANYUAN WAN[‡]

ABSTRACT. This paper proposes a unified approach to derive sharp bounds on all conventional policy parameters when the instrumental variables (IVs) are potentially invalid. Using a *Vine Copula* approach, we propose a novel characterization of the identified sets for the marginal treatment effect (MTE) and the policy-relevant treatment effect (PRTE) parameters. Our method has various advantages: First, it explicitly demonstrates how imposing different IV-related assumptions with different credibility levels affects the MTE and PRTE's identified set. Second, it can be used to test model specifications and hypotheses about various imperfect IV-related assumptions. Third, it provides a tractable way to inform policy choices in the presence of uncertainty of the validity of identifying assumptions. Our approach enlarges the MTE framework's scope by showing how it can be used to inform policy decisions even when valid instruments are not available.

Keywords: Desegregated MTE, Vine Copula, Identified Set, Policy-relevant treatment effect, Specification Test.

JEL Classification: C12, C14, C21 and C26

This version: Sunday 4th April, 2021

We are grateful to Peter Hull, Désiré Kédagni, Vitor Possebom, Pedro Sant'Anna, Thomas Russell, Alex Torgovitsky for valuable comments and/or discussions. We have greatly benefited from insightful comments from Jim Heckman. The research was conducted in part when Ismael Mourifié was visiting the Becker Friedman Institute (BFI) at the University of Chicago. Mourifié thanks his hosts for its hospitality and support. Wan and Mourifié thank the support from the SSHRC Insight Grant #435190500.

[†] Corresponding author. Department of Economics, University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada. ismael.mourifie@utoronto.ca

[‡] Department of Economics, University of Toronto. yuanyuan.wan@utoronto.ca.

1. INTRODUCTION

Evaluating the impact of an intervention is fundamental for policymakers. It generates knowledge about a program's effectiveness and determines whether it should be scaled up, down, or discontinued. However, the program (treatment) effects may vary widely across economic agents, and expectations about individual treatment effects may trigger strategic participation. In such an environment, uncovering aggregate treatment effect parameters and using them as baseline information to evaluate new policies is challenging. Heckman and Vytlacil (2005, HV05 hereafter) propose a key causal parameter: the *marginal treatment effect* (MTE). The identification of the MTE allows researchers to recover conventional causal parameters of interest, such as average treatment effect (ATE), Local ATE (LATE), and the ATE on treated/untreated (ATT/ATUT). It also allows researchers to evaluate new policies through the policy-relevant treatment effect parameter (PRTE). Since its introduction, various approaches have been proposed to identify the MTE and then the PRTE. HV05 requires the treatment selection to be defined by a single threshold crossing model —which imposed a monotonicity restriction, see (see Vytlacil, 2002)—and a continuous instrument. Recently, Brinch, Mogstad, and Wiswall (2017), Mogstad, Santos, and Torgovitsky (2018) shows that the MTE can be recovered even in the presence of discrete instruments but at the cost of imposing some parametric or shape restrictions. Lee and Salanié (2018) relaxes the single threshold selection rule and shows the identification of the MTE in the presence of multiple thresholds.

However, all existing MTE identification strategies strongly rely on the availability of valid instruments. The valid instruments assumption often creates a great deal of controversy amongst economists, see discussions in Deaton (2009) and Deaton, Heckman, and Imbens (2010). Manski (2011) questioned the "credibility" of policy predictions based on parameters obtained under doubtful, contestable, or non-testable restrictions and asserted that it would be harmful to policy choice. Then, there is a clear tension between the strength of the assumptions used to recover the MTE and the "credibility" of any policy recommendations based on it. One way to resolve this tension, as advocated by Manski (2011), is what he referred as *layered policy analysis*. *Layered policy analysis* demands researchers to visit various assumptions at different levels of credibility and analyze how this affects policy predictions.[1]

[1]Regarding the *layered analysis*, Manski (2011, F289) said: "*A researcher who performs an instructive layered policy analysis and exposits work clearly may see himself as having accomplished the objective of informing choice.*"

This paper's first main contribution is to show how one can use a modified version of the MTE to perform informative and credible policy analyses on conventional policy parameters, specifically the PRTE. To accommodate possibly invalid instruments, we introduce a modified MTE parameter, namely the *disaggregated marginal treatment effect* (DMTE). To fix the idea, consider the model $Y = Y_1 D + Y_0(1 - D)$ and $D = 1\{P(Z) \geq V\}$, where Y_1 and Y_0 are potential outcomes, D is the treatment, Y is the observed outcome, Z are (possibly invalid) instrument variables, and V is independent with $P(Z)$ and is normalized to have $U[0,1]$ distribution. We define the DMTE as the expectation of treatment effect conditioning on V and $P(Z)$ —the propensity score, that is, $\text{DMTE}(v, p) \equiv \mathbb{E}[Y_1 - Y_0 | V = v, P(Z) = p]$.[2] We show that all the aforementioned conventional policy parameters, including the MTE, can be expressed as a weighted average of the DMTE under the threshold-crossing treatment selection rule only, making the DMTE a more primitive parameter than the MTE. Unlike the MTE, the mapping between the DMTE and other policy parameters does not require any element of Z to be a valid instrument. Furthermore, the weights are directly identifiable from the data. Therefore, we can partially identify any of the conventional policy parameters as long as the identified set for DMTE is available.

Secondly, we propose a *Vine Copula* approach to partially identify the DMTE. We show that under the assumption that V is independent with $P(Z) \equiv P$, the dependence structure among variables (Y_d, P, V) is fully captured by two copula functions. The first one is the conditional copula of Y_d and V given P, i.e. $C_{Y_d,V|P}(.,.)$. This copula characterizes the endogenous selection in the model. If $C_{Y_d,V|P}(.,.)$ takes a product form, then we have "selection on observables". There is no issue of endogenous selection once P is controlled; otherwise, "selection on unobservables" exists. The other copula function is $C_{Y_d,P}(.,.)$, which captures the dependence between the potential outcomes and the propensity score. This copula measures the "quality of instruments" and is a key function that we investigate in this paper. For example, if $P(Z)$ is a valid instrument, as assumed in existing literature, then it must be the case that $C_{Y_d,P}(x_1, x_2) = x_1 x_2$. Therefore, we can view the IV independence assumption as a shape restriction on the unknown function $C_{Y_d,P}(x_1, x_2)$. Under the copula formulation, we show that calculating the identified set of the DMTE boils down to finding

[2] Like the LATE, the DMTE itself is an instrument-dependent parameter. It is not used here for causal interpretation but rather as an intermediary quantity to recover conventional policy parameters.

the set of conditional bivariate copulas that respect a set of *equality constraints* and any additional constraints that researchers would like to impose on $C_{Y_d,P}(x_1,x_2)$.

In our *Vine Copula* characterization, the identified set of DMTE depends on restrictions that we impose on $C_{Y_d,P}(x_1,x_2)$ in an explicit way. At one extreme, when we impose the IV-independence assumption, the set of equality constraints pins down a unique copula $C_{Y_d,V|P}(.,.)$, which therefore allows the point identification of the MTE and then other policy parameters. At the other extreme, where we impose no restrictions on $C_{Y_d,P}(.,.)$, our characterization recovers the sharp bounds on the DMTE under only the single threshold crossing assumption. One can also impose restrictions that are weaker than IV-independence. For instance, we show that imposing the Monotone IV assumption—see Manski and Pepper (2000)— is equivalent to considering only the set of copulas $C_{Y_d,P}(x_1,x_2)$ that are concave in x_2. In this case, we recover the sharp identified set under monotone IV. As we demonstrate in more detail in the main text, our approach, in general, provides empirical researchers a very flexible way to derive the identified set on the DMTE under any dependence restrictions she is willing to impose between the IV and the potential outcomes. From this perspective, our method shares the same spirit as the *layered policy analysis* discussed in Manski (2011).

To make our approach more operational, we also consider a *semi-parametric* version of our characterization by imposing parametric assumptions on the copulas but leaving the marginals entirely nonparametric. In this case, we show that the identification analysis boils down to constructing identified sets for only the copula function's finite-dimensional parameters. As discussed in Chen, Fan, and Tsyrennikov (2006), using this type of semi-parametric approach to study multivariate distributions has gained popularity in diverse fields for its flexibility and ability to circumvents the curse of dimensionality.

It is worth noting that the assumptions — IV-independence and monotonicity assumptions— imposed by HV05 to identify the MTE have testable implications (see HV05 Appendix A). As a by-product contribution, we provide a more tractable characterization of MTE assumptions' sharp testable implications. We show that the proposed characterization is sufficient to screen all possible observable violations of the MTE assumptions, and one can test it using existing inferential methods, e.g., Hsu, Liu, and Shi (2019). The rejection of the testable implication demands one to relax some of the MTE assumptions, and in such cases, our proposed layered analysis provides a possible solution. Our approach can also be used to perform specification tests for various alternative imperfect IV-related assumptions a researcher would like to consider.

Finally, we show that our identification strategy for MTE and PRTE also applies to models in which multiple thresholds on multiple unobserved heterogeneities define the selection equation. Therefore, this current paper (i) extends the Lee and Salanié (2018) identification approach to the case when none of the existing continuous covariates satisfies the IV-independence assumption; and (ii) also applies to the so-called "actual monotonicity" recently discussed in Mogstad, Torgovitsky, and Walters (2019).

We organize the rest of the paper as follows. In Section 2, we introduce the intermediate quantity DMTE and build its connection with other policy parameters. We characterize the identified set for DMTE under various assumptions on the IV and also discuss the implementation of our method in a semi-parametric setup in Section 3. We extend the analysis to a two-threshold model in Section 4. Section 5 concludes the paper.

2. Policy Parameters and disaggregated MTE

We adopt the framework of the potential outcomes model: $Y = Y_1 D + Y_0(1-D)$, where $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the observed outcome taking values from the support $\mathcal{Y}$, $D \in \{0,1\}$ is the observed treatment indicator, and (Y_1, Y_0) are potential outcomes. Heckman and Vytlacil (1999) trace the genealogy of this model, and we refer to them for terminology and attribution. Let Z be a vector of covariates taking values from the support $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ for $d_z \geq 1$. The following Assumptions 1 and 2 are required for the point identification of the MTE nonparametrically:

Assumption 1 (Single Threshold-Crossing: STC)**.** *The selection mechanism is governed by the following threshold crossing model $D = 1\{\nu(Z) \geq V\}$ for some measurable and non-trivial function ν, where V follows a uniform distribution over the interval $[0,1]$ and is statistically independent of the vector of covariates Z, i.e., $Z \perp V$.*[3]

Hereafter, we use the shorthand notation $P(Z)$ for $\mathbb{P}(D = 1|Z)$, a quantity directly recoverable from data. We can see that under Assumption 1, $\nu(\cdot)$ is identified over the support of Z since $P(z) = \mathbb{P}(V \leq \nu(z)|Z = z) = \nu(z)$ for all $z \in \mathcal{Z}$. When it cause no confuse, we will use the

[3] V follows uniform distribution is a normalization. Please see Vytlacil (2002) for primitive conditions under which this normalization is without loss of generality.

shorthand notation P or p to denote $P(Z)$ or $P(z)$, respectively. Let $\mathcal{P} \subseteq [0,1]$ denote the support of $P(Z)$ and $P(z)$ a generic element of $\mathcal{P}$. For the main text, we restrict our attention to the cases where the set of limit points of $\mathcal{P}$, denoted by $L(\mathcal{P})$, is non-empty. This excludes the case where P is discrete.

Assumption 2 (IV Independence)**.** *Conditioning on the first stage unobserved variable V, the propensity score P is statistically independent with the potential outcomes, i.e. $P(Z) \perp Y_d|V$ for $d = 0,1$.*

Assumptions 1 and 2 and the whole remaining analysis, distributional assumptions and theoretical results, are understood to be conditional on a set of observed covariates X, which will be omitted from the notation for sake of simplicity. In the following, we will first review why MTE, or the marginal treatment response (MTR), is not point identified without Assumption 2. Then we will examine the restrictions that can be used for partial identification. Let $g : \mathcal{Y} \to \mathbb{R}$ be a real integrable function such that $\mathbb{E}[|g(Y_d)|] < \infty$. Taking $d = 1$ as illustration and following the identification strategy of HV05, for all $p \in L(\mathcal{P})$:

$$\begin{aligned}\mathbb{E}[g(Y)D|P=p] &= \mathbb{E}[g(Y_1)|D=1,P=p]\mathbb{P}(D=1|P=p)\\ &= \mathbb{E}[g(Y_1)|V\le p,P=p]\mathbb{P}(V\le p|P=p) = \mathbb{E}[g(Y_1)|V\le p,P=p]p\\ &= \int_0^p \mathbb{E}[g(Y_1)|V=v,P=p]dF_{V|P=p} = \int_0^p \mathbb{E}[g(Y_1)|V=v,P=p]dv,\end{aligned}$$

where all equalities holds only under Assumption 1. The key equation is then:

$$\mathbb{E}[g(Y)D|P=p] = \int_0^p \mathbb{E}[g(Y_1)|V=v,P=p]dv. \tag{1}$$

By taking the derivative of the previous equation respect with p we obtain:

$$\frac{\partial}{\partial p}\mathbb{E}[g(Y)D|P=p] = \mathbb{E}[g(Y_1)|V=p,P=p] + \int_0^p \frac{\partial}{\partial p}\mathbb{E}[g(Y_1)|V=v,P=p]dv. \tag{2}$$

It can be seen that the left hand side of the equation, also known as the the Local IV (LIV) estimand proposed by HV05, can no longer identify the MTR (and MTE) because (i) $\frac{\partial}{\partial p}\mathbb{E}[g(Y_1)|V=v,P=p] \neq 0$, and (ii) $\mathbb{E}[g(Y_1)|V=p,P=p]$ is in general different from $\mathbb{E}[g(Y_1)|V=p]$ when P and Y_1 are not independent conditioning on V. Nevertheless, Equation (1) still contains useful information about the quantity

$$\theta_g^d(v,p) \equiv \mathbb{E}[g(Y_d)|V=v, P=p], \quad d=0,1,(v,p)\in[0,1]\times\mathcal{P},$$

which we refer in the rest of paper as the *disaggregated marginal treatment responses* with respect to the function g and abbreviate it as DMTR_g. Analogous to the relationship between MTR and MTE, we define another intermediate quantity *disaggregated marginal treatment effect* (DMTE) as:

$$\text{DMTE}(v,p) \equiv \mathbb{E}[Y_1 - Y_0|V=v, P=p], \ \ \forall (v,p)\in[0,1]\times\mathcal{P}. \tag{3}$$

DMTR implies DMTE since $\text{DMTE}(v,p) = \theta_g^1(v,p) - \theta_g^0(v,p)$ with $g(\cdot)$ being chosen as the identity function.[4] [5] It is apparent from Equation (3) that the identification of MTR (hence MTE) is readily available once DMTR is recovered since

$$\mathbb{E}[g(Y_d)|V=v] = \int_0^1 \mathbb{E}[g(Y_d)|V=v, P=t] f_{P|V}(t|v)dt = \int_{\mathcal{P}} \theta_g^d(v,t) f_P(t)dt,$$

and

$$\text{MTE}(v) \equiv \mathbb{E}[Y_1 - Y_0|V=v] = \int_{\mathcal{P}} \text{DMTE}(v,t) f_P(t)dt,$$

where $f_{P|V} = f_P$ by Assumption 1 and the density f_P of P is directly identified from data. Note that when $P(Z) \perp Y_d|V$ as in HV05, the DMTE is exactly equal to the MTE and we have

$$\text{DMTE}(v,p) = \text{DMTE}(v,p') = \text{MTE}(v), \tag{4}$$

for all $(p,p') \in \mathcal{P}\times\mathcal{P}$ and $v\in[0,1]$. Therefore, although the DMTR and DMTE are not necessarily parameters of direct interest—for being instrument-dependent—they do serve as useful intermediate quantities to identify the MTR, MTE, and other useful policy parameters such as the ATE, ATT, ATUT. Specifically, we will show in Theorem 1 below that all these mentioned parameters can be expressed as a weighted average of DMTE under Assumption 1 only.

Another useful parameter that often draws interest is the policy relevant treatment effect (PRTE_g):

[4] One can also recover its distributional version by choosing $g(Y_d) = \mathbf{1}[Y_d \le y]$, i.e. $\mathbb{P}(Y_1 \le y|V=v, P=p) - \mathbb{P}(Y_0 \le y|V=v, P=p), \forall (v,p)\in[0,1]\times\mathcal{P}$.

[5] It is worth-noting that our DMTE shares a superficial resemblance with the Redefined MTE ($\widetilde{MTE}$) introduced in Zhou and Xie (2019). In presence of a vector of exogenous covariates X, $\widetilde{MTE}(v,p) = \mathbb{E}[Y_1 - Y_0|V=v, P(Z,X)=p]$ while $\text{DMTE}(v,p,x) = \mathbb{E}[Y_1 - Y_0|V=v, P(Z,X)=p, X=x]$. Unlike the DMTE, when $X|P(X,Z)$ has a degenerate distribution, $\widetilde{MTE}$ is equal to the MTE, please see Zhou and Xie (2019, Page 3076, eq 8).

$$\text{PRTE}_g \equiv \frac{\mathbb{E}[g(Y)|\text{a}'] - \mathbb{E}[g(Y)|\text{a}]}{\mathbb{E}[D|\text{a}'] - \mathbb{E}[D|\text{a}]},$$

where a' and a denote the alternative policy under consideration and the baseline policy, respectively. Please refer to Heckman and Vytlacil (2001), HV05, and Carneiro, Heckman, and Vytlacil (2010) for a detailed discussion about the PRTE.

Hereafter, Let Y_d^a denote the potential outcome when the agent is externally set to treatment d under policy a. In HV05, the identification of PRTE rely on both Assumptions 1 and 2 and the following policy invariant assumption:

Assumption 3 (HV05, Policy Invariance). *$(Y_d^{a'}, V^{a'}) \sim (Y_d^a, V^a)$ for $a \neq a'$.*

Under Assumptions 1 to 3, HV05 shows that

$$\text{PRTE}_g = \int_0^1 \underbrace{\frac{F_{P^a}(v) - F_{P^{a'}}(v)}{\mathbb{E}_{F_{P^{a'}}}[P] - \mathbb{E}_{F_{P^a}}[P]}}_{w^{PRTE}(v)} \text{MTE}_g(v)dv.$$

The PRTE is designed by HV05 to evaluate a new policy that induces a change in P but keeping the full joint distribution of latent variable unchanged from the baseline policy to the targeting alternative policy, when Assumptions 1 to 3 hold. In the following, we will propose an alternative policy invariance assumption and show that the PRTE_g can be recovered even when Assumption 2 fails to hold.

Assumption 4 (Conditional Policy Invariance). *$Y_d^{a'}|V^{a'}, P^{a'} \sim Y_d^a|V^a, P^a$ with $V^{a'} \sim U[0,1]$, $V^a \sim U[0,1]$ and $Y_d^{a'} \sim Y_d^a$ for $a \neq a'$.*

In Assumption 4, we maintain the same normalization for the distribution of V in both baseline policy and alternative policy environments. Notice that under Assumption 2, Assumption 4 holds if and only if Assumption 3 holds. Therefore, if the independence assumption holds, our conditional policy-invariance becomes equivalent to the HV05's policy-invariance assumption. The difference is that without the IV-independence assumption, Assumption 4 requires the DMTE_g^a to be invariant from a policy a to an alternative policy a', namely $\text{DMTE}_g^{a'}$=DMTE_g^a. Unlike HV05, our conditional policy-invariance assumption does not impose the invariance of the MTE_g^a.

Theorem 1 below shows that the PRTE_g can be written as a weighted average of the DMTE_g under Assumptions 1 and 4. We now summarize the above discussion in the following Theorem.

Theorem 1. *Suppose that Assumption 1 is satisifed, then*

(i) $MTE_g(v) = \int_0^1 DMTE_g(v,p) f_P(p) dp$;

(ii) for any $\mathbf{s} \in \{ATE_g, LATE_g(u,u'), ATT_g, ATUT_g\}$[6] *and weights* $\omega^s(v,p)$ *listed in Table 1 below, we have*

$$\mathbf{s} = \int_0^1 \int_0^1 \omega^s(v,p) DMTE(v,p) dv dp. \tag{5}$$

(iii) If in addition Assumption 4 holds, Equation (5) holds with $\mathbf{s} = PRTE$.

TABLE 1. Policy Parameters and DMTE

Parameters	*weights* $\omega^s(v,p)$
ATE_g	$f_P(p)$
ATT_g	$\frac{f_P(p) 1\{v<p\}}{\mathbb{E}[P]}$
$ATUT_g$	$\frac{f_P(p) 1\{v>p\}}{\mathbb{E}[1-P]}$
$LATE_g(u,u')$	$\frac{f_P(p) 1_{\{u<v\leq u'\}}}{u'-u}$
$PRTE_g$	$\frac{[f_{pa'}(p) - f_{pa}(p)] 1\{v\leq p\}}{\mathbb{E}_{F_{pa'}}[P] - \mathbb{E}_{F_{pa}}[P]}$

Proof. See Appendix A.1. □

Notice that one can easily verify that when Assumption 2 holds, $\text{DMTE}_g(v,p) = \text{DMTE}_g(v)$ and then $\int_0^1 w^s(v,p)dp = w^s(v)$ for any $\mathbf{s} \in \{\text{ATE}_g, \text{LATE}_g(u,u'), \text{ATT}_g, \text{ATUT}_g, \text{PRTE}_g\}$, with $w^s(v)$ being exactly the weights derived in HV05. Although DMTE_g itself may or may not be the main parameter of interest, Theorem 1 shows that it plays an important role in the identification of many common parameters of interest. Note that the weights are known and can be estimated for each value $(v,p) \in [0,1] \times \mathcal{P}$. Thus, we can readily recover the identified sets for any of the conventional policy parameters once we have the identified set for the DMTR_g (hence DMTE_g). Therefore, our main goal will be to provide a tractable characterization of the identified set for the DMTRs.

[6] Here $\text{LATE}_g(u,u')$ represents the average treatment effect for the group of compliers when P is externally changed from u to u'.

Definition 1. *Suppose that Assumption 1 is satisfied. For any integrable real function $g(.)$, the identified set $\Theta_{I,g}$ for DMTR$_g$ is defined as follows:*

$$\Theta_{I,g} = \Bigg\{ (\theta_g^0, \theta_g^1) : [0,1] \times [0,1] \to \mathbb{R}^2 \textit{ such that}$$
$$\mathbb{E}[g(Y)1\{D = d\}|P = p] = \int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} \theta_g^d(v,p) dv \textit{ for } d \in \{0,1\} \Bigg\}.$$

In the main text, we focus on the half-interval class $\mathcal{G} \equiv \{g(\cdot) = \mathbf{1}[\cdot \leq y], y \in \mathcal{Y}\}$ when identifying DMTRs. Under the half interval class, the DMTR$_g^d$ can then be expressed as

$$\text{DMTR}_g^d(v,p) = \mathbb{P}[Y_d \leq y | V = v, P = p] \equiv F_{Y_d|V,P}(y|v,p) \quad d \in \{0,1\}.$$

Here, the DMTR$_g^d$ is just the conditional distribution function of Y_d given P and V. Therefore, the identification under the interval class recovers DMTRs for other classes of g functions.

3. Identification

In the previous section, we show that there exist intermediate quantities DMTE$_g$ or DMTR$_g$, which facilitate the partial identification of the MTE$_g$ and the MTR$_g$ without imposing the IV-independence Assumption 2. In this section, we will take the STC structure (Assumption 1) as given and characterize identified sets for the MTE$_g$ under a sequence of assumptions on the copula of (P, Y_d). In this sequence of assumptions, we have "no restrictions"[7] and the "IV-independence assumption" representing two extreme cases.

3.1. **Identification under a single threshold-crossing selection rule.** Once we focus on the half-interval class of g functions, the primitive parameter of interest, as defined in Definition 1, is the conditional distribution of Y_d given V and P. While we do not know precisely the full joint distribution, the STC structure (Assumption 1) does provide some restrictions. For instance, we know (or can directly identify from data) the distribution of two of three marginal distributions, i.e., V and P, and we know that they are independent. This feature suggests it is convenient to use copulas decomposition to study the joint distribution of interest. Copula theory is useful to separate marginal properties from properties related to the dependence structure. Here we cite Sklar (1959)'s result:

[7]We do, however, impose certain regularity conditions, as will be made clear later.

Lemma 1 (Sklar (1959)'s Theorem)**.** *There exist a copula* $\mathrm{C} : [0,1]^3 \to [0,1]$ *such that*

$$\mathbb{P}(Y_d \leq y, V \leq v, P \leq p) = C_{Y_d,V,P}(F_{Y_1}(y), F_V(v), F_P(p)), \quad \textit{for } y, v, p \in [-\infty, \infty].$$

Moreover, if the margins are continuous, then C *is unique; otherwise it is uniquely determined on* $RanF_{Y_d} \times RanF_V \times RanF_P$ *where* $RanF_X = F_X([-\infty, \infty])$ *is the range of* F_X*.*

Using the Sklar's result, we can decompose the joint trivariate distribution into three univariate distributions and one trivariate copula $C_{Y_d,V,P}(F_{Y_1}(y), F_V(v), F_P(p))$. However, working directly with the trivariate copula is not very convenient. Unlike the bivariate copula, the dependence of trivariate copula can be less intuitive to interpret. Also, the number of multivariate (> 2) parametric copula families with flexible dependence is limited. Furthermore, the STC assumption (Assumption 1) already provides some restrictions, such as the independence of P and V, and their known marginals. To fully take advantage of that information, we consider the *Vine Copula* approach, which was introduced by Joe (1994) to break down the dependence structure of a multivariate copula into a sequence of bivariate copulas and conditional bivariate copulas. The Vine copula approach has proven to be useful in various existing problems such as (constrained) sampling of correlation matrices, building non-parametric continuous Bayesian networks, and various applications in finance. Here, we will make use of the Vine copula in our treatment effect context. To this end, we consider the following regularity assumption:

Assumption 5. *The joint distribution of* (Y_d, V, P) *is absolutely continuous respect to the Lebesgue measure.*

Here we make Assumption 5 only for the ease of notation. V is continuous by definition, and P is continuous if Z contains a continuous element. It is worth noting that we do not require Z be valid instruments in the sense of being independent of $Y_d, d \in \{0,1\}$. Then the presence of any continuous exogenous covariates in the selection equation ensures P is continuous. Therefore, the main restriction of Assumption 5 is to focus the analysis on applications with continuous outcomes. Lemma 2 and theorem 2 below can be straightforwardly extended to the case of discrete outcome variables at the cost of additional notation, see Appendix A.9 for details.

Let $C_{Y_d,V|P=p}(F_{Y_d|P}(y|p), F_{V|P}(v|p); p) = F_{Y_d,V|P}(y, v|p)$ be the conditional copula of (Y_d, V) given $P = p$. Note that given our assumption that $V|P = p \sim U[0,1]$ for all p, the second term in the parenthesis can be simplified to v since $F_{V|P}(v|p) = v$, that is, $C_{Y_d,V|P=p}(F_{Y_d|P}(y|p), v; p)$. Under Assumption 5, the following Lemma expresses all the conditional distributions in terms of copulas.

Lemma 2 (Vine Copula). *Under Assumptions 1 and 5, for $d \in \{0,1\}$ we have for each $y \in \mathcal{Y}$ and $p \in \mathcal{P}$,*

$$F_{Y_d|P}(y|p) = \frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2)\Big|_{x_1=F_{Y_d}(y), x_2=F_P(p)} \equiv c_{d,F_P(p)}(F_{Y_d}(y)), \tag{6}$$

$$F_{Y_d|V,P}(y|v,p) = \frac{\partial}{\partial x_2} C_{Y_d,V|P=p}(x_1, x_2)\Big|_{x_1=F_{Y_d|P}(y|p), x_2=v} \tag{7}$$

and there exists strictly increasing mappings $\Psi_{1,p}$ and $\Psi_{0,p}$ such that for each $y \in \mathcal{Y}$ and $p \in \mathcal{P}$,

$$\mathbb{P}[Y \le y, D = 1|P = p] = \Psi_{1,p}(F_{Y_1}(y)) \equiv C_{Y_1,V|P=p}\Big(c_{1,F_P(p)}(F_{Y_1}(y)), p; p\Big). \tag{8}$$

$$\begin{aligned}\mathbb{P}[Y \le y, D = 0|P = p] &= \Psi_{0,p}(F_{Y_0}(y)) \\ &\equiv c_{0,F_P(p)}(F_{Y_0}(y)) - C_{Y_0,V|P}\Big(c_{0,F_P(p)}(F_{Y_0}(y)), p; p\Big).\end{aligned} \tag{9}$$

That is, the observed probability $\mathbb{P}[Y \le y, D = d|P = p]$ depends on y only through $F_{Y_d}(y)$.

Proof. See Appendix A.2. □

We have some remarks on the usefulness of the Lemma 2. First, by inserting Equation (6) to Equation (7), we observe that the parameter of interest DMTR (which equals $F_{Y_d|V,P}(y|v,p)$ under the half-interval class) can be considered as a composite mapping from $F_{Y_d}(y)$, where the mapping depends only on the (partial derivatives of) two bivariate copula functions: $C_{Y_d,V|P}$ and $C_{Y_d,P}$. Instead of working with $F_{Y_d|P,V}$, which involves with dependence structure of three variables, we can now focus on two bivariate copulas. To be more specific, let $\mathcal{F}^c$ be the set of continuous CDFs, $\mathcal{C}_d^c$ be the set of conditional copulas for $C_{Y_d,V|P}$, and $\mathcal{C}_d$ be the set of copulas for $C_{Y_d,P}$, the identification of DMTR is equivalent to the identification of $\tilde{\theta} \equiv \big(F_{Y_1}, F_{Y_0}, C_{Y_1,V|P}, C_{Y_0,V|P}, C_{Y_1,P}, C_{Y_0,P}\big) \in \widetilde{\Theta} \equiv \mathcal{F}^c \times \mathcal{F}^c \times \mathcal{C}_1^c \times \mathcal{C}_0^c \times \mathcal{C}_1 \times \mathcal{C}_0$ under only the STC structure (Assumption 1) and the regularity condition (Assumption 5). This is important not only because bivariate copula are easier to model, but also because $C_{Y_d,P}$ and $C_{Y_d,V|P}$ have appropriate economic interpretations, as we will see in the next sessions.

Secondly, Equations (8) and (9) in the second part of Lemma 2 provide a link between $\tilde{\theta}$ and the observed data distribution. Interestingly, given $\Psi_{d,p}$ is invertible (see Appendix A.3), one can "solve" $F_{Y_d}(y) = \Psi^{-1}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p])$ from Equations (8) and (9). So once the two bivariate copula are fixed, F_{Y_d} is uniquely determined. Meanwhile, since $F_{Y_d}(y)$ does not depends on p, it should be the case that $\forall p \neq p'$ and $d = 0, 1$,

$$\Psi^{-1}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p]) = \Psi^{-1}_{d,p'}(\mathbb{P}[Y \leq y, D = d|P = p']), \tag{10}$$

for all y in the intersection of the conditioning supports of $Y|D = d, P = p$ and $Y|D = d, p = p'$.[8] Notice that the intersection is always non-empty when the IV is valid. One can show any pair of copulas $C_{Y_d,V|P}$ and $C_{Y_d,P}$ that satisfy Equations (8) and (9) can be rationalized by the data and the model structure. This indeed characterizes the identified set for $\tilde{\theta}$, as we summarize in the following theorem.

Theorem 2. *Under Assumptions 1 and 5, the identified set $\Theta_{I,g}$ in Definition 1 can be equivalently expressed by the identified set Θ_I of $\tilde{\theta} \equiv \left(F_{Y_1}, F_{Y_0}, C_{Y_1,V|P}, C_{Y_0,V|P}, C_{Y_1,P}, C_{Y_0,P}\right)$, which is characterized as follows:*

$$\begin{aligned}\Theta_I \quad = \quad & \Big\{\tilde{\theta} \in \widetilde{\Theta} : \textit{For } d \in \{0,1\}, (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}^c_d \times \mathcal{C}_d \textit{ satisfies Equation } (10) \\ & \textit{and } \forall y \in \mathcal{Y}_{d,p}, F_{Y_d}(y) = \Psi^{-1}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p])\Big\},\end{aligned}$$

where $\mathcal{Y}_{d,p}$ is the conditioning support of $Y|D = d, P = p$.

Proof. See Appendix A.3. □

Theorem 2 characterizes the identified set for $\tilde{\theta}$ under the STC restriction imposed on the treatment selection alone. It says that any pair of copulas $(C_{Y_d,V|P}, C_{Y_d,P})$ such that the mapping $\Psi^{-1}_{d,p}$ produces a flat function in p, can be rationalized by the observed data and the STC model. Meanwhile, the theorem also provides a convenient characterization of the identified set for subvectors of

[8] It is possible that for all pair of $p \neq p'$, the conditioning supports of $Y|D = d, P = p$ and $Y|D = d, p = p'$ do not overlap. For instance, the distribution of $Y|D = d, P = p$ is degenerate. In this case, we lose all the identification power. We exclude such pathological scenarios throughout the paper and implicitly assume there exists at least one pair of (p, p') on which the support of $Y|D = d, P = p$ and $Y|D = d, p = p'$ has an overlap.

the parameters. For instance, the projection of the identified set for copulas are determined by Equation (10), and once $(C_{Y_d,V|P}, C_{Y_d,P})$ are fixed, F_{Y_d} is point identified. In particular, Equation (10) essentially uses the fact that the marginal distribution of potential outcome is invariant to the propensity score. This identification approach has some similarity with the identification restriction of Arellano and Bonhomme (2017, Lemma 1) in their study of the sample selection model. Since the marginal distributions of potential outcomes are uniquely determined given $(C_{Y_d,V|P}, C_{Y_d,P})$, one can expect that the assumptions that one imposes on the dependence structure among these variables largely determine the identification power of the model. Not surprisingly, Theorem 2 reduces to the identification equation in HV05 when Assumption 2 holds, that is, when P is independent with Y_d given V. The following corollary summarizes this observation.

Corollary 1. *Suppose Assumptions 1, 2 and 5 hold, then the identification equation postulated in Theorem 2 coincides with the identification result of HV05, that is,*

$$\frac{\partial \mathbb{P}[Y \leq y, D = 1 | P = p]}{\partial p} = \mathbb{P}[Y_1 \leq y | V = p].$$

and

$$-\frac{\partial \mathbb{P}[Y \leq y, D = 0 | P = p]}{\partial p} = \mathbb{P}[Y_0 \leq y | V = p].$$

Proof. See Appendix A.4.

Remark 1. *(Selection on observables) In our framework, the selection on observables assumption boils down to imposing that $C_{Y_d,V|P}(x_1, x_2) = x_1 x_2$. In such a case, we can easily see through Lemma 2 that we recover the well known identification result under selection on observables, i.e. $F_{Y_d|P}(y|p) = P(Y \leq y | D = d, P = p)$.*

3.2. **Identification with Imperfect IVs.** We previously characterized the identified set for the DMTE by only imposing the STC assumption. We also showed that if we additionally impose the independence assumption, we recover the HV05 point identification results for the MTE and subsequent parameters. However, the validity of the independence assumption is very often subject to significant controversy, see, for instance, Deaton, Heckman, and Imbens (2010). Therefore, Manski and Pepper (2000) proposed to relax the IV assumption in order to have more credible and trustworthy results.

In this subsection, we will further explore in this direction and study the identification of DMTE under imperfect IVs. Here, we refer to "imperfect IVs" as any covariates in the selection equation that could be dependent on the potential outcomes, with the type of dependence being restricted by the economic theory or the empirical context under study. We will show that these restrictions can be easily implemented in our approach to derive sharp bounds on the DMTE and, therefore, on all conventional policy parameters. Indeed, Manski and Pepper (2000) provide sharp bounds on the ATE under the monotone IV (MIV) assumption. However, their approach is not immediately transferrable in the sense that if an applied researcher is interested in other parameters or other deviations of the IV assumptions, she has to derive the specific sharp bounds for this parameter of interest. Using DMTE as a bridge, our unified approach allows the researcher to recover sharp bounds on a variety of parameters of interest under a various IV dependence assumptions, a nice feature inherited from the classical MTE framework. This will free applied researchers from case-by-case constructions.

Before visiting various imperfect IV restrictions, let us consider two empirical cases for illustration:

3.2.1. *Violation of the exclusion restriction.* Let us consider the following simple model:

$$
\begin{aligned}
Y &= \alpha D + \gamma P + \epsilon, \\
D &= 1\{P > V\}
\end{aligned}
$$

where $P \perp (\alpha, \epsilon, V)$ and γ is a constant. As can be seen, the usual MTE assumptions are violated as soon as $\gamma \neq 0$. This potential violation of the exclusion restriction is an important concern in many empirical applications. We argue that even if $\gamma \neq 0$, it is possible to provide informative bounds on MTE and PRTE parameters. Indeed, it can be easily shown that $P(Y_d \leq y | P = p) = F_{\epsilon+\alpha d}(y - \gamma p)$ for $d \in \{0, 1\}$ which implies a monotone IV restriction, i.e. $Y_d | P = p \succeq_{FSD} Y_d | P = p'$ for all $p' \geq p$ or $Y_d | P = p \succeq_{FSD} Y_d | P = p'$ for all $p' \leq p$, depending on the sign of γ.

3.2.2. *Misspecification in presence of multiple treatments.* One recent and growing empirical application using the MTE identification strategy is the Judge leniency IV designs.[9] Consider a model where two simultaneous treatments determine the outcome while researchers focus only on one treatment and overlook the second one. This is a paramount concern that appears in the Judge

[9]See for instance Kling (2006); Aizer and Doyle Jr (2015); Di Tella and Schargrodsky (2013); Mueller-Smith (2015); Dobbie, Goldin, and Yang (2018), and Bhuller, Dahl, Loken, and Mogstad (2019) among others.

leniency IV design literature. In this literature, researchers are interested in the causal effect of incarceration decisions on future outcomes such as recidivism, making abstraction of other potential treatments. However, trial decisions are multidimensional, with judges deciding on incarceration, fines, community service, etc.[10] Let us consider the following model:

$$Y = \underbrace{[Y_{11}D_2 + Y_{10}(1-D_2)]}_{Y_1} D_1 + \underbrace{[Y_{01}D_2 + Y_{00}(1-D_2)]}_{Y_0}(1-D_1),$$
$$D_1 = 1\{P_1 > V_1\}, \quad D_2 = 1\{P_2 > V_2\}.$$

where D_1 denotes the incarceration decision and D_2 is a second binary treatment that denotes if the agent receives a fine or not. $Y_{d_1 d_2}$ denotes the potential outcome when the two treatments are externally set to $D_1 = d_1$ and $D_2 = d_2$. P_1 and P_2 are measures of the judge's stringency level for a different punishment. Assuming that the judge's assignment to cases is entirely random, we might expect the following IV-independence assumption $(V_1, V_2, Y_{d_1 d_2}) \perp (P_1, P_2)$ to hold. When D_2 is neglected, researchers essentially adopt the following model:

$$Y = Y_1 D_1 + Y_0(1-D_1), \tag{11}$$

$$D_1 = 1\{P_1 > V_1\}. \tag{12}$$

where $Y_d \equiv [Y_{d1}D_2 + Y_{d0}(1-D_2)]$ for $d \in \{0,1\}$. When considering this simplified model, we must be cautious about the causal interpretation of $\mathbb{E}[Y_1 - Y_0 | V_1 = v_1]$. Indeed, the latter quantity does not capture the direct causal effect of incarceration (D_1), but rather the total causal effect of the incarceration —for an individual at the margin— that is mediated by the fines effect on the recidivism. Please see Pearl (2013) for a detailed discussion. Now, let's presume that we are interested in the identification of the total effect. Under the IV independence $(V_1, V_2, Y_{d_1 d_2}) \perp (P_1, P_2)$, where $d_1 \in \{0,1\}, d_2 \in \{0,1\}$, we can show that for $p_1 \leq p_1'$ (without loss of generality) and an arbitrary $y \in \mathcal{Y}$,

$$\mathbb{P}(Y_1 \leq y | P_1 = p_1) - \mathbb{P}(Y_1 \leq y | P_1 = p_1') = \mathbb{E}_{P_2|P_1=p_1}[H_y(P_2)] - \mathbb{E}_{P_2|P_1=p_1'}[H_y(P_2)],$$

where $H_y(p_2) \equiv \int_0^{p_2} \Big\{ \mathbb{P}(Y_{11} \leq y | V_2 = v_2) - \mathbb{P}(Y_{10} \leq y | V_2 = v_2) \Big\} dv_2$. This leads to the following result:

[10]Please see Bhuller, Dahl, Loken, and Mogstad (2019)'s section 5.5) for a detailed discussion.

Lemma 3. *Considering the model (11, 12), where* $Y_d \equiv [Y_{d1}D_2 + Y_{d0}(1 - D_2)]$ *for* $d \in \{0, 1\}$, $D_2 = 1\{P > V_2\}$ *and* $(V_1, V_2, Y_{d_1 d_2}) \perp (P_1, P_2)$. *We have,*

(i) If $\mathbb{P}(Y_{d1} \leq y | V_2 = v_2) = \mathbb{P}(Y_{d0} \leq y | V_2 = v_2)$ *for all* v_2 *and* y*, then* $Y_d \perp P_1$.

(ii) If $P_1 \perp P_2$*, then* $Y_d \perp P_1$.

(iii) If for each $v_2 \in [0, 1]$ *the conditional distribution of* Y_{d0} *given* $V_2 = v_2$ *first order stochastically dominates the distribution of* Y_{d1} *given* $V_2 = v_2$*, i.e.* $Y_{d0}|V_2 = v_2 \succeq_{FSD} Y_{d1}|V_2 = v_2$*, and if for any pair* $(p_1', p_1) \in \mathcal{P}_1^2$ *such that* $p_1' \geq p_1$*, we have* $P_2|P_1 = p_1' \succeq_{FSD} P_2|P = p_1$*, then* $Y_d|P_1 = p_1 \succeq_{FSD} Y_d|P_1 = p_1'$.

Proof. See Appendix A.5

Lemma 3 (i) and (ii) provide two sufficient conditions under which ignoring D_2 does not cause the failure of IV-independence assumption (with respect to P_1). In the misspecified model, Y_1 is essentially a mixture of Y_{11} and Y_{10}—two random variables that are independent with P_1. Condition (i) says that these two random variables have the same distribution conditioning on V_2; hence any mixing between them does not change the distribution. The condition in (ii) says the mixing weights are independent of P_1, so the mixture of Y_{11} and Y_{10} is independent of P_1 as well. However, in applications, it is hard to justify Y_{11} and Y_{10} have the same distribution conditioning on V_2, and we also observe that P_1 and P_2 tend to be positively correlated. Therefore, if receiving a fine or not has a direct causal impact on recidivism and if a judge's stringency indexes for both treatments are correlated, then the IV independence assumption in the misspecified model is violated. Bhuller, Dahl, Loken, and Mogstad (2019) propose some suggestive ideas to screen for such violations; however, it is worth noting their suggestive tests are not very informative to screen violations of the MTE assumptions. In Appendix B, we propose a nonparametric specification test of the MTE assumptions. We show that the test is sharp in the sense that it is the most informative one to screen all detectable violations of the MTE assumptions.

Lemma 3 (iii) shows that while the IV independence assumption is violated, we can invoke a monotone IV assumption under some reasonable restrictions. More precisely, $Y_{10}|V_2 = v_2 \succeq_{FSD} Y_{11}|V_2 = v_2$ means that conditionally on V_2, being externally assigned to both punishments (incarceration and fines), make someone less likely to reoffend than someone who is externally assigned to incarceration but with no fines. The second condition $P_2|P_1 = p_1' \succeq_{FSD} P_2|P = p_1$ for $p_1' > p_1$

suggests that the level of Judge's stringency is positively dependent for two types of punishments. Under those conditions, Lemma 3 (iii) suggests that P_1 can be considered as monotone IV; that is, being more stringent on incarceration leads to less recidivism. Below, we will show how we could bound the MTE in such an empirical context using the monotone IV assumption.

3.2.3. *Imperfect IVs.* We will consider the following list of imperfect IVs:

Assumption 6 (Imperfect IVs)**.**

(a) **Affiliated IV**: *We say the propensity score P is an Affiliated IV if the joint density f_{P,Y_d} of P and Y_d satisfies $f_{P,Y_d}(p,y)f_{P,Y_d}(p',y') \geq f_{P,Y_d}(p,y')f_{P,Y_d}(p',y)$ for any $p > p'$ and $y > y'$, where (p,y) and (p',y') belong to the joint support of (P,Y_d)*[11].

(b) **IHRD IV**: *We say the propensity score P is an Inverse Hazard Rate Decreasing IV if $\frac{F_{Y_d|P}(y|p)}{f_{Y_d|P}(y|p)}$ is non-increasing in p for all y.*

(c) **MIV**: *We say P is a monotone IV if for any pair $(p',p) \in \mathcal{P} \times \mathcal{P}$ such that $p' \geq p$, $P = p$, i.e. $Y_d|P = p' \succeq_{FSD} Y_d|P = p$. In other terms, $\mathbb{P}(Y_d > y|P = p)$ is non-decreasing in p for all $y \in \mathcal{Y}$.*[12]

(d) **RTI-IV**: *We say the propensity score P is a right tail increasing IV, if Y_d is right tail increasing in P, i.e. $RTI(Y_d|P)$ meaning that $\mathbb{P}(Y_d > y|P > p)$ is non-decreasing in p for all $y \in \mathcal{Y}$.*

(e) **LTD-IV**: *We say the propensity score P is a left tail decreasing IV, if Y_d is left tail decreasing in P, i.e. $LTD(Y_d|P)$ meaning that $\mathbb{P}(Y_d \leq y|P \leq p)$ is non-increasing in p for all $y \in \mathcal{Y}$.*

(f) **PQD-IV**: *We say the propensity score P is a positive quadrant dependent IV, if $\mathbb{P}(Y_d > y, P > p) \geq \mathbb{P}(Y_d > y)\mathbb{P}(P > p)$ for all $(y,p) \in \mathcal{Y} \times \mathcal{P}$.*

We define the imperfect IV assumptions relative to the propensity score P. However, we can also define it with respect to the vector of instruments.[13] Affiliated IV means that it is more likely that the pair of realizations of Y_d and P simultaneously take high values or low values than for Y_d to take a high (resp. low) realization while P take a low (resp. high) realization. The MIV is a positive

[11]It is also referred as $f_{P,Y_d}(y,p)$ being TP2 (*Totally Positive of Order 2*)

[12]This property is also referred as Y_d being positively regression dependent on P.

[13]For instance, let "$\succeq$" denotes the component-wise partial order when comparing vectors, then we can state: for any pair $z' \succeq z$ in the support of a vector of observable variables Z, the conditional distribution of $Y_d, d \in \{0,1\}$ given $Z = z'$ first order stochastically dominates the distribution of Y_d given $Z = z$, i.e. $Y_d|Z = z' \succeq_{FSD} Y_d|Z = z$. Then the partial ordering on $\mathcal{Z}$ induces an ordering on $\mathcal{P}$ which is what we considered in Assumption 6-(c).

dependence restriction, which means that Y_d is more likely to take on larger values when P increases. The RTI-IV captures the fact that Y_d is more likely to take on larger values when P takes high values as well, while the LTD captures that Y_d is more likely to take lower values when P takes low values as well. Finally, the PQD-IV suggests that Y_d and P are more likely to take large values together or to be smaller together compare to the case if they were independent. The PQD type of assumption has been discussed in Bhattacharya, Shaikh, and Vytlacil (2012). The relation between these positive dependence concepts could be summarized as follows:[14]

$$\text{Affiliated IV} \Rightarrow \text{IHRD-IV} \Rightarrow \text{MIV} \Rightarrow \text{LTD-IV} \Rightarrow \text{PQD-IV},$$

$$\text{Affiliated IV} \Rightarrow \text{IHRD-IV} \Rightarrow \text{MIV} \Rightarrow \text{RTI-IV} \Rightarrow \text{PQD-IV}.$$

The next result shows that all these imperfect IV restrictions can be equivalently written in terms of restrictions only on the copula $C_{Y_d,P}(.,.)$.

Lemma 4. *Let Y_d and P be two continuous variables satisfying Assumption 5, then*

(a) *P is an affiliated IV if and only if*

$$c_{Y_d,P}(x_1,x_2)c_{Y_d,P}(x_1',x_2') \geq c_{Y_d,P}(x_1',x_2)c_{Y_d,P}(x_1,x_2')$$

for all $(x_1,x_2) \in [0,1]$ and $(x_1',x_2') \in [0,1]$ such that $x_1 \geq x_1'$ and $x_2 \geq x_2'$, where $c_{Y_d,P}(x_1,x_2) = \frac{\partial^2 C_{Y_d,P}(x_1,x_2)}{\partial x_1 \partial x_2}$ is the copula density.

(b) *P is an IHRD IV if and only if for any $x_1 \in [0,1]$,*

$$\frac{\partial^2 \log \frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_2}}{\partial x_1 \partial x_2} \geq 0.$$

(c) *P is an MIV if and only if $C_{Y_d,P}(x_1,x_2)$ is concave in x_2 for all $x_1 \in [0,1]$,*

$$\frac{\partial^2 C_{Y_d,P}(x_1,x_2)}{\partial x_2^2} \leq 0. \tag{13}$$

(d) *P is an RTI-IV if and only if for any $x_2 \in [0,1]$ and almost all x_1,*

$$\frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_1} \geq \frac{[x_2 - C_{Y_d,P}(x_1,x_2)]}{1-x_1}. \tag{14}$$

[14]The proof of those implications can be found in Joe (1997, Theorem 2.3).

(e) *P is an LTD-IV if and only if for any $x_2 \in [0,1]$ and almost all x_1,*

$$\frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_1} \leq \frac{C_{Y_d,P}(x_1,x_2)}{x_1}. \tag{15}$$

(f) *P is an PQD-IV if and only if for all $(x_1,x_2) \in [0,1]^2$,*

$$C_{Y_d,P}(x_1,x_2) \geq x_1 x_2. \tag{16}$$

Proof. See Appendix A.6.

This proposition has a significant practical advantage since it allows us to see how the identified set for the DMTRs under the STC assumption shrinks with each of the above assumptions. We do not need to recompute the identified set for the DMTRs for each of these assumptions; instead, we only have to intersect the set of copulas that rationalize the model under the STC with the set of copulas that respect the IV restrictions that researchers would like to maintain. Notice that in this layered analysis, only the copula $C_{Y_d,P}(.,.)$ is involved. Let us denote by $\mathcal{C}_d^{\mathbf{r}}$ the set of copulas $C_{Y_d,P}(.,.)$ that respect a given restriction $\mathbf{r}$. More precisely, $\mathcal{C}_d^{\mathbf{r}} \equiv \{C_{Y_d,P}(.,.) \in \mathcal{C}_d$ such that the restriction $\mathbf{r}$ is satisfied$\}$.

Theorem 3. *Suppose that Assumptions 1 and 5 are satisfied.*

$$\Theta_I^{\mathbf{r}} = \Bigg\{\tilde{\theta} \in \widetilde{\Theta} : \textit{For } d \in \{0,1\}, (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}_d^c \times \mathcal{C}_d^{\mathbf{r}} \textit{ satisfies Equation (10)}$$
$$\textit{and } \forall y \in \mathcal{Y}_{d,p}, F_{Y_d}(y) = \Psi_{d,p}^{-1}(\mathbb{P}[Y \leq y, D = d | P = p])\Bigg\},$$

where $\mathcal{Y}_{d,p}$ is the conditioning support of $Y|D=d, P=p$, for $\mathbf{r} \in \{$Affiliated IV, MIV, RTI-IV, LTD-IV, PQD-IV$\}$.

Remark 2. *While we restrict the statement of Theorem 3 to the restrictions $\mathbf{r} \in \{$Affiliated IV, MIV, RTI-IV, LTD-IV, PQD-IV$\}$ for the copula $C_{Y_d,P}(.,.)$, it applies to any type of restrictions the researcher would like to impose on either $C_{Y_d,V|P}(.,.)$ or $C_{Y_d,P}(.,.)$. For instance, the HV05 identification assumptions, i.e. Assumptions 1 and 2 impose in our context that $C_{Y_d,V|P}(.,.) = C_{Y_d,V}(.,.)$ and $C_{Y_d,P}(x_1,x_2) = x_1 x_2$.*

Notice that if we have a sequence of restrictions $(\mathbf{r}_1, ..., \mathbf{r}_J)$ on the copula $C_{Y_d,P}(.,.)$ such that $\mathbf{r}_j$ is more restrictive than $\mathbf{r}_l$ for $l < j$, we have the following: $\Theta_I^{\mathbf{r}_J} \subseteq .... \subseteq \Theta_I^{\mathbf{r}_1}$. For instance, we have

$\Theta_I^{\text{Affiliated IV}} \subseteq \Theta_I^{\text{MIV}} \subseteq \Theta_I^{\text{RTI-IV}} \subseteq \Theta_I^{\text{PQD-IV}}$. An interesting feature of our approach is that the set of equality restrictions that characterize the identified set does not change with $\mathcal{C}_d^c$ nor $\mathcal{C}_d^{\mathfrak{r}}$. This feature will ease the computation of the identified set of the DMTRs under different layers of assumptions.

As previously discussed, the identified set for $\left(F_{Y_1}, F_{Y_0}, C_{Y_1,V|P}, C_{Y_0,V|P}, C_{Y_1,P}, C_{Y_0,P}\right)$ has a particular structure that once the copulas are fixed, the marginal distributions of potential outcomes are uniquely determined. The "size" or "volume" of the projected identified set for (F_{Y_1}, F_{Y_0}) is then determined by how many or what kind of restrictions one would like to impose on the copulas. For example, if we assume $Y_d \perp P|V$ as in HV05, (F_{Y_1}, F_{Y_0}) become point-identified, see Corollary 1. If we do not make any assumptions on the dependence between Y_d and P, either conditioning on V or not, then we obtain the identified set as shown in Theorem 2. If we are willing to take a middle ground on the "perfectness" of the instrument P or have prior information on the type of selection into treatment, we can use an analogous version of Theorem 3 that applies to the context and then obtain directly the identified set that corresponds to it.

3.3. **Semi-parametric Identification.** The identification results in Theorems 2 and 3 are fully nonparametric. In this subsection, we will consider an alternative approach by parametrizing the copulas with a finite-dimensional parameter θ. However, we will leave the marginals fully nonparametric. From this perspective, we consider semi-parametric identification in this subsection. As discussed in Chen, Fan, and Tsyrennikov (2006), such a semi-parametric approach has gained popularity in studying some features of multivariate distributions in diverse fields. It is flexible and circumvents the curse of dimensionality. It is worth-noting that the copula-based (partial) identification approach we propose below significantly differs from the one proposed in Chen, Fan, and Tsyrennikov (2006) and subsequent papers. In their models, all the marginals are nonparametrically (point) identified, while in our case, the marginal potential outcomes $F_{Y_d}(y), d \in \{0,1\}$ are not (point) identified. We make the following assumption.

Assumption 7 (Parametric Copula)**.** *There exists $\theta = (\alpha^0, \alpha^1, \delta^1, \delta^0) \in \Theta_0^c \times \Theta_1^c \times \Theta_0 \times \Theta_1 \equiv \Theta \subseteq \mathbb{R}^T$ with $T < \infty$ such that $C_{Y_d,P}(x_1, x_2) = C_{Y_d,P}(x_1, x_2; \alpha^d)$ and $C_{Y_d,V|P=p}(x_1, x_2) = C_{Y_d,V}(x_1, x_2; \sigma_d(p))$ where $\sigma_d(p)$ is known up to a finite number of parameters δ^d, $d \in \{0,1\}$.*

With this copula parametrization, our key unknown parameters of interest are θ and $F_{Y_d}(y), d \in \{0,1\}$. Let $\mathcal{F}^c$ define the set of continuous CDFs. The mapping $\Psi_{d,p}^{-1}(\mathbb{P}[Y \leq y, D = d|P = p]; \theta)$

is now known up to the finite dimensional parameter θ. The following theorem characterizes the identified set for $\tilde{\theta} \equiv \big(\theta, F_{Y_1}(\cdot), F_{Y_0}(\cdot)\big) \in \Theta \times \mathcal{F}^c \times \mathcal{F}^c$ under the semi-parametric specification.

Theorem 4. *Under Assumptions 1, 5 and 7, the semi-parametric identified set Θ_I^{SP} of $\tilde{\theta} \equiv \big(\theta, F_{Y_1}(\cdot), F_{Y_0}(\cdot)\big)$ is characterized as follows:*

$$\Theta_I^{SP} = \Bigg\{ \text{For } d \in \{0,1\}, \theta \in \Theta \text{ satisfying Equation (10)}, \\ \text{and for all } \forall y \in \mathcal{Y}_{d,p}, F_{Y_d}(y) = \Psi_{d,p}^{-1}(\mathbb{P}[Y \leq y, D = d | P = p]; \theta) \Bigg\}.$$

If in addition, a restriction $\mathbf{r}$ *is imposed on the copula $C_{Y_d,P}(.,.)$, we have then:*

$$\Theta_I^{SP,\mathbf{r}} = \Bigg\{ \text{For } d \in \{0,1\}, \theta \in \Theta_0^c \times \Theta_1^c \times \Theta_0^{\mathbf{r}} \times \Theta_1^{\mathbf{r}} \text{ satisfying Equation (10)}, \\ \text{for all } \forall y \in \mathcal{Y}_{d,p}, F_{Y_d}(y) = \Psi_{d,p}^{-1}(\mathbb{P}[Y \leq y, D = d | P = p]; \theta) \Bigg\},$$

for $\mathbf{r} \in$ *{Affiliated IV, MIV, RTI-IV, LTD-IV, PQD-IV}.*

Remark 3. *As Theorem 4 shows, the identification of the infinite-dimensional parameter F_{Y_d} boils down to the identification of a finite-dimensional parameter θ. The sharp identification region of θ is characterized by a set of equality constraints, which are easy to work with because they only contain finite-dimensional parameters and known quantities.*

Remark 4 (Specification tests)**.** *Recently, there have been an increasing number of papers that develop specifications tests for the assumptions usually maintained to identify causal effects, see for instance Kitagawa (2015), Huber and Mellace (2015), Mourifié and Wan (2017),* and *Kédagni and Mourifié (2017), etc. Our approach provides a unified way to do specification tests for the assumptions the researcher is willing to maintain. Indeed, each of the identified sets proposed in Theorems 2 to 4 can be empty if we cannot find copulas that respects the equality constraints. The "largest" identified set that imposes the minimum structure so far is the one derived in Theorem 3. If empty, this means imposing the STC specification for treatment selection is too stringent for the data.*

In Section 4 below, we relax this assumption to allow for multi-dimensional unobserved heterogeneity in the treatment selection.

3.3.1. *Choice of Copulas: Frank copulas.* In many applications, it is unknown ex-ante if there is positive or negative selection into the treatment. In such contexts, it is essential to consider a comprehensive copula family for $C_{Y_d,V|P}$. Comprehensive parametric copulas are copulas that (i) approach the countermonotonicity copula (resp. comonoticity copula), i.e., Fréchet Lower Bound (resp. Fréchet Upper Bound copula) for certain values of their parameters in their permissible range, (ii) and cover the entire domain between the Fréchet lower and upper copula bounds including the product copula as special case. Using these copulas, we may test the absence of selection by checking if the confidence region of $\sigma_d(p)$ excludes the value that corresponds to the product copula, which corresponds to the independence case. Comprehensive copulas, for instance Gaussian and Frank copulas, parameterize the full range of dependence. On the other hand, non-comprehensive copulas such as Farlie-Gumbel-Morgenstern (FGM), Clayton, Gumbel, and Joe copulas, are only able to capture dependence in a limited manner. In practice, it will be useful to use a different family of copula to analyze how sensitive the results are depending on the copula parametrization.

In Corollary 2 below, we show how the characterization of the identified set simplifies when considering the Frank Copula. [15]

Assumption 8 (Frank Copula)**.** *There exists* $\theta = (\alpha_0, \alpha_1, \delta_1, \delta_0) \in \Theta \subseteq \mathbb{R}^T$ *with* $T < \infty$ *such that* $C_{Y_d,P}(x_1, x_2) = -\frac{1}{\alpha_d} \ln\left[1 + \frac{(e^{-\alpha_d x_1}-1)(e^{-\alpha_d x_2}-1)}{(e^{-\alpha_d}-1)}\right]$ *for* $\alpha_d \in (-\infty, 0) \cup (0, \infty)$ *and* $C_{Y_d,V|P=p}(x_1, x_2) = -\frac{1}{\sigma_d(p)} \ln\left[1 + \frac{(e^{-\sigma_d(p)x_1}-1)(e^{-\sigma_d(p)x_2}-1)}{(e^{-\sigma_d(p)}-1)}\right]$ *for* $\sigma_d(p) \in (-\infty, 0) \cup (0, +\infty)$, $d \in \{0, 1\}$, *where* $\sigma_d(p)$ *is known up to a finite number of parameters* δ_d.

Corollary 2. *Under Assumptions 1, 5, 6 and 8, the identified set* Θ_I^{SP} *of* $\tilde{\theta} \equiv \big(\theta, (F_{Y_d}(y) : d \in \{0,1\}, y \in \mathbb{R})\big)$ *is characterized as follows:*

$$\Theta_I^{SP} = \Bigg\{ \textit{For } d \in \{0,1\}, F_{Y_d}(y;\theta) = -\frac{1}{\alpha_d} \ln\left[1 + \frac{H_d(y, p, \sigma_d(p))(e^{-\alpha_d} - 1)}{e^{-\alpha_d F_P(p)} - H_d(y, p, \sigma_d(p))(e^{-\alpha_d F_P(p)} - 1)}\right]$$
$$\textit{and } \theta = (\alpha_d, \delta_d) \textit{ satisfies } \alpha_d f_P(p)(1 - H_d)H_d + \frac{\partial H_d}{\partial p} = 0 \textit{ for all } p \textit{ and } y. \Bigg\}$$

[15] In Appendix C we derive a similar characterization for the FGM copula.

where

$$H_1(y,p,\sigma_1(p)) = -\frac{1}{\sigma_1(p)} \ln\Big[1 + \frac{(e^{-\sigma_1(p)F_{Y,D|P}(y,1|p)} - 1)(e^{-\sigma_1(p)} - 1)}{(e^{-\sigma_1(p)p} - 1)}\Big]$$

and

$$H_0(y,p,\sigma_0(p)) = \frac{1}{\sigma_0(p)} \ln\Big[1 + \frac{(e^{\sigma_0(p)F_{YD|P}(y,0|p)} - 1)(e^{-\sigma_0(p)} - 1)}{e^{-\sigma_0(p)} - e^{-\sigma_0(p)p}}\Big]$$

and $\sigma_d(p)$ *is parameterized by* δ_d*:* $\sigma_d(p) \equiv \sigma_d(p;\delta_d)$*.*

Proof. See Appendix C.3.

Remark 5. *When the function* σ_d *(or its finite dimensional parameter* δ_d*) is given,* α_d *is uniquely determined. To see this, fixing an arbitrary* y*, and then integrating both side of* $\alpha_d f_P(p)(1 - H_d)H_d + \frac{\partial H_d}{\partial p} = 0$ *from* $\underline{p}$ *to* $\overline{p}$ *with respect to* p *gives*

$$\alpha_d \mathbb{E}\Big[\{H_d(y,P,\sigma_d(P;\delta_d))(1 - H_d(y,P,\sigma_d(P;\delta_d)))\}1\{\underline{p} < P \le \overline{p}\}\Big] + \int_{\underline{p}}^{\overline{p}} \frac{\partial H_d}{\partial p} dp = 0$$

$$\Rightarrow \alpha_d = -\frac{H_d(y,\overline{p},\sigma_d(\overline{p};\delta_d)) - H_d(y,\underline{p},\sigma_d(\underline{p};\delta_d))}{\mathbb{E}\Big[\{H_d(y,P,\sigma_d(P;\delta_d))(1 - H_d(y,P,\sigma_d(P;\delta_d)))\}1\{\underline{p} < P \le \overline{p}\}\Big]}$$

where (with a abusing of notation) $\frac{\partial H_d(p)}{\partial p}$ *denotes the total derivative of* $H_d(y,p,\sigma_d(p;\delta_d)))$ *with respect to* p*. Then* δ_d *can be estimated using a minimum distance estimator by inserting pre-estimated* $\hat{p}$ *and* $\hat{f}_P$*. When inserting the true* $\sigma_d(\cdot)$ *into the right hand side of the above equation, since* $H_d(y,P,\sigma_d(P;\delta_d)) = F_{Y_d|P}(y|P)$*, we have*

$$\alpha_d = -\frac{F_{Y_d|P}(y|\overline{p}) - F_{Y_d|P}(y|\underline{p})}{\mathbb{E}\Big[F_{Y_d|P}(y|P)(1 - F_{Y_d|P}(y|P))1\{\underline{p} < P \le \overline{p}\}\Big]}.$$

This states that α_d *is positive iff* $Y_d|P = \overline{p} \succeq_{FSD} Y_d|P = \underline{p}$ *for* $\overline{p} \ge \underline{p}$*.*

3.3.2. *Choice of Copulas: Bernstein Copula.* We now turn our attention to the Bernstein copula. More precisely, let's introduce the following assumption:

Assumption 9 (Bernstein Copulas)**.**

$$(i) C_{Y_d,V|P}(x_1,x_2) = C_{Y_d,V|P}(x_1,x_2;\alpha^d) = K_d L_d \sum_{k=1}^{K_d} \sum_{l=1}^{L_d} \alpha^d_{kl} B_{k-1,K_d-1}(x_1) B_{l-1,L_d-1}(x_2)$$

$$(ii) C_{Y_d,P}(x_1,x_2) = C_{Y_d,P}(x_1,x_2;\beta^d) = R_d S_d \sum_{r=1}^{R_d}\sum_{s=1}^{S_d} \beta_{rs}^d B_{r-1,R_d-1}(x_1) B_{s-1,S_d-1}(x_2)$$

where $B_{i,I}(u) = \int_0^u b_{i,I}(t)dt$ *for* $u \in [0,1]$, $b_{i,I}(u) = \binom{I}{i} u^i (1-u)^{I-i}$, *and* $\alpha_{kl}^d \geq 0$ *and* $\beta_{rs}^d \geq 0$ *are unknown parameters that satisfy* $K_d \sum_{l=1}^{L_d} \alpha_{kl}^d = 1$, $L_d \sum_{k=1}^{K_d} \alpha_{kl}^d = 1$, $S_d \sum_{r=1}^{R_d} \beta_{rs}^d = 1$, $R_d \sum_{s=1}^{S} \beta_{rs}^d = 1$.[16]

The Bernstein copula a is very useful and an important copula since because of the Weirstrass-approximation theorem, we known that any copula can be approximated uniformly over $[0,1]^2$ for K_d and L_d sufficiently large, see Kingsley (1951). Note that in Assumption 9-(i), the parameters $\alpha_{k\ell}$ do not depend on P, so it implicitly assumes that the joint distribution of Y_d and V depends on P through the marginal distribution $F_{Y_d|P}$. One special case of Assumption 9-(i) is "selection on observable", which happens when $K_d = L_d = 1$, so that $F_{Y_d,V|P} = C_{Y_d,V|P}(F_{Y_d|P}, F_{V|P}) = F_{Y_d|P} \times F_{V|P}$. On the other hand, if we take $R_d = S_d = 1$, then Assumption 9-(ii) implies $F_{Y_d,P} = C_{Y_d,P}(F_{Y_d}, F_P) = F_{Y_d} \times F_P$. In this case, we have a valid instrument.

In Theorem 7 relegated in Appendix C.2, we characterize the identified set of the DMTR$_g$ for arbitrary fixed values of R_d, S_d, K_d, L_d, $d \in \{0,1\}$. For sake of simplicity, we present below the special case where $R_d = S_d = K_d = L_d = 2$, for $d \in \{0,1\}$ and $g(Y) = Y$.

Let's consider the following notations:

$$\begin{aligned}
\psi_{10}^d(p) &= \left(4\beta_{11}^d + 2(1-4\beta_{11}^d)F_P(p)\right) 4\alpha_{11}^d,\\
\psi_{11}^d(p) &= (4\beta_{11}^d - 1)(4F_P(p)-2)4\alpha_{11}^d + (4\beta_{11}^d + 2F_P(p)(1-4\beta_{11}^d))^2(1-4\alpha_{11}^d),\\
\psi_{12}^d(p) &= (12\beta_{11}^d + 6(1-4\beta_{11}^d)F_P(p))(1-4\alpha_{11}^d)(4\beta_{11}^d-1)(2F_P(p)-1),\\
\psi_{13}^d(p) &= 2(4\beta_{11}^d-1)^2(2F_P(p)-1)^2(1-4\alpha_{11}^d),
\end{aligned}$$

with $\boldsymbol{\psi}_1^d(p) = (\psi_{10}^d(p), \psi_{11}^d(p), \psi_{12}^d(p), \psi_{13}^d(p))'$, and

$$\begin{aligned}
\psi_{20}^d(p) &= -2\left(4\beta_{11}^d + 2(1-4\beta_{11}^d)F_P(p)\right),\\
\psi_{21}^d(p) &= 4(4\beta_{11}^d + 2F_P(p)(1-4\beta_{11}^d))^2 - 2(4\beta_{11}^d-1)(4F_P(p)-2),\\
\psi_{22}^d(p) &= (48\beta_{11}^d + 24(1-4\beta_{11}^d)F_P(p))(4\beta_{11}^d-1)(2F_P(p)-1),\\
\psi_{23}^d(p) &= 8(4\beta_{11}^d-1)^2(2F_P(p)-1)^2,
\end{aligned}$$

[16] Please see Dou et al (2016) for this formulation of the Bernstein copula.

with $\boldsymbol{\psi}_2^d(p) = (\psi_{20}^d(p), \psi_{21}^d(p), \psi_{22}^d(p), \psi_{23}^d(p))'$. In addition, let's define $\boldsymbol{\gamma}^d \equiv (\gamma_0^d, \gamma_1^d, \gamma_2^d, \gamma_3^d)$ and then define the following sets:

$$\Theta^{BC} \equiv \left\{(\alpha_{11}^1, \beta_{11}^1, \alpha_{11}^0, \beta_{11}^0) : 0 \leq \alpha_{11}^d \leq \frac{1}{2}, 0 \leq \beta_{11}^d \leq \frac{1}{2}, \text{ for } d \in \{0,1\}\right\}$$

$$\Gamma \equiv \left\{(\boldsymbol{\gamma}^1, \boldsymbol{\gamma}^0) : \liminf \mathcal{Y} \leq \gamma_j^d \leq \limsup \mathcal{Y}, \text{ for } j \in \{0,1,2,3\}, \text{ for } d \in \{0,1\}\right\}.$$

In this special case the identified set of the DMTRs can be characterized as follows:

$$\Theta_I = \Bigg\{ \Bigg(\mathbb{E}[Y_1|V = v, P = p; \bar{\theta}], \mathbb{E}[Y_0|V = v, P = p; \bar{\theta}]\Bigg) \text{ where}$$

$$\mathbb{E}[Y_d|V = v, P = p; \bar{\theta}] = \boldsymbol{\psi}_1^d(p) \cdot \boldsymbol{\gamma}^d + (\boldsymbol{\psi}_2^d(p) \cdot \boldsymbol{\gamma}^d)v,$$

$$\forall \;\; \bar{\theta} \equiv \left(\alpha_{11}^1, \beta_{11}^1, \alpha_{11}^0, \beta_{11}^0, \boldsymbol{\gamma}^1, \boldsymbol{\gamma}^0\right) \in \Theta^{BC} \times \Gamma \text{ such that the following equations are satisfied}$$

$$\mathbb{E}[YD|P = p] = (\boldsymbol{\psi}_1^1(p) \cdot \boldsymbol{\gamma}^1)p + (\boldsymbol{\psi}_2^1(p) \cdot \boldsymbol{\gamma}^1)\frac{p^2}{2}$$

$$\mathbb{E}[Y(1 - D)|P = p] = (\boldsymbol{\psi}_1^0(p) \cdot \boldsymbol{\gamma}^0)(1 - p) + (\boldsymbol{\psi}_2^0(p) \cdot \boldsymbol{\gamma}^0)(\frac{1}{2} - \frac{p^2}{2})\Bigg\}.$$

Notice that in this special case, imposing a MIV restriction is equivalent to simply restrict the range of β_{11}^d, from $0 \leq \beta_{11}^d \leq \frac{1}{2}$ to $\frac{1}{4} \leq \beta_{11}^d \leq \frac{1}{2}$.

Remark 6. *Our semi-parametric characterization of the DMTR$_g$ identified set using Bernstein copula — Theorem 7 relegated in Appendix C.2, shows that when $R_d = S_d = 1$ (i.e. the IV is valid), we have $\mathbb{E}[g(Y_d)|V = v] = \sum_{l=1}^{L_d} \theta_{dl}^g b_{l-1,L_d-1}(v)$ and recover the parametric form Mogstad, Santos, and Torgovitsky (2018) imposed on the MTR. Mogstad, Santos, and Torgovitsky (2018) approach imposes $\mathbb{E}[g(Y_d)|V = v] = \sum_{l=1}^{L_d} \theta_{dl}^g b_{l-1,L_d-1}(v)$ as a primitive. Here we show that under a valid IV assumption —Assumption 2, imposing such a structure on the MTRs is equivalent to model the "selection on unobservables" component of the dependence — $C_{Y_d,V|P}(x_1, x_2; \alpha^d)$— using a Bernstein copula of order L_d. This result has an important practical implication. As pointed out in Sancetta and Satchell (2004), a limitation of the Bernstein copula is that it cannot be used to model strong tail dependence behavior. So one might be concerned in using Mogstad, Santos, and Torgovitsky (2018) approach when there are strong tails dependence. However, our method offers more flexibility in that respect since we have a complete flexibility in our copula choice, we may consider the Gumbel* *copula that is well suited to capture strong tails dependence or the Composite Bernstein Copula —introduced in Yang et al (2014), that is comprehensive and also capture extreme tails dependence.*

3.4. **Implementation and Numerical illustration.** We conclude this section with two numerical examples on semiparametric identification.

3.4.1. *DGP1: IV-independence Assumptions fails to hold.* Let the marginals be specified as $Y_1 \sim N(1,1)$, $Y_0 \sim N(0,1)$, $V \sim U[0,1]$, $Z = P \sim U[0,1]$, and $D = 1[V \leq P]$. We specify the dependence among (Y_d, P, V) using Frank copula:

$$C_{Y_d,P}(x_1,x_2) = -\frac{1}{\alpha_d} \ln \Big[1 + \frac{(e^{-\alpha_d x_1} - 1)(e^{-\alpha_d x_2} - 1)}{(e^{-\alpha_d} - 1)}\Big],$$

$$C_{Y_d,V|P}(x_1,x_2;\sigma_d(p)) = -\frac{1}{\sigma_d(p)} \ln \Big[1 + \frac{(e^{-\sigma_d(p)x_1} - 1)(e^{-\sigma_d(p)x_2} - 1)}{(e^{-\sigma_d(p)} - 1)}\Big],$$

where true parameter values are $\alpha_1 = 2$, $\sigma_1 = 3$, $\alpha_0 = \sigma_0(p) = 0$. In this case, P is not a valid instrument because $\alpha_1 \neq 0$. The endogeniety issue exists because Y_1 is not independent with V given P as $\sigma_1(p) \neq 0$.

To evaluate the PRTE, we follow HV05 and consider a hypothetical policy intervention where the new policy "subsidizes" large propensity: if $Z > t$, $D = 1[Z(1+t) - V \geq 0]$; else $D = 1[Z - V \geq 0]$. For this exercise, we choose $t = 0.2$. The true parameter values are given by the following table:

TABLE 2. True Values of Parameters in DGP1

Parameters	True value
ATE	1.00
ATT	0.94
ATUT	1.06
PRTE	1.42
LATE(0.2,0.5)	0.78

We first demonstrate our main identification result in Theorems 2 and 4. Figure 1 plots the inverse mapping $\Psi^{-1}_{1,p}(\cdot, p; \alpha, \sigma)$ as a function of y at different values of $p \in \{0.2, 0.3, \cdots, 0.8\}$ (each

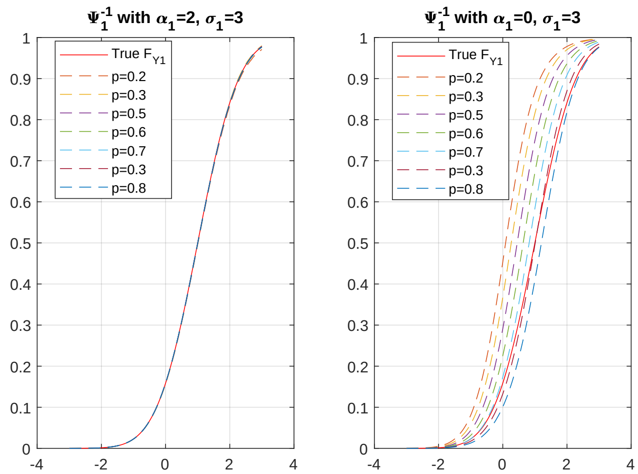

FIGURE 1. Plots of $\Psi^{-1}_{1,p}(\cdot, p; \alpha, \sigma)$ at different values of p and the true F_{Y_1} (DGP1)

dashed lines) as well as the true marginal CDF of Y_1 (solid red line).[17] In the right panel, we use a false parameter value $\alpha_1 = 0$ (other parameters fixed at their true values). As we can see, when we set α_1 at a false value and vary p, the "identified" marginal CDF of Y_1 varies. This shows that $\alpha_1 = 0$ can not be the true value. In the left panel, we set the parameter value $\alpha_1 = 2$. Now, regardless which p value that we insert into the mapping $\Psi^{-1}_{1,p}(\cdot, p; \alpha, \sigma)$, its shape remains unchanged and is the same as the true CDF of Y_0. How $\Psi^{-1}_{1,p}$ responds to the change of p provides the identification power for different parameter values. The left panel provides evidence that $\alpha_1 = 0$ should not be included in the identified set, while the right panel suggests $\alpha_1 = 2$ should.

Next, we demonstrate how our identification strategy can be operationalized in practice. Let $\mathcal{P}^I = \{p^1, p^2, \cdots, p^I\}$ and $\mathcal{Y}^J = \{y^1, y^2, \cdots, y^J\}$ be grid points in the support of P and Y, chosen by researchers. Define

$$\kappa_d(y, p; \alpha_d, \sigma_d) = \Psi^{-1}_{d,p}(\mathbb{P}[Y \le y, D = d | P = p]; \alpha_d, \sigma_d).$$

As shown by Theorem 4 and illustrated by Figure 1, when parameter are set to be the true values, $\kappa_d(y, p; \alpha_d, \sigma_d)$ is a flat function in p for any value of y. Therefore, the "sampling standard deviation" of $\{\kappa_d(y, p^1; \alpha_d, \sigma_d), \cdots, \kappa_d(y, p^I; \alpha_d, \sigma_d)\}$, denoted by $S_d(y; \alpha_d, \sigma_d)$, must be zero when evaluated

[17]The inverse mapping $\Psi^{-1}_{1,p}$ depends on $\mathbb{P}(Y \le y, D = 1 | P = p)$, which we do not have the analytic solution. So we approximate $P(Y \le y, D = 1 | P = p)$ using a kernel estimator and a very large sample size.

at the true parameter values. Hence at the true values, we must have

$$L_d(\alpha_d, \sigma_d) \equiv \sum_{j=1}^{J} S_d(y^j; \alpha_d, \sigma_d) = 0.$$

This motivates another representation of (the outer set) of the identified set for $\theta = (\alpha_1, \sigma_1, \alpha_0, \sigma_0)$ as[18]

$$\{\theta \in \Theta : L_1(\alpha_1, \sigma_1) + L_0(\alpha_0, \sigma_0) = 0\}.$$

Equivalently, we can represent the identified set for (α_d, σ_d) separately as $A_d = \{(\alpha, \sigma) : L_d(\alpha, \sigma) = 0\}$.

Remark 7. *The above discussion motivates a set estimator as*

$$\hat{A}_d = \{(\alpha, \sigma) : L_{d,n}(\alpha, \sigma) \leq \epsilon_n\},$$

where $L_{d,n}$ is the sample analog of L_d, and $\epsilon_n \downarrow 0$ is a tuning sequence converges to zero. In this numerical illustration, we set $\xi_n = cn^{-1}$ and consider a wide range of constant c. The results are similar. The specific choice of ξ_n and the asymptotic behavior of set estimator $\hat{A}_d$ can be studied in the general framework of Chernozhukov, Hong, and Tamer (2007). We leave these for future research.

Figure 2 plots the approximated contour sets of L_1 and the identified sets.[19] The red dot is where the true parameter value locates. The right panel plots the set of (α_1, σ_1) at which $L_1(\alpha_1, \sigma_1)$ is ϵ-away from its minimal value, where ϵ is a very small constant. Hence, this set can be viewed as an approximation of the identified set (or its outer set). In this example, for the choice of the infinitesimal constant ϵ, we can have a singleton as the approximated identified set. This is not surprising because the parameters are point-identified in this DGP.

Next, we turn to the identified set of treatment parameters that are recovered from the identified set of the copula parameters. Figure 3 draws the identified set for the MTE (again, it is a singleton

[18] This set may be the outer set because we only consider finite grid points in p and y.

[19] Again, because we do not have an analytical expression for L_d, we generated a huge sample and approximated L_d by its sample analog $L_{d,n}(\alpha_d, \sigma_d) = \sum_{j=1}^{J} \hat{S}_d(y^j; \alpha_d, \sigma_d)$, where $\hat{S}_d$ is the sample variance of $\{\hat{\kappa}_d(y, p^1; \alpha_d, \sigma_d), \cdots, \hat{\kappa}_d(y, p^I; \alpha_d, \sigma_d)\}$, and

$$\hat{\kappa}_d(y, p; \alpha_d, \sigma_d) = \Psi_{d,p}^{-1}(\widehat{\mathbb{P}}[Y \leq y, D = d | P = p]; \alpha_d, \sigma_d).$$

with the kernel estimator $\widehat{\mathbb{P}}[Y \leq y, D = d | P = p]$ occurs in the place of the true conditional probability $\mathbb{P}[Y \leq y, D = d | P = p]$.

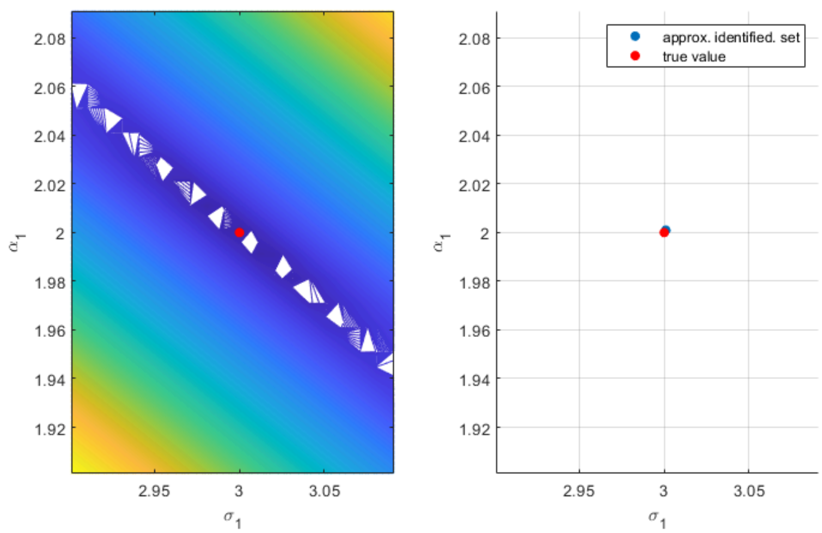

FIGURE 2. Contour set of L_1 and Approximated Identified Set (DGP1)

since the identified set for σ_d and α_d is a singleton). The red line is the true MTE, and the blue line is the identified MTE based on our copula approach. The green line what LIV would identify under the (false) IV-independence assumption. Apparently, there is a notable bias.

Finally, Table 3 compares the treatment parameters estimates using the copula-based approach vs. the LIV estimand (that assumes the IV-independence). We consider four different layers of the IV-related assumption when using the copula-based approach: (a) We impose the IV-independence assumption, i.e., $(\alpha_1 = 0)$; (b) We impose that $\alpha_1 \leq 0$, this restriction relaxes the IV-independence assumption but imposes a negative regression dependence between the IV and the potential outcome, i.e., $\mathbb{P}(Y_d > y|P = p)$ is non-increasing in p, we denote MIV^-; (c) We assume the MIV assumption (positive regression dependence); and (d) we leave the dependence structure captured by α_1 entirely unrestricted. As can be seen, in the two first cases, the copula-based approach can detect that the two related IV assumptions (IV-independence and MIV^-) are not compatible with the observed data. In the two latter cases, the copula-based approach can point-identify all our policy parameters of interests even when the IV is not valid.

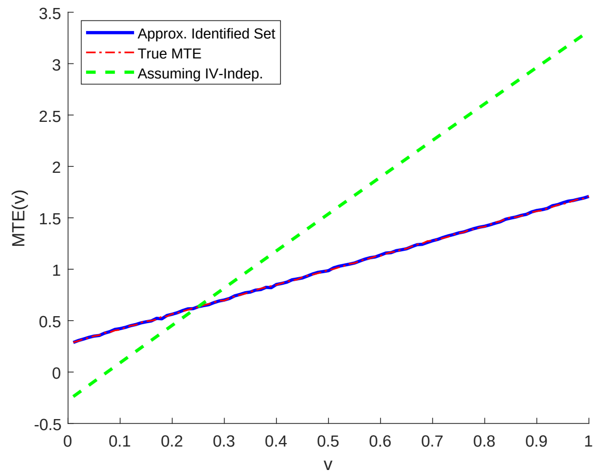

FIGURE 3. Identified Set for MTE (DGP1)

On the other hand, the LIV estimand shows a considerable positive bias for various policy parameters. Indeed, since the "MTE" identified from LIV has positive bias over most parts of the unit interval, it is not surprising that the treatment parameters identified under the IV-independence assumption have a positive bias. To the best of our knowledge, all the empirical papers that have used the LIV approach have never implemented the specification test related to it; see Theorem 6 in Appendix B. In the case that the IV-independence assumption is indeed violated, their policy recommendations could be significantly biased. An advantage of the copula-based approach is that the identification strategy and the specification tests are implemented simultaneously. So, suppose the IV-independence assumption is indeed not compatible with the observed data. In that case, the copula-based approach will not return a biased estimate but will return an empty set that suggests a relaxation of the IV-independence assumption is needed.

3.4.2. *DGP2: Misspecified Copula.* In this subsection, we would like to investigate the copula-based approach's performance in the presence of a misspecified copula. We consider a DGP2 in which the IV-independence assumption holds, and the observed data is not generated using a Frank copula. In other words, LIV would correctly identify MTE in this setup, and our copula-based method is subject to the problem of misspecification. To be more specific, we consider the DGP entertained in HV05 (page 683). The true parameter values and those identified from the copula-based approach are summarized in Table 4.

TABLE 3. Identified Values (DGP1)

Parameters	True value	Identified by Copula-based Approach IV ($\alpha_1 = 0$)	MIV$^-$($\alpha_1 \leq 0$)	MIV$^+$($\alpha_1 \geq 0$)	($\alpha_1 \in \mathbb{R}$)	LIV
ATE	1.00	Empty	Empty	1.00	1.00	1.52
ATT	0.94	Empty	Empty	0.94	0.94	0.94
ATUT	1.06	Empty	Empty	1.06	1.06	2.10
PRTE	1.42	Empty	Empty	1.42	1.42	2.55
LATE(0.2,0.5)	0.78	Empty	Empty	0.78	0.78	1.04

TABLE 4. Parameter Values (DGP2)

Parameters	True value	Identified by Copula-based Approach
ATE	0.200	0.200
ATT	0.235	0.248
ATUT	0.157	0.158
PRTE	0.155	0.158
LATE(0.2,0.5)	0.225	0.225

We can see from Table 4 that, while our semiparametric model is misspecified, the copula-based approach has very small biases. Figure 4 shows why this is the case. In this example, we set the parameter space for σ_1 as $[-20, 20]$. In HV05's example, Y_1 and V are negatively correlated, and the correlation coefficient equals -1. For this, our identification approach would push σ_1 to $-\infty$. In HV05's example, Y_1 and V exhibits a perfect negative dependence; their dependence structure is captured by the Fréchet lower bound copula. Since the Frank copula is comprehensive, it could approximate this specific dependence when σ_1 converges to $-\infty$.

In the implementation, the search for true parameters ends at the lower boundary (-20) of the parameter space of σ_1. At the same time, the identified values of α_1 and α_0 are close to zeros, which is also reflected in Figure 4. The α_1 is not exactly zero because it needs to compensate for the fact that we can not set σ_1 to $-\infty$. This example shows that even though we consider a semi-parametric

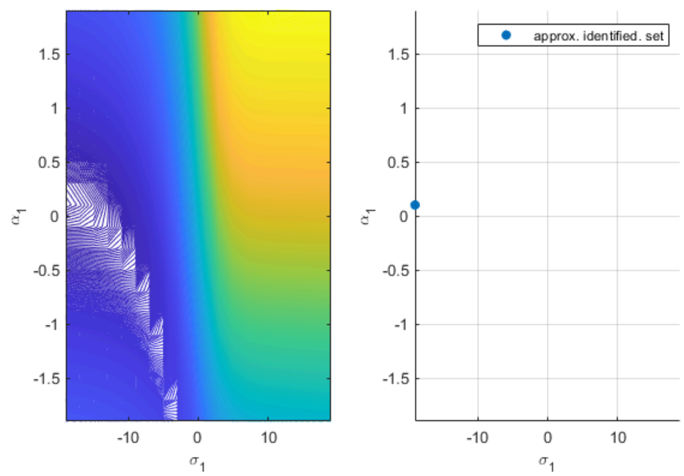

FIGURE 4. Contour set of L_1 and Approximated Identified Set (DGP2)

identification approach, the copula can still be flexible enough to capture the essential part of the dependence structure among the latent variables.

Figure 5 plots the MTE that we construct based upon the identified (α_d, σ_d) and the true MTE. Except at the two boundaries, the semi-parametrically identified MTE is very close to the true function. Again, the discrepancy at the two boundaries is because we can not set σ_d as $\pm\infty$ in practice. However, we should expect a smaller discrepancy when we allow a larger parameter space.

4. EXTENSION: MULTIPLE THRESHOLD-CROSSING MODELS

There are many empirical applications where a model imposing only STC model cannot adequately model the selection to the treatment. Various examples are given in Heckman and Pinto (2018). In presence of multiple potential instruments, one way to relax the “strong" monotonicity assumption is to consider the “AM monotonicity" —in the language of Mogstad, Torgovitsky and Walters (2019), which can be modelled by considering the multiple hurdle model entertained in Lee and Salanié (2018). Our approach can be applied to the case where selection into treatment is defined by a

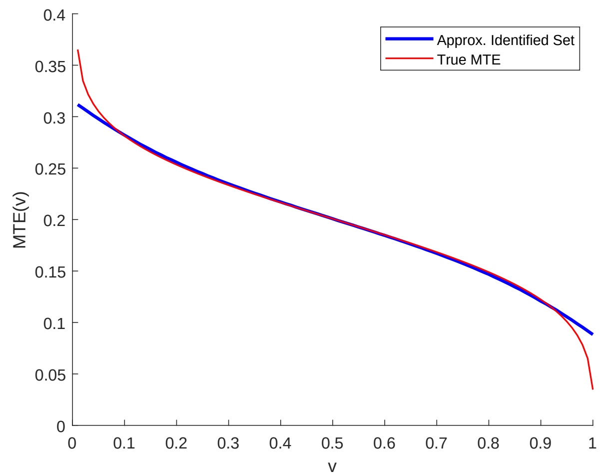

FIGURE 5. Identified Set for MTE (DGP2)

finite number of thresholds. However, for the sake of simplicity, we will consider the case with two thresholds.

Assumption 10 (Double Hurdle (DH) model)**.** *The selection mechanism is governed by* $D = 1\{Q_1(Z) > V_1, Q_2(Z) > V_2\}$ *for some measurable and non-trivial function* (Q_1, Q_2)*, where* (V_1, V_2) *has a joint continuous distribution over interval* $[0,1]^2$ *with marginal uniform distributions and are statistically independent of the vector of* $\big(Q_1(Z), Q_2(Z)\big)$*, i.e.* $\Big(Q_1(Z), Q_2(Z)\Big) \perp (V_1, V_2)$.

Unlike in the STC model, $Q_1(Z)$ and $Q_2(Z)$ are not readily identified from the choice probability $\mathbb{P}(D = 1|Z)$. Theorem 4.2 in Lee and Salanié (2018) provides conditions under which $Q_1(Z)$, $Q_2(Z)$ and the joint distribution $F_{V_1,V_2}(v_1, v_2)$ are non-parametrically identified from the propensity score $\mathbb{P}(D = 1|Z)$. Their non-parametric identification approach requires two continuous “exogenous" covariates that generate all possible values of the thresholds. In our current approach, the exogeneity refers only to the selection equation, i.e., $(Z \perp (V_1, V_2))$; Z could be correlated with the potential outcomes. Without loss of generality, we use Z_1 and Z_2 to denote those exogenous covariates such that $Q_1(Z) \equiv Q_1(Z_1)$ does not depend Z_2 and $Q_2(Z) \equiv Q_2(Z_2)$ does not depend on Z_1.[20] Hereafter, we will assume that the regularity conditions of Lee and Salanié (2018, Theorem 4.2) are valid and that $F_{V_1,V_2}(v_1, v_2)$, $Q_1(Z_1)$, and $Q_2(Z_2)$ are identifiable from the data. To ease the notation, we will write $\mathbf{V} = (V_1, V_2)$ and $\mathbf{Q} = (Q_1, Q_2)$. In the presence of multiple unobserved heterogeneity in the selection mechanism, we define the DMTR as follows:

$$\text{DMTR}_g^d(\mathbf{v}, \mathbf{q}) = \mathbb{P}[Y_d \leq y | \mathbf{V} = \mathbf{v}, \mathbf{Q} = \mathbf{q}] \equiv F_{Y_d|\mathbf{V},\mathbf{Q}}(y|\mathbf{v}, \mathbf{q}),$$

for $\mathbf{v} \in [0,1]^2$, $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$, and $d \in \{0,1\}$. First, we can show that all conventional policy parameters can also be written as a weighted average of the DMTR_g^d even in the presence of multiple thresholds. Before doing so, let's introduce the following assumption:

Assumption 11 (Conditional Policy Invariance)**.** *$Y_d^{a'} | \mathbf{V}^{a'}, \mathbf{Q}^{a'} \sim Y_d^a | \mathbf{V}^a, \mathbf{Q}^a$ with $\mathbf{V}^{a'} \sim \mathbf{V}^a$ and $Y_d^{a'} \sim Y_d^a$ for $a \neq a'$.*

We have the following results for the double hurdle model.

Theorem 5. *Suppose that Assumption 10 is satisfied, then*

(i) $MTE(\mathbf{v}) = \int_{\mathbf{q}} f_{\mathbf{Q}}(\mathbf{q}) DMTE(\mathbf{v}, \mathbf{q}) d\mathbf{q}$;

(ii) For any $\mathbf{s} \in \{$ATE, LATE($\mathbf{u}, \mathbf{u}'$), ATT, ATUT$\}$[21] and weights $\omega^s(\mathbf{v}, \mathbf{q})$ listed in Table 5 below, we have

$$\mathbf{s} = \int_{\mathbf{v}} \int_{\mathbf{q}} \omega^s(\mathbf{v}, \mathbf{q}) DMTE(\mathbf{v}, \mathbf{q}) d\mathbf{v} d\mathbf{q}. \tag{17}$$

(iii) If in addition Assumption 11 holds, Equation (17) holds with $\mathbf{s} = PRTE$.

(iv) If in addition, Lee and Salanié (2018, Assumption 4.2) holds, then all the weights $\omega^s(\mathbf{v}, \mathbf{q})$ are point identified.

Proof. See Appendix A.7. □

Remark 8. *In presence of valid IVs, i.e. $(Z_1, Z_2) \perp Y_d | \mathbf{V}$ our weights $\omega^s(\mathbf{v}, \mathbf{q})$ for any $\mathbf{s} \in \{$ATE, ATT, ATUT, PRTE$\}$ collapse to the weights proposed by Lee and Salanié (2018) for the DH*

[20] For the entire list of requirements please see Assumption 4.2 in Lee and Salanié (2018). They also discussed identification under weaker conditions.

[21] Here $\text{LATE}(\mathbf{u}, \mathbf{u}')$ represents the average treatment effect for the group of compliers when P is externally changed from $\mathbf{u}$ to $\mathbf{u}'$: $\text{LATE}_g(\mathbf{u}, \mathbf{u}') \equiv \mathbb{E}[g(Y_1) - g(Y_0) | u_1 < V_1 \leq u_1', u_2 < V_2 \leq u_2']$.

TABLE 5. Policy Parameters and DMTE in the multiple thresholds case.

Parameters	weights $\omega^s(\mathbf{v}, \mathbf{q})$
ATE	$f_{\mathbf{Q}}(\mathbf{q})f_{\mathbf{V}}(\mathbf{v})$
ATT	$\frac{f_{\mathbf{Q}}(\mathbf{q})f_{\mathbf{V}}(\mathbf{v})1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}}}{\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q})]}$
ATUT	$\frac{f_{\mathbf{Q}}(\mathbf{q})f_{\mathbf{V}}(\mathbf{v})1_{\{\mathbf{v}\notin[0,q_1]\times[0,q_2]\}}}{\mathbb{E}[1-F_{\mathbf{V}}(\mathbf{Q})]}$
LATE($\mathbf{u}$,$\mathbf{u}'$)	$\frac{f_{\mathbf{Q}}(\mathbf{q})f_{\mathbf{V}}(\mathbf{v})1_{\{\mathbf{v}\in[u_1,u_1']\times[u_2,u_2']\}}}{F_{\mathbf{V}}\left(\mathbf{v}\in[u_1,u_1']\times[u_2,u_2']\right)}$
PRTE	$\frac{[f_{\mathbf{Q}^{a'}}(\mathbf{q})-f_{\mathbf{Q}^a}(\mathbf{q})]f_{\mathbf{V}^a}(\mathbf{v})1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}}}{\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^{a'})]-\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^a)]}$

model. If in addition both $\mathbf{V}$ *and* $\mathbf{Q}$ *are scalar-valued random variable, then the weights in Table 5 reduce to the weights in Table 1.*

Notice that $\text{LATE}_g(\mathbf{u}, \mathbf{u}') \equiv \mathbb{E}[g(Y_1) - g(Y_0)|u_1 < V_1 \leq u_1', u_2 < V_2 \leq u_2']$ is a generalization of the LATE defined in Imbens and Angrist (1994) when the selection into treatment is defined by two thresholds. This type of parameter has recently received attention from empirical researchers, e.g. Arteaga (2018).

Assumption 12. *The joint distribution of* $(Y_d, \mathbf{V}, \mathbf{Q})$ *is absolutely continuous respect to the Lebesgue measure.*

Lemma 5. *[Vine Copula] Under Assumptions 10 and 12, for* $d \in \{0, 1\}$ *we have for each* $y \in \mathcal{Y}$ *and* $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

$$F_{Y_d|Q_1}(y|q_1) = \frac{\partial}{\partial x_2} C_{Y_d,Q_1}(x_1, x_2)\Big|_{x_1=F_{Y_d}(y), x_2=F_{Q_1}(q_1)} \equiv c^I_{d,F_{Q_1}(q_1)}(F_{Y_d}(y)), \tag{18}$$

$$F_{Y_d|\mathbf{Q}}(y|\mathbf{q}) = \frac{\partial}{\partial x_2} C_{Y_d,Q_1|Q_2=q_2}(x_1, x_2)\Big|_{x_1=F_{Y_d|Q_1}(y|q_1), x_2=F_{Q_1|Q_2}(q_1|q_2)} \equiv c^{II}_{d,F_{Q_1|Q_2}}(F_{Y_d|Q_1}(y|q_1)), \tag{19}$$

$$F_{Y_d|V_2,\mathbf{Q}}(y|v_2, \mathbf{Q}) = \frac{\partial}{\partial x_2} C_{Y_d,V_2|\mathbf{Q}=\mathbf{q}}(x_1, x_2)\Big|_{x_1=F_{Y_d|\mathbf{Q}}(y|\mathbf{q}), x_2=v_2} \equiv c^{III}_{d,v_2}(F_{Y_d|\mathbf{Q}}(y|\mathbf{q})), \tag{20}$$

$$F_{Y_d|\mathbf{V},\mathbf{Q}}(y|\mathbf{v}, \mathbf{Q}) = \frac{\partial}{\partial x_2} C_{Y_d,V_1|V_2=v_2,\mathbf{Q}=\mathbf{q}}(x_1, x_2)\Big|_{x_1=F_{Y_d|V_2,\mathbf{Q}}(y|v_2,\mathbf{q}), x_2=F_{V_1|V_2}(v_1|v_2)}, \tag{21}$$

and there exists monotone mapping $\Psi_{1,\mathbf{q}}$ and $\Psi_{0,\mathbf{q}}$ such that for each $y \in \mathcal{Y}$ and $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

$$\mathbb{P}[Y \leq y, D = 1 | \mathbf{Q} = \mathbf{q}] = \Psi_{1,\mathbf{q}}(F_{Y_1}(y)), \tag{22}$$

and

$$\mathbb{P}[Y \leq y, D = 0 | \mathbf{Q} = \mathbf{q}] = \Psi_{0,\mathbf{q}}(F_{Y_0}(y)) \tag{23}$$

where the expressions for $\Psi_{d,\mathbf{q}}$ are collected in Appendix A.8.

Proof. See Appendix A.8. □

Given Lemma 5, the identified set for DMTR or DMTE can be constructed as in Theorems 2 to 4.

5. Discussion and Future Work

This paper shows how to use the MTE framework to perform the layered policy analysis when the potential IVs are not necessarily valid. We focus on the case where the propensity score is continuous, as assumed in HV05. While this condition is easier to be satisfied since we allow for imperfect IVs, there are empirical applications in which all potential IVs are discrete. In the follow-up research, we will extend our framework to incorporate discrete propensity score using the sub-copula approach we entertained in Appendix A.9. Finally, this paper focuses on population-level analysis. We briefly discuss estimation and inference in Section 3.4 and leave the complete analysis for future work.

Appendix A. Proofs of results in the main text.

A.1. **Proof of Theorem 1.** It is easy to see that

$$MTE_g(v) \equiv \mathbb{E}[g(Y_1) - g(Y_0)|V = v] = \int_0^1 f_P(p) DMTE_g(v,p)dp,$$

and

$$ATE_g \equiv \mathbb{E}[g(Y_1) - g(Y_0)] = \int_0^1 \int_0^1 \underbrace{f_P(p)}_{w^{ATE}(v,p)} DMTE_g(v,p)dpdv.$$

Regarding LATE,

$$LATE_g(u,u') \equiv \mathbb{E}[g(Y_1) - g(Y_0)|u < V \leq u'] = \int_0^1 \int_0^1 \underbrace{\frac{f_P(p) 1_{\{u<v\leq u'\}}}{u'-u}}_{w^{LATE(u,u')}(v,p)} DMTE_g(v,p)dvdp$$

For ATT, we have

$$ATT_g \equiv \int_0^1 \mathbb{E}[g(Y_1) - g(Y_0)|D = 1, P = p]dF_{P|D=1}(p)$$
$$= \int_0^1 \mathbb{E}[g(Y_1) - g(Y_0)|V \leq p, P = p]dF_{P|D=1}(p) = \int_0^1 \frac{1}{p} \int_0^p \mathbb{E}[g(Y_1) - g(Y_0)|V = v, P = p]dvdF_{P|D=1}(p)$$
$$= \int_0^1 \frac{1}{p} \int_0^p DMTE(v,p)dv \frac{p}{\mathbb{P}(D=1)} f_P(p)dp = \int_0^1 \int_0^1 DMTE(v,p)dv \frac{f_P(p)1\{v \leq p\}}{\mathbb{P}(D=1)}dp$$
$$= \int_0^1 \int_0^1 \underbrace{\frac{f_P(p)1\{v \leq p\}}{\mathbb{E}[P]}}_{w^{ATT}(v,p)} DMTE_g(v,p)dvdp,$$

where $dF_{P|D=1}(p) = \frac{p}{\mathbb{P}(D=1)} f_P(p)dp$ by Bayesian rule and $\mathbb{P}(D = 1) = \mathbb{E}[\mathbb{E}[D|P]] = \mathbb{E}[P]$.

Following the similar derivation as ATT, we can show that

$$ATUT_g \equiv \int_0^1 \mathbb{E}[g(Y_1) - g(Y_0)|D = 0, P = p]dF_{P|D=0}(p)$$
$$= \int_0^1 \int_0^1 \underbrace{\frac{f_P(p)1\{v > p\}}{\mathbb{E}[1-P]}}_{w^{ATUT}(v,p)} DMTE_g(v,p)dvdp$$

Concerning the PRTE_g, under Assumption 1 only, we have:

$$\mathbb{E}[g(Y^a)] = \int_0^1 \mathbb{E}[g(Y^a)|P^a = p]dF_{P^a(p)} = \int_0^1 \mathbb{E}[(g(Y_1^a) - g(Y_0^a))D^a|P^a = p]dF_{P^a(p)} + \mathbb{E}[g(Y_0^a)]$$
$$= \int_0^1 \int_0^1 1\{v \leq p\} f_{P^a(p)} \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, P^a = p]dpdv + \mathbb{E}[g(Y_0^a)]$$
$$= \int_0^1 \int_0^1 1\{v \leq p\} f_{P^a(p)} \text{DMTE}_g^a(v,p)dpdv + \mathbb{E}[g(Y_0^a)]$$

Since we have $\text{DMTE}_g^{a'} = \text{DMTE}_g^a$ and $\mathbb{E}[g(Y_0^{a'})] = \mathbb{E}[g(Y_0^a)]$ under Assumption 4, then under both Assumptions 1 and 4 we have:

$$\mathbb{E}[g(Y^{a'}) - g(Y^a)] = \int_0^1 \int_0^1 [f_{P^{a'}}(p) - f_{P^a(p)}]1\{v \le p\}\text{DMTE}_g^a(v,p)dpdv$$

Therefore,

$$\text{PRTE}_g = \int_0^1 \int_0^1 \underbrace{\frac{[f_{P^{a'}}(p) - f_{P^a(p)}]1\{v \le p\}}{\mathbb{E}_{F_{p^{a'}}}[P] - \mathbb{E}_{F_{P^a}}[P]}}_{w^{PRTE}(v,p)} \text{DMTE}_g(v,p)dpdv.$$

A.2. **Proof of Lemma 2.** To show Equation (6), note that first that

$$f_{Y_d,P}(y,p) = \frac{\partial^2 F_{Y_d,P}(t_1,t_2)}{\partial t_1 \partial t_2}|_{t_1=y,t_2=p} = \frac{\partial^2 C_{Y_d,P}(x_1,x_2)}{\partial x_1 \partial x_2}|_{x_1=F_{Y_d}(y),x_2=F_P(p)} f_{Y_d}(y) f_P(p)$$

Therefore,

$$\begin{aligned}F_{Y_d|P}(y|p) &= \int_{-\infty}^{y} f_{Y_d|P}(t|p)dt = \int_{-\infty}^{y} \frac{f_{Y_d,P}(t,p)}{f_P(p)}dt \\ &= \int_{-\infty}^{y} \frac{\frac{\partial^2 C_{Y_d,P}(x_1,x_2)}{\partial x_1 \partial x_2}|_{x_1=F_{Y_d}(t),x_2=F_P(p)} f_{Y_d}(t) f_P(p)}{f_P(p)}dt = \frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_2}|_{x_1=F_{Y_d}(y),x_2=F_P(p)} \\ &\equiv c_{d,F_P(p)}(F_{Y_d}(y)),\end{aligned}$$

where we write the RHS as $c_{d,F_P(p)}(F_{Y_d}(y))$ since the RHS depends on y only through $F_{Y_d}(y)$ and the mapping $c_{d,F_P(p)}$ depends on the joint distribution of Y_d and P. Note that for any given p, both $F_{Y_d|P}(y|p)$ and $F_{Y_d}(y)$ are strictly increasing in y. Therefore, the mapping $c_{d,F_P(p)}(\cdot)$ is strictly increasing and we can express

$$F_{Y_d}(y) = c_{d,F_P(p)}^{-1}\left(F_{Y_d|P}(y|p)\right). \tag{24}$$

To see Equation (7), note that

$$\begin{aligned}f_{Y_d|V,P}(y|v,p) &= \frac{f_{Y_d,V|P}(y,v|p)}{f_{V|P}(v|p)} = \frac{\frac{\partial^2 C_{Y_d,V|P}(x_1,x_2;p)}{\partial x_1 \partial x_2}|_{x_1=F_{Y_d}(y|p),x_2=F_{V|P}(v|p)} f_{Y_d|P}(y|p) f_{V|P}(v|p)}{f_{V|P}(v|p)} \\ &= \frac{\partial^2 C_{Y_d,V|P}(x_1,x_2;p)}{\partial x_1 \partial x_2}|_{x_1=F_{Y_d}(y|p),x_2=F_{V|P}(v|p)} f_{Y_d|P}(y|p)\end{aligned}$$

Therefore,

$$\begin{aligned}F_{Y_d|V,P}(y|v,p) &= \int_{-\infty}^{y} f_{Y_d|V,P}(t|v,p)dt \\ &= \int_{-\infty}^{y} \frac{\partial^2 C_{Y_d,V|P}(x_1,x_2;p)}{\partial x_1 \partial x_2}|_{x_1=F_{Y_d}(t|p),x_2=F_{V|P}(v|p)} f_{Y_d|P}(t|p)dt = \frac{\partial C_{Y_d,V|P}(x_1,x_2;p)}{\partial x_2}|_{x_1=F_{Y_d}(y|p),x_2=v}\end{aligned}$$

where again we use $F_{V|P}(v|p) = v$.

At last, we consider Equations (8) and (9). Suppose $d = 1$

$$\mathbb{P}[Y \leq y, D = 1|P = p] = \mathbb{P}[Y_1 \leq y, V \leq p|P = p] = C_{Y_1,V|P}(c_{1,F_P(p)}(F_{Y_1}(y)), p; p)$$

where we inserting Equation (24) to obtain the result. As discussed earlier $u \mapsto c_{1,F_P(p)}(u)$ is strictly increasing and $x_1 \mapsto C_{Y_1,V|P}(x_1, x_2; p)$ is also strictly increasing, therefore $u \mapsto C_{Y_1,V|P=p}\Big(c_{1,F_P(p)}(u), p; p\Big) \equiv \Psi_{1,p}(u)$ is strictly increasing. For $d = 0$,

$$\begin{aligned}\mathbb{P}[Y \leq y, D = 0|P = p] = \mathbb{P}[Y_0 \leq y, V > p|P = p] &= \mathbb{P}[Y_0 \leq y|P = p] - \mathbb{P}[Y_0 \leq y, V \leq p|P = p] \\ &= c_{0,F_P(p)}(F_{Y_0}(y)) - C_{Y_0,V|P}(c_{0,F_P(p)}(F_{Y_0}(y)), p; p) \equiv \Psi_{0,p}(F_{Y_0}(y)),\end{aligned}$$

where the mapping $\Psi_{0,p}(u)$ is strictly increasing in u because the left hand side of the equation above is strictly increasing in y (by the definition of conditioning probability), and $F_{Y_0}(y)$ is strictly increasing in y.

A.3. **Proof of Theorem 2.** Let $\mathbb{P}(Y \leq y, D = d|P = p)$ be the distribution of observables. It is apparent from Definition 1, $(C_{Y_d,V|P}, C_{Y_d,P}, F_d)$ satisfy Equations (8) and (9), then they can rationalize the data and model; on the other hand, if $(C_{Y_d,V|P}, C_{Y_d,P}, F_d)$ are the true model parameters, they they must connect with the implied data distribution through Equations (8) and (9). In this sense, the set defined in Definition 1 is sharp.

To verify the set defined in Theorem 2 is also sharp, it is sufficient to show that Equations (8) and (9) and Equation (10) in Theorem 2 are equivalent. First, it is straightforward to see that Equations (8) and (9) imply Equation (10). Second, suppose Equation (10) hold, that is, $\Psi^{-1}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p])$ is flat in p and only varies as a function of y. Note that under Assumption 5, F_{Y_d} and F_P are continuous and strictly increasing, and both $C_{Y_d,V|P}$ and $\frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_2}$ are increasing in their first arguments. Therefore, $\Psi^{-1}_{d,p}$ is strictly increasing in y by construction.

Next from the definitions in Equations (8) and (9) we know that for any $p > 0$

$$p = C_{Y_1,V|P=p}(1, p; p) \quad 0 = C_{Y_1,V|P=p}(0, p; p),$$

and

$$1 = c_{d,F_P(p)}(1), \quad 0 = c_{d,F_P(p)}(0).$$

Therefore it is easy to see that $\Psi^{-1}_{d,p}(\mathbb{P}[Y \leq -\infty, D = d|P = p]) = \Psi^{-1}_{d,p}(0) = 0$ and $\Psi^{-1}_{d,p}(\mathbb{P}[Y \leq \infty, D = d|P = p]) = \Psi^{-1}_{d,p}(\mathbb{P}[D = d|P = p]) = 1$. This says that $\Psi^{-1}(\mathbb{P}[Y \leq \cdot, D = d|P = p])$, as a function of y, is a valid distribution function, which we can choose as the counterfactual distribution F_{Y_d}. This completes the proof.

A.4. **Proof to Corollary 1.** We take $d = 1$ as an example; the case for $d = 0$ is similar. Suppose for simplicity the mapping Ψ^{-1} is differentiable with respect to p, and all inverse functions below are properly defined. Then the restriction is equivalent to

$$\frac{\partial\Big\{\Psi^{-1}_{1,p}(\mathbb{P}[Y \leq y, D = 1|P = p])\Big\}}{\partial p} = 0.$$

On the other hand, $P \perp Y_1|V$ and $P \perp V$ implies $(Y_1, V) \perp P$. By the definition in Equation (8), the mapping $\Psi_{1,p}$ then reduces to the following simple form:

$$\Psi_{1,p}(F_{Y_1}(y)) = C_{Y_1,V}\Big(F_{Y_1}(y), p\Big).$$

Let $C^{-1}_{1,Y_1,V}(\cdot, x_2)$ be the inverse of $C_{Y_1,V}(\cdot, x_2)$ with respect to the first argument, then,

$$\begin{aligned}
0 = \frac{\partial\Big\{\Psi^{-1}_{1,p}(\mathbb{P}[Y \le y, D = 1|P = p])\Big\}}{\partial p} &= \frac{\partial\Big\{C^{-1}_{1,Y_1,V}\Big(\mathbb{P}[Y \le y, D = 1|P = p], p\Big)\Big\}}{\partial p} \\
&= \frac{\partial C^{-1}_{1,Y_1,V}\Big(y, x_2\Big)}{\partial y}|_{y=\mathbb{P}[Y\le y,D=1|P=p],x_2=p} \times \frac{\partial \mathbb{P}[Y \le y, D = 1|P = p]}{\partial p} \\
&\quad + \frac{\partial C^{-1}_{1,Y_1,V}\Big(y, x_2\Big)}{\partial x_2}|_{y=\mathbb{P}[Y\le y,D=1|P=p],x_2=p}.
\end{aligned}$$

This implies

$$\begin{aligned}
\frac{\partial \mathbb{P}[Y \le y, D = 1|P = p]}{\partial p} &= -\frac{\frac{\partial C^{-1}_{1,Y_1,V}\left(y,x_2\right)}{\partial x_2}|_{y=\mathbb{P}[Y\le y,D=1|P=p],x_2=p}}{\frac{\partial C^{-1}_{1,Y_1,V}\left(y,x_2\right)}{\partial y}|_{y=\mathbb{P}[Y\le y,D=1|P=p],x_2=p}} \\
&= \frac{\partial C_{Y_1,V}\Big(x_1, x_2\Big)}{\partial x_2}|_{x_1=F_1(y),x_2=p} = \mathbb{P}[Y_1 \le y|V = p],
\end{aligned}$$

where the first equality solves the previous displayed equation, the third equality is due to the definition of a copula and $V \sim U[0, 1]$. For the second equality, note first that $C^{-1}_{1,Y_1,V}(C_{Y_1,V}(x_1, x_2), x_2) = x_1$ and differentiate both sides with respect to x_2 yields

$$\frac{\partial C^{-1}_{1,Y_1,V}\Big(C_{Y_1,V}(x_1, x_2), x_2\Big)}{\partial x_2} + \frac{\partial C^{-1}_{1,Y_1,V}\Big(C_{Y_1,V}(x_1, x_2), x_2\Big)}{\partial y}\frac{\partial C_{Y_1,V}\Big(x_1, x_2\Big)}{\partial x_2} = 0,$$

then the second equality holds by noticing $C_{Y_1,V}(F_1(y), p) = \mathbb{P}[Y \le y, D = 1|P = p]$ under the independence Assumption 2.

A.5. **Proof of Lemma 3.** (i) and (ii) hold obviously, in which cases $\mathbb{P}(Y_1 \le y|P_1 = p_1) = \mathbb{P}(Y_1 \le y|P_1 = p'_1)$ for all y, p_1 and p'_1. To see (iii) holds, note first $Y_{d0}|V_2 = v_2 \succeq_{FSD} Y_{d1}|V_2 = v_2$ implies $H_y(\cdot)$ is an increasing function. Since $P_2|P_1 = p'_1 \succeq_{FSD} P_2|P = p_1$, it follows that $\mathbb{E}_{P_2|P_1=p_1}[H_y(P_2)] - \mathbb{E}_{P_2|P_1=p'_1}[H_y(P_2)] \le 0$. Therefore, $\mathbb{P}(Y_1 \le y|P_1 = p_1) - \mathbb{P}(Y_1 \le y|P_1 = p'_1) \le 0$ for any $p'_1 \ge p_1$, that is, $Y_d|P_1 = p_1 \succeq_{FSD} Y_d|P_1 = p'_1$ for any $p'_1 \ge p_1$. □

A.6. **Proof of Lemma 4.** For Property (c) see Nelsen (2007, Theorem 5.2.10). For Properties (d) and (e) see Nelsen (2007, Corollary 5.2.6).

Properties (a) and (f) are obvious. For property (b), note first that $F_{Y_d|P}(y|p) = \frac{\partial C_{Y_d,P}(F_{Y_d}(y),F_P(p))}{\partial x_2}$ and $f_{Y_d|P}(y|p) = \frac{\partial^2 C_{Y_d,P}(F_{Y_d}(y),F_P(p))}{\partial x_1 \partial x_2} f_{Y_d}(y)$. Therefore,

$$\frac{f_{Y_d|P}(y|p)}{F_{Y_d|P}(y|p)} = \frac{\frac{\partial^2 C_{Y_d,P}(F_{Y_d}(y),F_P(p))}{\partial x_1 \partial x_2} f_{Y_d}(y)}{\frac{\partial C_{Y_d,P}(F_{Y_d}(y),F_P(p))}{\partial x_2}} = \frac{\partial \log \frac{\partial C_{Y_d,P}(F_{Y_d}(y),F_P(p))}{\partial x_2}}{\partial x_1} f_{Y_d}(y).$$

If P is an IHRD IV, then $\frac{f_{Y_d|P}(y|p)}{F_{Y_d|P}(y|p)}$ is non-decreasing in p for all y. Since $f_{Y_d}(y) > 0$ and $F_P(p)$ is increasing in p, it is equivalent to say $\frac{\partial \log \frac{\partial C_{Y_d,P}(x_1,x_2)}{\partial x_2}}{\partial x_1}$ is non-decreasing in x_2.

A.7. **Proof of Theorem 5.** under Assumption 10 we have: $f_{\mathbf{Q},\mathbf{V}}(\mathbf{q},\mathbf{v}) = f_{\mathbf{Q}}(\mathbf{q}) f_{\mathbf{V}}(\mathbf{v})$, this latter equality will be directly used in all the derivations below. For (i) to (iii) we have:

$$MTE_g(\mathbf{v}) \equiv \mathbb{E}[g(Y_1) - g(Y_0)|\mathbf{V} = \mathbf{v}] = \int_{\mathbf{q}} f_{\mathbf{Q}}(\mathbf{q}) DMTE_g(\mathbf{v},\mathbf{q}) d\mathbf{q}$$

$$ATE_g \equiv \mathbb{E}[g(Y_1) - g(Y_0)] = \int_{\mathbf{v}} \int_{\mathbf{q}} \underbrace{f_{\mathbf{Q}}(\mathbf{q}) f_{\mathbf{V}}(\mathbf{v})}_{w^{ATE}(\mathbf{v},\mathbf{q})} DMTE_g(\mathbf{v},\mathbf{q}) d\mathbf{q} d\mathbf{v}$$

$$\begin{aligned} LATE_g(\mathbf{u},\mathbf{u}') &\equiv \mathbb{E}[g(Y_1) - g(Y_0)|u_1 < V_1 \leq u_1', u_2 < V_2 \leq u_2'] \\ &= \int_{\mathbf{v}} \int_{\mathbf{q}} \underbrace{\frac{f_{\mathbf{Q}}(\mathbf{q}) f_{\mathbf{V}}(\mathbf{v}) 1_{\{\mathbf{v} \in [u_1,u_1'] \times [u_2,u_2']\}}}{F_{\mathbf{V}}\big(\mathbf{v} \in [u_1,u_1'] \times [u_2,u_2']\big)}}_{w^{LATE(\mathbf{u},\mathbf{u}')}(\mathbf{v},\mathbf{q})} DMTE_g(\mathbf{v},\mathbf{q}) d\mathbf{q} d\mathbf{v} \end{aligned}$$

For vector $\mathbf{a}$ and $\mathbf{b}$, let $\mathbf{a} \leq \mathbf{b}$ denote "component-wise smaller or equal to". Then,

$$\begin{aligned} ATT_g &\equiv \int_{\mathbf{q}} \mathbb{E}[g(Y_1) - g(Y_0)|D = 1, \mathbf{Q} = \mathbf{q}] dF_{\mathbf{Q}|D=1}(\mathbf{q}) \\ &= \int_{\mathbf{q}} \mathbb{E}[g(Y_1) - g(Y_0)|\mathbf{V} \leq \mathbf{q}, \mathbf{Q} = \mathbf{q}] dF_{\mathbf{Q}|D=1}(\mathbf{q}) \\ &= \int_{\mathbf{q}} \int_{\mathbf{v} \leq \mathbf{q}} \frac{1}{F_{\mathbf{V}}(\mathbf{q})} \mathbb{E}[g(Y_1) - g(Y_0)|\mathbf{V} = \mathbf{v}, \mathbf{Q} = \mathbf{q}] dF_{\mathbf{Q}|D=1}(\mathbf{q}) \\ &= \int_{\mathbf{v}} \int_{\mathbf{q}} \underbrace{\frac{f_{\mathbf{Q}}(\mathbf{q}) f_{\mathbf{V}}(\mathbf{v}) 1_{\{\mathbf{v} \in [0,q_1] \times [0,q_2]\}}}{\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q})]}}_{w^{ATT}(\mathbf{v},\mathbf{q})} DMTE_g(\mathbf{v},\mathbf{q}) d\mathbf{q} d\mathbf{v} \end{aligned}$$

where $dF_{\mathbf{Q}|D=1}(\mathbf{q}) = \frac{f_{\mathbf{Q}}(\mathbf{q})F_{\mathbf{V}}(\mathbf{q})}{\mathbb{P}(D=1)}$ by Bayesian rule and $\mathbb{P}(D=1) = \mathbb{E}[\mathbb{E}[D|\mathbf{Q}]] = \mathbb{E}[F_{\mathbf{V}}(\mathbf{Q})]$. Likewise, we can derive the ATUT weights as follows:

$$ATUT_g \equiv \int_{\mathbf{q}} \mathbb{E}[g(Y_1) - g(Y_0)|D=0, \mathbf{Q} = \mathbf{q}]dF_{\mathbf{Q}|D=0}(\mathbf{q})$$
$$= \int_{\mathbf{v}}\int_{\mathbf{q}} \underbrace{\frac{f_{\mathbf{Q}}(\mathbf{q})f_{\mathbf{V}}(\mathbf{v})1_{\{\mathbf{v}\notin[0,q_1]\times[0,q_2]\}}}{\mathbb{E}[1-F_{\mathbf{V}}(\mathbf{Q})]}}_{w^{ATUT}(\mathbf{v},\mathbf{q})} DMTE_g(\mathbf{v},\mathbf{q})d\mathbf{q}d\mathbf{v}$$

Concerning the $PRTE_g$, under Assumption 10 only, we have:

$$\mathbb{E}[g(Y^a)] = \int_{\mathbf{q}} \mathbb{E}[g(Y^a)|\mathbf{Q}^a = \mathbf{q}]dF_{\mathbf{Q}^a}(\mathbf{q}) = \int_{\mathbf{q}} \mathbb{E}[(g(Y_1^a) - g(Y_0^a))D^a|\mathbf{Q}^a = \mathbf{q}]dF_{\mathbf{Q}^a}(\mathbf{q}) + \mathbb{E}[g(Y_0^a)]$$
$$= \int_{\mathbf{v}}\int_{\mathbf{q}} 1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}} f_{\mathbf{Q}^a}(\mathbf{q})f_{\mathbf{V}^a}(\mathbf{v})\mathbb{E}[g(Y_1^a) - g(Y_0^a)|\mathbf{V}^a = \mathbf{v}, \mathbf{Q}^a = \mathbf{p}]d\mathbf{p}d\mathbf{v} + \mathbb{E}[g(Y_0^a)]$$
$$= \int_{\mathbf{v}}\int_{\mathbf{q}} 1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}} f_{\mathbf{Q}^a}(\mathbf{q})f_{\mathbf{V}^a}(\mathbf{v})DMTE_g^a d\mathbf{q}d\mathbf{v} + \mathbb{E}[g(Y_0^a)]$$

Since we have $DMTE_g^{a'} = DMTE_g^a$, $f_{\mathbf{V}^{a'}} = f_{\mathbf{V}^a}$, and $\mathbb{E}[g(Y_0^{a'})] = \mathbb{E}[g(Y_0^a)]$ under Assumption 11, then under both Assumptions 10 and 11 we have:

$$\mathbb{E}[g(Y^{a'}) - g(Y^a)] = \int_{\mathbf{v}}\int_{\mathbf{q}} [f_{\mathbf{Q}^{a'}}(\mathbf{q}) - f_{\mathbf{Q}^a}(\mathbf{q})]f_{\mathbf{V}^a}(\mathbf{v})1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}} DMTE_g^a d\mathbf{q}d\mathbf{v}$$

Therefore,

$$PRTE_g = \int_{\mathbf{v}}\int_{\mathbf{q}} \underbrace{\frac{[f_{\mathbf{Q}^{a'}}(\mathbf{q}) - f_{\mathbf{Q}^a}(\mathbf{q})]f_{\mathbf{V}^a}(\mathbf{v})1_{\{\mathbf{v}\in[0,q_1]\times[0,q_2]\}}}{\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^{a'})] - \mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^a)]}}_{w^{PRTE}(\mathbf{v},\mathbf{q})} DMTE_g^a d\mathbf{q}d\mathbf{v}.$$

For (iv): $Q_1(Z_1)$, $Q_2(Z_1)$ and $F_{\mathbf{V}}(\mathbf{v})$ are shown to be identified in Lee and Salanié (2018, Theorem 4.2). The remaining point that we need to show is that the joint distribution $F_{\mathbf{Q}}(\mathbf{q})$ is also point identified. Indeed, we have for $\mathbf{Z} = (Z_1, Z_2)$,

$$F_{\mathbf{Q}}(\mathbf{q}) = \mathbb{P}(Q_1(Z_1) \le q_1, Q_2(Z_2) \le q_2)$$
$$= \mathbb{P}\big(Z_1 \le Q_1^{-1}(q_1), Z_2 \le Q_2^{-1}(q_2)\big) = F_{\mathbf{Z}}\big(Q_1^{-1}(q_1), Q_2^{-1}(q_2)\big)$$

where the joint distribution $F_{\mathbf{Z}}(.,.)$ is directly observed from the data. The invertibility of Q_1 and Q_2 is ensured by Lee and Salanié (2018, Assumption 4.2).

A.8. **Proof of Lemma 5.** Equations (18) to (21) are direct applications of Joe (1996, Property 2), with some simplifications due to the fact that Assumption 10 imposes that $\mathbf{V} \perp \mathbf{Q}$.

Next we prove Equation (22). Given $d = 1$, $y \in \mathcal{Y}$, and $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

$$\mathbb{P}[Y \le y, D = 1|\mathbf{Q} = \mathbf{q}] = \mathbb{P}[Y_1 \le y, \mathbf{V} \le \mathbf{q}|\mathbf{Q} = \mathbf{q}] = \mathbb{P}[Y_1 \le y, V_1 \le q_1|V_2 \le q_2, \mathbf{Q} = \mathbf{q}]q_2$$
$$= \int_0^{q_2} \mathbb{P}[Y_1 \le y, V_1 \le q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}]dF_{V_2|\mathbf{Q}}(v_2|\mathbf{q}) = \int_0^{q_2} \mathbb{P}[Y_1 \le y, V_1 \le q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}]dv_2 \quad (25)$$

For a given v_2, the integrand $\mathbb{P}[Y_1 \le y, V_1 \le q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}]$ can be handled in a similar way as in the STC case:

$$\mathbb{P}[Y_1 \le y, V_1 \le q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}] = C_{Y_1,V_1|V_2,\mathbf{Q}}\left(F_{Y_1|V_2,\mathbf{Q}}(y|v_2,\mathbf{q}), F_{V_1|V_2,\mathbf{Q}}(q_1|v_2,\mathbf{q}); v_2, \mathbf{q}\right)$$
$$= C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2}(F_{Y_1|\mathbf{Q}}(y|\mathbf{q})), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right) = C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2} \circ c^{II}_{1,F_{Q_1|Q_2}}(F_{Y_1|Q_1}(y|q_1)), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right)$$
$$= C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2} \circ c^{II}_{1,F_{Q_1|Q_2}} \circ c^{I}_{1,F_{Q_1}(q_1)}(F_{Y_1}(y)), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right) \equiv \widetilde{\Psi}_{1,\mathbf{q}}(F_{Y_1}(y), v_2). \quad (26)$$

where the equalities hold by Equations 18, 19, and 20 and the assumption that $\mathbf{Q} \perp \mathbf{V}$, and the "$\circ$" denotes composite functions. Since for every y, $\mathbf{q}$, and v_2, the functions c^I, c^{II} and c^{III} are monotone, and $C_{Y_1,V_1|V_2,\mathbf{Q}}$ is also monotone in its first argument, then it follows that for each given $y \in \mathcal{Y}$ and $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

$$\mathbb{P}[Y \le y, D = 1|\mathbf{Q} = \mathbf{q}] = \int_0^{q_2} \widetilde{\Psi}_{1,\mathbf{q}}(F_{Y_1}(y), v_2)dv_2 \equiv \Psi_{1,\mathbf{q}}(F_{Y_1}(y))$$

is also a monotone function in $F_{Y_1}(y)$.

Remark 9. *If there was no V_2 and Q_2, that is, if the model is single threshold crossing model, then we would not have the additional integration in Equation (25), and we do not need to use the two layers of vine-copula operation c^{III} and c^I for V_2 and Q_2, respectively. In this case, the expression for $\Psi_{1,\mathbf{q}}$ exactly reduces to the expression for $\Psi_{1,p}$ in Lemma 2.*

For the case of $d = 0$, note that

$$\mathbb{P}[Y \le y, D = 0|\mathbf{Q} = \mathbf{q}] = \mathbb{P}[Y_0 \le y|\mathbf{Q} = \mathbf{q}] - \mathbb{P}[Y_0 \le y, D = 1|\mathbf{Q} = \mathbf{q}]$$
$$= c^{II}_{0,F_{Q_1|Q_2}} \circ c^{I}_{0,F_{Q_1}(q_1)}(F_{Y_0}(y)) - \int_0^{q_2} \widetilde{\Psi}_{0,\mathbf{q}}(F_{Y_0}(y), v_2)dv_2 \equiv \Psi_{0,\mathbf{q}}(F_{Y_0}(y)),$$

where $\widetilde{\Psi}_{0,\mathbf{q}}$ is defined in the same way as $\widetilde{\Psi}_{1,\mathbf{q}}$ with index "1" being replaced by "0".

A.9. **Discrete Outcome Variables.** In this subsection, we drop Assumption 5 and show how to extend Theorem 2 to the case of discrete outcome variables.

Assumption 13. *The joint density $f_{(V,P)|Y_d}(v, p|y)$of (V, P) given $Y_d = y$, $d = 0, 1$, exists and is positive for all $(v, p) \in [0, 1] \times [0, 1]$ and all $y \in \mathcal{Y} = \{y_1, y_2, \cdots, y_K\}$. Without loss of generality, assume the set $\mathcal{Y}$ is ordered: $y_j < y_\ell$ for $j < \ell$.*

Assumption 13 says that the marginal distribution of Y_d has finite support. Furthermore, the joint support of (Y_d, P, V) is "rectangular". Let $\mathcal{T}_d = \{F_{Y_d}(y_1), F_{Y_d}(y_2), \cdots, F_{Y_d}(y_K)\}$ be the set of values that F_{Y_d} can take. Similarly, defined $\mathcal{T}_d^p = \{F_{Y_d|P}(y_1|p), F_{Y_d|P}(y_2|p), \cdots, F_{Y_d|P}(y_K|p)\}$ be the set of values that $F_{Y_d|P}(\cdot|p)$ can take for each given p. Also define $\mathcal{T}_d^{Dp} = \{\mathbb{P}(Y \leq y_1, D = d|P = p), \cdots, \mathbb{P}(Y \leq y_K, D = d|P = p)\}$. Again, let $C_{Y_d,V|P=p}$ and $C_{Y_d,P}$ be the true copulas that generate the data. By Sklar's theorem, they must be strictly increasing in the first argument over $\mathcal{T}_d^p$ and $\mathcal{T}_d$, respectively. Let $C^{sub}_{Y_d,V|P=p}$ and $C^{sub}_{Y_d,P}$ be two sub-copulas that coincide with the true copulas $C_{Y_d,V|P=p}$ and $C_{Y_d,P}$ over $\mathcal{T}_d^p$ and $\mathcal{T}_d$, respectively.

Lemma 6 (Vine Copula)**.** *Under Assumptions 1 and 13, for each $y \in \mathcal{Y}$,*

$$F_{Y_d|P}(y|p) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,P}(x_1, x_2)\Big|_{x_1=F_{Y_d}(y), x_2=F_P(p)} \equiv c^{sub}_{d,F_P(p)}(F_{Y_d}(y)), \tag{27}$$

$$F_{Y_d|V,P}(y|v,p) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,V|P=p}(x_1, x_2)\Big|_{x_1=F_{Y_d|P}(y|p), x_2=v} \tag{28}$$

Also, for each given p, there exists strictly increasing mappings $\Gamma_{d,p}$: $\mathcal{T}_d \to \mathcal{T}_d^{Dp}$ such that

$$\mathbb{P}[Y \leq y, D = 1|P = p] = \Gamma_{1,p}(F_{Y_1}(y)) \equiv C^{sub}_{Y_1,V|P=p}\left(c^{sub}_{1,F_P(p)}(F_{Y_1}(y)), p; p\right), \tag{29}$$

$$\mathbb{P}[Y \leq y, D = 0|P = p] = \Gamma_{0,p}(F_{Y_0}(y)) \equiv c^{sub}_{0,F_P(p)}(F_{Y_0}(y)) - C^{sub}_{Y_0,V|P}\left(c^{sub}_{0,F_P(p)}(F_{Y_0}(y)), p; p\right). \tag{30}$$

That is, the observed probability $\mathbb{P}[Y \leq y, D = d|P = p]$ depends on y only through $F_{Y_d}(y)$.

Furthermore, fixing p, let $\Gamma^{(-1)}_{d,p}$ be defined as

$$\Gamma^{(-1)}_{d,p}(t) = \{u \in \mathcal{T}_d : \Gamma_{d,p}(u) = t\},$$

then $\Gamma^{(-1)}_{d,p}(t)$ is singleton for any $t \in \mathcal{T}_d^{Dp}$. Furthermore,

$$\Gamma^{(-1)}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p]) = \Gamma^{(-1)}_{d,p'}(\mathbb{P}[Y \leq y, D = d|P = p']), \tag{31}$$

Finally, the identified set is characterized by

$$\Theta_I = \Big\{\tilde{\theta} \in \widetilde{\Theta} : \textit{For } d \in \{0,1\}, (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}_d^c \times \mathcal{C}_d \textit{ who admits subcopulas satisfying Equation (31)}$$
$$\textit{and } \forall y \in \mathcal{Y}, F_{Y_d}(y) = \Gamma^{(-1)}_{d,p}(\mathbb{P}[Y \leq y, D = d|P = p])\Big\}.$$

Proof. First, we show that Equations (27) and (28) hold. By the Sklar (1959)'s theorem we known that there exists a copula $C_{Y_d,P}(x_1, x_2)$ such that $\mathbb{P}(Y_d \leq y, P \leq p) = C_{Y_d,P}(F_{Y_d}(y), F_P(p))$. Note that the copula $C_{Y_d,P}$ may not be unique, but the subcopula $C^{sub}_{Y_d,P}$, which defined on $\mathcal{T}_d \times [0,1]$, is uniquely determined. Then

$$\begin{aligned} F_{Y_d|P}(y|p) = \lim_{\delta \to 0} \frac{\mathbb{P}(Y_d \leq y, P \leq p + \delta) - \mathbb{P}(Y_d \leq y, P \leq p - \delta)}{\mathbb{P}(p - \delta \leq P \leq p + \delta)} \\ = \lim_{\delta \to 0} \frac{C_{Y_d,P}(F_{Y_d}(y), F_P(p + \delta)) - C_{Y_d,P}(F_{Y_d}(y), F_P(p - \delta))}{F_P(p + \delta) - F_P(p - \delta)} \\ = \frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2)\Big|_{x_1=F_{Y_d}(y), x_2=F_P(p)} \end{aligned}$$

where by Nelsen (2007, Theorem 2.2.7) the partial derivative $\frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2)$ exists and is non-decreasing for almost all x_1 on $[0, 1]$. Because $C^{sub}_{Y_d,P}$ coincide with $C_{Y_d,P}$ over $\mathcal{T}_d \times [0, 1]$, we must have $\frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,P}(x_1, x_2)$ for any $x_1 \in \mathcal{T}_d$ and $x_2 \in [0, 1]$. Furthermore, $\frac{\partial}{\partial x_2} C^{sub}_{Y_d,P}(x_1, x_2)$ must be strictly increasing in the first argument over $\mathcal{T}_d$ because $\frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2)$ is. This verifies Equation (27). Similarly, for almost all $x_1 \in [0, 1]$ there exists a partial derivative $\frac{\partial}{\partial x_2} C^{sub}_{Y_d,V|P}(x_1, x_2)$ that is non-decreasing x_1 such that the following holds

$$F_{Y_d|V,P}(y|v, p) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,V|P=p}(x_1, x_2)\Big|_{x_1=F_{Y_d|P}(y|p), x_2=F_{V|P}(v|p)} = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,V|P=p}(x_1, x_2)\Big|_{x_1=F_{Y_d|P}(y|p), x_2=v}$$

where the last equality holds because $F_{V|P}(v|p) = v$. This verifies Equation (7).

Now, fixing y, for $d = 1$

$$\begin{aligned}\mathbb{P}[Y \le y, D = 1|P = p] = \mathbb{P}[Y_1 \le y, V \le p|P = p] &= C^{sub}_{Y_1,V|P}(F_{Y_d|P}(y|p), p; p) \\ &= C^{sub}_{Y_1,V|P}(c^{sub}_{1,F_P(p)}(F_{Y_1}(y)), p; p),\end{aligned}$$

where the last equality holds by using Equation (27). As discussed earlier, over $\mathcal{T}_d$, $u \mapsto c^{sub}_{1,F_P(p)}(u)$ is strictly increasing, and over $\mathcal{T}^p_d$, $x_1 \mapsto C^{sub}_{Y_1,V|P}(x_1, x_2; p)$ is also strictly increasing, therefore $u \mapsto C_{Y_1,V|P=p}\left(c_{1,F_P(p)}(u), p; p\right) \equiv \Psi_{1,p}(u)$ is strictly increasing over $\mathcal{T}_d$. Similarly, for $d = 0$, then

$$\begin{aligned}\mathbb{P}[Y \le y, D = 0|P = p] = \mathbb{P}[Y_0 \le y, V > p|P = p] &= \mathbb{P}[Y_0 \le y|P = p] - \mathbb{P}[Y_0 \le y, V \le p|P = p] \\ &= c^{sub}_{0,F_P(p)}(F_{Y_0}(y)) - C^{sub}_{Y_0,V|P}(c^{sub}_{0,F_P(p)}(F_{Y_0}(y)), p; p) \equiv \Gamma_{0,p}(F_{Y_0}(y)),\end{aligned}$$

where the mapping $\Gamma_{0,p}(u)$ is strictly increasing in u over $\mathcal{T}_d$ because the left hand side of the equation above is increasing in y over $\mathcal{Y}$ (by the definition of conditioning probability), and $F_{Y_0}(y)$ is increasing in y over $\mathcal{Y}$. Because $\Gamma_{d,p}(u)$ is strictly increasing over $\mathcal{T}_d$, its inverse, as a subset of $\mathcal{T}_d$, must be a singleton. In the next step of the proof, we will show that the identified set is characterized by Equation (31). To verify the set defined in Theorem 2 is sharp, it is sufficient to show that Equations (29) and (30) and Equation (31) are equivalent. It is straightforward to see that Equations (29) and (30) imply Equation (31), we will verify the reverse.

Take a pair of candidate copula functions $C_{Y_d,V|P=p}$ and $C_{Y_d,P}$ (that respect the support condition) and suppose their subcopulas satisfy Equation (31), that is, $\Gamma^{(-1)}_{d,p}(\mathbb{P}[Y \le y, D = d|P = p])$ is flat in p for any $y \in \mathcal{Y}$. Note by construction and the definition of copula, $\Gamma^{(-1)}_{d,p}$ is strictly increasing in y over $\mathcal{Y}$ by construction.

Next from the definitions in Equations (29) and (30) we know that because $1 \in \mathcal{T}_d$ and $1 \in \mathcal{T}^p_d$, we have

$$c^{sub}_{1,F_P(p)}(1) = c_{1,F_P(p)}(1) = 1, \quad C^{sub}_{Y_1,V|P=p}(1, p; p) = C_{Y_1,V|P=p}(1, p; p) = p \Rightarrow \Gamma^{-1}_{1,p}(\mathbb{P}[Y \le y_K, D = 1|P = p]) = 1.$$

Also,

$$c^{sub}_{1,F_P(p)}(F_{Y_1}(y_1)) = c_{1,F_P(p)}(F_{Y_d}(y_1)) > 0, \quad C^{sub}_{Y_1,V|P=p}(F_{Y_1|P}(y_1|p), p; p) = C_{Y_1,V|P=p}(F_{Y_1|P}(y_1|p), p; p) > 0$$
$$\Rightarrow \Gamma^{-1}_{1,p}(\mathbb{P}[Y \leq y_1, D = 1|P = p]) > 0,$$

This says that $\Gamma^{-1}_{1,p}(\mathbb{P}[Y \leq \cdot, D = 1|P = p])$, as a function of y, is positive, strictly increasing, and no bigger than 1 over the set $\mathcal{Y}$. Therefore, it is valid distribution function for a discrete random variable that takes values from $\mathcal{Y}$, which we can choose as the counterfactual distribution F_{Y_1}. Similar argument applies to F_{Y_0}. This completes the proof.

APPENDIX B. TESTING MTE ASSUMPTIONS

B.1. Testable Implication.

Theorem 6 (Sharp characterization of the MTE assumptions)**.** *Let $Y, D, Y_1, Y_0, P(Z)$ define a potential outcome model $Y = Y_1 D + Y_0(1 - D)$. (i) If Assumptions 1 and 2 hold, then for all $y, y' \in \mathcal{Y}$, $\mathbb{P}(y < Y \leq y', D = 1|P = p)$ and $-\mathbb{P}(y < Y \leq y', D = 0|P = p)$ are non-decreasing in p for all $p \in \mathcal{P}$. (ii) If for all $y, y' \in \mathcal{Y}$ $\mathbb{P}(y < Y \leq y', D = 1|P = p)$ and $-\mathbb{P}(y < Y \leq y', D = 0|P = p)$ are non-decreasing in p for all $p \in \mathcal{P}$, there exists a joint distribution of $(\tilde{V}, \tilde{Y}_1, \tilde{Y}_0, P(Z))$ such that Assumptions 1 and 2 hold, and $(\tilde{Y}, \tilde{D}, P(Z))$ has the same distribution as $(Y, D, P(Z))$.*

Proof. Theorem 6-(i) is a direct application of HV05 testable implications where $G(Y) = 1\{Y \in [y, y']\}$ for $y \leq y'$. We show (ii) is true. We will assume that $\frac{\partial \mathbb{P}(y<Y\leq y', D=1|P=p)}{\partial p}$ is continuous over the set of limit points of $\mathcal{P}$. First, we construct $\tilde{V}$ and $\tilde{D}$ as follows:

$$\mathbb{P}(\tilde{V} \leq t|P = p) = t \;\; \forall t \in [0, 1] \text{ and } \forall p \in \mathcal{P}. \tag{32}$$

$$\tilde{D} = 1\{P(Z) \geq \tilde{V}\}. \tag{33}$$

Note that by construction, Assumption 1 is satisfied. Let $L(\mathcal{P})$ be the set of limit points of $\mathcal{P}$, $L^o(\mathcal{P})$ be a set of interior point of $\mathcal{P}$, and $C(\mathcal{P})$ be the closure of $\mathcal{P}$. Furthermore, let $I(\mathcal{P}) = C(\mathcal{P})/L^o(\mathcal{P})$ be the complement of $L^o(\mathcal{P})$ in the closure of $\mathcal{P}$. So $I(\mathcal{P})$ also contains isolation points. Note that $L^o(\mathcal{P})$ can be written as a union of countable or finite exclusive open intervals: $L^o(\mathcal{P}) = \cup_{j=1}^{J}(a_j, b_j)$, where $(a_j, b_j) \subseteq \mathcal{P}$, $b_j < a_{j+1}$, and J can be infinity. Let $\Omega(\mathcal{P})$ be a collection of intervals belonging to $(0, 1]$ defined as follows:

$$\Omega(\mathcal{P}) \equiv \{(p, p'] : p, p' \in I(\mathcal{P}) \cup \{0, 1\} \text{ and } \nexists \tilde{p} \in \mathcal{P}, \text{ such that } p < \tilde{p} < p'\}.$$

Then we again knows that Ω contains a generic element $(c_k, d_k]$, where $c_k, d_k \in I(\mathcal{P})$, $d_k \leq c_{k+1}$, $k = 1, 2, \cdots, K$ with K possibly taking ∞. Note that with above notation, for any $v \in (0, 1]$, v must belongs to one of the following categories: (i) an element of $L^o(\mathcal{P})$ so that $v \in (a_j, b_j)$ for some j, (ii) $v \in L(\mathcal{P})/L^o(\mathcal{P})$, and (iii) there exist an integer k such that $v \in (c_k, d_k]$. The following figure illustrates the partition of the unit interval.

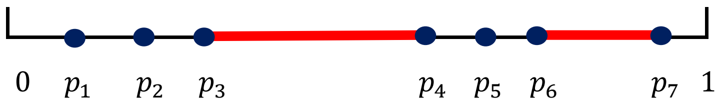

FIGURE 6. An illustration: $\mathcal{P} = \{p_1, p_2, p_5\} \cup [p_3, p_4] \cup [p_6, p_7]$, $L^o(\mathcal{P}) = (p_3, p_4) \cup (p_6, p_7)$, and $\Omega(\mathcal{P}) = \{(0, p_1], (p_1, p_2], (p_4, p_5], (p_5, p_6], (p_7, 1]\}$.

Next, we propose the following distribution for $\tilde{Y}_1|\tilde{V}, P$. For any arbitrary $p \in \mathcal{P}$ and $v \in (0, 1]$, we define

$$\mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p) \quad = \quad \begin{cases} \frac{\partial}{\partial t}\mathbb{P}(Y \le y, D = 1|P = t)|_{t=v} & \text{if } v \in L^o(\mathcal{P}) \\ \lim_{\tilde{v}\to v} \frac{\partial}{\partial t}\mathbb{P}(Y \le y, D = 1|P = t)|_{t=\tilde{v}} & \text{if } v \in L(\mathcal{P})/L^o(\mathcal{P}) \\ \frac{\mathbb{P}(Y \le y, D=1|P=d_k) - \mathbb{P}(Y \le y, D=1|P=c_k)}{d_k - c_k} & \text{if } v \notin L(P) \text{ but } v \in (c_k, d_k] \in \Omega(\mathcal{P}). \end{cases}$$

$$\mathbb{P}(\tilde{Y}_0 \le y|\tilde{V} = v, P = p) \quad = \quad \begin{cases} -\frac{\partial}{\partial v}\mathbb{P}(Y \le y, D = 0|P = t)|_{t=v} & \text{if } v \in L^o(\mathcal{P}) \\ -\lim_{\tilde{v}\to v} \frac{\partial}{\partial v}\mathbb{P}(Y \le y, D = 0|P = t)|_{t=\tilde{v}} & \text{if } v \in L^o(\mathcal{P}) \\ \frac{\mathbb{P}(Y \le y, D=0|P=c_k) - \mathbb{P}(Y \le y, D=0|P=d_k)}{d_k - c_k} & \text{if } v \notin L^o(P) \text{ but } v \in (c_k, d_k] \in \Omega(\mathcal{P}). \end{cases}$$

Note that the conditioning on $\tilde{V} = v$, the distribution of $\tilde{Y}_1$ does not depend on P. Hence, Assumption 2 is satisfied by construction.

We now show that the distribution function constructed above is well defined. We focus on $\mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)$ and the verification for $\mathbb{P}(\tilde{Y}_0 \le y|\tilde{V} = v, P = p)$ is analogous.

(1) $\mathbb{P}(\tilde{Y}_1 < \underline{y} - \epsilon|\tilde{V} = v, P = p) = 0$ for all $v \in [0, 1]$ and for any arbitrarily small $\epsilon > 0$. To see this, suppose $v \notin L(\mathcal{P})$, then there exists $[c_k, d_k] \in \Omega(\mathcal{P})$ such that $v \in (c_k, d_k]$, therefore,

$$\mathbb{P}(\tilde{Y}_1 \le \underline{y} - \epsilon|\tilde{V} = v, P = p)$$
$$= \frac{\mathbb{P}(Y \le \underline{y} - \epsilon, D = 1|P = d_k) - \mathbb{P}(Y \le \underline{y} - \epsilon, D = 1|P = c_k)}{d_k - c_k} = \frac{0 - 0}{d_k - c_k} = 0.$$

On the other hand, if $v \in L^o(\mathcal{P})$, then $\mathbb{P}(Y \le \underline{y} - \epsilon, D = 1|P = \tilde{v}) = 0$ for all $\tilde{v}$ in a small neighborhood of v, which implies $\frac{\partial}{\partial v}\mathbb{P}(Y \le \underline{y} - \epsilon, D = 1|P = v) = 0$. The case that $v \in L^o(\mathcal{P})$ follows straightforwardly.

(2) $\mathbb{P}(\tilde{Y}_1 \le \overline{y}|\tilde{V} = v, P = p) = 1$. First, if $v \in L^o(\mathcal{P})$, then $\mathbb{P}(Y \le \overline{y}, D = 1|P = v) = \mathbb{P}(D = 1|P = v) = v \Rightarrow \frac{\partial}{\partial v}\mathbb{P}(Y \le \overline{y}, D = 1|P = v) = 1$. On the other hand, if $v \notin L(\mathcal{P})$, then

$$\mathbb{P}(\tilde{Y}_1 \le \overline{y}|\tilde{V} = v, P = p)$$
$$= \frac{\mathbb{P}(Y \le \overline{y}, D = 1|P = d_k) - \mathbb{P}(Y \le \overline{y}, D = 1|P = c_k)}{p' - p} = \frac{d_k - c_k}{d_k - c_k} = 1.$$

(3) $\mathbb{P}(\tilde{Y}_1 \le y | \tilde{V} = v, P = p)$ is nondecreasing in y. For $y < y'$ we have

$$\mathbb{P}(\tilde{Y}_1 \le y' | \tilde{V} = v, P = p) - \mathbb{P}(\tilde{Y}_1 \le y | \tilde{V} = v, P = p) =$$

$$\begin{cases} \frac{\partial}{\partial t}\mathbb{P}(y < Y \le y', D = 1 | P = t)|_{t=v} \ge 0 & \text{if } v \in L^o(\mathcal{P}), \\ \lim_{\tilde{v} \to v} \frac{\partial}{\partial t}\mathbb{P}(y < Y \le y, D = 1 | P = t)|_{t=\tilde{v}} \ge 0 & \text{if } v \in L(\mathcal{P})/L^o(\mathcal{P}) \\ \frac{\mathbb{P}(y<Y\le y', D=1|P=d_k) - \mathbb{P}(y<Y\le y', D=1|P=c_k)}{d_k - c_k} \ge 0 \text{ if } v \notin L^o(P) \text{ but } v \in [c_k, d_k] \in \Omega(\mathcal{P}), \end{cases}$$

where the last inequalities hold whenever the testable implications hold, i.e. $\mathbb{P}(y < Y \le y', D = 1 | P = p)$ is a non-decreasing function for all $p \in \mathcal{P}$ and all $y < y'$, and by the continuously differentiability of $\mathbb{P}(y < Y \le y', D = 1 | P = p)$ over $L(\mathcal{P})$.

Finally, we show that $(\tilde{V}, \tilde{Y}_d, P(Z))$, $d \in \{0, 1\}$ is observationally equivalent to $(V, Y_d, P(Z))$ $d \in \{0, 1\}$. For this, we show that the joint distribution of $(\tilde{Y}, \tilde{D}, P(Z))$ is the same as the joint distribution of $(Y, D, P(Z))$. Take an arbitrary $z \in \mathscr{Z}$ and let $p = p(z) \in \mathcal{P}$.

Suppose first $p \notin L^o(\mathcal{P})$, then $(0, p]$ can be expressed as unions of exclusive intervals $\left(\cup_{j=1}^{J^*}(a_j, b_j)\right) \cup \left(\cup_{k=1}^{K^*}(c_k, d_k]\right)$ for some J^* and K^*, where $(a_j, b_j)s$ are connected subsets of $\mathcal{P}$. Therefore,

$$\begin{aligned} \mathbb{P}(\tilde{Y} \le y, \tilde{D} = 1 | P = p) &= \mathbb{P}(\tilde{Y}_1 \le y, \tilde{V} \le p | P = p) = \int_0^p \mathbb{P}(\tilde{Y}_1 \le y | \tilde{V} = v, P = p) dv \\ &= \sum_{j=1}^{J^*} \int_{a_j}^{b_j} \mathbb{P}(\tilde{Y}_1 \le y | \tilde{V} = v, P = p) dv + \sum_{k=1}^{K^*} \int_{c_k}^{d_k} \mathbb{P}(\tilde{Y}_1 \le y | \tilde{V} = v, P = p) dv \\ &= \sum_{j=1}^{J^*} \Big(\mathbb{P}(Y \le y, D = 1 | P = b_j) - \mathbb{P}(Y \le y, D = 1 | P = a_j)\Big) \\ &+ \sum_{k=1}^{K^*} (\mathbb{P}(Y \le y, D = 1 | P = d_k) - \mathbb{P}(Y \le y, D = 1 | P = c_k)) \\ &= \mathbb{P}(Y \le y, D = 1 | P = p) - \mathbb{P}(Y \le y, D = 1 | P = 0) = \mathbb{P}(Y \le y, D = 1 | P = p), \end{aligned}$$

where the first equality is by definition, the second equality is by construction that $\tilde{V}$ satisfies Assumption 1, the fourth equality is obtained by inserting the constructed counterfactural distributions, and the second last equality holds because $(0, p]$ can be expressed as unions of exclusive intervals $\left(\cup_{j=1}^{J^*}(a_j, b_j)\right) \cup \left(\cup_{k=1}^{K^*}(c_k, d_k]\right)$.

Suppose that $p \in (a_{j^*}, b_{j^*}) \subseteq L^0(\mathcal{P})$ for some j^*, then the right hand side equals to

$$
\begin{aligned}
\mathbb{P}(\tilde{Y} \leq y, \tilde{D} = 1 | P = p) = \mathbb{P}(\tilde{Y}_1 \leq y, \tilde{V} \leq p | P = p) &= \int_0^p \mathbb{P}(\tilde{Y}_1 \leq y | \tilde{V} = v, P = p) dv \\
&= \int_0^{a_{j^*}} \mathbb{P}(\tilde{Y}_1 \leq y | \tilde{V} = v, P = p) dv + \int_{a_{j^*}}^p \mathbb{P}(\tilde{Y}_1 \leq y | \tilde{V} = v, P = p) dv \\
&= \mathbb{P}(Y \leq y, D = 1 | P = a_{j^*}) + \int_{a_{j^*}}^p \frac{\partial}{\partial v} \mathbb{P}(Y \leq y, D = 1 | P = v) dv \\
&= \mathbb{P}(Y \leq y, D = 1 | P = a_{j^*}) + \mathbb{P}(Y \leq y, D = 1 | P = p) \\
&\quad - \mathbb{P}(Y \leq y, D = 1 | P = a_{j^*}) = \mathbb{P}(Y \leq y, D = 1 | P = p),
\end{aligned}
$$

where the fourth equality holds by the above argument and the fifth equality holds by inserting the constructed counterfactural distributions. This completes the proof. □

It is worth noting that testable implications of the MTE assumptions has been previously derived in Heckman and Vytlacil (2005, Appendix A). More precisely, HV05 derived testable implications of Assumptions 1 and 2 are: for any non-negative measurable function, i.e. $G^+(.) : \mathcal{Y} \to \mathbb{R}^+$ we have

$$
\frac{\partial}{\partial p} \mathbb{E}[G^+(Y) D | P = p] \geq 0, \tag{34}
$$

$$
\frac{\partial}{\partial p} \mathbb{E}[G^+(Y)(1 - D) | P = p] \leq 0. \tag{35}
$$

The main contributions of Theorem 6 are (i) it shows that the testable implication is sharp, and (ii) it shows that we do not need to visit all those non-negative measurable functions, but we can restricts our attention to a tractable sub-class which is sufficient to screen all possible observable violations of the MTE assumptions. This tractable characterization has the direct advantage to propose a formal statistical test to screen the violation of the MTE assumptions.

B.2. **Testing Procedures.** The sharp testable implications of the MTE assumptions can be summarized as follows:

$$
\mathbb{P}(y < Y \leq y', D = 1 | P = \tilde{p}) \geq \mathbb{P}(y < Y \leq y', D = 1 | P = p), \tag{36}
$$

$$
-\mathbb{P}(y < Y \leq y', D = 0 | P = \tilde{p}) \geq -\mathbb{P}(y < Y \leq y', D = 0 | P = p), \tag{37}
$$

for all $\tilde{p} \geq p$ with $(\tilde{p}, p) \in \mathcal{P}^2$ and $y, y' \in \mathcal{Y}$.

Recently, Hsu et al (2018) propose a uniform a valid test for null hypothesis of the type:

$$
\begin{aligned}
&\mathbb{E}[f^{(1)}(W, \tau) | S = \tilde{s}, X = x] \geq \mathbb{E}[f^{(2)}(W, \tau) | S = s, X = x], \\
&\qquad \tilde{s} \geq s \text{ with } (\tilde{s}, s) \in \mathcal{S}^2, x \in \mathcal{X}, \text{ and } \tau \in \Omega,
\end{aligned} \tag{38}
$$

where W is a vector of observed random variables that contains S and X as elements, $f^{(1)}(W, \tau), f^{(2)}(W, \tau)$ are known real valued functions indexed by $\tau \in \Omega$ —which can be finite or infinite. By defining the following $\tau \equiv (d, k) \in \Omega \equiv \{(d, k) : d \in \{0, 1\}, k = (y, y') : -\infty \leq y \leq y' \leq \infty\}$, $f^{(1)}(W, (1, k)) = f^{(2)}(W, (1, k)) = 1\{Y \in C_k\} 1\{D = 1\}$,

$f^{(1)}(W,(0,k)) = f^{(2)}(W,(0,k)) = -1\{Y \in C_k\}1\{D=0\}$ with $C_k \equiv [y,y'] \cap \mathcal{Y}$, and $P = S$; we see that the MTE sharp testable implications, i.e. eqs (36,37) could be recasted into Hsu, Liu, and Shi (2019) testing framework if P was an observed random variable. In our context, P is not observed but could be consistently estimated. To test the MTE testable implications one can proceed in two steps: (a) the first step consist in finding a consistent estimator for P, i.e. $\hat{P}$, (b) second consist in using Hsu, Liu, and Shi (2019) testing procedure with $\hat{P}$. A theoretical challenge is to ensure that the pre-estimation does not affect the statistical properties of Hsu, Liu, and Shi (2019) testing procedure.

Appendix C. Additional results on Copulas

C.1. Farlie-Gumbel-Morgenstern (FGM) Copula.

Assumption 14 (FGM Copula)**.** *There exists $\theta = (\alpha_0, \alpha_1, \delta_1, \delta_0) \in \Theta \subseteq \mathbb{R}^T$ with $T < \infty$ such that $C_{Y_d,P}(x_1,x_2) = x_1x_2\big(1+\alpha_d(1-x_1)(1-x_2)\big)$ and $C_{Y_d,V|P=p}(x_1,x_2) = x_1x_2\big(1+\sigma_d(p)(1-x_1)(1-x_2)\big)$ for $\sigma_d(p) \in [-1,1]$, $d \in \{0,1\}$, where $\sigma_d(p)$ is known up to a finite number of parameters δ_d.*

Corollary 3. *Under Assumptions 1, 5, 6 and 14, the identified set Θ_I^{SP} of $\tilde{\theta} \equiv \big(\theta, (F_{Y_d}(y) : d \in \{0,1\}, y \in \mathbb{R})\big)$ is characterized as follows:*

$$\begin{aligned}\Theta_I^{SP} \quad = \quad & \Bigg\{ \textit{For } d \in \{0,1\}, F_{Y_d}(y;\theta) = \frac{\alpha_d(1-2F_P(p))+1-\sqrt{(\alpha_d(1-2F_P(p))+1)^2-4\alpha_d(1-2F_P(p))H_d}}{2\alpha_d(1-2F_P(p))} \\ & \theta = (\alpha_d, \delta_d) \textit{ satisfies } (\Delta_d - \Delta_d^{\frac{1}{2}})\frac{\partial B_d}{\partial p} = B_d \frac{\partial \Delta_d}{\partial p} \textit{ for all } p \Bigg\},\end{aligned}$$

where $B_d = \alpha_d(1-2F_P(p))$, $\Delta_d = (B_d+1)^2 - 4B_dH_d$, and

$$H_1(y,p,\sigma_1(p)) = \frac{\sigma_1(p)p(1-p)+p-\sqrt{(\sigma_1(p)p(1-p)+p)^2-4\sigma_1(p)p(1-p)F_{Y,D|P}(y,1|p)}}{2\sigma_1(p)p(1-p)}.$$

and

$$H_0(y,p,\sigma_0(p)) = \frac{\sigma_0(p)p(1-p)+p+1+\sqrt{(\sigma_0(p)p(1-p)+p+1)^2+4\sigma_0(p)p(1-p)F_{Y,D|P}(y,0|p)}}{2\sigma_0(p)p(1-p)}.$$

Proof. Consider $d=1$. Note first by equation (8) and assumption 14,

$$F_{Y,D|P}(y,1|p) = c_{1,F_P(p)}(F_{Y_1}(y))p\big(1+\sigma_1(p)(1-c_{1,F_P(p)}(F_{Y_1}(y)))(1-p)\big).$$

Let $A = \sigma_1(p)p(1-p)$. We focus on the case $A \neq 0$, that is, we do not consider $p=0$ or $p=1$ or values of p such that $\sigma_1(p) = 0$, because the solution is straightforward for those values. From the above equation, we have two possible solutions for $c_{1,F_P(p)}(F_{Y_1}(y))$,

$$\frac{A+p+\sqrt{(A+p)^2-4AF_{Y,D|P}(y,1|p)}}{2A} \quad \text{or} \quad \frac{A+p-\sqrt{(A+p)^2-4AF_{Y,D|P}(y,1|p)}}{2A}$$

The expression in the square root sign is always non-negative. This is obviously true when $A<0$. When $A>0$,

$$(A+p)^2-4AF_{Y,D|P}(y,1|p)=(A-p)^2+4A(p-F_{Y,D|P}(y,1|p))\geq 4A(p-F_{Y,D|P}(y,1|p))\geq 0,$$

where the last inequality holds because $p=\mathbb{P}(D=1|P=p)\geq\mathbb{P}(Y\leq y,D=1|P=p)=F_{Y,D|P}(y,1|p)$.

Although both solutions are well defined, the first solution is not valid for this context. When $\sigma_1(p)>0$, we have $p>A>0$ (unless $p=0$), and,

$$\frac{A+p+\sqrt{(A+p)^2-4AF_{Y,D|P}(y,1|p)}}{2A}\geq\frac{A+p}{2A}>1.$$

When $\sigma_1(p)<0$, we have $A<0$ and $p+A>0$, and,

$$\frac{A+p+\sqrt{(A+p)^2-4AF_{Y,D|P}(y,1|p)}}{2A}<0$$

because the numerator and denominator have opposite sign. Therefore, the first solution can not be valid since $c_{1,F_P(p)}(F_{Y_1}(y))=F_{Y_d|P}(y,p)$ must takes values from $[0,1]$.

It remains to verify the second solution is valid. First consider the case $A>0$. Let $\Delta=(A+p)^2-4AF_{Y,D|P}(y,1|p)$. Then,

$$\begin{aligned}0\leq F_{Y,D|P}(y,1|p)\leq p\Rightarrow(A-p)^2\leq(A+p)^2-4AF_{Y,D|P}(y,1|p)\leq(A+p)^2&\\ \Rightarrow 0\leq A+p-\sqrt{\Delta}\leq 2A\Rightarrow 0\leq\frac{A+p-\sqrt{\Delta}}{2A}\leq 1&\end{aligned}$$

Next consider the case $A<0$,

$$\begin{aligned}0\leq F_{Y,D|P}(y,1|p)\leq p\Rightarrow(p+A)^2\leq(A+p)^2-4AF_{Y,D|P}(y,1|p)\leq(p-A)^2&\\ \Rightarrow 2A\leq A+p-\sqrt{\Delta}\leq 0\Rightarrow 0\leq\frac{A+p-\sqrt{\Delta}}{2A}\leq 1&\end{aligned}$$

To summarize, we have one valid solution

$$c_{1,F_P(p)}(F_{Y_1}(y))=\frac{\sigma_1(p)p(1-p)+p-\sqrt{(\sigma_1(p)p(1-p)+p)^2-4\sigma_1(p)p(1-p)F_{Y,D|P}(y,1|p)}}{2\sigma_1(p)p(1-p)}\equiv H_1(y,p,\sigma_1(p)).$$

The right hand side only depends on unknown finite dimensional parameters and quantities that can be identified from data.

Next, recall that

$$\begin{aligned}H_1(y,p,\sigma_1(p))=c_{1,F_P(p)}(F_{Y_1}(y))&=\frac{\partial C_{Y_1,P}(x_1,x_2)}{\partial x_2}|_{x_1=F_{Y_1}(y),x_2=F_P(p)}\\ &=x_1+\alpha_1x_1(1-x_1)(1-2x_2)|_{x_1=F_{Y_1}(y),x_2=F_P(p)}=F_{Y_1}(y)+\alpha_1F_{Y_1}(y)(1-F_{Y_1}(y))(1-2F_P(p))\end{aligned}$$

Solve $F_{Y_1}(y)$ from the above equation, we get again two possible solutions.

$$\frac{B_1+1+\sqrt{(B_1+1)^2-4B_1H_1}}{2B_1}, \quad \text{or} \quad \frac{B_1+1-\sqrt{(B_1+1)^2-4B_1H_1}}{2B_1}$$

where $B_1=\alpha_1(1-2F_P(p))$. Again, we restrict our attention to the case $B_1\neq 0$, otherwise $F_{Y_1}(y)$ has a unique solution which equals to $H_1(y,p,\sigma_1(p))$. Note also that $B_1\in[-1,1]$, so $B_1+1\geq 2B_1$ and $B_1+1\geq 0$. Following similar argument as above (and use the fact that $0\leq H_1\leq 1$, we can show that the first solution is not valid while the second solution is. Therefore, it must be the case that

$$F_{Y_1}(y)=\frac{\alpha_1(1-2F_P(p))+1-\sqrt{(\alpha_1(1-2F_P(p))+1)^2-4\alpha_1(1-2F_P(p))H_1(y,p,\sigma_1(p))}}{2\alpha_1(1-2F_P(p))},$$

where

$$H_1=\frac{\sigma_1(p)p(1-p)+p-\sqrt{(\sigma_1(p)p(1-p)+p)^2-4\sigma_1(p)p(1-p)F_{Y,D|P}(y,1|p)}}{2\sigma_1(p)p(1-p)}.$$

Now consider $d=0$. Note first by equation (9) and assumption 14,

$$F_{Y,D|P}(y,0|p)=c_{0,F_P(p)}(F_{Y_0}(y))-c_{0,F_P(p)}(F_{Y_0}(y))p\big(1+\sigma_0(p)(1-c_{0,F_P(p)}(F_{Y_0}(y)))(1-p)\big).$$

From the above equation, we have two possible solutions for $c_{0,F_P(p)}(F_{Y_0}(y))$,

$$\frac{A_0+p-1+\sqrt{(A_0+p-1)^2+4A_0F_{Y,D|P}(y,0|p)}}{2A_0} \quad \text{or} \quad \frac{A_0+p-1-\sqrt{(A_0+p-1)^2+4A_0F_{Y,D|P}(y,0|p)}}{2A_0}$$

where $A_0=\sigma_0(p)p(1-p)$, so $A_0+p-1=(1-p)(\sigma_0(p)p-1)<0$. The term inside the square root is always non-negative. It is obvious when $A_0>0$. When $A_0<0$, we have

$$\begin{aligned}(A_0+p-1)^2+4A_0F_{Y,D|P}(y,0|p)&=(A_0-p+1)^2+4A_0(F_{Y,D|P}(y,0|p)-(1-p))\\&\geq 4A_0(F_{Y,D|P}(y,0|p)-(1-p))=4A_0(P(Y>y,D=1|P=p)-P(Y>y|P=p))\geq 0.\end{aligned}$$

Although both solutions are well-defined, we will argue the second solution is not valid. To see this, when $A_0>0$, the numerator is negative and denominator is positive, implies

$$\frac{A_0+p-1-\sqrt{(A_0+p-1)^2+4A_0F_{Y,D|P}(y,0|p)}}{2A_0}<0.$$

When $A_0<0$, since $A_0+1-p>0$ so $2A_0>A_0+p-1$, we have

$$\frac{A_0+p-1-\sqrt{(A_0+p-1)^2+4A_0F_{Y,D|P}(y,0|p)}}{2A_0}\geq\frac{A_0+p-1}{2A_0}>\frac{2A_0}{2A_0}=1.$$

On the other hand, the first solution always falls between 0 and 1. When $\sigma_0(p) > 0$, $A_0 > 0$, we have

$$\frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2 + 4A_0 F_{Y,D|P}(y,0|p)}}{2A_0} = \frac{A_0 + p - 1 + \sqrt{(A_0 - p + 1)^2 + 4A_0(F_{Y,D|P}(y,0|p) - (1-p))}}{2A_0}$$
$$\leq \frac{A_0 + p - 1 + \sqrt{(A_0 - p + 1)^2}}{2A_0} = \frac{2A_0}{2A_0} = 1,$$

and

$$\frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2 + 4A_0 F_{Y,D|P}(y,0|p)}}{2A_0} \geq \frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2}}{2A_0} \geq 0.$$

When $\sigma_0(p) < 0$, $A_0 < 0$, we have

$$\frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2 + 4A_0 F_{Y,D|P}(y,0|p)}}{2A_0} \geq \frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2}}{2A_0} = \frac{0}{2A_0} = 0,$$

$$\frac{A_0 + p - 1 + \sqrt{(A_0 + p - 1)^2 + 4A_0 F_{Y,D|P}(y,0|p)}}{2A_0} = \frac{A_0 + p - 1 + \sqrt{(A_0 - p + 1)^2 + 4A_0(F_{Y,D|P}(y,0|p) - (1-p))}}{2A_0}$$
$$\leq \frac{A_0 + p - 1 + \sqrt{(A_0 - p + 1)^2}}{2A_0} = \frac{2A_0}{2A_0} = 1,$$

To conclude, we must have

$$c_{0,F_P(p)}(F_{Y_0}(y)) = \frac{\sigma_0 p(1-p) + p + 1 + \sqrt{(\sigma_0 p(1-p) + p + 1)^2 + 4\sigma_0 p(1-p) F_{Y,D|P}(y,0|p)}}{2\sigma_0 p(1-p)} \equiv H_0(y, p, \sigma_0(p)).$$

Finally, repeating what we did for $F_{Y_1}(y)$ and $c_{1,F_P(p)}(F_{Y_1}(y))$, we can do the same and obtain

$$F_{Y_0}(y) = \frac{\alpha_0(1 - 2F_P(p)) + 1 - \sqrt{(\alpha_0(1 - 2F_P(p)) + 1)^2 - 4\alpha_0(1 - 2F_P(p)) H_0(y, p, \sigma_0(p))}}{2\alpha_0(1 - 2F_P(p))}.$$

□

C.2. **Bernstein Copulas.** The density of the Bernstein copula is given by:

$$c_{Y_d,V|P}(x_1, x_2; \alpha^d) = K_d L_d \sum_{k=1}^{K_d} \sum_{l=1}^{L_d} \alpha^d_{kl} b_{k-1,K_d-1}(x_1) b_{l-1,L_d-1}(x_2)$$

Note that $b_{i,I}(u)$ has an alternative representation:

$$b_{i,I}(u) = \sum_{j=i}^{I} (-1)^{j-i} \binom{I}{j} \binom{j}{i} u^j \tag{39}$$

First let assume that Assumptions 1 and 5 hold, then under Assumption 9(i), we have

$$\begin{aligned}\mathbb{E}[g(Y_d)|V=v,P=p] &= \int_{\mathcal{Y}} g(y)K_dL_d\sum_{k=1}^{K_d}\sum_{l=1}^{L_d}\alpha_{kl}^d b_{k-1,K_d-1}\left(F_{Y_d|P}(y|p)\right)b_{l-1,L_d-1}(v)f_{Y_d|P}(y|p)dy \\ &= K_dL_d\sum_{k=1}^{K_d}\sum_{l=1}^{L_d}\alpha_{kl}^d b_{l-1,L_d-1}(v)\int_{\mathcal{Y}} g(y)b_{k-1,K_d-1}(F_{Y_d|P}(y|p))f_{Y_d|P}(y|p)dy.\end{aligned}$$

In addition, under Assumption 9(ii) we can derive the following:

$$F_{Y_d|P}(y|p) = R_dS_d\sum_{r=1}^{R_d}\sum_{s=1}^{S_d}\underbrace{\beta_{rs}^d B_{r-1,R_d-1}(F_{Y_d}(y))b_{s-1,S_d-1}(F_P(p))}_{\chi_{rs}^d},$$

$$f_{Y_d|P}(y|p) = f_{Y_d}(y)R_dS_d\sum_{r=1}^{R_d}\sum_{s=1}^{S_d}\underbrace{\beta_{rs}^d b_{r-1,R_d-1}(F_{Y_d}(y))b_{s-1,S_d-1}(F_P(p))}_{\zeta_{rs}^d}.$$

To ease the notation, when there is no confusion we will make the following abuse of notation $F_P \equiv F_P(p)$ and $F_{Y_d} \equiv F_{Y_d}(y)$.

$$\begin{aligned}b_{k-1,K_d-1}(F_{Y_d|P}(y|p)) &= \sum_{j=k-1}^{K_d-1}(-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}\left(F_{Y_d|P}(y|p)\right)^j \\ &= \sum_{j=k-1}^{K_d-1}(-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}(R_dS_d)^j\left(\sum_{r=1}^{R_d}\sum_{s=1}^{S_d}\chi_{rs}^d\right)^j \\ &= \sum_{j=k-1}^{K_d-1}(-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}(R_dS_d)^j\sum_{n_{11}+n_{12}+...+n_{R_dS_d}=j}\binom{j}{n_{11},n_{12},...,n_{R_dS_d}}(\chi_{11}^d)^{n_{11}}(\chi_{12}^d)^{n_{12}}\cdots(\chi_{R_dS_d}^d)^{n_{R_dS_d}}\end{aligned}$$

and,

$$\begin{aligned}b_{k-1,K_d-1}(F_{Y_d|P}(y|p))f_{Y_d|P}(y|p) = \sum_{r=1}^{R_d}\sum_{s=1}^{S_d}\Bigg\{&\sum_{j=k-1}^{K_d-1}(-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}(R_dS_d)^{j+1} \\ &\sum_{n_{11}+n_{12}+...+n_{R_dS_d}=j}\binom{j}{n_{11},n_{12},...,n_{R_dS_d}}(\chi_{11}^d)^{n_{11}}(\chi_{12}^d)^{n_{12}}\cdots(\chi_{R_dS_d}^d)^{n_{R_dS_d}}\Bigg\}\zeta_{rs}^d f_{Y_d}(y)\end{aligned}$$

We have the following factorization:

$$\begin{aligned}&(\chi_{11}^d)^{n_{11}}(\chi_{12}^d)^{n_{12}}\cdots(\chi_{R_dS_d}^d)^{n_{R_dS_d}}\zeta_{rs}^d \\ &\quad = \beta_{rs}^d b_{s-1,S_d-1}(F_P)\prod_{e=1,f=1}^{R_d,S_d}\left(\beta_{ef}^d\right)^{n_{ef}}\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{n_{e\cdot}}\prod_{f=1}^{S_d}\left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}}b_{r-1,R_d-1}(F_{Y_d})\end{aligned}$$

where $n_{e\cdot} = \sum_{f=1}^{S_d} n_{ef}$ and $n_{\cdot f} = \sum_{e=1}^{R_d} n_{ef}$. Then, we have

$$\int_{\mathcal{Y}} g(y)(\chi_{11}^d)^{n_{11}}(\chi_{12}^d)^{n_{12}} \cdots (\chi_{R_d S_d}^d)^{n_{R_d S_d}} \zeta_{rs}^d f_{Y_d}(y)dy$$
$$= \beta_{rs}^d b_{s-1,S_d-1}(F_P) \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \prod_{e=1,f=1}^{R_d,S_d} (\beta_{ef}^d)^{n_{ef}} \underbrace{\mathbb{E}\Big[g(Y_d) \prod_{e=1}^{R_d} \left(B_{e-1,R_d-1}(F_{Y_d})\right)^{n_{e\cdot}} b_{r-1,R_d-1}(F_{Y_d})\Big]}_{\gamma_{r-1,g}^{d,n_{e\cdot}}}$$

Therefore, we can write:

$$\int_{\mathcal{Y}} g(y) b_{k-1,K_d-1}(F_{Y_d|P}(y|p)) f_{Y_d|P}(y|p)dy = \sum_{r=1}^{R_d} \sum_{s=1}^{S_d} \beta_{rs}^d b_{s-1,S_d-1}(F_P) \Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_d-1}{j} \binom{j}{k-1} (R_d S_d)^{j+1} \times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_d S_d}=j} \binom{j}{n_{11}, n_{12}, \ldots, n_{R_d S_d}} \prod_{e=1,f=1}^{R_d,S_d} (\beta_{ef}^d)^{n_{ef}} \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{d,n_{e\cdot}} \Bigg\}$$

Finally, we have

$$\mathbb{E}[g(Y_d)|V=v, P=p] = K_d L_d \sum_{k=1}^{K_d} \sum_{l=1}^{L_d} \alpha_{kl}^d b_{l-1,L_d-1}(v) \sum_{r=1}^{R_d} \sum_{s=1}^{S_d} \beta_{rs}^d b_{s-1,S_d-1}(F_P) \Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_d-1}{j} \binom{j}{k-1} (R_d S_d)^{j+1} \times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_d S_d}=j} \binom{j}{n_{11}, n_{12}, \ldots, n_{R_d S_d}} \prod_{e=1,f=1}^{R_d,S_d} (\beta_{ef}^d)^{n_{ef}} \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{d,n_{e\cdot}} \Bigg\} \quad (40)$$

Remark that when $R_d = S_d = 1$, $C_{Y_d,P}(x_1, x_2; \beta^d) = x_1 x_2$ which is equivalent to $Y_d \perp P$, in such a case Equation (40) simplifies to

$$\mathbb{E}[g(Y_d)|V=v, P=p] = K_d L_d \sum_{k=1}^{K_d} \sum_{l=1}^{L_d} \alpha_{kl}^d b_{l-1,L_d-1}(v) \underbrace{\int_{\mathcal{Y}} g(y) b_{k-1,K_d-1}(F_{Y_d}(y)) f_{Y_d}(y)dy}_{\tau_{g,k}^d}$$
$$= \sum_{l=1}^{L_d} \underbrace{\left(K_d L_d \sum_{k=1}^{K_d} \alpha_{kl}^d \tau_{g,k}^d\right)}_{\theta_{dl}^g} b_{l-1,L_d-1}(v) = \sum_{l=1}^{L_d} \theta_{dl}^g b_{l-1,L_d-1}(v) = \mathbb{E}[g(Y_d)|V=v]. \quad (41)$$

As can be seen we recover the parametric form Mogstad, Santos, and Torgovitsky (2018) imposed on the MTR. Mogstad, Santos, and Torgovitsky (2018) approach imposes $\mathbb{E}[g(Y_d)|V=v] = \sum_{l=1}^{L_d} \theta_{dl}^g b_{l-1,L_d-1}(v)$ as a primitive, while in contrast we show that under a valid IV assumption —Assumption 2, imposing such a structure on the MTRs is equivalent to parametrize the "selection on unobservables" dependence — $C_{Y_d,V|P}(x_1, x_2; \alpha^d)$— using a Bernstein Copula of order L_d.

Going back to the general context, and by integrating the DMTEs we obtain the following model restriction:

$$\mathbb{E}[g(Y)1\{D=d\}|P=p] = \int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} \mathbb{E}[g(Y_d)|V=v,P=p]dv$$
$$= K_d L_d \sum_{k=1}^{K_d}\sum_{l=1}^{L_d} \alpha_{kl}^d \int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} b_{l-1,L_d-1}(v)dv \sum_{r=1}^{R_d}\sum_{s=1}^{S_d} \beta_{rs}^d b_{s-1,S_d-1}(F_P)\Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}(R_d S_d)^{j+1}\times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_d S_d}=j} \binom{j}{n_{11},n_{12},\ldots,n_{R_d S_d}} \prod_{e=1,f=1}^{R_d,S_d} \left(\beta_{ef}^d\right)^{n_{ef}} \prod_{f=1}^{S_d}\left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{d,n_{e\cdot}}\Bigg\}$$

Remark, we can show that $B_{l-1,L_d-1}(1) \equiv \int_0^1 b_{l-1,L_d-1}(v)dv = 1/L_d = \underbrace{\int_0^p b_{l-1,L_d-1}(v)dv}_{B_{l-1,L_d-1}(p)} + \int_p^1 b_{l-1,L_d-1}(v)dv.$

Therefore, we have:

$$\mathbb{E}[g(Y)D|P=p] = K_1 L_1 \sum_{k=1}^{K_1}\sum_{l=1}^{L_1} \alpha_{kl}^1 B_{l-1,L_1-1}(p) \sum_{r=1}^{R_1}\sum_{s=1}^{S_1} \beta_{rs}^1 b_{s-1,S_1-1}(F_P)\Bigg\{ \sum_{j=k-1}^{K_1-1} (-1)^{j-k+1}\binom{K_1-1}{j}\binom{j}{k-1}(R_1 S_1)^{j+1}\times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_1 S_1}=j} \binom{j}{n_{11},n_{12},\ldots,n_{R_1 S_1}} \prod_{e=1,f=1}^{R_1,S_1} \left(\beta_{ef}^1\right)^{n_{ef}} \prod_{f=1}^{S_1}\left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{1,n_{e\cdot}}\Bigg\} \quad (42)$$

and

$$\mathbb{E}[g(Y)(1-D)|P=p] = K_0 L_0 \sum_{k=1}^{K_0}\sum_{l=1}^{L_0} \alpha_{kl}^0 (1/L_0 - B_{l-1,L_0-1}(p)) \sum_{r=1}^{R_0}\sum_{s=1}^{S_0} \beta_{rs}^0 b_{s-1,S_0-1}(F_P)\Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1}\binom{K_0-1}{j}\binom{j}{k-1}(R_0 S_0)^{j+1}\times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_0 S_0}=j} \binom{j}{n_{11},n_{12},\ldots,n_{R_0 S_0}} \prod_{e=1,f=1}^{R_0,S_0} \left(\beta_{ef}^0\right)^{n_{ef}} \prod_{f=1}^{S_0}\left(b_{f-1,S_0-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{0,n_{e\cdot}}\Bigg\}. \quad (43)$$

Restrictions on $\gamma_{r-1,g}^{d,j_e}$. Recall $\gamma_{r-1,g}^{d,j_e} \equiv \mathbb{E}\Big[g(Y_d)\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big]$ for $j_e \in \{1,\ldots,L_d\}$. $\gamma_{r-1,g}^{d,j_e}$ is an unknown parameter to estimate, but the set of potential values it can take is restricted by the model.

In fact, each choice of $g(.)$ imposes a restriction on $\gamma_{r-1,g}^{d,j_e}$, for instance for $g(\cdot) = 1\{\cdot \le y\}$, $g(.) = 1\{y < \cdot \le y'\}$, $g^+ : \mathcal{Y} \to \mathbb{R}^+$, or $g^- : \mathcal{Y} \to \mathbb{R}^-$ we have respectively

$$\gamma_{r-1,y}^{d,j_e} = \mathbb{E}\Big[1\{Y \le y\}\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \ge 0, \quad (44)$$

$$\gamma_{r-1,y'}^{d,j_e} - \gamma_{r-1,y}^{d,j_e} = \mathbb{E}\Big[1\{y < Y \le y'\}\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \ge 0 \;\; \forall y' > y, \quad (45)$$

$$\gamma_{r-1,g^+}^{d,j_e} = \mathbb{E}\Big[g^+(Y_d)\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \ge 0, \quad (46)$$

$$\gamma_{r-1,g^-}^{d,j_e} = \mathbb{E}\Big[g^-(Y_d)\prod_{e=1}^{R_d}\left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \le 0. \quad (47)$$

Remark that, while the dimensionality of $\gamma^{d,j_e}_{r-1,g}$ depends on the complexity of the $g(.)$, the set of unknown parameters $\theta = (\alpha^0, \alpha^1, \beta^1, \beta^0) \in \Theta^{BC}$ where

$$\Theta^{BC} \equiv \Big\{\alpha^d_{kl} \geq 0, \beta^d_{rs} \geq 0, 1 \leq l \leq L_d, 1 \leq k \leq K_d, 1 \leq r \leq R_d, 1 \leq s \leq S_d, \text{ such that}$$
$$K_d \sum_{l=1}^{L_d} \alpha^d_{kl} = 1, L_d \sum_{k=1}^{K_d} \alpha^d_{kl} = 1, S_d \sum_{r=1}^{R} \beta^d_{rs} = 1, R_d \sum_{s=1}^{S_d} \beta^d_{rs} = 1, \text{ for } d \in \{0,1\}\Big\}$$

is invariant to the choice of $g(.)$. So choosing a more informative class of $g(.)$, will provide a tighter identified set of the copula parameters Θ^{BC}_I. To do so, we will consider the half-interval class $\mathcal{G} \equiv \{g(\cdot) = \mathbf{1}[\cdot \leq y], y \in \mathcal{Y}\}$ which allow us to recover the distributional DMTR $F_{Y_d|V,P}(y|v,p)$. We then consider $(\gamma^{j_e}_{r-1,y})_{y\in\mathcal{Y}} \equiv (\gamma^{0,j}_{r-1,y}, \gamma^{1,j}_{r-1,y})_{y\in\mathcal{Y}} \in \Gamma^{BC}$ where

$$\Gamma^{BC} \equiv \Big\{\gamma^{d,j_e}_{r-1,y} \geq 0 \quad \forall y \in \mathcal{Y} \text{ such that } \gamma^{d,j_e}_{r-1,y'} - \gamma^{d,j_e}_{r-1,y} \geq 0 \quad \forall \infty \geq y' > y \geq -\infty, \text{ for } d \in \{0,1\}\Big\}.$$

Theorem 7. *Under Assumptions 1, 5 and 9, the identified set of the Bernstein copulas parameters* $\overline{\Theta}^{BC}_I$ *of* $\left(\theta, (\gamma^{j_e}_{r-1,y})_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC}$ *is characterized as follows:*

$$\overline{\Theta}^{BC}_I = \left\{\left(\theta, (\gamma^{j_e}_{r-1,y})_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC} \textit{ that satisfies Equations } (42) \textit{and } (43), \textit{ for all } g(.) \in \mathcal{G}\right\},$$

and for any integrable real function $g(.)$*, the identified set* $\Theta_{I,g}$ *for DMTR*$_g$ *is defined as follows:*

$$\Theta_{I,g} = \Bigg\{ \left(\mathbb{E}[g(Y_1)|V=v, P=p; \bar{\theta}], \mathbb{E}[g(Y_0)|V=v,P=p;\bar{\theta}]\right) \textit{ such that}$$
$$\mathbb{E}[g(Y_d)|V=v, P=p; \bar{\theta}] = K_d L_d \sum_{k=1}^{K_d} \sum_{l=1}^{L_d} \alpha^d_{kl} b_{l-1,L_d-1}(v) \sum_{r=1}^{R_d} \sum_{s=1}^{S_d} \beta^d_{rs} b_{s-1,S_d-1}(F_P) \Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_d-1}{j} \binom{j}{k-1} (R_d S_d)^{j+1} \times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_dS_d}=j} \binom{j}{n_{11}, n_{12}, \ldots, n_{R_dS_d}} \prod_{e=1,f=1}^{R_d,S_d} \left(\beta^d_{ef}\right)^{n_{ef}} \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma^{d,n_{e\cdot}}_{r-1,g} \Bigg\}$$
$$\forall \ \bar{\theta} \equiv \left(\theta, (\gamma^{j_e}_{r-1,y})_{y\in\mathcal{Y}}\right) \in \overline{\Theta}^{BC}_I \Bigg\}.$$

Example 1. *Now we consider an example with* $K_d = L_d = R_d = S_d = 2$*. In this cases*

$$b_{0,1}(u) = \binom{1}{0} u^0 (1-u)^{1-0} = 1-u; \quad b_{1,1}(u) = \binom{1}{1} u^1 (1-u)^{1-1} = u.$$

$$B_{0,1}(u) = u - \frac{u^2}{2}; \quad B_{1,1}(u) = \frac{u^2}{2}.$$

Therefore, we have

$$C_{Y_d,V|P}(x_1, x_2; \alpha^d) = 4\{\alpha^d_{11} B_{0,1}(x_1) B_{0,1}(x_2) + \alpha^d_{12} B_{0,1}(x_1) B_{1,1}(x_2) + \alpha^d_{21} B_{1,1}(x_1) B_{0,1}(x_2) + \alpha^d_{22} B_{1,1}(x_1) B_{1,1}(x_2)\},$$

where $\alpha^d_{kl} \geq 0$ *and* $\beta^d_{rs} \geq 0$ *satisfy* $2(\alpha^d_{11}+\alpha^d_{12})=1$, $2(\alpha^d_{21}+\alpha^d_{22})=1$, $2(\alpha^d_{11}+\alpha^d_{21})=1$, $2(\alpha^d_{12}+\alpha^d_{22})=1$. *In this case, we can express other* α*s in terms of* α^d_{11}, *that is,* $\alpha^d_{22}=\alpha^d_{11}$, $\alpha^d_{21}=\alpha^d_{12}=\frac{1}{2}-\alpha^d_{11}$. *To ensures all the parameters are greater or equal to zero, we need to have* $0 \leq \alpha^d_{11} \leq \frac{1}{2}$. *Therefore, we can write*

$$C_{Y_d,V|P}(x_1,x_2;\alpha^d) = 4\left\{\alpha^d_{11}(x_1-\frac{x_1^2}{2})(x_2-\frac{x_2^2}{2})+(\frac{1}{2}-\alpha^d_{11})(x_1-\frac{x_1^2}{2})\frac{x_2^2}{2}+(\frac{1}{2}-\alpha^d_{11})(x_2-\frac{x_2^2}{2})\frac{x_1^2}{2}+\alpha^d_{11}\frac{x_1^2}{2}\frac{x_2^2}{2}\right\}$$
$$= 4\alpha^d_{11}x_1x_2+(1-4\alpha^d_{11})(x_1x_2^2+x_2x_1^2)+(4\alpha^d_{11}-1)x_1^2x_2^2.$$

$$c_{Y_d,V|P}(x_1,x_2;\alpha^d) = 4\alpha^d_{11}+(1-4\alpha^d_{11})(2x_1+2x_2)+4(4\alpha^d_{11}-1)x_1x_2.$$

Note that if we define $4\alpha^d_{11}-1 \equiv \delta$, *then*

$$C_{Y_d,V|P}(x_1,x_2;\alpha^d) = x_1x_2\big(1+\delta(1-x_1)(1-x_2)\big)$$

which is the FGM copula with dependence parameter δ. *If we impose* $\alpha_{11}=\frac{1}{4}$, *then we are imposing the selection-on-observable assumption.*

Likewise, when $R_d = S_d = 2$, *we have*

$$C_{Y_d,P}(x_1,x_2;\beta^d) = 4\beta^d_{11}x_1x_2+(1-4\beta^d_{11})(x_1x_2^2+x_2x_1^2)+(4\beta^d_{11}-1)x_1^2x_2^2.$$

and

$$c_{Y_d,P}(x_1,x_2;\beta^d) = 4\beta^d_{11}+(1-4\beta^d_{11})(2x_1+2x_2)+4(4\beta^d_{11}-1)x_1x_2.$$

We need $0 \leq \beta^d_{11} \leq \frac{1}{2}$. *When we impose* $\beta^d_{11}=\frac{1}{4}$, *it follows that* $C_{Y_d,P}(x_1,x_2;\beta^d)=x_1x_2$, *that is, the IV independence assumption is satisfied.*

Recall that $f_{Y_d|V,P}(y|v,p) = c_{Y_d,V|P}(F_{Y_d|P}(y|p),F_{V|P}(v|p);\alpha^d)f_{Y_d|P}f_{V|P}(v|p) = c_{Y_d,V|P}(F_{Y_d|P}(y|p),F_{V|P}(v|p);\alpha^d)f_{Y_d|P}$, *we have*

$$\begin{aligned}\mathbb{E}[g(Y_d)|V=v,P=p] &= \int_{\mathcal{Y}} g(y)f_{Y_d|V,P}(y|v,p)dy \\ &= \int_{\mathcal{Y}} g(y)\left\{4\alpha^d_{11}+(1-4\alpha^d_{11})(2F_{Y_d|P}(y|p)+2v)+4(4\alpha^d_{11}-1)F_{Y_d|P}(y|p)v\right\}f_{Y_d|P}(y|p)dy \\ &= \underbrace{\int_{\mathcal{Y}} g(y)f_{Y_d|P}(y|p)\left(4\alpha^d_{11}+(1-4\alpha^d_{11})F_{Y_d|P}\right)dy}_{RHS1^d(p)} + v\underbrace{\int_{\mathcal{Y}} g(y)f_{Y_d|P}(y|p)\left(4\alpha^d_{11}-1\right)\left(4F_{Y_d|P}-2\right)dy}_{RHS2^d(p)} \end{aligned} \quad (48)$$

So the RHS of the Equation (48) *is linear in* v. *To proceed, recall that*

$$F_{Y_d|P}(y|p) = \frac{C_{Y_d,P(x_1,x_2)}}{\partial x_2}|_{x_1=F_{Y_d}(y),x_2=F_P(p)} = 4\beta^d_{11}F_{Y_d}(y)+(1-4\beta^d_{11})(2F_{Y_d}F_P+F^2_{Y_d})+2(4\beta^d_{11}-1)F^2_{Y_d}F_P$$

and

$$f_{Y_d|P}(y|p) = f_{Y_d}\left(4\beta^d_{11}+(1-4\beta^d_{11})(2F_{Y_d}+2F_P)+4(4\beta^d_{11}-1)F_{Y_d}F_P\right)$$

Insert the above two equations into the first RHS term of Equation (48) and re-arrange, we have

$$RHS1^d(p) = \psi^d_{10}\underbrace{\mathbb{E}[g(Y_d)]}_{\gamma^d_0(g)} + \psi^d_{11}\underbrace{\mathbb{E}[g(Y_d)F_{Y_d}(Y_d)]}_{\gamma^d_1(g)} + \psi^d_{12}\underbrace{\mathbb{E}[g(Y_d)F^2_{Y_d}(Y_d)]}_{\gamma^d_2(g)} + \psi^d_{13}\underbrace{\mathbb{E}[g(Y_d)F^3_{Y_d}(Y_d)]}_{\gamma^d_3(g)} \equiv \boldsymbol{\psi}^d_1 \cdot \boldsymbol{\gamma}^d(g),$$

where $\boldsymbol{\psi}^d_1(p) = (\psi^d_{10}(p), \psi^d_{11}(p), \psi^d_{12}(p), \psi^d_{13}(p))'$,

$$\psi^d_{10}(p) = \left(4\beta^d_{11} + 2(1-4\beta^d_{11})F_P(p)\right)4\alpha^d_{11}$$

$$\psi^d_{11}(p) = (4\beta^d_{11}-1)(4F_P(p)-2)4\alpha^d_{11} + (4\beta^d_{11} + 2F_P(p)(1-4\beta^d_{11}))^2(1-4\alpha^d_{11})$$

$$\psi^d_{12}(p) = (12\beta^d_{11} + 6(1-4\beta^d_{11})F_P(p))(1-4\alpha^d_{11})(4\beta^d_{11}-1)(2F_P(p)-1)$$

$$\psi^d_{13}(p) = 2(4\beta^d_{11}-1)^2(2F_P(p)-1)^2(1-4\alpha^d_{11})$$

The γ *parameters are indexed by function* g*, and* $\boldsymbol{\gamma}^d(g) = (\gamma^d_0(g), \gamma^d_1(g), \gamma^d_2(g), \gamma^d_3(g))'$.

For the RHS2 term, we have

$$RHS2^d(p) = \psi^d_{20}\gamma^d_0(g) + \psi^d_{21}\gamma^d_1(g) + \psi^d_{22}\gamma^d_2(g) + \psi^d_{23}\gamma^d_3(g) \equiv \boldsymbol{\psi}^d_2 \cdot \boldsymbol{\gamma}^d(g)$$

where $\boldsymbol{\psi}^d_2(p) = (\psi^d_{20}(p), \psi^d_{21}(p), \psi^d_{22}(p), \psi^d_{23}(p))'$,

$$\psi^d_{20}(p) = -2\left(4\beta^d_{11} + 2(1-4\beta^d_{11})F_P(p)\right)$$

$$\psi^d_{21}(p) = 4(4\beta^d_{11} + 2F_P(p)(1-4\beta^d_{11}))^2 - 2(4\beta^d_{11}-1)(4F_P(p)-2)$$

$$\psi^d_{22}(p) = (48\beta^d_{11} + 24(1-4\beta^d_{11})F_P(p))(4\beta^d_{11}-1)(2F_P(p)-1)$$

$$\psi^d_{23}(p) = 8(4\beta^d_{11}-1)^2(2F_P(p)-1)^2$$

Finally, consider

$$\begin{aligned}\mathbb{E}[g(Y)1\{D=d\}|P=p] &= \int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} \mathbb{E}[g(Y_d)|V=v, P=p]dv \\ &= RHS1^d(p)\int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} dv + RHS2^d(p)\int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} vdv\end{aligned}$$

Setting $d=1$ *and* $d=0$*, respectively, we have*

$$\mathbb{E}[g(Y)1\{D=1\}|P=p] = (\boldsymbol{\psi}^1_1(p)\cdot\boldsymbol{\gamma}^1(g))p + (\boldsymbol{\psi}^1_2(p)\cdot\boldsymbol{\gamma}^1(g))\frac{p^2}{2} \tag{49}$$

$$\mathbb{E}[g(Y)1\{D=0\}|P=p] = (\boldsymbol{\psi}^0_1(p)\cdot\boldsymbol{\gamma}^0(g))(1-p) + (\boldsymbol{\psi}^0_2(p)\cdot\boldsymbol{\gamma}^0(g))(\frac{1}{2}-\frac{p^2}{2}) \tag{50}$$

In this case, the set

$$\Theta^{BC} \equiv \left\{(\alpha^1_{11}, \beta^1_{11}, \alpha^0_{11}, \beta^0_{11}) : 0 \le \alpha^d_{11} \le \frac{1}{2}, 0 \le \beta^d_{11} \le \frac{1}{2}, \textit{for } d \in \{0,1\}\right\}$$

If we consider the class of g functions to be $\mathcal{G} \equiv \{g(\cdot) = \mathbf{1}[\cdot \leq y], y \in \mathcal{Y}\}$*, then the set*

$$\Gamma^{BC} \equiv \Big\{(\boldsymbol{\gamma}^1(y), \boldsymbol{\gamma}^0(y)) : \gamma_j^d(y) \geq 0, \gamma_j^d(y) \geq \gamma_j^d(y'), \textit{ for all } y \geq y', \textit{ for } j \in \{0,1,2,3\}, \textit{ for } d \in \{0,1\}\Big\}.$$

The identified set of the Bernstein copulas parameters $\overline{\Theta}_I^{BC}$ *of* $\left(\alpha_{11}^d, \beta_{11}^d, \gamma_j^d(y)_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC}$ *is characterized as follows:*

$$\overline{\Theta}_I^{BC} \quad = \quad \left\{\left(\alpha_{11}^d, \beta_{11}^d, \gamma_j^d(y)_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC} \textit{ that satisfies Equations } (49) \textit{and } (50)\right\},$$

and for any integrable real function $g(.)$*, the identified set* $\Theta_{I,g}$ *for* $DMTR_g$ *is defined as follows:*

$$\Theta_{I,g} = \Bigg\{ \left(\mathbb{E}[g(Y_1)|V = v, P = p; \bar{\theta}], \mathbb{E}[g(Y_0)|V = v, P = p; \bar{\theta}]\right) \textit{ such that}$$

$$\mathbb{E}[g(Y_d)|V = v, P = p; \bar{\theta}] = \boldsymbol{\psi}_1^d(p) \cdot \boldsymbol{\gamma}^d(g) + (\boldsymbol{\psi}_2^d(p) \cdot \boldsymbol{\gamma}^d(g))v, \quad \forall \; \bar{\theta} \equiv \left(\alpha_{11}^1, \beta_{11}^1, \alpha_{11}^0, \beta_{11}^0, \boldsymbol{\gamma}^1(y), \boldsymbol{\gamma}^0(y)\right) \in \overline{\Theta}_I^{BC} \Bigg\}.$$

Remark 10. *It is possible to additionally impose some conditions listed in Lemma 4. In Example 1, further assuming MIV is amount to assume the second derivative of* $C_{Y_d,P}$ *with respect to* x_2 *is non-positive, that is*

$$\frac{\partial^2 C_{Y_d,P}(x_1, x_2)}{\partial x_2^2} = 2(4\beta_{11}^d - 1)(x_1^2 - x_1) \leq 0, \forall x_1 \in [0,1]$$

This is satisfied if we assume $\frac{1}{2} \geq \beta_{11}^d \geq \frac{1}{4}$.

C.3. **Frank Copula: Proof of Corollary 2.** Consider $d = 1$. Note first,

$$F_{Y,D|P}(y, 1|p) = -\frac{1}{\sigma_1(p)} \ln\left[1 + \frac{(e^{-\sigma_1(p) c_{1,F_P(p)}(F_{Y_1}(y))} - 1)(e^{-\sigma_1(p)p} - 1)}{(e^{-\sigma_1(p)} - 1)}\right]$$

Solve $c_{1,F_P(p)}(F_{Y_1}(y))$ from the above equation we have

$$c_{1,F_P(p)}(F_{Y_1}(y)) = -\frac{1}{\sigma_1(p)} \ln\left[1 + \frac{(e^{-\sigma_1(p) F_{Y,D|P}(y,1|p)} - 1)(e^{-\sigma_1(p)} - 1)}{(e^{-\sigma_1(p)p} - 1)}\right] \equiv H_1(y, p, \sigma_1(p))$$

Here, the $H_1(y, p, \sigma_1(p))$ only depends on quantities that are directly identifiable from the data and the finite dimensional parameters. Then recall that

$$H_1(y, p, \sigma_1(p)) = c_{1,F_P(p)}(F_{Y_1}(y)) = \frac{\partial C_{Y_1,P}(x_1, x_2)}{\partial x_2}\Big|_{x_1 = F_{Y_1}(y), x_2 = F_P(p)}$$

$$= \frac{(e^{-\sigma_1 F_{Y_1}(y)} - 1)e^{-\sigma_1 F_P(p)}}{(e^{-\sigma_1} - 1) + (e^{-\sigma_1 F_{Y_1}(y)} - 1)(e^{-\sigma_1 F_P(p)} - 1)},$$

Again, solving $F_{Y_1}(y)$ from it yields

$$F_{Y_1}(y;\theta) = -\frac{1}{\alpha_1}\ln\Big[1+\frac{H_1(y,p,\sigma_1(p))(e^{-\alpha_1}-1)}{e^{-\alpha_1 F_P(p)} - H_1(y,p,\sigma_1(p))(e^{-\alpha_1 F_P(p)}-1)}\Big]$$

Next consider $d=0$. We know that

$$F_{YD|P}(y,0|p) = c_{0,F_P(p)}(F_{Y_0}(y)) + \frac{1}{\sigma_0(p)}\ln\Big[1+\frac{(e^{-\sigma_0(p)c_{0,F_P(p)}(F_{Y_0}(y))}-1)(e^{-\sigma_0(p)p}-1)}{(e^{-\sigma_0(p)}-1)}\Big]$$

Solving $c_{0,F_P(p)}(F_{Y_0}(y))$ from the above equation we have

$$c_{0,F_P(p)}(F_{Y_0}(y)) = \frac{1}{\sigma_0(p)}\ln\Big[1+\frac{(e^{\sigma_0(p)F_{YD|P}(y,0|p)}-1)(e^{-\sigma_0(p)}-1)}{e^{-\sigma_0(p)} - e^{-\sigma_0(p)p}}\Big] \equiv H_0(y,p,\sigma_0(p))$$

Again, recall that

$$H_0(y,p,\sigma_0(p)) = c_{0,F_P(p)}(F_{Y_0}(y)) = \frac{\partial C_{Y_0,P}(x_1,x_2)}{\partial x_2}|_{x_1=F_{Y_0}(y),x_2=F_P(p)} = \frac{(e^{-\sigma_0 F_{Y_0}(y)}-1)e^{-\sigma_0 F_P(p)}}{(e^{-\sigma_0}-1)+(e^{-\sigma_0 F_{Y_0}(y)}-1)(e^{-\sigma_0 F_P(p)}-1)}$$

Solve $F_{Y_0}(y)$ from the above equation yields

$$F_{Y_0}(y;\theta) = -\frac{1}{\alpha_0}\ln\Big[1+\frac{H_0(y,p,\sigma_0(p))(e^{-\alpha_0}-1)}{e^{-\alpha_0 F_P(p)} - H_0(y,p,\sigma_0(p))(e^{-\alpha_0 F_P(p)}-1)}\Big].$$

Finally, since $F_d(y;\theta)$ does not depends on p, its partial derivative with respect to p must be flat at 0 for all value of p, and so does the right hand side of the equation. Therefore, for all p, noticing $e^{-\alpha_d F_P(p)} \neq 0$ and $H_d \neq 0$,

$$\frac{\partial\left\{\frac{e^{-\alpha_d F_P(p)} - H_d(y,p,\sigma_d(p))(e^{-\alpha_d F_P(p)}-1)}{H_d(y,p,\sigma_d(p))}\right\}}{\partial p} = 0 \Rightarrow e^{-\alpha_d F_P(p)}\frac{\alpha_d f_P(p)(H_d-1)H_d - \frac{\partial H_d}{\partial p}}{H_d^2} = 0$$
$$\Rightarrow \alpha_d f_P(p)(1-H_d)H_d + \frac{\partial H_d}{\partial p} = 0.$$

To obtain the identified set for the distributional DMTR, simply note that $H_d(y,p,\sigma_d(p)) = c_{d,F_P(p)}(F_{Y_d}(y)) = F_{Y_d|P}(y|p)$, hence

$$F_{Y_d|P,V}(y|p,v) = \frac{\partial C_{Y_d,V|P}(x_1,x_2)}{\partial x_2}|_{x_1=H_d,x_2=v} = \frac{(e^{-\sigma_d H_d}-1)e^{-\sigma_d v}}{(e^{-\sigma_d}-1)+(e^{-\sigma_d H_d}-1)(e^{-\sigma_d v}-1)},$$

and restrict θ taking values from Θ_I^{SP} and

$$\mathbb{E}[Y_d|P=p,V=v] = \int y f_{Y_d|P,V}(y|p,v)dy = \int y c_{Y_d,V|P}(y,v|p) f_{Y_d|P}(y|p)dy$$
$$= \int y \frac{-\sigma_d(p)(e^{-\sigma_d(p)}-1)e^{-\sigma_d(p)(H_d+v)}}{(e^{-\sigma_d}-1)+(e^{-\sigma_d H_d}-1)(e^{-\sigma_d v}-1))^2}\frac{\partial H_d(y,p,\sigma_d(p))}{\partial y}dy$$

References

AIZER, A., AND J. J. DOYLE JR (2015): "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 130(2), 759–803.

ARELLANO, M., AND S. BONHOMME (2017): "Quantile selection models with an application to understanding changes in wage inequality," *Econometrica*, 85(1), 1–28.

ARTEAGA, C. (2018): "The Cost of Bad Parents: Evidence from the Effects of Parental Incarceration on Children's Education," *Working Papers*.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2012): "Treatment effect bounds: An application to Swan–Ganz catheterization," *Journal of Econometrics*, 168(2), 223–243.

BHULLER, M., G. B. DAHL, K. V. LOKEN, AND M. MOGSTAD (2019): "Incarceration, Recidivism, and Employment," .

BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125(4), 985–1039.

CARNEIRO, P., J. J. HECKMAN, AND E. VYTLACIL (2010): "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 78(1), 377–394.

CHEN, X., Y. FAN, AND V. TSYRENNIKOV (2006): "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association*, 101(475), 1228–1240.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

DEATON, A. S. (2009): "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development," Discussion paper, National Bureau of Economic Research.

DEATON, A. S., J. J. HECKMAN, AND G. W. IMBENS (2010): "Forum on the Estimation of Treatment Effects," *The Journal of Economic Literature*, 48(2), 356–455.

DI TELLA, R., AND E. SCHARGRODSKY (2013): "Criminal recidivism after prison and electronic monitoring," *Journal of Political Economy*, 121(1), 28–73.

DOBBIE, W., J. GOLDIN, AND C. S. YANG (2018): "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges," *American Economic Review*, 108(2), 201–40.

HECKMAN, J. J., AND R. PINTO (2018): "Unordered monotonicity," *Econometrica*, 86(1), 1–35.

HECKMAN, J. J., AND E. VYTLACIL (2001): "Policy-relevant treatment effects," *American Economic Review*, 91(2), 107–111.

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73(3), 669–738.

HECKMAN, J. J., AND E. J. VYTLACIL (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the national Academy of Sciences*, 96(8), 4730–4734.

HSU, Y.-C., C.-A. LIU, AND X. SHI (2019): "Testing generalized regression monotonicity," *Econometric Theory*, 35(6), 1146–1200.

HUBER, M., AND G. MELLACE (2015): "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 97(2), 398–411.

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

JOE, H. (1994): "Multivariate extreme-value distributions with applications to environmental data," *Canadian Journal of Statistics*, 22(1), 47–64.

——— (1996): "Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters," *Lecture Notes-Monograph Series*, pp. 120–141.

——— (1997): *Multivariate models and multivariate dependence concepts*. CRC Press.

KÉDAGNI, D., AND I. MOURIFIÉ (2017): "Generalized Instrumental Inequalities: Testing IV Independence Assumption," .

KITAGAWA, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83(5), 2043–2063.

KLING, J. R. (2006): "Incarceration length, employment, and earnings," *American Economic Review*, 96(3), 863–876.

LEE, S., AND B. SALANIÉ (2018): "Identifying effects of multivalued treatments," *Econometrica*, 86(6), 1939–1963.

MANSKI, C. F. (2011): "Policy analysis with incredible certitude," *The Economic Journal*, 121(554), F261–F289.

MANSKI, C. F., AND J. V. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86(5), 1589–1619.

MOGSTAD, M., A. TORGOVITSKY, AND C. WALTERS (2019): "Identification of causal effects with multiple instruments: Problems and some solutions," *NBER Working Paper*, (w25691).

MOURIFIÉ, I., AND Y. WAN (2017): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99(2), 305–313.

MUELLER-SMITH, M. (2015): "The criminal and labor market impacts of incarceration," *Unpublished Working Paper*, 18.

NELSEN, R. B. (2007): *An introduction to copulas*. Springer Science & Business Media.

PEARL, J. (2013): "Direct and indirect effects," *arXiv preprint arXiv:1301.2300*.

SANCETTA, A., AND S. SATCHELL (2004): "The Bernstein copula and its applications to modeling and approximations of multivariate distributions," *Econometric theory*, 20(3), 535–562.

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.