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 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 provides a basis for testing 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.

JEL Classification: C01, C14, C21 and C26

This version: Saturday 30th November, 2024

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 and the Simon Institute at the University of California Berkeley. Mourifié thanks his hosts for their hospitality and support. Wan and Mourifié thank the support from the SSHRC Insight Grant #435190500.

[†] Department of Economics, Washington University in St Louis, One Brookings Drive St. Louis, MO, 63130-4899. ismaelm@wustl.edu.

[‡] Department of Economics, University of Toronto. 150 St George Street, Toronto, Ontario, M5S 3G7, Canada, 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 much controversy among 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 harm policy choice. Furthermore, Coulibaly, Hsu, Mourifié, and Wan (2024) recently proposed a sharp test for the MTE assumptions, illustrating that they can fail to hold in some empirical applications. Hence, 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. As advocated by Manski (2011), one way to resolve this tension is

what he referred to 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]

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 $D = 1\{\nu(Z) \geqslant \tilde{V}\}$, where Y_1 and Y_0 are potential outcomes, D is the treatment, Y is the observed outcome, Z are (possibly invalid) instrument variables so that Z and $(Y_1, Y_0, \tilde{V})$ can be correlated. We define the DMTE as the expectation of treatment effect conditioning on the transformed error term $V \equiv F_{\tilde{V}|Z}(\tilde{V})$ and the propensity score $P(Z) \equiv P(D=1|Z)$, that is, $\text{DMTE}(v,p) \equiv \mathbb{E}[Y_1 - Y_0 | V = v, P(Z) = p]$.[2] We show that all the conventional policy parameters mentioned above, including the MTE, can be expressed as a weighted integral of the DMTE under the single 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 independent with $(Y_1, Y_0, \tilde{V})$. 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 focus on the joint distribution of (Y_d, P, V) because these variables appear in the key parameter DMTE, and it is convenient to work with the one dimensional "propensity score" P. We further show that 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 a version of "selection on observables": The potential outcome and the treatment are independent once the propensity score P is controlled; otherwise, "selection on unobservables" exists.[3] The other copula function is $C_{Y_d,P}(.,.)$, which captures the dependence between the potential outcomes and

[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."*

[2] Note that when $\tilde{V}$ is independent with Z, as often assumed in the literature, we have $V \equiv F_{\tilde{V}}(\tilde{V})$, and $P(z) = \nu(z)$.

[3] Masten and Poirier (2018) discussed partial identification of treatment effect parameters when the conditional independence between Y_d and D given covariates X are relaxed to the "Conditional c-Dependence". It is different from the MTE framework studied in our paper, where instrumental variables Z contribute the identification.

the propensity score. This copula measures the "quality of instruments" and is a key function that we investigate in this paper. For example, if Z is independent with (Y_0, Y_1), 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 $C_{Y_d,P}(.,.)$. Under the copula formulation, we show that calculating the identified set of the DMTE boils down to finding the set of conditional bivariate copulas that respect a set of *equality constraints* and any additional constraints that researchers would like to impose on $C_{Y_d,P}(.,.)$.

In our *Vine Copula* characterization, the identified set of DMTE depends on restrictions that we impose on $C_{Y_d,P}(.,.)$ 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 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 a version of 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. We recover the sharp identified set under such monotone IV assumption in this case. As we demonstrate in more detail in the main text, our approach, in general, provides empirical researchers with a very flexible way to derive the identified set on the DMTE under any dependence restrictions they are 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).

While our main identification result is nonparametric, it can also accommodate parametric assumptions on the copula functions when researchers have good reason to impose them. For example, one may assume the copula belongs to a particular parametric class and still leave the marginal distributions unspecified. 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 circumvent the curse of dimensionality. Our semiparametric identification result also shares a similar idea as Han and Vytlacil (2017), which employed parametric copula with unknown marginals to study the identification and estimation of Bivariate Probit models.

On the other hand, it still offers more flexibility than the classical parametric Roy model of Heckman and Honoré (1990), where the distribution of (Y_d, V) is assumed to be joint normal.

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 provide general identification results for DMTR (and hence DMTE) under various assumptions on the IV in Section 3. In Section 4, we discuss two approaches to implement the identification results: by approximating the unknown copula nonparametrically with Bernstein's copula series or by parameterizing the dependence structure. A numerical illustration is provided in Section 5. Section 6 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 $d_z \geqslant 1$. 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 $D = 1\{\nu(Z) \geqslant \widetilde{V}\}$ for some measurable and non-trivial function ν, where the conditional distribution of $\widetilde{V}|Z=z$ is absolute continuous for all $z \in \mathcal{Z}$.*

Assumption 2 (IV Independence)**.** *Let Z be a vector of covariates is statistically independent of the unobservables in the models, i.e. $Z \perp (Y_d, \widetilde{V})$ for $d = 0, 1$.*

Our main goal in this paper will be to develop a (partial) identification approach for the policy-relevant parameters when Assumption 2 fails to hold. In practice, it is possible that the STC assumption is also violated. This can happen when there is a heterogenous effect in the treatment selection (see Gautier and Hoderlein, 2015) or when there are multiple thresholds (Lee and Salanié, 2018). In Appendix C.1, we discuss how to extend our analysis to multiple threshold-crossing models. To this end, we follow Vytlacil (2002), and more recently Zhou and Xie (2019), to make the

following useful normalization:

$$D = 1\{\nu(Z) \geqslant \widetilde{V}\} = 1\{\underbrace{F_{\widetilde{V}|Z}(\nu(Z))}_{P(Z)} \geqslant \underbrace{F_{\widetilde{V}|Z}(\widetilde{V})}_{V}\} = 1\{P(Z) \geqslant V\}, \tag{1}$$

where V is independent of Z and uniformly distributed over [0,1] by construction (see Chen and Xie, 2021, for a formal proof). Therefore, $P(Z) \equiv \mathbb{P}(D = 1|Z) = F_{\widetilde{V}|Z}(\nu(Z))$ and $P(Z)$ is directly recoverable from data. Note that the independence between V and $P(Z)$ or between V and Z is induced by the normalization; the structural unobservable $\widetilde{V}$ can still be dependent on Z. When it causes no confusion, we will use the shorthand notation P or p to denote $P(Z)$ or $P(z)$, respectively. Let $\mathcal{P} \subseteq [0,1]$ denote the support of $P(Z)$. We restrict our attention to cases where P is continuous for the main text. The case with discrete P is analyzed in Appendix C.2.2 at the expense of extra notation. Here, we omit observed covariates X. Our analysis can be considered as conditioning on the realizations of X.

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 $p \in \mathcal{P}$:

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

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

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

By taking the derivative of both sides of Equation (2) respect with p, we obtain:

$$\frac{\partial}{\partial p}\mathbb{E}[g(Y)D|P = p] = \mathbb{E}[g(Y_1)|V = p, P = p] + \int_0^p \frac{\partial}{\partial p}\mathbb{E}[g(Y_1)|V = v, P = p]dv. \tag{3}$$

It can be seen that the left-hand side of Equation (3), also known as 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 (3) 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:

$$\text{DMTE}(v,p) \equiv \mathbb{E}[Y_1 - Y_0|V = v, P = p], \ \ \forall (v,p) \in [0,1] \times \mathcal{P}. \tag{4}$$

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] One can also recover the distributional version of the DMTE by choosing $g(Y_d) = \mathbf{1}[Y_d \leqslant y]$ to obtain $\mathbb{P}(Y_1 \leqslant y|V = v, P = p) - \mathbb{P}(Y_0 \leqslant y|V = v, P = p), \forall (v,p) \in [0,1] \times \mathcal{P}$. It is apparent from Equation (4) 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 $f_{P|V} = f_P$ due to the normalization made when Assumption 1 holds and the density f_P of P is directly identified from data. When $P(Z) \perp Y_d|V$ as in HV05, the DMTE is exactly equal to the MTE, and we have

$$\text{DMTE}(v,p) = \text{DMTE}(v,p') = \text{MTE}(v), \tag{5}$$

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, they do serve as useful intermediate quantities to identify the MTR,

[4] It is worth-noting that our DMTE shares a superficial resemblance with the Redefined MTE ($\widetilde{MTE}$) introduced in the analysis of Zhou and Xie (2019), where the IV-independence assumption is assumed to hold. In presence of a vector of covariates X, $\widetilde{MTE}(v,p) = \mathbb{E}[Y_1 - Y_0|V = v, P(Z,X) = p]$ while $\text{DMTE}(v,p,x) = \mathbb{E}[Y_1 - Y_0|V = v, P(Z,X) = p, X = x]$. Note also that when X is fully exogenous and is independent of all other variables, $\widetilde{MTE}(v,p) = MTE(v)$ once conditioning on a subpopulation of $X = x$ because $P(Z,x)$ is independent with (Y_d, V) in Zhou and Xie (2019). This is not the case for DMTE in our setup.

MTE, and other useful policy parameters such as the ATE, ATT, and 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.

Remark 1. *In our model, $V \equiv F_{\tilde{V}|Z}(\tilde{V})$, where $\tilde{V}$ is the structural error term, e.g., the disutility of taking the treatment. If Assumption 2 holds, then it is without loss of generality to normalize the distribution of $\tilde{V}$ to uniform. In this case, $\tilde{V} = V$ can be interpreted as the ranking of the dis-utility and $MTE(v)$ represents the average treatment effect for the sub-population whose disutility of taking treatment ranks at the v-percentile. When Assumption 2 fails to hold, $MTE(v)$ does not have such an interpretation but can still serve as a useful intermediate quantity to recover other policy-relevant parameters.*

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

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

where a' and a denote the alternative policy regimes 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.

For $\ell \in \{a, a'\}$, let A^ℓ be a generic random variable A under policy regime ℓ. For example, Y_d^ℓ is the potential outcome under policy regime ℓ. HV05 assumes that the joint distributions of (Y_d^ℓ, V^ℓ) are the same across different policies, which is referred as the policy invariant assumption (PI): $(Y_d^{a'}, V^{a'}) \sim (Y_d^a, V^a)$ for $a \neq a'$. Under Assumptions 1 and 2, and the PI, HV05 shows that

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

where again w^{PRTE} is a weighting function that can be directly recovered from the observed distribution of P. Under PI, PRTE can be used to evaluate the effect of a new policy that induces a change in P, i..e $P^a \neq P^{a'}$ but keeping the full joint distribution of latent variable unchanged from the baseline policy to the targeting alternative policy. For our context, we propose a generalized policy invariance

assumption and show that the PRTE_g can be recovered from DMTE even when Assumption 2 fails to hold.

Assumption 3 (Generalized Policy Invariance, GPI)**.** *Let* $U^{P^\ell} \equiv F_{P^\ell}(P^\ell)$ *for* $\ell \in \{a, a'\}$. *Then* $(Y_d^{a'}, V^{a'}, U^{P^{a'}}) \sim (Y_d^a, V^a, U^{P^a})$.

Note that when the independence Assumption 2 holds, our GPI is equivalent to the HV05's PI assumption. However, GPI provides a more convenient venue to study the PRTE when Assumption 2 fails to hold. To better understand the intuition behind Assumption 3, suppose Y_d^ℓ and P^ℓ are continuous. By Sklar's Theorem, the PI is equivalent to have the same copula and the same marginal distributions for the latent variables across policies, i.e., $C_{Y_d^a, V^a}(x_1, x_2) = C_{Y_d^{a'}, V^{a'}}(x_1, x_2)$ and $Y_d^{a'} \sim Y_d^a$ for $a \neq a'$ (where $V^a \sim V^{a'}$ by construction). This requirement was built under the underlying assumption that the propensity score P^ℓ was independent of (Y_d^ℓ, V^ℓ). In our setup, because we allow such dependence, we therefore requires a generalized version of the PI that $C_{Y_d^a, V^a, P^a}(x_1, x_2, x_3) = C_{Y_d^{a'}, V^{a'}, P^{a'}}(x_1, x_2, x_3)$ and $Y_d^{a'} \sim Y_d^a$, for $a \neq a'$.[5] In Assumption 3, $V^a \sim V^{a'}$ and $U^{P^a} \sim U^{P^{a'}}$ hold because they follow the uniform distribution by construction.[6] Beyond the invariance of the marginal distribution of potential outcomes, what we require here is essentially the invariance of the joint dependence structure among potential outcomes, dis-utility of taking treatment, and propensity score, but we leave the marginal distribution of the propensity score to freely change from the environment a to a', i.e. $F_{P^a}(\cdot) \neq F_{P^{a'}}(\cdot)$.

Theorem 1(iii) below shows that the PRTE_g can be written as a weighted average of the DMTE_g under Assumptions 1 and 3.

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

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

[5] By Proposition 4(2) of Embrechts and Hofert (2013), we have $C_{Y_d^\ell, V^\ell, P^\ell}(x_1, x_2, x_3) = C_{Y_d^\ell, V^\ell, U^{P^\ell}}(x_1, x_2, x_3)$.

[6] Notice, that U^{P^ℓ} is a uniform distribution since we consider that P^ℓ is continuous, in the discrete case U^{P^ℓ} is not necessarily uniformly distributed and then $U^{P^a} \sim U^{P^{a'}}$ is no longer the result of a normalization but imposes a further restriction.

(ii) for any $\boldsymbol{s} \in \{ATE_g, LATE_g(u,u'), ATT_g, ATUT_g\}$[7] *and weights* $\omega^s(v,p)$ *listed in Table 1 below, we have*

$$\boldsymbol{s} = \int_0^1 \int_0^1 \omega^s(v,p) DMTE(v,p) dv dp. \tag{6}$$

(iii) If in addition Assumption 3 holds, and the cumulative distribution function $F_{P^\ell}(\cdot)$ *for* $l \in \{a, a'\}$ *is continuous and strictly increasing, then Equation (6) holds with* $\boldsymbol{s} = PRTE$.

TABLE 1. Policy Parameters and DMTE

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

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. As in the HV05 framework, 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).

[7] 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'.

3. IDENTIFICATION

In the previous section, we show that the intermediate quantities DMTE_g or DMTR_g can uniquely recover MTE_g and MTR_g under the single threshold crossing (STC) condition only. In this section, our main goal will be to provide a tractable characterization of the identified set for the DMTRs.

Definition 1. *For any integrable real function $g(\cdot)$, the identified set Θ_I for DMTR_g under the Assumption 1 (STC) is defined as follows:*

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

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

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

Here, the DMTR_g^d is just the conditional distribution function of Y_d given P and V. Therefore, the identification under the half-interval class also enables us to recover DMTRs for other classes of g functions. Note that for a given distribution of (Y, D, P), the identified set for DMTRs is not the entire parameter space. It can be better seen from the key identification Equation (3):

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

Suppose we observe in the data that $\frac{\partial}{\partial p}\mathbb{E}[g(Y)D|P=p] < 0$, then because $\mathbb{E}[g(Y_1)|V=p, P=p] > 0$ for indicator class of function g, the joint distribution of (Y_1, V, P) must make $\int_0^p \frac{\partial}{\partial p}\mathbb{E}[g(Y_1)|V=v, P=p]dv < 0$. In other words, $\frac{\partial}{\partial p}\mathbb{E}[g(Y_1)|V=v, P=p]$ can not be positive for all value of $v \in [0,p]$. Of course, under single-threshold crossing conditions only, the identified set for DMTR and other causal parameters can be too big to be practical. Nevertheless, this characterization and its equivalent copula characterization (see Theorem 2 below) serve as a desirable starting point for the layered analysis.

Before we proceed with the main identification results, we provide examples in which the IV-independence can fail to hold.

Example 1. ***Violation of the IV independence restriction.*** *Let us consider the following model used in Carneiro, Heckman, and Vytlacil (2010) to estimate the return to education:*

$$\begin{aligned} Y &= Y_1 D + Y_0(1-D), \\ D &= 1\{P(Z) - V \geqslant 0\}, \end{aligned}$$

where D is a binary indicator for college education, Y_d represents potential wage, $P(Z)$ measures the predicted gain of return to college, V denotes the unobserved cost of education, which is assumed to be independent of Z. However, finding a valid IV, i.e., $Y_d \perp Z$, to point-identify the return to college has always been a difficult task for labor economists. Willis and Rosen (1979) proposed the parental education as an instrument. Unfortunately, the validity of this instrument is doubtful, as parental education level may be correlated with unobserved individual productivity and hence with potential outcomes, i.e., $Y_d \not\perp Z$. In fact, Cunha, Heckman, and Schennach (2010) argues that cognitive and noncognitive unobserved skills are determined in great part by parental environment and investment, which in turn are highly correlated with parental education. Distance to college is another "tainted" instrument for returns to education, as discussed in Card (2001) and Carneiro, Lokshin, and Umapathi (2017). A similar concern also applies to local labor market conditions used in Carneiro, Heckman, and Vytlacil (2010), which may drive endogenous location choices. So, many popular instruments in the literature are potentially contestable.

Example 2. ***Misspecification in presence of multiple treatments.*** *One recent and growing empirical application using the MTE identification strategy is the Judge leniency IV designs.*[8] *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 in the Judge 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,*

[8]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.

etc.[9] *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 propensity scores that measure the judge's stringency level for a different punishment, which are generated by two instruments, Z_1 and Z_2, respectively. 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:

$$Y = Y_1 D_1 + Y_0(1-D_1),$$
$$D_1 = 1\{P_1 > V_1\}.$$

where $Y_d \equiv [Y_{d1}D_2 + Y_{d0}(1-D_2)]$ for $d \in \{0,1\}$. In this case, two conditions ensure IV-independence: $Y_{d1}|V_2 = v_2 \sim Y_{d0}|V_2 = v_2$ for all v_2 or $P_1 \perp P_2$, then $Y_d \perp P_1$. As we shown in Appendix A.6, in this misspecified model, Y_1 is essentially a mixture of Y_{11} and Y_{10}—two random variables that are independent with P_1. This first condition 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 second condition 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.

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

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

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 out of the three marginal distributions, i.e., V and P, and we know that they are independent because of the normalization. This feature suggests it is convenient to use copula decomposition to study the joint distribution of interest. Copula theory is useful for separating marginal properties from properties related to the dependence structure. Here we cite Sklar (1959)'s result:

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

$$\mathbb{P}(Y_d \leqslant y, V \leqslant v, P \leqslant p) = C_{Y_d,V,P}(F_{Y_1}(y), F_V(v), F_P(p)), \quad \text{for all } y, v, p.$$

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 Sklar's result, we can decompose the joint trivariate distribution into three univariate distributions and one trivariate copula $C_{Y_d,V,P}(F_{Y_d}(y), F_V(v), F_P(p))$. 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 4. *The joint distribution of* (Y_d, V, P) *is absolutely continuous with respect to the Lebesgue measure.*

Here, we make Assumption 4 only for the ease of notation. V is continuous by construction as long as the dis-utility of taking the treatment $\tilde{V}$ is a continuous random variable, which is commonly

assumed in the literature. P is continuous if Z contains a continuous element and $\nu(\cdot)$ in the treatment selection equation is continuous in that element. It is worth noting that we do not require Z to be independent of potential outcomes and $\tilde{V}$. Therefore, the presence of any continuous exogenous covariates in the selection equation can ensure the continuity of P. Hence, the main restriction of Assumption 4 is to focus the analysis on applications with continuous outcomes. It implies that both marginal distribution and the conditioning distribution of the potential outcomes are continuous with respect to the Lebesgue measure. 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 C.2.1 for details. In Appendix C.2.2, we also discuss how to extend our method to the discrete propensity score case.

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, $C_{Y_d,V|P=p}(F_{Y_d|P}(y|p), v; p)$. Let $\mathcal{Y}_d$ be the conditioning support of Y_d given $(D,P) = (d,p)$. We assume it is invariant to p to simplify notation, and the proof in Lemma 2 follows without assuming it.

Lemma 2 (Vine Copula)**.** *Let $d \in \{0,1\}$ and consider a $p \in \mathcal{P}$. Under Assumptions 1 and 4, we have for all $y \in \mathcal{Y}_d$,*

$$F_{Y_d|P}(y|p) = \frac{\partial}{\partial x_2} C_{Y_d,P}(x_1,x_2)\Big|_{x_1=F_{Y_d}(y),x_2=F_P(p)} \equiv c_{d,F_P(p)}(F_{Y_d}(y)), \tag{7}$$

$$F_{Y_d|V,P}(y|v,p) = \frac{\partial}{\partial x_2} C_{Y_d,V|P=p}(x_1,x_2)\Big|_{x_1=F_{Y_d|P}(y|p),x_2=v} \tag{8}$$

Furthermore, there exists strictly increasing mappings $\Psi_{1,p}$ and $\Psi_{0,p}$ such that

$$\mathbb{P}[Y \leqslant y, D = 1|P = p] = \Psi_{1,p}(F_{Y_1}(y)) \equiv C_{Y_1,V|P=p}\Big(c_{1,F_P(p)}(F_{Y_1}(y)), p\Big). \tag{9}$$

$$\begin{aligned}\mathbb{P}[Y \leqslant y, D = 0|P = p] &= \Psi_{0,p}(F_{Y_0}(y)) \\ &\equiv c_{0,F_P(p)}(F_{Y_0}(y)) - C_{Y_0,V|P}\Big(c_{0,F_P(p)}(F_{Y_0}(y)), p\Big).\end{aligned} \tag{10}$$

That is, the observed probability $\mathbb{P}[Y \leqslant y, D = d|P = p]$ 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, the parameter of interest DMTR can be explicitly characterized by a mapping from observed data distribution $\mathbb{P}[Y \leqslant y, D = d|P = p]$, where the mapping only depends on the copula $C_{Y_d,V|P=p}$. To see this, let $\tau_{(2)}(x_1, x_2)$ be the derivative of $C_{Y_d,V|P=p}(x_1, x_2)$ with respect to the second argument, and let $r^{-1}_{(1)}$ be the inverse of $C_{Y_d,V|P=p}(x_1, x_2)$ with respect to the first argument. The inverse is well-defined by Assumption 4. Then by using Equations (8) and (9) we have

$$F_{Y_d|V,P}(y|v, p) = \tau_{(2)}\left(r^{-1}_{(1)}(\mathbb{P}[Y \leqslant y, D = d|P = p], p), v\right) \equiv \Xi_{d,p,v}(\mathbb{P}[Y \leqslant y, D = d|P = p]). \tag{11}$$

Therefore, once we know $C_{Y_d,V|P=p}$, we can uniquely recover the DMTR.

Secondly, Equations (9) and (10) in the second part of Lemma 2 provide a link between $C_{Y_d,V|P=p}$ and $C_{Y_d,P}$, and the observed data distribution. Interestingly, given $\Psi_{d,p}$ is invertible (see Appendix A.3), one can "solve" $F_{Y_d}(y) = \Psi^{-1}_{d,p}(\mathbb{P}[Y \leqslant y, D = d|P = p])$ from Equations (9) and (10). Meanwhile, since $F_{Y_d}(y)$ does not depends on p, it should be the case that for any $y \in \mathcal{Y}_d$,

$$\Psi^{-1}_{d,p}(\mathbb{P}[Y \leqslant y, D = d|P = p]) = \Psi^{-1}_{d,p'}(\mathbb{P}[Y \leqslant y, D = d|P = p']), \quad \forall (p, p') \in \mathcal{P}^2. \tag{12}$$

One can show any pair of copulas $C_{Y_d,V|P}$ and $C_{Y_d,P}$ that satisfy Equations (9) and (10) can be rationalized by the data and the model structure.

The above discussions imply that we can now focus on characterizing the identified set for the two bivariate copulas. Once this identified set is given, we can obtain the identified set for DMTR and F_{Y_d} immediately from the two mappings $\Xi_{d,p,v}$ and $\Psi^{-1}_{d,p}$. To be more specific, let $\mathcal{C}^c_d$ be the set of conditional copulas containing the true value $C_{Y_d,V|P}$, and $\mathcal{C}_d$ be the set of copulas containing the true $C_{Y_d,P}$. These sets can be collections of all copula intersects with any prior restrictions researchers may impose. The identified set for DMTRs is summarized by the following theorem.

Theorem 2. *Under Assumptions 1 and 4, the identified set Θ_I in Definition 1 can be equivalently expressed as:*

$$\Theta_I = \left\{ F_{Y_d}(y|p, v) : F_{Y_d}(y|p, v) = \Xi_{d,p,v}(\mathbb{P}[Y \leqslant y, D = d|P = p]), (C_{Y_d,V|P}, C_{Y_d,P}) \in \Lambda_d \right\},$$

where $\Xi_{d,p,v}$ is defined in Equation (11) and Λ_d is the identified set for the copula functions, given by

$$\Lambda_d = \left\{ (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}_d^c \times \mathcal{C}_d : \textit{Equation (12) holds } \forall y \in \mathcal{Y}_d \right\}.$$

Proof. See Appendix A.3. □

Theorem 2 characterizes the identified set for DMTR 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 $\Psi_{d,p}^{-1}$ 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 the parameters. For instance, the projection of the identified set for copulas are determined by Equation (12), and once $(C_{Y_d,V|P}, C_{Y_d,P})$ are fixed, DMTRs and F_{Y_d} are point identified. In particular, Equation (12) 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 4 hold, then the identification equation postulated in Theorem 2 coincides with the identification result of HV05, that is,*

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

and

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

Proof. See Appendix A.4.

Remark 2. *Suppose a researcher hopes to impose the selection on observables assumption such that $Y_d \perp \tilde{V}|Z$. In our framework, this assumption implies that $Y_d \perp V|Z$ and together with $V \perp Z$ (by construction), we have $Y_d \perp V|P$. Therefore, "selection on observables" boils down to imposing a*

functional form restriction on the copula: $C_{Y_d,V|P}(x_1,x_2) = x_1x_2$. *We can easily see from the proof of Lemma 2 that it recovers the well-known identification result under selection on observables, i.e.,* $F_{Y_d|P}(y|p) = P(Y \leqslant y|D = d, P = p)$.

3.2. **Nonparametric identification with Imperfect IVs.** In Theorem 2, we characterize the identified set for the DMTR and DMTE by only imposing the STC assumption. We also showed in Corollary 1 that if we additionally impose the independence assumption, we recover the HV05 point identification results for the MTE and subsequent parameters. While the IV-independence assumption can be controversial in many empirical contexts (see, for instance, Deaton, Heckman, and Imbens, 2010), a relaxed and more credible IV-independence assumption can be credible and can be used to identify trustworthy results that are more informative than not using IVs. For example, Manski and Pepper (2000) proposed the monotone IV assumption and derived tighter bounds for the ATE. However, their approach is not easily extendable to other parameters of interest, especially the PRTE. In this subsection, we will further explore in this direction and study the identification of DMTR 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 borrow the idea of monotone IV (Manski and Pepper, 2000, **MIV**) as the main motivating example and show that these restrictions can be easily implemented in our approach to derive sharp bounds on the DMTR and, therefore, on all conventional policy parameters. To this end, the discussion in this subsection also contributes to Manski and Pepper (2000); indeed, using the DMTR as a bridge, our unified approach allows the researcher to recover sharp bounds on a variety of parameters of interest under various IV dependence assumptions, a nice feature inherited from the classical MTE framework. This will free applied researchers from case-by-case constructions.

In the absence of valid IVs, applied researchers often consider a parametric model (functional forms and/or parametric distributions) and look for point identification (by functional form), especially in the presence of exogenous covariates X. This practice is usually entertained by considering variants of Heckman's (1976, 1979) classic two-step ("Heckit") estimator that impose a joint normality assumption between Y_d and $\tilde{V}$ given covariates X.[10] In these scenarios, point or set identification is often obtained because of an over-restriction of the individual treatment effect, i.e., $Y_1 - Y_0$ and also

[10]Please see Sartori (2003) and Wilde (2000) for a discussion.

the joint normality assumption imposed on the unobservables (see Kline and Walters, 2019). Our current approach is more flexible since it keeps the individual treatment effect entirely unrestricted and allows for a wider class of dependence structures.[11]

Definition 2 (Monotone IV)**.** *We say P is a monotone IV (***MIV***) if for any pair $(p', p) \in \mathcal{P} \times \mathcal{P}$ such that $p' \geqslant p$, and for $d \in \{0,1\}$, we have $Y_d|P = p' \succeq_{FSD} Y_d|P = p$, where $\succeq_{FSD}$ denotes "first order stochastic dominate". In other terms, $\mathbb{P}(Y_d \leqslant y|P = p)$ is non-increasing in p for all $y \in \mathcal{Y}$. This property is also referred as Y_d being positively regression dependent on P.*

Note that our Definition 2 differs from the classical monotone IV definition of Manski and Pepper (2000) in two aspects. First, ours is defined on the conditional distribution of Y_d given P, instead of Y_d given Z. This definition is easier to work with, particularly when Z is multi-dimensional. In Appendix A.5, we provide primitive conditions under which the FSD of the conditional distribution of Y_d given P can be implied by the FSD of the conditional distribution of Y_d given Z. Second, our definition is the conditional distribution, whereas Manski and Pepper (2000) focuses on the conditional mean. Despite these differences, the two definitions share the same spirit. We, therefore, still refer to our assumptions as monotone IV (MIV). Before presenting our main result, we revisit Examples 1 and 2 to further motivate the MIV assumption.

Example 1 Cont'd. *Mourifie, Henry, and Meango (2020) argued that even if the parental education is not independent of the children's unobserved skills, an increase in parental education cannot worsen potential labor market outcomes; therefore, it would be reasonable to consider that $P(Y_d > y|P = p)$ is non-decreasing in p. Similar shape restrictions could be invoked for other potentially "tainted" instruments: the distance to college and college fees. The FSD is also used in other scenarios. For example, Blundell, Gosling, Ichimura, and Meghir (2007) uses it to model the positive selection into the labor market by assuming that the wage distribution of workers first-order stochastically dominates those who do not work.*

[11] Our approach could also be used to analyze how much identification is obtained by restricting the individual treatment effect, i.e., $Y_1 - Y_0$ versus the selection on unobservables.

Example 2 Cont'd. *Now suppose the researcher indeed ignores the second treatment and focuses on the following model:*

$$Y = Y_1 D_1 + Y_0(1 - D_1), \tag{13}$$

$$D_1 = 1\{P_1 > V_1\}. \tag{14}$$

where again $Y_d \equiv [Y_{d1} D_2 + Y_{d0}(1 - D_2)]$ *for* $d \in \{0, 1\}$. *In this case, the quantity* $\mathbb{E}[Y_1 - Y_0 | V_1 = v_1]$ *measures the total causal effect of incarceration on recidivism for an individual at the margin that is mediated by fining. Please see Pearl (2013) for a detailed discussion. Let's presume that we are interested in identifying 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 have the following results.*

Lemma 3. *Considering the model (13, 14), 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* $(V_1, V_2, Y_{d_1 d_2}) \perp (P_1, P_2)$. *Then:*

(i) If $Y_{d1} | V_2 = v_2 \sim Y_{d0} | V_2 = v_2$ *for all* v_2 *or* $P_1 \perp P_2$, *then* $Y_d \perp P_1$.

(ii) If $Y_{d0} | V_2 = v_2 \succeq_{FSD} Y_{d1} | V_2 = v_2$ *for all* $v_2 \in [0, 1]$ *and*

- *(a)* $P_2 | P_1 = p_1' \succeq_{FSD} P_2 | P = p_1$ *for all* $p_1' \geqslant p_1$, *then* $Y_d | P_1 = p_1 \succeq_{FSD} Y_d | P_1 = p_1'$ *for all* $p_1' \geqslant p_1$.
- *(b)* $P_2 | P_1 = p_1 \succeq_{FSD} P_2 | P = p_1'$ *for all* $p_1' \geqslant p_1$, *then* $Y_d | P_1 = p_1' \succeq_{FSD} Y_d | P_1 = p_1$ *for all* $p_1' \geqslant p_1$.

(iii) If $Y_{d1} | V_2 = v_2 \succeq_{FSD} Y_{d0} | V_2 = v_2$ *for all* $v_2 \in [0, 1]$ *and*

- *(a)* $P_2 | P_1 = p_1' \succeq_{FSD} P_2 | P = p_1$ *for all* $p_1' \geqslant p_1$, *then* $Y_d | P_1 = p_1' \succeq_{FSD} Y_d | P_1 = p_1$ *for all* $p_1' \geqslant p_1$.
- *(b)* $P_2 | P_1 = p_1 \succeq_{FSD} P_2 | P = p_1'$ *for all* $p_1' \geqslant p_1$, *then* $Y_d | P_1 = p_1 \succeq_{FSD} Y_d | P_1 = p_1'$ *for all* $p_1' \geqslant p_1$.

Proof. See Appendix A.6.

Lemma 3 (i) reiterates the sufficient conditions under which the IV-independence assumption remains to hold even when the model is misspecified. Lemma 3 (ii) and (iii) show that while the IV-independence assumption is violated, we can invoke a monotone IV assumption under some reasonable restrictions. For instance, if the researcher is willing to assume:(i) $Y_{10} | V_2 = v_2 \succeq_{FSD} Y_{11} | V_2 = v_2$ *which means that conditionally on* V_2, *being externally assigned to both punishments*

(incarceration and fines), make someone less likely to re-offend than someone who is externally assigned to incarceration but with no fines; and (ii) $P_2|P_1 = p_1' \succeq_{FSD} P_2|P = p_1$ *for* $p_1' > p_1$ *—suggesting that the level of Judge's stringency is positively dependent for two types of punishments; then she could invoke a specific direction for the MIV. Notice that while the second condition is directly testable, the first one is not, but alternative directions could also be investigated, as shown by Lemma 3(iii). Below, we will show how we could bound the MTE in such an empirical context using the monotone IV assumption.* □

Here, we state the main result of this section.

Theorem 3. *P is an* **MIV** *in Definition 2 if and only if* $C_{Y_d,P}(x_1, x_2)$ *is concave in* x_2 *for all* $x_1 \in [0,1]$*, that is,*

$$\frac{\partial^2 C_{Y_d,P}(x_1, x_2)}{\partial x_2^2} \leqslant 0. \tag{15}$$

If in addition Assumptions 1 and 4 are satisfied, then the identified set under the **MIV** *restriction is given by*

$$\Theta_I^{\boldsymbol{MIV}} = \left\{ F_{Y_d}(y|p,v) : F_{Y_d}(y|p,v) = \Xi_{d,p,v}(\mathbb{P}[Y \leqslant y, D = d|P = p]), (C_{Y_d,V|P}, C_{Y_d,P}) \in \Lambda_d^{\boldsymbol{MIV}} \right\},$$

where $\Lambda_d^{\boldsymbol{MIV}}$ *is the identified set for the copula functions under* ***MIV*** *restrictions, given by*

$$\Lambda_d^{\boldsymbol{MIV}} = \left\{ (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}_d^c \times \mathcal{C}_d : \textit{ Equations } (12) \textit{and} (15) \textit{ holds, } \forall y \in \mathcal{Y}_{d,p} \right\}.$$

Theorem 3 has a significant practical advantage since it allows us to see how the identified set for the DMTRs under the STC assumption shrinks under **MIV**. We only need to intersect the set of copulas that rationalize the model under the STC with the set of copulas that respect the **MIV** restriction. The set of equality restrictions that are used to identify DMTRs do not change with how we specify $C_{Y_d,P}$. For instance, 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 would have $\Theta_I^{\mathbf{r}_J} \subseteq .... \subseteq \Theta_I^{\mathbf{r}_1}$. This provides a convenient way to operationalize Manski's "*Layered Policy Analysis*" when applied researchers want to impose different layers of assumptions on $C_{Y_d,P}$.

Specifically, motivated by **MIV**, one may consider different levels of positive dependence. The **MIV** is a particular type of positive dependence restriction, which requires that Y_d is more likely to

take on larger values when P increases. We may also consider Affiliated IV (**AIV**), which 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. When researchers have knowledge regarding the tail behaviors, possible choices also include the right tail increasing IV (**RTI-IV**), which captures the fact that Y_d is more likely to take on larger values when P takes high values as well, and the left tail decreasing IV (**LTD-IV**), which captures that Y_d is more likely to take lower values when P takes low values. One can also define a positive quadrant dependent IV (**PQD-IV**, see Bhattacharya, Shaikh, and Vytlacil, 2012, for discussions about positive quadrant dependence in treatment effect analysis). As we shown in Appendix A.7, each of these positive dependence assumptions is equivalent to a certain shape restriction on $C_{Y_d,P}$, and they nest each other as follows (Joe, 1997, Theorem 2.3):

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

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

The analysis in Theorem 3 immediately implies $\Theta_I^{\textbf{AIV}} \subseteq \Theta_I^{\textbf{MIV}} \subseteq \Theta_I^{\textbf{RTI-IV}}(\text{or } \Theta_I^{\textbf{LTD-IV}}) \subseteq \Theta_I^{\textbf{PQD-IV}}$. 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 IV-independence 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.

Remark 3. *While we focus the statement of Theorem 3 and the following discussion on the copula $C_{Y_d,P}(.,.)$, a similar analysis also applies to any restrictions the researcher would like to impose on either $C_{Y_d,V|P}(.,.)$. For instance, the HV05 identification assumptions can be transformed into our context by assuming $C_{Y_d,V|P}(.,.) = C_{Y_d,V}(.,.)$ and $C_{Y_d,P}(x_1,x_2) = x_1x_2$.*

3.3. **Semiparametric identification with unknown marginals.** 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.[12] 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 if one further (i) assumes independence between (Y_d, V) and P, (ii) restricts the copula between (Y_d, V) to be Gaussian copula, and (iii) assumes the marginal distribution of Y_d to be normal, then our model recovers the classical normal Roy selection model discussed by Heckman and Honoré (1990). Our semi-parametric identification strategy thus offers additional flexibility over the full parametric one.

To fix the idea, suppose there exists finite dimensional vector $\theta_d \equiv (\beta_d, \delta_d)$ such that $C_{Y_d,P}(x_1, x_2) = C_{Y_d,P}(x_1, x_2; \beta_d)$ 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_d and $F_{Y_d}(y), d \in \{0, 1\}$. The mapping $\Psi^{-1}_{d,p}(\mathbb{P}[Y \leqslant y, D = d|P = p]; \beta_d, \delta_d)$ is now known up to the finite dimensional parameter (β_d, δ_d). The identification of the infinite-dimensional parameter F_{Y_d} boils down to the identification of a finite-dimensional parameter (β_d, δ_d). The sharp identification region of (β_d, δ_d) is characterized by a set of equality constraints, which are easy to work with because they only contain finite-dimensional parameters and known quantities.

Assumption 5 (Parametric Copula)**.** *There exists finite dimensional copula parameters* $(\beta_d, \delta_d) \in \widetilde{\Lambda}_d \subset \mathbb{R}^T$ *such that* $C_{Y_d,P}(x_1, x_2; \beta_d)$ *is known up to* β_d *and* $C_{Y_d,V|P=p}(x_1, x_2; \sigma_d(p; \delta_d))$ *is known up to a finite number of parameters* δ_d*.*

Corollary 2. *Under Assumptions 1, 4 and 5, the identified set for DMTR is characterized as follows:*

$$\Theta_I^{\mathbf{SP}} = \left\{ F_{Y_d}(y|p, v) : F_{Y_d}(y|p, v) = \Xi_{d,p,v}(\mathbb{P}[Y \leqslant y, D = d|P = p]; (\beta_d, \delta_d)), (\beta_d, \delta_d) \in \Lambda_d^{\mathbf{SP}} \right\},$$

[12] In principle, one can consider a fully parametric model in which the joint distribution of (Y_d, P, V) is set to know up to a finite-dimensional parameter. We do not consider this approach.

where $\Xi_{d,p,v}$ *is defined in Equation* (11) *and* $\Lambda_d^{\mathbf{SP}}$ *is the identified set for the copula parameters, given by*

$$\Lambda_d^{\mathbf{SP}} = \left\{ (\tilde{\beta}_d, \tilde{\delta}_d) \in \widetilde{\Lambda}_d : \Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d | P = p]; \tilde{\beta}_d, \tilde{\delta}_d) \text{ is flat in } p \right\}.$$

Corollary 2 is a straightforward extension of Theorem 2. In the next section, we will provide a concrete example of how to implement Corollary 2.

3.4. **Discussion: Specification test.** 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), Kédagni and Mourifié (2017), and Machado, Shaikh, and Vytlacil (2019). Our approach provides a unified way to do specification tests for the assumptions the researcher is willing to maintain in the MTE framework. Indeed, each of the identified sets proposed in Theorems 2 and 3 and Corollary 2 (and those with other layers of IV assumptions) can be empty if we cannot find copulas that respect the equality constraints. The "largest" identified set that imposes the minimum structure so far is the one derived in Theorem 2. If empty, this means imposing the STC specification for treatment selection is too stringent for the data.

4. Computation of the Identified Set

Theorem 3 and corollary 2 provide general nonparametric identification results, where we either impose non-parametric shape restrictions on copulas or semi-parametric restrictions. In this section, we propose two concrete approaches to operationalize the identification results. For the first approach, we approximate the unknown copula functions nonparametrically by Bernstein copula. It has the advantage of approximating unknown copula when the order is large and being able to represent the identified set of copula parameters by polynomial constraints. For the semi-parametric approach, we show some parametrization that allows us to derive a closed-form solution for the identified set of potential outcome distributions as a function of copula parameters.

4.1. **Approximation using Bernstein Copula.** The following condition assumes that $C_{Y_d,V|P}$ and $C_{Y_d,P}$ takes a form of Bernstein copula used in Dou, Kuriki, Lin, and Richards (2021).

Assumption 6 (Bernstein Copula)**.**

$$(i) C_{Y_d,V|P}(x_1,x_2;p) = C_{Y_d,V|P}(x_1,x_2;\boldsymbol{\alpha}^d(p)) = K_d L_d \sum_{k=1}^{K_d}\sum_{\ell=1}^{L_d} \alpha_{k\ell}^d(p) B_{k-1,K_d-1}(x_1) B_{\ell-1,L_d-1}(x_2)$$

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

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* $\alpha_{k\ell}^d \geqslant 0$ *and* $\beta_{rs}^d \geqslant 0$ *are unknown parameters that satisfy* $K_d \sum_{\ell=1}^{L_d} \alpha_{k\ell}^d(p) = 1$, $L_d \sum_{k=1}^{K_d} \alpha_{k\ell}^d(p) = 1$, $S_d \sum_{r=1}^{R_d} \beta_{rs}^d = 1$, $R_d \sum_{s=1}^{S} \beta_{rs}^d = 1$.

The Bernstein copula is very useful since it can approximate any copula uniformly over $[0,1]^2$ for K_d and L_d being sufficiently large, see Sancetta and Satchell (2004, Theorem 4). When parameters $\alpha_{k\ell}^d(p)$ do not depend on p, then the joint distribution of Y_d and V depends on P through the marginal distribution $F_{Y_d|P}$. One special case of Assumption 6-(i) is analogous to "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 6-(ii) implies $F_{Y_d,P} = C_{Y_d,P}(F_{Y_d}, F_P) = F_{Y_d} \times F_P$. In this case, the IV is valid and is independent with the potential outcomes.

In Theorem 4 relegated in Appendix B.1, we characterize the identified set of the DMTR_g for arbitrary fixed values of R_d, S_d, K_d, L_d, $d \in \{0,1\}$. If we fix orders K_d, S_d, R_d, and L_d, then the Bernstein copula is essentially parametric; however, as shown in Sancetta and Satchell (2004), it can uniformly approximate any copula (subject to certain smoothness conditions) if we allow the orders to increase. Hence, the result in Theorem 4 can be viewed as a nonparametric approximation of the identified set if one is willing to consider larger orders.

We present below the special case where $R_d = S_d = K_d = L_d = 2$, for $d \in \{0,1\}$. Because $\alpha_{k\ell}^d$ and $\beta_{k\ell}^d$ are positive and must satisfy $K_d \sum_{\ell=1}^{L_d} \alpha_{k\ell}^d(p) = 1$, $L_d \sum_{k=1}^{K_d} \alpha_{k\ell}^d(p) = 1$, $S_d \sum_{r=1}^{R_d} \beta_{rs}^d = 1$, $R_d \sum_{s=1}^{S} \beta_{rs}^d = 1$, it turns out that we only need to redefine parameters $\sigma_d(p) \equiv 4\alpha_{11}^d(p) - 1 \in [-1,1]$ and $\beta_d \equiv 4\beta_{11}^d - 1 \in [-1,1]$. This reparameterization is convenient because $\sigma_d(p) = 0$ represents "selection on observables" and $\beta_d = 0$ represents valid IV. Imposing the MIV assumption is equivalent to narrowing the range of β_d from $[-1,1]$ to $[0,1]$ because **MIV** requires

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

This is equivalent to assume $\beta_{11}^d \geqslant \frac{1}{4}$. We also assume $\sigma_d(p)$ is known up to finite-dimensional parameter δ_d.[13] Let $\widetilde{\Lambda}_d = \{(\tilde{\beta}, \tilde{\delta}_d) : \sigma_d(p; \tilde{\delta}_d) \in [-1,1], \tilde{\beta}_d \in [-1,1]\}$ be the parameter space of copula parameters.

Corollary 3. *Suppose Assumptions 1 and 4 hold. Consider Assumption 6 is satisfied with* $R_d = S_d = K_d = L_d = 2$, *and the reparameterization of* $\sigma_d(p;\delta_d) \equiv 4\alpha_{11}^d(p) - 1$ *and* $\beta_d \equiv 4\beta_{11}^d - 1$. *Then, the DMTR is characterized by*

$$F_{Y_d|P,V}(y|p,v) = H_d + \sigma_d(p;\delta_d)H_d(1-H_d - 2\sigma_d(p;\delta_d)H_d(1-H_d)v,$$

where

$$\begin{aligned} H_1 &= \frac{A_1 - p + \sqrt{(A_1-p)^2 + 4A_1F_{Y,D|P}(y,1|p)}}{2A_1}, \quad A_1 = \sigma_1(p;\delta_1)p(1-p) \\ H_0 &= \frac{A_0 + 1 - p - \sqrt{(A_0+1-p)^2 + 4A_0F_{Y,D|P}(y,0|p)}}{2A_0}, \quad A_0 = \sigma_0(p;\delta_0)p(1-p), \end{aligned}$$

and H_d *depends on* y *only through* $F_{Y,D|P}(y,1|p) = \mathbb{P}(Y \leqslant y, D = d|P = p)$. *The marginal distribution of* Y_d *is given by*

$$\begin{aligned} F_{Y_d}(y) = \Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d|P = p]; \beta_d, \delta_d) \\ \equiv \frac{\beta_d B(p) + 1 - \sqrt{(\beta_d B(p)+1)^2 - 4\beta_d B(p) H_d}}{2\beta_d B(p)}, \end{aligned} \tag{16}$$

where $B(p) = 1 - 2F_P(p)$ *and the copula parameters* (β_d, δ_d) *take value from the following identified set:*

$$\Lambda_d^{\boldsymbol{BC}} = \left\{ (\tilde{\beta}_d, \tilde{\delta}_d) \in \widetilde{\Lambda}_d : \Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d|P = p]; \tilde{\beta}_d, \tilde{\delta}_d) \text{ is flat in } p \right\}.$$

Proof. See Appendix B.2.

[13] One may consider the *simplified pair-copula* specification to assume $\sigma_d(p)$ be a constant, so that the dependence structure between Y_d and V remains the same for different values of p. Haff, Aas, and Frigessi (2010) provides sufficient conditions for the validity of using *simplified pair-copula* and argues that it often provides a good approximation in practice. We can also approximate $\sigma_d(p)$ by a sequence of functions $\sigma_d(p;\delta_d)$ with the dimension of parameter δ_d increase to infinity. In this case, we need to impose restrictions on δ_d such that $\sigma_d(p;\delta_d) \in [-1,1]$ for all $p \in [0,1]$. For instance, if $\sigma_d(p;\delta_d) = \delta_d^0 + \delta_d^1 p$, then we need $\delta_d^0 \geqslant 0$ and $\delta_d^0 + \delta_d^1 \leqslant 1$ if $\delta_1^d \geqslant 0$, and $\delta_d^0 \leqslant 1$ and $\delta_d^0 + \delta_d^1 \geqslant 0$ if $\delta_d^1 < 0$.

In the statement of Corollary 3, when a denominator equals to zero, for example, when $\sigma_1(p) = 0$ or $F_P(p) = \frac{1}{2}$, the statement still holds by using L'Hospital's rule to take the limits on both numerator and denominator.

Under the second-order Bernstein copula, the identified set for DMTRs is a set of linear functions in v with intercept and slope coefficient depending on p. It is easy to analyze in practice. While Corollary 3 only reports the results of second-order Bernstein copula, it already echoes some interesting results in the literature on MTE. In fact, Corollary 3 (resp. Theorem 4 in Appendix B.1) generalizes the parametrization of the MTR proposed in Brinch, Mogstad, and Wiswall (2017) (resp. Mogstad, Santos, and Torgovitsky, 2018) to the cases: (i) where the IV is potentially invalid and (ii) where the researcher is interested in the distributional treatment effect instead of only the average as entertained in Brinch, Mogstad, and Wiswall (2017) (resp. Mogstad, Santos, and Torgovitsky, 2018). Indeed, for the generalized case discussed in Theorem 4, we showed that when $R_d = S_d = 1$ (i.e. the IV is independent with the potential outcome), 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. Then, our result provides a "micro foundation" for Mogstad, Santos, and Torgovitsky (2018)'s specification: it can be generated by a Bernstein copula $C_{Y_d,V|P}(x_1, x_2; \alpha^d)$ of order L_d.[14]

4.2. **Parametric Copula.** 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 $\sigma_d(p; \delta_d)$ excludes the value that corresponds to the product copula, which corresponds to the independence case. Comprehensive copulas, such as Gaussian and Frank copulas, parameterize the

[14] An appealing property of the Mogstad, Santos, and Torgovitsky (2018)'s approach is that under the IV independence assumption, their parametrization of the MTR is linear in the parameters, allowing them to estimate the model using a linear programming approach. However, this linearity breaks down in the presence of an invalid IV. It could be tempting to impose a linear structure on the DMTR in order to make use of the linear programming approach for estimation; however, such a parametrization (i) could significantly restrict the individual treatment effect heterogeneity, and (ii) may not be compatible with the model structure, i.e., it may not exist copulas that imply a linear structure for the DMTR.

full range of dependence. On the other hand, non-comprehensive copulas such as Farlie-Gumbel-Morgenstern (FGM), Clayton, Gumbel, and Joe copulas can only capture dependence in a limited manner. In practice, using a different family of copula to analyze how sensitive the results are depending on the copula parametrization will be useful. In this paper, we provide results based on Frank copula as the leading example because (i) Frank copula is comprehensive, and (ii) it offers a close-form solution to the identified set of $F_{Y_d}(y)$, as shown in by Corollary 4 below.[15]

Assumption 7 (Frank Copula)**.** *There exists copula parameters* $(\beta_d, \delta_d) \in \widetilde{\Lambda}_d \subseteq \mathbb{R}^T$ *with* $T < \infty$ *such that* $C_{Y_d,P}(x_1, x_2) = -\frac{1}{\beta_d} \ln\left[1 + \frac{(e^{-\beta_d x_1} - 1)(e^{-\beta_d x_2} - 1)}{(e^{-\beta_d} - 1)}\right]$ *for* $\beta_d \in (-\infty, 0) \cup (0, \infty)$ *and* $C_{Y_d,V|P=p}(x_1, x_2) = -\frac{1}{\sigma_d(p;\delta_d)} \ln\left[1 + \frac{(e^{-\sigma_d(p;\delta_d)x_1} - 1)(e^{-\sigma_d(p;\delta_d)x_2} - 1)}{(e^{-\sigma_d(p;\delta_d)} - 1)}\right]$ *for* $\sigma_d(p;\delta_d) \in (-\infty, 0) \cup (0, +\infty)$, $d \in \{0, 1\}$, *where* $\sigma_d(p;\delta_d)$ *is known up to a finite number of parameters* δ_d.

The following corollary provides a closed-form solution for DMTR when using the Frank copula.

Corollary 4. *Under Assumptions 1, 4 and 7, the DMTR is characterized as follows:*

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

where

$$\begin{aligned} H_1 &= -\frac{1}{\sigma_1(p;\delta_1)} \ln\left[1 + \frac{(e^{-\sigma_1(p;\delta_1)F_{Y,D|P}(y,1|p)} - 1)(e^{-\sigma_1(p;\delta_1)} - 1)}{(e^{-\sigma_1(p;\delta_1)p} - 1)}\right] \\ H_0 &= \frac{1}{\sigma_0(p;\delta_0)} \ln\left[1 + \frac{(e^{\sigma_0(p;\delta_0)F_{YD|P}(y,0|p)} - 1)(e^{-\sigma_0(p;\delta_0)} - 1)}{e^{-\sigma_0(p;\delta_0)} - e^{-\sigma_0(p;\delta_0)p}}\right], \end{aligned}$$

and the marginal distribution of Y_d *is given by*

$$\begin{aligned} F_{Y_d}(y) &= \Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d|P = p]; \beta_d, \delta_d) \\ &\equiv -\frac{1}{\beta_d} \ln\left[1 + \frac{H_d(y, p, \sigma_d(p;\delta_d))(e^{-\beta_d} - 1)}{e^{-\beta_d F_P(p)} - H_d(y, p, \sigma_d(p;\delta_d))(e^{-\beta_d F_P(p)} - 1)}\right], \end{aligned} \quad (17)$$

where the copula parameters (β_d, δ_d) *take value from the following identified set:*

$$\Lambda_d^{\boldsymbol{FC}} = \left\{(\tilde{\beta}_d, \tilde{\delta}_d) \in \widetilde{\Lambda}_d : \Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d|P = p]; \tilde{\beta}_d, \tilde{\delta}_d) \text{ is flat in } p\right\}.$$

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

Proof. See Appendix B.3.

Remark 4. *When the parameter $\sigma_d(p;\delta_d)$ or its finite dimensional parameter δ_d is given, β_d is uniquely determined. To see this, fixing an arbitrary y, and then integrating both side of $\beta_d f_P(p)(1-H_d)H_d + \frac{\partial H_d}{\partial p} = 0$ from $\underline{p}$ to $\overline{p}$ with respect to p gives*

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

$$\Rightarrow \beta_d = -\frac{H_d(y,\overline{p},\sigma_d(\overline{p};\delta_d)) - H_d(y,\underline{p},\sigma_d(\underline{p};\delta_d))}{\mathbb{E}\left[\{H_d(y,P,\sigma_d(P;\delta_d))(1-H_d(y,P,\sigma_d(P;\delta_d)))\}1\{\underline{p}<P\leqslant\overline{p}\}\right]}$$

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

$$\beta_d = -\frac{F_{Y_d|P}(y|\overline{p}) - F_{Y_d|P}(y|\underline{p})}{\mathbb{E}\left[F_{Y_d|P}(y|P)(1-F_{Y_d|P}(y|P))1\{\underline{p}<P\leqslant\overline{p}\}\right]}.$$

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

4.3. **Estimation.** To estimate the identified set, let $\mathcal{P}^M = \{p^1,p^2,\cdots,p^M\}$ and $\mathcal{Y}^J = \{y^1,y^2,\cdots,y^J\}$ be grid points in the support of P and Y, chosen by researchers. For generic $(\tilde{\beta}_d,\tilde{\delta}_d)$, define,

$$\kappa_d(y,p;\tilde{\beta}_d,\tilde{\delta}_d) = \Psi_{d,p}^{-1}(\mathbb{P}[Y\leqslant y, D=d|P=p];\tilde{\beta}_d,\tilde{\delta}_d).$$

The form of $\Psi_{d,p}^{-1}$ has been given in Equations (16) and (17) for Bernstein copula with order 2 and Frank copula, respectively. As illustrated by Figure 1 in a numerical example, when parameters are set to be the true values, $\kappa_d(y,p;\beta_d,\delta_d)$ is a flat function in p for any value of y. Therefore, the "sample standard deviation" of

$$\{\kappa_d(y,p^1;\beta_d,\delta_d),\cdots,\kappa_d(y,p^M;\beta_d,\delta_d)\},$$

denoted by $S_d(y;\beta_d,\delta_d)$, must be zero when evaluated at the true parameter values. Hence at the true values, we must have

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

This leads the outer set (with abuse of notation) of the identified set Λ_d for (β_d,δ_d) as[16]

$$\Lambda_d \equiv \left\{(\tilde{\beta}_d,\tilde{\delta}_d) \in \widetilde{\Lambda}_d : L_d(\tilde{\beta}_1,\tilde{\delta}_d) = 0\right\} = \underset{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d}{\operatorname{argmin}} L_d(\tilde{\beta}_d,\tilde{\delta}_d).$$

The above discussion motivates a set estimator as

$$\widehat{\Lambda}_d = \left\{(\tilde{\beta}_d,\tilde{\delta}_d) \in \widetilde{\Lambda}_d : L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d) \leqslant \epsilon_n\right\},$$

where $L_{d,n} = \sum_{j=1}^{J} \hat{S}_d(y^j;(\tilde{\beta}_d,\tilde{\delta}_d))$ is the sample analog of L_d, $\hat{S}_d(y^j;(\tilde{\beta}_d,\tilde{\delta}_d))$ is the sample variance of $\{\hat{\kappa}_d(y,p^1;\beta_d,\delta_d),\cdots,\hat{\kappa}_d(y,p^M;\beta_d,\delta_d)\}$, with

$$\hat{\kappa}_d(y,p;\tilde{\beta}_d,\tilde{\delta}_d) = \Psi_{d,p}^{-1}(\widehat{\mathbb{P}}[Y\leqslant y, D=d|\hat{P}=p];\tilde{\beta}_d,\tilde{\delta}_d),$$

and $\widehat{\mathbb{P}}[Y\leqslant y, D=d|\hat{P}=p]$ is a consistent non-parametric estimator of $\mathbb{P}[Y\leqslant y, D=d|P=p]$, $\epsilon_n \downarrow 0$ is a tuning sequence converges to zero. One can choose the rate of ϵ_n and examine the asymptotic behavior of set estimator $\widehat{\Lambda}_d$ under the general framework of Chernozhukov, Hong, and Tamer (2007). In our context, it also depends on the convergence rate of the generated regressor $\hat{P}$ and the nonparametric estimator $\widehat{\mathbb{P}}[Y\leqslant y, D=d|\hat{P}=p]$, for which we provide more details in Appendix A.8. Let $d_H(A,B)$ be the Hausdorff distance between two generic subsets A and B of the parameter space $\widetilde{\Lambda}_d$, that is, $d_H(A,B) = \max\{h(A,B),h(B,A)\}$, where

$$h(A,B) = \sup_{a\in A}\inf_{b\in B}\|a-b\|,$$

and

$$h(B,A) = \sup_{b\in B}\inf_{a\in A}\|a-b\|.$$

The following proposition shows the consistency of the (outer) set estimator.

Proposition 1. *Suppose Assumption 1 and Assumptions 8 to 12 in Appendix A.8 are satisfied, and ϵ_n converges to zero at a rate specified by Assumption 13, then $d_H(\widehat{\Lambda}_d,\Lambda_d) \xrightarrow{p} 0$.*

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

Proof. See Appendix A.8.

5. NUMERICAL ILLUSTRATION

In this subsection, we provide two numerical examples to illustrate identification results under Frank copula. In DGP1, the IV-independence assumption fails to hold, and the Frank copula is correctly specified. In this case, we demonstrate the LIV estimand fails to identify MTRs and treatment effect parameters while our approach can. The IV-independence assumption holds in DGP2 (borrowed from HV05), but the Frank copula is mis-specified. The LIV estimand identifies the MTRs. Surprisingly, the identified set for treatment effect parameters based on Frank copula is very close to the true value, perhaps due to the fact that Frank copula is comprehensive.

5.1. **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 $D = 1[V \leqslant P]$. We specify the dependence among (Y_d, P, V) using Frank copula:

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

where true parameter values are $\beta_1 = 2$, $\sigma_1 = 3$, $\beta_0 = \sigma_0(p) = 0$. In this case, P is not a valid instrument because $\beta_1 \neq 0$. The endogeneity 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 $Z > t$, $D = 1[Z(1+t) - V \geqslant 0]$; else $D = 1[Z - V \geqslant 0]$. For this exercise, we choose $t = 0.2$. The true parameter values are given by the following table:

We first demonstrate the identification result in Corollary 4. Figure 1 plots the inverse mapping $\Psi^{-1}_{1,p}(\cdot, p; \beta, \sigma)$ as a function of y at different values of $p \in \{0.2, 0.3, \cdots, 0.8\}$ (each 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 $\beta_1 = 0$ (other parameters fixed at their true values). As we can see, when we set β_1 at a false

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

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

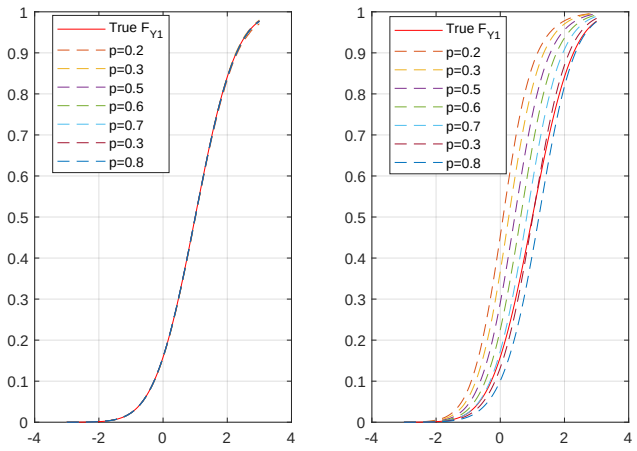

FIGURE 1. Plots of $\Psi_{1,p}^{-1}(\cdot, p; \beta, \sigma)$ at different values of p and the true F_{Y_1} (DGP1)
Left panel: parameters at true values. Right panel: β_1 is set to zero.

value and vary p, the "implied" marginal CDF of Y_1 varies. This shows that $\beta_1 = 0$ can not be the true value. In the left panel, we set the parameter value $\beta_1 = 2$. Now, regardless of which p value we insert into the mapping $\Psi_{1,p}^{-1}(\cdot, p; \beta, \sigma)$, its shape remains unchanged and is the same as the true CDF of Y_0. How $\Psi_{1,p}^{-1}$ responds to the change of p provides the identification power for different parameter values. The right panel provides evidence that $\beta_1 = 0$ should not be included in the identified set, while the left panel suggests $\beta_1 = 2$ should.

Once we obtain the approximation of the identified set for copula parameters, we turn to the identified set of the treatment effect parameters. Figure 2 draws the identified set for the MTE. In GDP1, the identified set for copula parameters is a singleton, which implies that the MTE is also

point-identified. The red line is the true MTE, and the blue line is the identified MTE based on our copula approach. The green line is what LIV would identify under the (false) IV-independence assumption. There is a notable bias when the IV-independence assumption fails to hold.

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., $(\beta_1 = 0)$; (b) We impose that $\beta_1 \leqslant 0$, 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 $\mathbf{MIV}^-$; (c) We assume the MIV assumption (positive regression dependence); and (d) we leave the dependence structure captured by β_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 $\mathbf{MIV}^-$) are not compatible with the observed data. In the two latter cases, the copula-based approach can point-identify all our policy parameters of interest even when the IV is not valid.

On the other hand, the LIV estimand shows a considerable positive bias for various policy parameters. Indeed, since the "MTE" identified from LIV has a positive bias over most parts of the unit interval, it is unsurprising that the treatment parameters identified under the IV-independence assumption have a positive bias. 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.

5.2. **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.

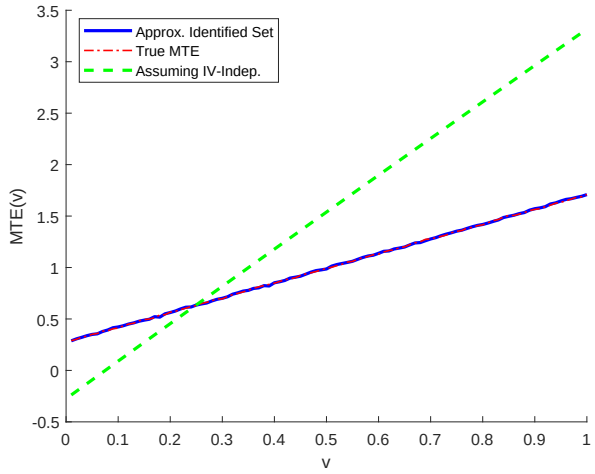

FIGURE 2. Identified Set for MTE (DGP1)

TABLE 3. Identified Values (DGP1)

Parameters	True value	Identified by Copula-based Approach				LIV
		IV ($\beta_1 = 0$)	MIV$^-$($\beta_1 \leqslant 0$)	MIV$^+$($\beta_1 \geqslant 0$)	($\beta_1 \in \mathbb{R}$)	
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

We can see from Table 4 that, while our semiparametric model is misspecified, the copula-based approach has very small biases. 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. This example shows that even though we consider

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

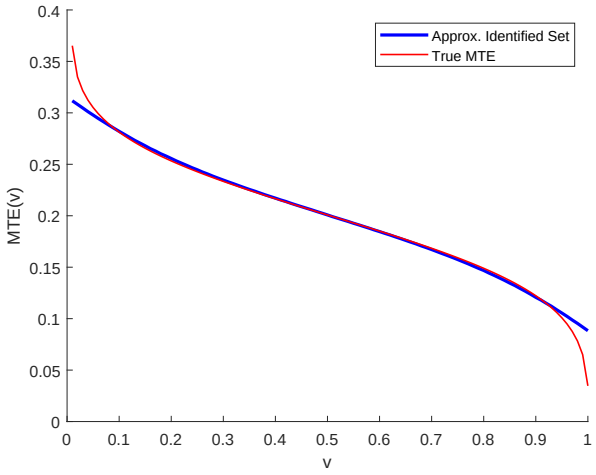

FIGURE 3. Identified Set for MTE (DGP2)

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

Figure 3 plots the MTE that we construct based upon the identified (β_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.

6. Discussion and Future Work

This paper shows how to use the MTE framework to perform layered policy analysis when the potential IVs are not necessarily valid. We propose a novel intermediate parameter, DMTE, and show it bridges the marginal treatment effect (MTE) and the policy-relevant treatment effect (PRTE) parameters even without the instrument's validity. We characterize the identified set using a Vine-copula framework, providing a unified way for researchers to impose additional IV-related assumptions of different layers of credibility.

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

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

Regarding LATE,

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

For ATT, we have

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

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

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

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

$$\mathbb{E}[g(Y^a)] = \int_0^1 \mathbb{E}[g(Y^a)|P^a = p]dF_{P^a(p)} = \int_0^1 \mathbb{E}[(g(Y_1^a) - g(Y_0^a))D^a|P^a = p]dF_{P^a(p)} + \mathbb{E}[g(Y_0^a)]$$
$$= \int_0^1 \int_0^1 1\{v \leqslant p\} f_{P^a(p)} \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, P^a = p]dpdv + \mathbb{E}[g(Y_0^a)]$$
$$= \int_0^1 \int_0^1 1\{v \leqslant p\} f_{P^a(p)} \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, U^{P^a} = F_{P^a}(p)]dpdv + \mathbb{E}[g(Y_0^a)]$$
$$= \int_0^1 \int_0^1 1\{v \leqslant F_{P^a}^{-1}(u)\} \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, U^{P^a} = u]dudv + \mathbb{E}[g(Y_0^a)]$$

The second equality holds because $U^{P^a} \equiv F_{P^a}(P^a)$ and the last equality holds by making the change of variable $u = F_{P^a}(p)$. Under Assumption 3 and the continuity of $F_{P^l}(.)$ we have that $\mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, U^{P^a} = u] = \mathbb{E}[g(Y_1^{a'}) - g(Y_0^{a'})|V^{a'} = v, U^{P^{a'}} = u]$ for all $u, v \in [0,1] \times [0,1]$ and $\mathbb{E}[g(Y_0^a)] = \mathbb{E}[g(Y_0^{a'})]$.

Then, we have:

$$\mathbb{E}[g(Y^{a'}) - g(Y^a)] = \int_0^1 \int_0^1 \left[1\left\{v \leqslant F_{P^{a'}}^{-1}(u)\right\} - 1\left\{v \leqslant F_{P^a}^{-1}(u)\right\}\right] \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, U^{P^a} = u]dudv$$
$$= \int_0^1 \int_0^1 \left[1\left\{v \leqslant F_{P^{a'}}^{-1}(F_{P^a}(p))\right\} - 1\{v \leqslant p\}\right] \mathbb{E}[g(Y_1^a) - g(Y_0^a)|V^a = v, P^a = p]dpdv$$

where the last equality holds by redoing the change of variable $u = F_{P^a}(p)$. Therefore,

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

A.2. **Proof of Lemma 2.** To show Equation (7), 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,

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

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

$$F_{Y_d}(y) = c^{-1}_{d,F_P(p)}\left(F_{Y_d|P}(y|p)\right). \tag{18}$$

To see Equation (8), 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 (9) and (10). Suppose $d = 1$

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

where we inserting Equation (18) 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 $u \mapsto C_{Y_1,V|P=p}\left(c_{1,F_P(p)}(u), p; p\right) \equiv \Psi_{1,p}(u)$ is strictly increasing. For $d = 0$,

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

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 $\mathbb{P}(Y \leqslant y, D = d|P = p)$ be the distribution of observables. It is apparent from Definition 1, $(C_{Y_d,V|P}, C_{Y_d,P}, F_d)$ satisfy Equations (9) and (10), 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 (9) and (10). 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 (9) and (10) and Equation (12) in Theorem 2 are equivalent. First, it is straightforward to see that Equations (9) and (10) imply Equation (12). Second,

suppose Equation (12) hold, that is, $\Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant y, D = d|P = p])$ is flat in p and only varies as a function of y. Note that under Assumption 4, 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, $\Psi_{d,p}^{-1}$ is strictly increasing in y by construction.

Next from the definitions in Equations (9) and (10) 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 $\Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant -\infty, D = d|P = p]) = \Psi_{d,p}^{-1}(0) = 0$ and $\Psi_{d,p}^{-1}(\mathbb{P}[Y \leqslant \infty, D = d|P = p]) = \Psi_{d,p}^{-1}(\mathbb{P}[D = d|P = p]) = 1$. This says that $\Psi^{-1}(\mathbb{P}[Y \leqslant \cdot, D = d|P = p])$, 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. First, $P \perp Y_1|V$ and $P \perp V$ implies $(Y_1, V) \perp P$. By the definition in Equation (9), the mapping $\Psi_{1,p}$ then reduces to the following simple form:

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

Since the joint distribution of (Y_d, V, P) is continuous, the mapping $\Psi_{1,p}$, its inverse $\Psi_{1,p}^{-1}$ and the population probability $\mathbb{P}[Y \leqslant y, D = 1|P = p])$ are continuously differentiable with respect to p. Next, note that the restriction is equivalent to

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

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

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

This implies

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

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 $C^{-1}_{1,Y_1,V}\big(C_{Y_1,V}(x_1,x_2),x_2\big) = x_1$ 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 $C_{Y_1,V}(F_1(y),p) = \mathbb{P}[Y \leqslant y, D = 1|P = p]$ under the independence Assumption 2.

A.5. **FSD of the conditional distribution of Y_d given P.** In this subsection, we provide conditions under which FSD of the conditional distribution of Y_d given Z implies FSD of the conditional distribution of Y_d given P. Let (Z_1, Z_2) be a vector of two instruments.

Lemma 4. *Let $P(z_1,z_2) \equiv \mathbb{P}(D = 1|Z_1 = z_1, Z_2 = z_2)$. Suppose the following conditions are satisfied:*

(i) The vector $Z = (Z_1', Z_2)'$ contains a monotone instrumental variable Z_2 in the sense that $Y_d|(Z_1 = z_1, Z_2 = z_2)$ first order stochastic dominates $Y_d|(Z_1 = z_1, Z_2 = \tilde{z}_2)$ for any z_1 and $z_2 > \tilde{z}_2$.

(ii) The function $\nu(z_1,z_2)$ is strictly increasing in z_2 for any z_1.

(iii) $\tilde{V}|(Z_1 = z_1, Z_2 = \tilde{z}_2)$ first order stochastic dominates $\tilde{V}|(Z_1 = z_1, Z_2 = z_2)$ for any z_1 and $z_2 > \tilde{z}_2$.

(iv) For any z_1 and p, there exists a z_2 such that $P(z_1,z_2) = p$.

Then $Y_d|P = p$ first order stochastic dominates $Y_d|P = \tilde{p}$ for any $p \geqslant \tilde{p}$.

Proof. First we verify that $P(z_1,z_2)$ in strictly increasing in z_2. Fixing z_1 and consider $z_2 > \tilde{z}_2$. Then

$$\begin{aligned} P(z_1,z_2) \equiv \mathbb{P}(D = 1|Z_1 = z_1, Z_2 = z_2) &= \mathbb{P}(\tilde{V} \leqslant \nu(z_1,z_2)|Z_1 = z_1, Z_2 = z_2) \\ &> \mathbb{P}(\tilde{V} \leqslant \nu(z_1,\tilde{z}_2)|Z_1 = z_1, Z_2 = z_2) \geqslant \mathbb{P}(\tilde{V} \leqslant \nu(z_1,\tilde{z}_2)|Z_1 = z_1, Z_2 = \tilde{z}_2) \\ &= \mathbb{P}(D = 1|Z_1 = z_1, Z_2 = \tilde{z}_2) \equiv P(z_1,\tilde{z}_2), \end{aligned}$$

where the first inequality holds by the strict monotonicity of ν, and the second inequality holds because $\tilde{V}|(Z_1 = z_1, Z_2 = \tilde{z}_2)$ first order stochastic dominates $\tilde{V}|(Z_1 = z_1, Z_2 = z_2)$.

Fixing z_1 and a generic p, let $\psi_2(z_1,p)$ be the inverse function of $P(z_1,z_2) = p$ with respect to the second argument. Since $P(z_1,z_2)$ is strictly increasing in z_2, ψ_2 is well-defined and is strictly increasing in p. Let $\phi(\cdot)$ be an increasing function mapping from $\mathcal{Y}_d$ to $\mathbb{R}$, then we have

$$\begin{aligned} \mathbb{E}[\phi(Y_d)|P(Z_1,Z_2) = p] = \mathbb{E}[\phi(Y_d)|Z_1 \in \mathcal{Z}_1(p), z_2 = \psi_2(Z_1,p)] = \\ \int_{\mathcal{Z}_1} \mathbb{E}[\phi(Y_d)|Z_1 = t, z_2 = \psi_2(t,p)]dF_{Z_1}(t), \end{aligned}$$

where the integration region is $\mathcal{Z}_1$ by (iv). Similarly, let $0 < \tilde{p} < p$, then we have

$$\mathbb{E}[\phi(Y_d)|P(Z_1,Z_2)=\tilde{p}] = \mathbb{E}[\phi(Y_d)|Z_1 \in \mathcal{Z}_1(\tilde{p}), z_2 = \psi_2(Z_1,\tilde{p})] = \int_{\mathcal{Z}_1} \mathbb{E}[\phi(Y_d)|Z_1 = t, z_2 = \psi_2(t,\tilde{p})] dF_{Z_1}(t).$$

Take the difference we have

$$\begin{aligned}\mathbb{E}[\phi(Y_d)|P(Z_1,Z_2)=p] - \mathbb{E}[\phi(Y_d)|P(Z_1,Z_2)=\tilde{p}] \\ = \int_{\mathcal{Z}_1} \{\mathbb{E}[\phi(Y_d)|Z_1 = t, z_2 = \psi_2(t,p)] - \mathbb{E}[\phi(Y_d)|Z_1 = t, z_2 = \psi_2(t,\tilde{p})]\} \, dF_{Z_1}(t) \geqslant 0.\end{aligned}$$

where the last inequality holds because $Y_d|(Z_1 = z_1, Z_2 = z_2)$ first order stochastic dominates $\tilde{Y}_d|(Z_1 = z_1, Z_2 = \tilde{z}_2)$. □

Condition (i) in Lemma 4 is a standard monotone IV assumption. Condition (ii) requires the benefit of taking treatment is strictly increasing in z_2. This type of monotonicity is often assumed and is reasonable in many empirical settings. In the return to college education example, if z_2 is the parent educations, it says that observed incentive of enroll in college is strictly increasing in parents education (perhaps because the parents with higher education can help children prepare a better application package). Condition (iii) would hold if individuals with higher parent education levels would prefer college life more than those with lower parent educations (given everything else equal) due to the role model effect. Finally, condition (iv) requires z_2 has a rich support and it can be verified from observed data distribution. Note that for these conditions to hold, Z_2 does not necessarily need be continuous if Z_1 is discrete.

Finally, note that when Z and $\tilde{V}$ are independent (when the potential treatment is independent with the instrument), the FSD based on P is directly induced by the FSD based on partial ordering of Z. 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 stochastic 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 Definition 2.

A.6. **Proof of Lemma 3.** Statement (1) holds straightforwardly, in which cases $\mathbb{P}(Y_1 \leqslant y|P_1 = p_1) = \mathbb{P}(Y_1 \leqslant y|P_1 = p_1')$ for all y, p_1 and p_1'.

Now consider (2)-(a). Let $H_y(p_2) \equiv \int_0^{p_2} \Big\{\mathbb{P}(Y_{11} \leqslant y|V_2 = v_2) - \mathbb{P}(Y_{10} \leqslant y|V_2 = v_2)\Big\} dv_2$. If $Y_{d0}|V_2 = v_2 \succeq_{FSD} Y_{d1}|V_2 = v_2$, then $H_y(\cdot)$ is an increasing function because the integrand is non-negative. Let $\boldsymbol{P} = (P_1, P_2)'$ and

$\boldsymbol{p} = (p_1, p_2)$, then,

$$\begin{aligned}
\mathbb{P}(Y_1 \leqslant y|\boldsymbol{P} = \boldsymbol{p}) &= \mathbb{P}(Y_{11}D_2 + Y_{10}(1 - D_2) \leqslant y|\boldsymbol{P} = \boldsymbol{p}) \\
&= \mathbb{P}(Y_{11}D_2 + Y_{10}(1 - D_2) \leqslant y, D_2 = 1|\boldsymbol{P} = \boldsymbol{p}) + \mathbb{P}(Y_{11}D_2 + Y_{10}(1 - D_2) \leqslant y, D_2 = 0|\boldsymbol{P} = \boldsymbol{p}) \\
&= \mathbb{P}(Y_{11} \leqslant y, V_2 \leqslant P_2|\boldsymbol{P} = \boldsymbol{p}) + \mathbb{P}(Y_{10} \leqslant y, V_2 > P_2|\boldsymbol{P} = \boldsymbol{p}) \\
&= \mathbb{P}(Y_{11} \leqslant y, V_2 \leqslant P_2|\boldsymbol{P} = \boldsymbol{p}) - \mathbb{P}(Y_{10} \leqslant y, V_2 \leqslant P_2|\boldsymbol{P} = \boldsymbol{p}) + \mathbb{P}(Y_{10} \leqslant y|\boldsymbol{P} = \boldsymbol{p}) \\
&= \int_0^{p_2} \mathbb{P}(Y_{11} \leqslant y|V_2 = v_2, \boldsymbol{P} = \boldsymbol{p})dv_2 - \int_0^{p_2} \mathbb{P}(Y_{10} \leqslant y|V_2 = v_2, \boldsymbol{P} = \boldsymbol{p})dv_2 + \mathbb{P}(Y_{10} \leqslant y) \\
&= \int_0^{p_2} \mathbb{P}(Y_{11} \leqslant y|V_2 = v_2)dv_2 - \int_0^{p_2} \mathbb{P}(Y_{10} \leqslant y|V_2 = v_2)dv_2 + \mathbb{P}(Y_{10} \leqslant y) = H_y(p_2) + \mathbb{P}(Y_{10} \leqslant y),
\end{aligned}$$

Therefore,

$$\mathbb{P}(Y_1 \leqslant y|P_1 = p_1) = \mathbb{E}_{P_2|P_1=p_1}[\mathbb{P}(Y_1 \leqslant y|\boldsymbol{P})] = \mathbb{E}_{P_2|P_1=p_1}[H_y(P_2)] + \mathbb{P}(Y_{10} \leqslant y).$$

Similarly, we have

$$\mathbb{P}(Y_1 \leqslant y|P_1 = p_1') = \mathbb{E}_{P_2|P_1=p_1'}[H_y(P_2)] + \mathbb{P}(Y_{10} \leqslant y),$$

Take the difference, we have

$$\mathbb{P}(Y_1 \leqslant y|P_1 = p_1) - \mathbb{P}(Y_1 \leqslant y|P_1 = p_1') = \mathbb{E}_{P_2|P_1=\boldsymbol{p}}[H_y(P_2)] - \mathbb{E}_{P_2|P_1=p_1'}[H_y(P_2)].$$

If $P_2|P_1 = p_1' \succeq_{FSD} P_2|P = p_1$, then it follows that $\mathbb{E}_{P_2|\boldsymbol{P}=\boldsymbol{p}_1}[H_y(P_2)] - \mathbb{E}_{P_2|P_1=p_1'}[H_y(P_2)] \leqslant 0$. Therefore, $\mathbb{P}(Y_1 \leqslant y|P_1 = p_1) - \mathbb{P}(Y_1 \leqslant y|P_1 = p_1') \leqslant 0$ for any $p_1' \geqslant p_1$, that is, $Y_d|P_1 = p_1 \succeq_{FSD} Y_d|P_1 = p_1'$ for any $p_1' \geqslant p_1$. For (2)-(b), we have $P_2|P_1 = p_1 \succeq_{FSD} P_2|P = p_1'$, then it must be the case that $\mathbb{E}_{P_2|\boldsymbol{P}=\boldsymbol{p}_1}[H_y(P_2)] - \mathbb{E}_{P_2|P_1=p_1'}[H_y(P_2)] \geqslant 0$, hence $\mathbb{P}(Y_1 \leqslant y|P_1 = p_1) - \mathbb{P}(Y_1 \leqslant y|P_1 = p_1') \geqslant 0$ for any $p_1' \geqslant p_1$, that is, $Y_d|P_1 = p_1' \succeq_{FSD} Y_d|P_1 = p_1$ for any $p_1' \geqslant p_1$.

Part (3) can be proved analogously as part (2). □

A.7. **Positive Dependence and Copula.** We first state the following definitions.

Definition 3 (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 $f_{P,Y_d}(p,y)f_{P,Y_d}(p',y') \geqslant f_{P,Y_d}(p,y')f_{P,Y_d}(p',y)$ for any $p > p'$ and $y > y'$, where (p,y) and (p',y') belong to the joint support of (P,Y_d)*[18].

(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 $p' \geqslant p$, $P = p$, 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 $y \in \mathcal{Y}$.*[19]

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

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

(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 $\mathbb{P}(Y_d \leqslant y|P \leqslant p)$ 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 $\mathbb{P}(Y_d > y, P > p) \geqslant \mathbb{P}(Y_d > y)\mathbb{P}(P > p)$ for all $(y, p) \in \mathcal{Y} \times \mathcal{P}$.*

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

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

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

for all $(x_1, x_2) \in [0,1]$ and $(x_1', x_2') \in [0,1]$ such that $x_1 \geqslant x_1'$ and $x_2 \geqslant x_2'$, 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 $x_1 \in [0,1]$,*

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

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

$$\frac{\partial^2 C_{Y_d,P}(x_1, x_2)}{\partial x_2^2} \leqslant 0. \tag{19}$$

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

$$\frac{\partial C_{Y_d,P}(x_1, x_2)}{\partial x_1} \geqslant \frac{[x_2 - C_{Y_d,P}(x_1, x_2)]}{1 - x_1}. \tag{20}$$

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

$$\frac{\partial C_{Y_d,P}(x_1, x_2)}{\partial x_1} \leqslant \frac{C_{Y_d,P}(x_1, x_2)}{x_1}. \tag{21}$$

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

$$C_{Y_d,P}(x_1, x_2) \geqslant x_1 x_2. \tag{22}$$

Proof. 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.8. **Proof of Proposition 1.** We make the following assumptions.

Assumption 8. *$\{Y_i, D_i, Z_i\}_{i=1}^n$ are i.i.d. observations.*

Assumption 9. *We Assume the following.*

i *P is continuous with continuously differentiable density function f_P.*

ii *There exists $\delta > 0$ such that $f_P(p) > \delta > 0$ all $p \in \mathcal{P}$.*

iii *For all $y \in \mathcal{Y}$, and $d = 0, 1$, $P(Y \leqslant y, D = d|P = p)$ is twice continuously differentiable in p.*

Assumption 9-i is made for the sake of notation simplicity. If P has a mass point at a given $p \in \mathcal{P}^M$, then we can estimate $P(Y \leqslant y, D = d|P = p)$ by counting the portion of the event $\{Y_i \leqslant y, D_i = d\}$ over the subsample of $P_i = p$. Assumption 9-ii ensures that there exist enough observations whose propensity score is in the neighborhood of p so that we do not condition on an empty event. Assumption 9-(iii) is needed for consistent estimation of the generated regressor $P(Z_i)$.

Assumption 10. *There exists a sequence $a_n \to \infty$ such that*

$$\sup_{y \in \mathcal{Y}^J, p \in \mathcal{P}^M, d \in \{0,1\}} a_n |\widehat{P}(Y \leqslant y, D = d|\hat{P} = p) - P(Y \leqslant y, D = d|P = p)| = O_p(1),$$

where the sequence a_n is polynomial, that is, it satisfies $a_n = n^\gamma$ for some $\gamma > 0$.

Assumption 10 is a high-level condition. It states that the quantity $P(Y \leqslant y, D = d|P = p)$ can be consistently estimated using generated regressor $\hat{P}$ at a polynomial rate uniformly over $\mathcal{P}^M$ and $\mathcal{Y}^J$. Note that the set of values $\mathcal{P}^M$ for the conditioning variable P is finite. The variable to be taken expectation, $1\{Y \leqslant y, D = d\}$, is an indicator function and is bounded. Hence, this condition holds for common nonparametric estimators, e.g., the two-step Nadaraya-Watson kernel estimator of Rilstone (1996) or the two-step local polynomial estimator of Mammen, Rothe, and Schienle (2012, Corollary 1). In Lemma 6 below, we provide primitive conditions for a Kernel estimator to satisfy Assumption 10.

Assumption 11. *The parameter space $\widetilde{\Lambda}_d$ for (β_d, δ_d) is compact. The identified set is in the interior of $\widetilde{\Lambda}_d$.*

Assumption 11 is a technical assumption that enables us to derive consistency results in Proposition 1. In practice, when one employs a comprehensive copula, such as the Frank copula, the parameter values are not bounded. In this case,

one can set parameter space to $[-B,-c]\cup[c,B]$ for some large B and small c, unless we believe the variables involved are perfectly correlated or independent.

Assumption 12. *The following conditions hold for the copula functions.*

i The copula functions The copulas $C_{Y_d,V|P=p}(\cdot,\cdot;\sigma_d(p;\delta_d))$ and $C_{Y_d,P}(\cdot,\cdot;\alpha_d)$ are continuously differentiable with respect to its arguments for any $\theta_d\in\widetilde{\Lambda}_d$. Furthermore, $C_{Y_d,V|P=p}(\cdot,\cdot;\sigma_d(p;\delta_d))$ is continuously differentiable with respect to p.

ii The copulas $C_{Y_d,V|P=p}(\cdot,\cdot;\sigma_d(p;\delta_d))$ and $C_{Y_d,P}(\cdot,\cdot;\alpha_d)$ are continuously differentiable with respect to $\theta_d=(\alpha_d,\delta_d)$ over $\widetilde{\Lambda}_d$.

Assumption 12 holds common parameterization of the copula functions, as well as for the Bernstein approximation. It ensures that the mapping $\Psi_{d,p}(\cdot)$ and its inverse are continuously differentiable.

Assumption 13. *Let a_n be specified in Assumption 10. Then ϵ_n is chosen such that $\epsilon_n\to 0$ and $\epsilon_n a_n\to\infty$.*

Assumption 13 specifies the "level" of the set estimator. A simple choice would be $\epsilon_n=\frac{\log n}{a_n}$.

Proof to Proposition 1. Our proof follows the same idea of Chernozhukov, Hong, and Tamer (2007, Theorem 3.1). We first show

$$\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d} a_n\left|L_d(\tilde{\beta}_d,\tilde{\delta}_d)-L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d)\right|=O_p(1). \tag{23}$$

Note first for each $(y,p)\in\mathcal{Y}^J\times\mathcal{P}^M$,

$$\begin{aligned}
&\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d}|\hat{\kappa}_d(y,p;\tilde{\beta}_d,\tilde{\delta}_d)-\kappa_d(y,p;\tilde{\beta}_d,\tilde{\delta}_d)|\\
&\quad=\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d}|\Psi_{d,p}^{-1}(\mathbb{P}[Y\leqslant y,D=d|P=p];\tilde{\beta}_d,\tilde{\delta}_d)-\Psi_{d,p}^{-1}(\hat{\mathbb{P}}[Y\leqslant y,D=d|P=p];\tilde{\beta}_d,\tilde{\delta}_d)|\\
&\leqslant\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d,t\in[0,1],p\in[0,1]}|\Psi_{d,p}^{-1'}(t;\tilde{\beta}_d,\tilde{\delta}_d)|\times|\hat{\mathbb{P}}[Y\leqslant y,D=d|P=p]-\mathbb{P}[Y\leqslant y,D=d|P=p)|=O_p(1/a_n),
\end{aligned} \tag{24}$$

where $\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d,t\in[0,1],p\in[0,1]}|\Psi_{d,p}^{-1'}(t;\tilde{\beta}_d,\tilde{\delta}_d)|<C_B$ for some constant C_B by Assumption 12 and the last equality holds by Assumption 10.

Recall that

$$L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d)=\frac{1}{MJ}\sum_{j=1}^{J}\sum_{m=1}^{I}(\hat{\kappa}_d(y^j,p^m;\tilde{\beta}_d,\tilde{\delta}_d)-\bar{\kappa}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d))^2=\frac{1}{MJ}\sum_{j=1}^{J}\hat{\boldsymbol{\kappa}}_d'(y^j;\tilde{\beta}_d,\tilde{\delta}_d)(I_M-P_\iota)\hat{\boldsymbol{\kappa}}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d)$$

where $\hat{\boldsymbol{\kappa}}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d)$ is the $M\times 1$ vector of $(\hat{\kappa}_d(y^j,p^m;\tilde{\beta}_d,\tilde{\delta}_d))_{m=1,2,\cdots,M}$, $\iota=(1,1,\cdots,1)'$ is a $M\times 1$ vector of ones, and P_ι is the projection matrix $\iota(\iota'\iota)^{-1}\iota'$, and I_M is the identity matrix with dimension M. $L_d(\tilde{\beta}_d,\tilde{\delta}_d)$ takes a similar form

with $\hat{\kappa}_d$ replaced by κ_d. Hence,

$$\begin{aligned}
&\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d} \left|L_d(\tilde{\beta}_d,\tilde{\delta}_d) - L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d)\right| \\
&\quad\leqslant \frac{1}{MJ}\sum_{j=1}^{J} \sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d} (\hat{\boldsymbol{\kappa}}'_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d) - \boldsymbol{\kappa}'_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d))(I_M - P_l)(\hat{\boldsymbol{\kappa}}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d) - \boldsymbol{\kappa}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d)) \\
&\qquad + \frac{2}{MJ}\sum_{j=1}^{J} \sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\widetilde{\Lambda}_d} |(\hat{\boldsymbol{\kappa}}'_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d) - \boldsymbol{\kappa}'_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d))(I_M - P_l)\boldsymbol{\kappa}_d(y^j;\tilde{\beta}_d,\tilde{\delta}_d)| \\
&\qquad\qquad = O_p(1/a_n^2) + O_p(1/a_n) = O_p(1/a_n).
\end{aligned}$$

where the final equality holds because Equation (24).

By definition of the identified set, $L_d(\tilde{\beta}_d,\tilde{\delta}_d) = 0$ for all $(\tilde{\beta}_d,\tilde{\delta}_d) \in \Lambda_d$, and the set estimator

$$\widehat{\Lambda}_d = \left\{(\tilde{\beta}_d,\tilde{\delta}_d) : a_n L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d) \leqslant a_n\epsilon_n\right\},$$

Because $\sup_{(\tilde{\beta}_d,\tilde{\delta}_d)\in\Lambda_d} a_n L_{d,n}(\tilde{\beta}_d,\tilde{\delta}_d) < a_n\epsilon_n$ w.p.a.1 by Assumption 13 and Equation (23), it follows that $\mathbb{P}(\Lambda_d \subseteq \widehat{\Lambda}_d) \to 1$, or equivalently, $\sup_{\theta_d\in\Lambda_d} d(\theta_d,\widehat{\Lambda}_d) \xrightarrow{p} 0$, where for a generic set A, $d(\theta_d,A) = \inf_{\theta'_d\in A}\|\theta_d - \theta'_d\|$.

To show the other direction, let $\eta > 0$ be an arbitrarily small positive number, and let Λ_d^η be the $\eta-$expansion of Λ_d: $\Lambda_d^\eta = \{\theta_d \in \widetilde{\Lambda}_d : d(\theta_d,\Lambda_d) \leqslant \eta\}$. By Equation (23), we have $\inf_{\theta_d\in\widetilde{\Lambda}_d/\Lambda_d^\eta} L_{n,d}(\theta_d) \geqslant \inf_{\theta_d\in\widetilde{\Lambda}_d/\Lambda_d^\eta} L_d(\theta_d) + o_p(1)$. By the identification condition, there exists a δ_η, such that $\inf_{\theta_d\in\widetilde{\Lambda}_d/\Lambda_d^\eta} L_d(\theta_d) \geqslant \delta_\eta > 0$. On the other hand, $\sup_{\widehat{\Lambda}_d} L_{n,d}(\theta_d) \leqslant \epsilon_n = o_p(1)$. This means $\widehat{\Lambda}_d \cap (\widetilde{\Lambda}_d/\Lambda_d^\eta) = \varnothing$ w.p.a.1, which in turn implies $\mathbb{P}(\widehat{\Lambda}_d \subseteq \Lambda_d^\eta) \to 1$, or $\sup_{\theta_d\in\widehat{\Lambda}_d} d(\theta_d,\Lambda_d) \leqslant \eta$ w.p.a.1. Since η is an arbitrarily small number, it follows that $\sup_{\theta_d\in\widehat{\Lambda}_d} d(\theta_d,\Lambda_d) \xrightarrow{p} 0$.

Combine both parts, and it follows that $d_H(\Lambda_d,\widehat{\Lambda}_d) \xrightarrow{p} 0$. □

Lemma 6. *For $p \in \mathcal{P}^M$ and $y \in \mathcal{Y}^J$, let*

$$\widehat{P}(Y \leqslant y, D = d|\hat{P} = p) = \frac{\sum_{i=1}^n 1\{Y_i \leqslant y, D_i = d\} K_p\left(\frac{\hat{P}_i - p}{h_p}\right)}{\sum_{i=1}^n K_p\left(\frac{\hat{P}_i - p}{h_p}\right)}, \tag{25}$$

where $h_p \downarrow 0$ is a bandwidth, K_p is a univariate kernel function, and

$$\hat{P}_i \equiv \hat{P}(Z_i) = \frac{\sum_{j\neq i} D_i K_z\left(\frac{Z_j - Z_i}{h_z}\right)}{\sum_{j\neq i} K_z\left(\frac{Z_j - Z_i}{h_z}\right)}$$

with $h_z \downarrow 0$ and K_z being another pair of bandwidth and kernel function in estimating $P_i \equiv P(Z_i)$. Suppose the following conditions hold:

(1) Let $K(\cdot)$ be a second order kernel defined on $[0,1]$ (Epanechnikov kernel or biweight kernel). Then we choose $K_p(\cdot) = K(\cdot)$, and $K_z(\cdot) = K^{d_z}(\cdot)$, where d_z is the dimension of z.

(2) For some positive constants ϵ_z and ϵ_p, we have

$$h_p = c_p n^{-\frac{1+\epsilon_p}{5}}, \quad h_z = c_z n^{-\frac{1+\epsilon_z}{4+d_z}}$$

where the undersmoothing constants ϵ_z and ϵ_p are positive and satisfies $nh_p \to \infty$ and $nh_z^{d_z} \to \infty$.

(3) Assumptions 8 and 9 hold.

then Assumption 10 holds with $a_n = \sqrt{nh_z^{d_z}}$.

Proof. The proof to the lemma is a direct application of the main proposition of Rilstone (1996) and therefore omitted. □

Appendix B. Additional results on Copulas

B.1. General Results using Bernstein Copula.

Theorem 4. *Under Assumptions 1, 4 and 6, the identified set of the Bernstein copulas parameters $\overline{\Theta}_I^{BC}$ of $\left(\theta, (\gamma_{r-1,y}^{j_e})_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC}$ is characterized as follows:*

$$\overline{\Theta}_I^{BC} = \left\{ \left(\theta, (\gamma_{r-1,y}^{j_e})_{y\in\mathcal{Y}}\right) \in \Theta^{BC} \times \Gamma^{BC} \text{ that satisfies Equations (29)and (30), for all } g(.) \in \mathcal{G} \right\},$$

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

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

$$\mathbb{E}[g(Y_d)|V=v,P=p;\bar{\theta}] = K_d L_d \sum_{k=1}^{K_d}\sum_{l=1}^{L_d} \beta_{kl}^d b_{l-1,L_d-1}(v) \sum_{r=1}^{R_d}\sum_{s=1}^{S_d} \beta_{rs}^d b_{s-1,S_d-1}(F_P) \Bigg\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_d-1}{j} \binom{j}{k-1} (R_d S_d)^{j+1} \times$$

$$\sum_{n_{11}+n_{12}+...+n_{R_d S_d}=j} \binom{j}{n_{11}, n_{12}, ..., n_{R_d S_d}} \prod_{e=1,f=1}^{R_d,S_d} \left(\beta_{ef}^d\right)^{n_{ef}} \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{d,n_{e\cdot}} \Bigg\}$$

$$\forall\ \bar{\theta} \equiv \left(\theta, (\gamma_{r-1,y}^{j_e})_{y\in\mathcal{Y}}\right) \in \overline{\Theta}_I^{BC} \Bigg\}.$$

Proof. The density of the Bernstein copula is given by:

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

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

$$b_{i,I}(u) = \sum_{j=i}^{I} (-1)^{j-i} \binom{I}{j}\binom{j}{i} u^j \tag{26}$$

First, we assume that Assumptions 1 and 4 hold, then under Assumption 6(i), we have

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

In addition, under Assumption 6(ii) we can derive the following:

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

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

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

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

and,

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

We have the following factorization:

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

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

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

Therefore, we can write:

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

Finally, we have

$$\mathbb{E}[g(Y_d)|V=v,P=p]=K_dL_d\sum_{k=1}^{K_d}\sum_{l=1}^{L_d}\alpha_{kl}^d b_{l-1,L_d-1}(v)\sum_{r=1}^{R_d}\sum_{s=1}^{S_d}\beta_{rs}^d b_{s-1,S_d-1}(F_P)\Bigg\{\sum_{j=k-1}^{K_d-1}(-1)^{j-k+1}\binom{K_d-1}{j}\binom{j}{k-1}(R_dS_d)^{j+1}\times$$
$$\sum_{n_{11}+n_{12}+\ldots+n_{R_dS_d}=j}\binom{j}{n_{11},n_{12},\ldots,n_{R_dS_d}}\prod_{e=1,f=1}^{R_d,S_d}(\beta_{ef}^d)^{n_{ef}}\prod_{f=1}^{S_d}\left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}}\gamma_{r-1,g}^{d,n_{e\cdot}}\Bigg\}\quad(27)$$

Remark that when $R_d=S_d=1$, $C_{Y_d,P}(x_1,x_2;\beta^d)=x_1x_2$ which is equivalent to $Y_d\perp P$, in such a case Equation (27) simplifies to

$$\mathbb{E}[g(Y_d)|V=v,P=p]=K_dL_d\sum_{k=1}^{K_d}\sum_{l=1}^{L_d}\alpha_{kl}^d b_{l-1,L_d-1}(v)\underbrace{\int_{\mathcal{Y}}g(y)b_{k-1,K_d-1}(F_{Y_d}(y))f_{Y_d}(y)dy}_{\tau_{g,k}^d}$$
$$=\sum_{l=1}^{L_d}\underbrace{\Big(K_dL_d\sum_{k=1}^{K_d}\alpha_{kl}^d\tau_{g,k}^d\Big)}_{\theta_{dl}^g}b_{l-1,L_d-1}(v)=\sum_{l=1}^{L_d}\theta_{dl}^g b_{l-1,L_d-1}(v)=\mathbb{E}[g(Y_d)|V=v].\quad(28)$$

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:

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

$$= K_d L_d \sum_{k=1}^{K_d}\sum_{l=1}^{L_d} \alpha_{kl}^d \int_{p1\{d=0\}}^{p+(1-p)1\{d=0\}} b_{l-1,L_d-1}(v)dv \sum_{r=1}^{R_d}\sum_{s=1}^{S_d} \beta_{rs}^d b_{s-1,S_d-1}(F_P) \left\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_d-1}{j} \binom{j}{k-1} (R_d S_d)^{j+1} \times \right.$$

$$\left. \sum_{n_{11}+n_{12}+...+n_{R_dS_d}=j} \binom{j}{n_{11},n_{12},...,n_{R_dS_d}} \prod_{e=1,f=1}^{R_d,S_d} (\beta_{ef}^d)^{n_{ef}} \prod_{f=1}^{S_d} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{d,n_{e\cdot}} \right\}$$

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:

$$\mathbb{E}[g(Y)D|P=p] = K_1 L_1 \sum_{k=1}^{K_1}\sum_{l=1}^{L_1} \alpha_{kl}^1 B_{l-1,L_1-1}(p) \sum_{r=1}^{R_1}\sum_{s=1}^{S_1} \beta_{rs}^1 b_{s-1,S_1-1}(F_P) \left\{ \sum_{j=k-1}^{K_1-1} (-1)^{j-k+1} \binom{K_1-1}{j} \binom{j}{k-1} (R_1 S_1)^{j+1} \times \right.$$

$$\left. \sum_{n_{11}+n_{12}+...+n_{R_1S_1}=j} \binom{j}{n_{11},n_{12},...,n_{R_1S_1}} \prod_{e=1,f=1}^{R_1,S_1} (\beta_{ef}^1)^{n_{ef}} \prod_{f=1}^{S_1} \left(b_{f-1,S_d-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{1,n_{e\cdot}} \right\} \quad (29)$$

and

$$\mathbb{E}[g(Y)(1-D)|P=p] = K_0 L_0 \sum_{k=1}^{K_0}\sum_{l=1}^{L_0} \alpha_{kl}^0 (1/L_0 - B_{l-1,L_0-1}(p)) \sum_{r=1}^{R_0}\sum_{s=1}^{S_0} \beta_{rs}^0 b_{s-1,S_0-1}(F_P) \left\{ \sum_{j=k-1}^{K_d-1} (-1)^{j-k+1} \binom{K_0-1}{j} \binom{j}{k-1} (R_0 S_0)^{j+1} \times \right.$$

$$\left. \sum_{n_{11}+n_{12}+...+n_{R_0S_0}=j} \binom{j}{n_{11},n_{12},...,n_{R_0S_0}} \prod_{e=1,f=1}^{R_0,S_0} (\beta_{ef}^0)^{n_{ef}} \prod_{f=1}^{S_0} \left(b_{f-1,S_0-1}(F_P)\right)^{n_{\cdot f}} \gamma_{r-1,g}^{0,n_{e\cdot}} \right\}. \quad (30)$$

Restrictions on $\gamma_{r-1,g}^{d,j_e}$

Recall $\gamma_{r-1,g}^{d,j_e} \equiv \mathbb{E}\left[g(Y_d) \prod_{e=1}^{R_d} \left(B_{e-1,R_d-1}(F_{Y_d})\right)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\right]$ for $j_e \in \{1,...,L_d\}$. $\gamma_{r-1,g}^{d,j_e}$ 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 $\gamma^{d,j_e}_{r-1,g}$, for instance for $g(\cdot) = 1\{\cdot \leqslant y\}$, $g(.) = 1\{y < \cdot \leqslant y'\}$, $g^+ : \mathcal{Y} \to \mathbb{R}^+$, or $g^- : \mathcal{Y} \to \mathbb{R}^-$ we have respectively

$$\gamma^{d,j_e}_{r-1,y} = \mathbb{E}\Big[1\{Y \leqslant y\} \prod_{e=1}^{R_d} \big(B_{e-1,R_d-1}(F_{Y_d})\big)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \geqslant 0, \tag{31}$$

$$\gamma^{d,j_e}_{r-1,y'} - \gamma^{d,j_e}_{r-1,y} = \mathbb{E}\Big[1\{y < Y \leqslant y'\} \prod_{e=1}^{R_d} \big(B_{e-1,R_d-1}(F_{Y_d})\big)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \geqslant 0 \ \ \forall y' > y, \tag{32}$$

$$\gamma^{d,j_e}_{r-1,g^+} = \mathbb{E}\Big[g^+(Y_d) \prod_{e=1}^{R_d} \big(B_{e-1,R_d-1}(F_{Y_d})\big)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \geqslant 0, \tag{33}$$

$$\gamma^{d,j_e}_{r-1,g^-} = \mathbb{E}\Big[g^-(Y_d) \prod_{e=1}^{R_d} \big(B_{e-1,R_d-1}(F_{Y_d})\big)^{j_e} b_{r-1,R_d-1}(F_{Y_d})\Big] \leqslant 0. \tag{34}$$

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

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

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 $\mathcal{G} \equiv \{g(\cdot) = \mathbf{1}[\cdot \leqslant y], y \in \mathcal{Y}\}$ 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

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

□

B.2. **Proof to Corollary 3.** Now we consider the Bernstein copula with $K_d = L_d = R_d = S_d = 2$. In this case

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

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

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 $\alpha^d_{kl} \geqslant 0$ and $\beta^d_{rs} \geqslant 0$ 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 αs 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 $0 \leqslant \alpha_{11}^d \leqslant \frac{1}{2}$. Therefore, we can write

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

Note that if we define $\sigma_d \equiv 4\alpha_{11}^d - 1 \in [-1,1]$, then

$$C_{Y_d,V|P}(x_1,x_2;\sigma_d) = x_1x_2\big(1+\sigma_d(p)(1-x_1)(1-x_2)\big)$$

If we impose $\alpha_{11}^d(p) = \frac{1}{4} \Leftrightarrow \sigma_d(p) = 0$, then we are imposing the selection-on-observable assumption.

Likewise, when $R_d = S_d = 2$, if we parameterize $\beta_d \equiv 4\beta_{11}^d - 1 \in [-1,1]$, then

$$C_{Y_d,P}(x_1,x_2;\beta^d) = x_1x_2\big(1+\beta_d(1-x_1)(1-x_2)\big).$$

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

Consider $d = 1$. Note first by Equation (9),

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

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 $c_{1,F_P(p)}(F_{Y_1}(y)) \equiv H_1(y,p,\sigma_1)$. It turns out the only that is valid is given by the following expression (the other one takes value outside of $[0,1]$):

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

Note take limit of $\sigma_1 \to 0$ yield $H_1 \to \frac{F_{Y,D|P}(y,1|p)}{p}$. Next, we express DMTR as a function of $(F_{Y,D|P},\sigma_1)$

$$\begin{aligned} F_{Y_1|P,V}(y|p,v) = \frac{\partial C_{Y_1,V|P}(x_1,x_2)}{\partial x_2}|_{x_1=H_1,x_2=v} &= H_1+\sigma_1H_1(1-H_1)(1-2v) \\ &= H_1+\sigma_1H_1(1-H_1)-2\sigma_1H_1(1-H_1)v \end{aligned}$$

where

$$H_1 = \frac{A-p+\sqrt{(A-p)^2+4AF_{Y,D|P}(y,1|p)}}{2A}, \quad A = \sigma_1p(1-p)$$

In this case, the DMTR is linear in v for any fixed value of p, but not linear in p for any fixed value of v. This is not a result that we expect ex-ante.

Next, we derive the possible range of σ_1. Recall that

$$H_1(y,p,\sigma_1(p)) = c_{1,F_P(p)}(F_{Y_1}(y)) = \frac{\partial C_{Y_1,P}(x_1,x_2)}{\partial x_2}|_{x_1=F_{Y_1}(y),x_2=F_P(p)}$$
$$= x_1 + \beta_1 x_1(1-x_1)(1-2x_2)|_{x_1=F_{Y_1}(y),x_2=F_P(p)} = F_{Y_1}(y) + \beta_1 F_{Y_1}(y)(1-F_{Y_1}(y))(1-2F_P(p))$$

Define $B_1 = \beta_1(1-2F_P(p))$. Solve $F_{Y_1}(y)$ from the above equation, we get again two possible solutions.

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

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 $B_1+1 \geqslant 2B_1$ and $B_1+1 \geqslant 0$. Following similar argument as above (and use the fact that $0 \leqslant H_1 \leqslant 1$, we can show that the first solution is not valid while the second solution is. Therefore, it must be the case that

$$F_{Y_1}(y) = \frac{\beta_1(1-2F_P(p))+1-\sqrt{(\beta_1(1-2F_P(p))+1)^2-4\beta_1(1-2F_P(p))H_1(y,p,\sigma_1(p))}}{2\beta_1(1-2F_P(p))},$$

where

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

The possible range of (β_1,σ_1) is such that the right hand side of $F_{Y_1}(y)$ being flat in p.

Now consider $d=0$. Note first by Equation (10),

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

From the above equation, we have two possible roots for $c_{0,F_P(p)}(F_{Y_0}(y))$. The valid one is given by

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

where $A_0 = \sigma_0(p)p(1-p)$. It turns out that the second root is valid, that is,

$$c_{0,F_P(p)}(F_{Y_0}(y)) = \frac{\sigma_0p(1-p)+1-p-\sqrt{(\sigma_0p(1-p)+1-p)^2+4\sigma_0p(1-p)F_{Y,D|P}(y,0|p)}}{2\sigma_0p(1-p)} \equiv H_0(y,p,\sigma_0(p)).$$

The DMTR is given by

$$F_{Y_0|P,V}(y|p,v) = \frac{\partial C_{Y_0,V|P}(x_1,x_2)}{\partial x_2}|_{x_1=H_0,x_2=v} = H_0 + \sigma_1 H_0(1-H_0)(1-2v)$$
$$= H_0 + \sigma_0 H_0(1-H_0) - 2\sigma_0 H_0(1-H_0)v.$$

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

$$F_{Y_0}(y) = \frac{\beta_0(1-2F_P(p))+1-\sqrt{(\beta_0(1-2F_P(p))+1)^2-4\beta_0(1-2F_P(p))H_0(y,p,\sigma_0(p))}}{2\beta_0(1-2F_P(p))}.$$

And the range of (β_0,σ_0) is such that the right hand side of $F_{Y_0}(y)$ being flat in p for all y.

B.3. **Frank Copula: Proof of Corollary 4.** Consider $d = 1$. Note first,

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

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

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

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

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

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

$$F_{Y_1}(y;\theta) = -\frac{1}{\beta_1} \ln\Big[1 + \frac{H_1(y,p,\sigma_1(p))(e^{-\beta_1} - 1)}{e^{-\beta_1 F_P(p)} - H_1(y,p,\sigma_1(p))(e^{-\beta_1 F_P(p)} - 1)}\Big]$$

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

$$F_{Y_0}(y;\theta) = -\frac{1}{\beta_0} \ln\Big[1 + \frac{H_0(y,p,\sigma_0(p))(e^{-\beta_0} - 1)}{e^{-\beta_0 F_P(p)} - H_0(y,p,\sigma_0(p))(e^{-\beta_0 F_P(p)} - 1)}\Big].$$

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 $e^{-\beta_d F_P(p)} \neq 0$ and $H_d \neq 0$,

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

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 $\Lambda_I^{\mathbf{F}}$ and

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

Appendix C. Extensions

In this appendix section, we discussed two extensions to our baseline model. In section C.1, we consider the case in which the treatment is determined by a double hurdle (DH) model. In section C.2, we consider the case in which the outcome variable or propensity score is discrete.

C.1. **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 finite number of thresholds. However, for the sake of simplicity, we will consider the case with two thresholds.

Assumption 14 (Double Hurdle 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:

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

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 15 (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 14 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 $s \in \{$*ATE, LATE*$(\mathbf{u}, \mathbf{u}')$*, ATT, ATUT*$\}$[21] *and weights* $\omega^s(\mathbf{v}, \mathbf{q})$ *listed in Table 5 below, we have*

$$s = \int_{\mathbf{v}} \int_{\mathbf{q}} \omega^s(\mathbf{v}, \mathbf{q}) DMTE(\mathbf{v}, \mathbf{q}) d\mathbf{v} d\mathbf{q}. \tag{35}$$

(iii) If in addition Assumption 15 holds, Equation (35) holds with $s = PRTE$.

(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. under Assumption 14 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:

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

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

[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 $\mathbf{u}'$: $\text{LATE}_g(\mathbf{u}, \mathbf{u}') \equiv \mathbb{E}[g(Y_1) - g(Y_0)|u_1 < V_1 \leqslant u_1', u_2 < V_2 \leqslant u_2']$.

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

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

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

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 14 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 15, then under both Assumptions 14 and 15 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 $\mathbf{Z} = (Z_1, Z_2)$,

$$\begin{aligned} F_{\mathbf{Q}}(\mathbf{q}) &= \mathbb{P}(Q_1(Z_1) \leqslant q_1, Q_2(Z_2) \leqslant q_2) \\ &= \mathbb{P}\left(Z_1 \leqslant Q_1^{-1}(q_1), Z_2 \leqslant Q_2^{-1}(q_2)\right) = F_{\mathbf{Z}}\left(Q_1^{-1}(q_1), Q_2^{-1}(q_2)\right) \end{aligned}$$

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). □

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	$\frac{[f_{\mathbf{Q}^{a'}}(\mathbf{q}) - f_{\mathbf{Q}^a}(\mathbf{q})] f_{\mathbf{V}^a}(\mathbf{v}) 1_{\{\mathbf{v} \in [0,q_1] \times [0,q_2]\}}}{\mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^{a'})] - \mathbb{E}[F_{\mathbf{V}}(\mathbf{Q}^a)]}$

Remark 5. *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 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 $\text{LATE}_g(\mathbf{u}, \mathbf{u}') \equiv \mathbb{E}[g(Y_1) - g(Y_0) | u_1 < V_1 \leqslant u_1', u_2 < V_2 \leqslant u_2']$ 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 16. *The joint distribution of* $(Y_d, \mathbf{V}, \mathbf{Q})$ *is absolutely continuous respect to the Lebesgue measure.*

Lemma 7. *[Vine Copula] Under Assumptions 14 and 16, for $d \in \{0,1\}$ we have for each $y \in \mathcal{Y}$ and $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,*

$$F_{Y_d|Q_1}(y|q_1) = \frac{\partial}{\partial x_2} C_{Y_d,Q_1}(x_1,x_2)\Big|_{x_1=F_{Y_d}(y),x_2=F_{Q_1}(q_1)} \equiv c^I_{d,F_{Q_1}(q_1)}(F_{Y_d}(y)), \tag{36}$$

$$F_{Y_d|\mathbf{Q}}(y|\mathbf{q}) = \frac{\partial}{\partial x_2} C_{Y_d,Q_1|Q_2=q_2}(x_1,x_2)\Big|_{x_1=F_{Y_d|Q_1}(y|q_1),x_2=F_{Q_1|Q_2}(q_1|q_2)} \equiv c^{II}_{d,F_{Q_1|Q_2}}(F_{Y_d|Q_1}(y|q_1)), \tag{37}$$

$$F_{Y_d|V_2,\mathbf{Q}}(y|v_2,\mathbf{Q}) = \frac{\partial}{\partial x_2} C_{Y_d,V_2|\mathbf{Q}=\mathbf{q}}(x_1,x_2)\Big|_{x_1=F_{Y_d|\mathbf{Q}}(y|\mathbf{q}),x_2=v_2} \equiv c^{III}_{d,v_2}(F_{Y_d|\mathbf{Q}}(y|\mathbf{q})), \tag{38}$$

$$F_{Y_d|\mathbf{V},\mathbf{Q}}(y|\mathbf{v},\mathbf{Q}) = \frac{\partial}{\partial x_2} C_{Y_d,V_1|V_2=v_2,\mathbf{Q}=\mathbf{q}}(x_1,x_2)\Big|_{x_1=F_{Y_d|V_2,\mathbf{Q}}(y|v_2,\mathbf{q}),x_2=F_{V_1|V_2}(v_1|v_2)}, \tag{39}$$

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

$$\mathbb{P}[Y \leqslant y, D = 1|\mathbf{Q} = \mathbf{q}] = \Psi_{1,\mathbf{q}}(F_{Y_1}(y)), \tag{40}$$

and

$$\mathbb{P}[Y \leqslant y, D = 0|\mathbf{Q} = \mathbf{q}] = \Psi_{0,\mathbf{q}}(F_{Y_0}(y)) \tag{41}$$

where the expressions for $\Psi_{d,\mathbf{q}}$ is defined in the proof below.

Proof. Equations (36) to (39) are direct applications of Joe (1996, Property 2), with some simplifications due to the fact that Assumption 14 imposes that $\mathbf{V} \perp \mathbf{Q}$.

Next we prove Equation (40). Given $d = 1$, $y \in \mathcal{Y}$, and $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

$$\begin{aligned}\mathbb{P}[Y \leqslant y, D = 1|\mathbf{Q} = \mathbf{q}] &= \mathbb{P}[Y_1 \leqslant y, \mathbf{V} \leqslant \mathbf{q}|\mathbf{Q} = \mathbf{q}] = \mathbb{P}[Y_1 \leqslant y, V_1 \leqslant q_1|V_2 \leqslant q_2, \mathbf{Q} = \mathbf{q}]q_2 \\ &= \int_0^{q_2} \mathbb{P}[Y_1 \leqslant y, V_1 \leqslant q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}]dF_{V_2|\mathbf{Q}}(v_2|\mathbf{q}) = \int_0^{q_2} \mathbb{P}[Y_1 \leqslant y, V_1 \leqslant q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}]dv_2 \end{aligned} \tag{42}$$

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

$$\begin{aligned}&\mathbb{P}[Y_1 \leqslant y, V_1 \leqslant q_1|V_2 = v_2, \mathbf{Q} = \mathbf{q}] = C_{Y_1,V_1|V_2,\mathbf{Q}}\left(F_{Y_1|V_2,\mathbf{Q}}(y|v_2,\mathbf{q}), F_{V_1|V_2,\mathbf{Q}}(q_1|v_2,\mathbf{q}); v_2, \mathbf{q}\right) \\ &= C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2}(F_{Y_1|\mathbf{Q}}(y|\mathbf{q})), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right) = C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2} \circ c^{II}_{1,F_{Q_1|Q_2}}(F_{Y_1|Q_1}(y|q_1)), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right) \\ &= C_{Y_1,V_1|V_2,\mathbf{Q}}\left(c^{III}_{1,v_2} \circ c^{II}_{1,F_{Q_1|Q_2}} \circ c^{I}_{1,F_{Q_1}(q_1)}(F_{Y_1}(y)), F_{V_1|V_2}(q_1|v_2); v_2, \mathbf{q}\right) \equiv \widetilde{\Psi}_{1,\mathbf{q}}(F_{Y_1}(y), v_2). \end{aligned} \tag{43}$$

where the equalities hold by Equations 36, 37, and 38 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 $\mathbf{q} \in \mathcal{Q}_1 \times \mathcal{Q}_2$,

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

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

Note that 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 (42), 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

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

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". □

Given Lemma 7, the identified set for DMTR or DMTE can be constructed as in Theorems 2 and 3, and as in Corollaries 3 and 4 when a specific approximation or parametrization is imposed on copulas.

C.2. **Discrete Variables.** In this subsection, we drop Assumption 4 and show how to extend Theorem 2 to the case of discrete outcome variables. We start with discrete outcome variable, and then add discreteness to propensity score.

C.2.1. *Discrete Outcome Variables.*

Assumption 17. *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 $y \in \mathcal{Y} \equiv \{y^1, y^2, \cdots, y^J\}$. Without loss of generality, assume the set $\mathcal{Y}$ is ordered: $y^j < y^\ell$ for $1 \leqslant j < \ell \leqslant J$.*

Assumption 17 says that the marginal distribution of Y_d has finite support. Furthermore, the joint support of (Y_d, P, V) is "rectangular". Let $\mathcal{T}_d = \{F_{Y_d}(y^1), F_{Y_d}(y^2), \cdots, F_{Y_d}(y^J)\}$ be the set of values that $F_{Y_d}(y)$ can take. Similarly, defined $\mathcal{T}_d^p = \{F_{Y_d|P}(y^1|p), F_{Y_d|P}(y^2|p), \cdots, F_{Y_d|P}(y^J|p)\}$ be the set of values that $F_{Y_d|P}(\cdot|p)$ can take for each given p. Also define $\mathcal{T}_d^{Dp} = \{\mathbb{P}(Y \leqslant y^1, D = d|P = p), \cdots, \mathbb{P}(Y \leqslant y^J, D = d|P = p)\}$. 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 $C^{sub}_{Y_d,V|P=p}$ be a sub-copula that coincides with the true copula $C_{Y_d,V|P=p}$ over $\mathcal{T}_d^p \times [0, 1]$. Let $C^{sub}_{Y_d,P}$ be a sub-copula that coincides with the true copula $C_{Y_d,P}$ over $\mathcal{T}_d \times \mathcal{P}$.

Lemma 8 (Vine Copula with discrete Y)**.** *Under Assumptions 1 and 17, for each $y \in \mathcal{Y}$,*

$$F_{Y_d|P}(y|p) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,P}(x_1, x_2)\Big|_{x_1 = F_{Y_d}(y), x_2 = F_P(p)} \equiv c^{sub}_{d,F_P(p)}(F_{Y_d}(y)), \tag{44}$$

$$F_{Y_d|V,P}(y|v, p) = \frac{\partial}{\partial x_2} C^{sub}_{Y_d,V|P=p}(x_1, x_2)\Big|_{x_1 = F_{Y_d|P}(y|p), x_2 = v} \tag{45}$$

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

$$\mathbb{P}[Y \leqslant y, D = 1|P = p] = \Gamma_{1,p}(F_{Y_1}(y)) \equiv C^{sub}_{Y_1,V|P=p}\Big(c^{sub}_{1,F_P(p)}(F_{Y_1}(y)), p; p\Big), \tag{46}$$

$$\mathbb{P}[Y \leqslant y, D = 0|P = p] = \Gamma_{0,p}(F_{Y_0}(y)) \equiv c^{sub}_{0,F_P(p)}(F_{Y_0}(y)) - C^{sub}_{Y_0,V|P}\Big(c^{sub}_{0,F_P(p)}(F_{Y_0}(y)), p; p\Big). \tag{47}$$

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

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

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

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

$$\Gamma_{d,p}^{(-1)}(\mathbb{P}[Y \leqslant y, D = d|P = p]) = \Gamma_{d,p'}^{(-1)}(\mathbb{P}[Y \leqslant y, D = d|P = p']), \tag{48}$$

Finally, the identified set for copula functions is characterized by

$$\Lambda_d \quad = \quad \left\{\tilde{\theta} \in \widetilde{\Theta} : \textit{For } d \in \{0,1\}, (C_{Y_d,V|P}, C_{Y_d,P}) \in \mathcal{C}_d^c \times \mathcal{C}_d \textit{ who admits subcopulas satisfying Equation (48)}\right\}.$$

Proof. First, we show that Equations (44) and (45) hold. By the Sklar (1959)'s theorem we known that there exists a copula $C_{Y_d,P}(x_1, x_2)$ such that $\mathbb{P}(Y_d \leqslant y, P \leqslant p) = C_{Y_d,P}(F_{Y_d}(y), F_P(p))$. Note that the copula $C_{Y_d,P}$ may not be unique, but the subcopula $C_{Y_d,P}^{sub}$, which defined on $\mathcal{T}_d \times [0,1]$, is uniquely determined. Then

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

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 $C_{Y_d,P}^{sub}$ coincide with $C_{Y_d,P}$ over $\mathcal{T}_d \times [0,1]$, we must have $\frac{\partial}{\partial x_2} C_{Y_d,P}(x_1, x_2) = \frac{\partial}{\partial x_2} C_{Y_d,P}^{sub}(x_1, x_2)$ for any $x_1 \in \mathcal{T}_d$ and $x_2 \in [0,1]$. Furthermore, $\frac{\partial}{\partial x_2} C_{Y_d,P}^{sub}(x_1, x_2)$ 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 (44). Similarly, for almost all $x_1 \in [0,1]$ there exists a partial derivative $\frac{\partial}{\partial x_2} C_{Y_d,V|P}^{sub}(x_1, x_2)$ that is strictly increasing in x_1 over $\mathcal{T}_d^p$ such that the following holds

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

where the last equality holds because $F_{V|P}(v|p) = v$. This verifies Equation (45).

Now, fixing y, for $d = 1$

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

where the last equality holds by using Equation (44). As discussed earlier, over $\mathcal{T}_d$, $u \mapsto c_{1,F_P(p)}^{sub}(u)$ is strictly increasing, and over $\mathcal{T}_d^p$, $x_1 \mapsto C_{Y_1,V|P}^{sub}(x_1, x_2; p)$ 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

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

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 (48). To verify the set defined in Theorem 2 is sharp, it is sufficient to show that Equations (46) and (47) and Equation (48) are equivalent. It is straightforward to see that Equations (46) and (47) imply Equation (48), 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 (48), that is, $\Gamma^{(-1)}_{d,p}(\mathbb{P}[Y \leqslant y, D = d|P = p])$ 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 (46) and (47) we know that because $1 \in \mathcal{T}_d$ and $1 \in \mathcal{T}^p_d$, we have

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

Also,

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

This says that $\Gamma^{-1}_{1,p}(\mathbb{P}[Y \leqslant \cdot, D = 1|P = p])$, 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. □

Given Lemma 8 characterizes the identified set for copula functions Λ_d, we can derove the identified set for the DMTR as in the main text. Let $\tau_{(2)}(x_1, x_2)$ be the derivative of $C^{sub}_{Y_d,V|P=p}(x_1, x_2)$ with respect to the second argument, and let $r^{-1}_{(1)}$ be the inverse of $C^{sub}_{Y_d,V|P=p}(x_1, x_2)$ with respect to the first argument. The inverse is well-defined over $\mathcal{T}^{Dp}_d$ for any given p. Then by using Equations (45) to (47) , we have

$$F_{Y_d|V,P}(y|v, p) = \tau_{(2)}\left(r^{-1}_{(1)}(\mathbb{P}[Y \leqslant y, D = d|P = p], p), v\right) \equiv \Xi_{d,p,v}(\mathbb{P}[Y \leqslant y, D = d|P = p]).$$

C.2.2. *Discrete Propensity Score.* In this section, we consider the case in which the propensity score P is discrete. We will argue that Lemma 8 still holds but with the definition of $c^{sub}_{d,F_P(p)}$ being properly modified due to the discreteness of P.

Assumption 18. *The density $f_{V|P,Y_d}(v|p,y)$ of V given $(Y_d,P)=(y,p)$, $d=0,1$, exists and is positive for all $v\in[0,1]$, all $y\in\mathcal{Y}$, and all $p\in\mathcal{P}=\{p^1,p^2,\cdots,p^M\}$. The conditioning support of Y_d given $P=p$ does not depends on p. $\mathcal{P}$ is ordered such that $0<p^j<p^\ell<1$ for all $1\leqslant j<\ell\leqslant I$.*

Let $C_{Y_d,V|P=p}$ and $C_{Y_d,P}$ be the true copulas that generate the data. As in Appendix C.2.1, Let $C^{sub}_{Y_d,V|P=p}$ be a sub-copula that coincides with the true copula $C_{Y_d,V|P=p}$ over $\mathcal{T}_d^p\times[0,1]$. Let $C^{sub}_{Y_d,P}$ be a sub-copula that coincides with the true copula $C_{Y_d,P}$ over $\mathcal{T}_d\times\mathcal{T}_p$, where $\mathcal{T}_p$ is the set of values $\mathcal{T}_p\equiv\{F_P(p):p\in\mathcal{P}\}$. If Y is discrete, then both $\mathcal{T}_d^p$ and $\mathcal{T}_d$ are sets contains finite elements. Under Assumption 18, $\mathcal{T}_p$ is a finite set of J elements. For a given y, define $\mathcal{T}_p^{DY}=\{\mathbb{P}(Y\leqslant y,D=d|P=p^1),\cdots,\mathbb{P}(Y\leqslant y^J,D=d|P=p^M)\}$.

First, Equation (45) still holds, as the argument in Appendix C.2.1 only requires $C_{Y_d,V|P}$ to be differentiable with respect to the dimension of V, which is ensured by Assumption 18. Equation (44) will take a different form. Define p_- be the largest value in $\mathcal{P}$ that is strictly smaller than p (in p is already the smallest element, then $p_-=0$). For each given $p\in\mathcal{P}$ and $y\in\mathcal{Y}$, we have

$$F_{Y_d|P}(y|p)=\frac{\mathbb{P}(Y_d\leqslant y,P=p)}{\mathbb{P}(P=p)}=\frac{C^{sub}_{Y_d,P}(F_{Y_d}(y),F_P(p))-C^{sub}_{Y_d,P}(F_{Y_d}(y),F_P(p_-))}{C^{sub}_{Y_d,P}(1,F_P(p))-C^{sub}_{Y_d,P}(1,F_P(p_-))}\equiv c^{sub}_{d,F_P(p)}(F_{Y_d}(y)),\tag{49}$$

where we still have a mapping $c^{sub}_{d,F_P(p)}$ that maps $\mathcal{T}_d$ to $\mathcal{T}_d^p$. Again, fixing p, when Y_d is discrete, both $\mathcal{T}_d$ and $\mathcal{T}_d^p$ has J elements under Assumption 17. When Y_d is continuous, they are both continuous sets. The difference is that when P is continuous, the mapping $c^{sub}_{d,F_P(p)}$ is defined as the partial derivative of $C^{sub}_{Y_d,P}$ with respect to its second argument. When P is discrete, it takes the form in Equation (49). By property of sub-copula, $c^{sub}_{d,F_P(p)}(\cdot)$ must be strictly increasing over $\mathcal{T}_d$ and admits an inverse function. Hence, the mapping $\Gamma_{d,p}$: $\mathcal{T}_d\to\mathcal{T}_d^{Dp}$, which was defined in Equations (46) and (47) is still strictly increasing and admits an inverse $\Gamma^{(-1)}_{d,p}$. The rest of the results follows from the same argument as in Lemma 8.

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," .

BLUNDELL, R., A. GOSLING, H. ICHIMURA, AND C. MEGHIR (2007): "Changes in the distribution of male and female wages accounting for employment composition using bounds," *Econometrica*, 75(2), 323–363.

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

CARD, D. (2001): "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69(5), 1127–1160.

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.

CARNEIRO, P., M. LOKSHIN, AND N. UMAPATHI (2017): "Average and marginal returns to upper secondary schooling in Indonesia," *Journal of Applied Econometrics*, 32(1), 16–36.

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.

CHEN, Y.-C., AND H. XIE (2021): "Global Representation of the Conditional LATE Model: A Separability Result," *Oxford Bulletin of Economics and Statistics*.

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

COULIBALY, M., Y.-C. HSU, I. MOURIFIÉ, AND Y. WAN (2024): "A Sharp Test for the Judge Leniency Design," Discussion paper, National Bureau of Economic Research.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78(3), 883–931.

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.

DOU, X., S. KURIKI, G. D. LIN, AND D. RICHARDS (2021): "Dependence properties of B-spline copulas," *Sankhya A*, 83(1), 283–311.

EMBRECHTS, P., AND M. HOFERT (2013): "A note on generalized inverses," *Mathematical Methods of Operations Research*, 77(3), 423–432.

GAUTIER, E., AND S. HODERLEIN (2015): "A triangular treatment effect model with random coefficients in the selection equation," .

HAFF, I. H., K. AAS, AND A. FRIGESSI (2010): "On the simplified pair-copula construction—simply useful or too simplistic?," *Journal of Multivariate Analysis*, 101(5), 1296–1310.

HAN, S., AND E. J. VYTLACIL (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199(1), 63–73.

HECKMAN, J. J., AND B. E. HONORÉ (1990): "The empirical content of the Roy model," *Econometrica*, pp. 1121–1149.

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.

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.

KLINE, P., AND C. R. WALTERS (2019): "On Heckits, LATE, and numerical equivalence," *Econometrica*, 87(2), 677–696.

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.

MACHADO, C., A. M. SHAIKH, AND E. J. VYTLACIL (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212(2), 522–555.

MAMMEN, E., C. ROTHE, AND M. SCHIENLE (2012): "Nonparametric regression with nonparametrically generated covariates," *The Annals of Statistics*, 40(2), 1132 – 1170.

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.

MASTEN, M. A., AND A. POIRIER (2018): "Identification of treatment effects under conditional partial independence," *Econometrica*, 86(1), 317–351.

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).

MOURIFIE, I., M. HENRY, AND R. MEANGO (2020): "Sharp bounds and testability of a roy model of stem major choices," *Journal of Political Economy*, 128(8), 3220–3283.

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*.

RILSTONE, P. (1996): "Nonparametric estimation of models with generated regressors," *International Economic Review*, pp. 299–313.

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.

SARTORI, A. E. (2003): "An estimator for some binary-outcome selection models without exclusion restrictions," *Political Analysis*, 11(2), 111–138.

SKLAR, M. (1959): "Fonctions de repartition an dimensions et leurs marges," *Publ. inst. statist. univ. Paris*, 8, 229–231.

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

WILDE, J. (2000): "Identification of multiple equation probit models with endogenous dummy regressors," *Economics letters*, 69(3), 309–312.

WILLIS, R. J., AND S. ROSEN (1979): "Education and self-selection," *Journal of political Economy*, 87(5, Part 2), S7–S36.

ZHOU, X., AND Y. XIE (2019): "Marginal treatment effects from a propensity score perspective," *Journal of Political Economy*, 127(6), 3070–3084.