arXiv:2403.05850v2 [econ.EM] 14 Dec 2024

Estimating Causal Effects of Discrete and Continuous Treatments with Binary Instruments*

Victor Chernozhukov Iván Fernández-Val
Sukjin Han Kaspar Wüthrich

December 17, 2024

Abstract

We propose an instrumental variable framework for identifying and estimating causal effects of discrete and continuous treatments with binary instruments. The basis of our approach is a local copula representation of the joint distribution of the potential outcomes and unobservables determining treatment assignment. This representation allows us to introduce an identifying assumption, so-called *copula invariance*, that restricts the local dependence of the copula with respect to the treatment propensity. We show that copula invariance identifies treatment effects for the entire population and other subpopulations such as the treated. The identification results are constructive and lead to practical estimation and inference procedures based on distribution regression. An application to estimating the effect of sleep on well-being uncovers interesting patterns of heterogeneity.

JEL Numbers: C14, C21, C31.
Keywords: Quantile treatment effects, endogeneity, binary instruments, copula.

*The authors respectively represent MIT, BU, U of Bristol, and UCSD. For helpful comments, we are grateful to Yingying Dong, Dalia Ghanem, Stella Hong, Desire Kedagni, Michael Knaus, Eric Mbakop, Ulrich Mueller, Whitney Newey, Ed Vytlacil, Martin Weidner, Daniel Wilhelm, and participants in seminars at WashU, Indiana, MSU, Yale, Zhejiang, HKUST, UCSD, Northwestern, Stanford, UW, and participants at the Asian Meeting of the Econometric Society, the Bristol Econometrics Study Group, Munich Econometrics Workshop, KL Leuven Summer Event, and California Conference; and Matt Hong for able research assistance.

1 Introduction

Endogeneity and heterogeneity are key challenges in causal inference. Endogeneity arises because most treatments and policies of interest are the result of decisions made by economic agents. Unobserved heterogeneity also arises naturally as many of the agents' characteristics are unobserved to the researcher. Accounting for endogeneity and heterogeneity in treatment effects estimation is crucial to answer policy questions, such as how to allocate social resources and combating inequalities. This paper proposes a flexible instrumental variable (IV) modeling framework for identifying heterogeneous treatment effects under endogeneity, which yields practical estimation and inference procedures.

Without additional assumptions, IV strategies cannot point-identify meaningful treatment effects in the presence of heterogeneity. The literature has proposed different solutions to deal with this challenge that exhibit trade-offs between adding structure to the treatment assignment mechanism and potential outcomes. One line of research has restricted the structure and heterogeneity of the potential outcomes while allowing for flexible treatment assignment mechanisms (e.g., Chernozhukov and Hansen (2005) with binary treatments and Newey and Powell (2003) with continuous treatments). Another line of research has shown the usefulness of restricting the treatment assignment while being flexible regarding how the potential outcomes are formed (e.g., Imbens and Angrist (1994) with binary treatments and Imbens and Newey (2009) with continuous treatments).

We explore an intermediate route that imposes structure on the relationship between the treatment assignment and the potential outcomes to achieve point identification of treatment effects. The basis of this approach is a local Gaussian representation of the copula capturing the dependence between the potential outcomes and the unobservable determinants of treatment assignment. This representation is fully *nonparametric*, that is, it does not require that potential outcomes and treatment unobservables are jointly or marginally Gaussian. Indeed, the bivariate Gaussian structure always holds locally by treating the correlation parameter as an implicit function that equates the bivariate Gaussian copula with the copula of the potential outcome and selection unobservables. We use this representation to introduce an assumption that has not been previously considered for identification of treatment effects. This assumption, so-called *copula invariance* (CI), restricts the local dependence of the copula with respect to treatment propensity, and thus it restricts the form of endogeneity.

We show that, even with a binary IV, CI identifies quantile and average treatment effects (QTE and ATE) of binary and ordered discrete treatments and quantile and average structural functions (QSF and ASF) of continuous treatments for the entire population

and subpopulations such as the treated. The results for ordered discrete and continuous treatments have particular empirical relevance, as previous identification results for global treatment parameters are scarce and typically rely on rich instrument variation; see below for the review.[1] Moreover, the same CI assumption applies without modification to continuous, discrete and mixed continuous-discrete outcomes. As a byproduct, we also identify the dependence function, which captures the direction and magnitude of endogenous selection and may be of interest in applications. When covariates are available, we impose CI conditional on these covariates, allowing for an additional source of heterogeneity in our model.

The CI-based identification strategy is constructive and leads to practical semiparametric estimation procedures based on distribution regression for both discrete and continuous treatments.[2] We establish the asymptotic Gaussianity of the estimators of the potential outcome distributions, dependence function, and functionals such as the unconditional QSF and QTE. We show that bootstrap is valid for estimating the limiting laws and provide an explicit algorithm for constructing uniform confidence bands.

The proposed method adds to the IV toolkit by expanding the directions of modeling trade-offs in identifying treatment effects. We allow for richer patterns of effect heterogeneity, compared to Chernozhukov and Hansen (2005) and Newey and Powell (2003), and more heterogeneity in the treatment assignment, compared to Imbens and Angrist (1994) and Imbens and Newey (2009), while imposing more restrictions on the dependence structure (i.e., the form of endogeneity), as we detail below.

We apply our method to estimating the distributional effects of sleep on well-being. In this case, sleep time is treated as a continuous treatment. We use the data from the experimental analysis of Bessone et al. (2021b), who studied the effects of randomized interventions to increase sleep time of low-income adults in India. A simple two-stage least squares analysis suggests that sleep has moderate average effects on well-being. Using our method, we document interesting patterns of heterogeneity across the distributions of sleep time and well-being, which are overlooked by standard analyses focusing on average effects. For example, the quantile treatment effects of increasing sleep on well-being at the lower tail, which are particularly policy-relevant in this context, are substantially larger than the corresponding average effect.

[1] There is a wide range of empirical studies estimating the effects of continuous and ordered endogenous variables using IVs with limited variation. Prominent examples include work on intergenerational mobility (e.g., Black and Devereux, 2011), studies of the returns to schooling (e.g., Angrist and Imbens, 1995), analyses of the impact of air pollution (e.g., Chay and Greenstone, 2005), demand analyses with discrete IVs (e.g., Angrist et al., 2000), and studies where discrete experimental treatments are used as IVs for ordered discrete and continuous treatments (e.g., Bessone et al., 2021b).

[2] We refer to these estimators as "semiparametric" because distribution regression models involve function-valued parameters.

1.1 Related Literature

The literature on identification of heterogeneous treatment effects with endogeneity is vast. We focus the review on approaches that do not impose parametric distributional assumptions to achieve point identification.

A first strand of literature focused on imposing assumptions on the generation of the potential outcomes, such as rank invariance and rank similarity. Rank invariance imposes that the outcome equation is strictly monotonic in a scalar unobservable such that there is a one-to-one mapping between the potential outcome and unobservable, and the unobservable is the same for all potential outcomes. This assumption is very convenient for the identification analysis. It allows for identification of QTE and ATE with discrete treatments (Chernozhukov and Hansen, 2005) and identification of the ASF with continuous treatments (Newey and Powell, 2003; Blundell et al., 2007).[3] Chernozhukov and Hansen (2005)'s rank similarity is slightly weaker than rank invariance as it does not necessarily restrict the unobservable to be the same for all the potential outcomes. However, both assumptions produce the same testable restriction (Chernozhukov and Hansen, 2013). In subsequent work, Vuong and Xu (2017) showed that rank invariance and strict monotonicity are powerful enough to identify individual treatment effects (under suitable regularity conditions) in addition to the QTE and ATE. This literature remains flexible about the treatment selection process.

The proposed CI assumption and rank similarity or invariance are non-nested. The former concerns the dependence between potential outcomes and selection unobservables, whereas the latter concerns the dependence between potential outcomes. Indeed, we show that rank similarity can be viewed as another form of copula invariance assumption. However, our CI allows for more general patterns of treatment effect heterogeneity than rank invariance and similarity. Also, approaches based on rank invariance and similarity rely on strict monotonicity on the outcome equation to achieve point identification. This assumption can only hold for continuous outcomes. Moreover, these approaches rely on completeness conditions on the relationship between the treatment and instrument that rule out, for example, an ordered discrete and continuous treatment when the instrument is binary. Furthermore, when treatments are continuous, completeness conditions typically lead to ill-posedness, which complicates estimation. CI does not rely on monotonicity nor completeness and therefore can accommodate discrete and mixed discrete-continuous outcomes and ordered and continuous treatments with a binary instrument.

A second strand of literature focuses on assumptions imposed on the treatment assignment. Imbens and Angrist (1994) and Heckman and Vytlacil (2005) assumed that treatment

[3] Ai and Chen (2003), Chen and Pouzo (2009, 2012, 2015) consider a general framework that nests these models.

assignment is determined by a scalar unobservable and combined this assumption with monotonicity of potential treatments with respect to a binary instrument to show identification of local average and marginal effects of binary treatments. Abadie et al. (2002) and Carneiro and Lee (2009) extended this approach to the corresponding local quantile effects, and Vytlacil (2006) extended it to the case of ordered discrete treatments. Newey et al. (1999) and Imbens and Newey (2009) imposed strict monotonicity of the treatment selection equation with respect to a scalar unobservable and large support of the instrument (ruling out discrete instruments) to identify global effects using a control variable approach. Due to the nature of the restrictions, their approach only applies to continuous treatments. Newey and Stouli (2021) avoided the large support requirement on the instrument of Imbens and Newey (2009) by assuming a parametric structure on the expectation of the treatment conditional on the instrument that enables extrapolation outside the instrument support. Compared to this strand of the literature, CI restricts the relationship between potential outcomes and treatment assignment, but allows for identifying global treatment effects of discrete and continuous treatments with discrete instruments without relying on a scalar unobservable in the treatment assignment mechanism.

There are also approaches that combine or modify the assumptions of the previous strands. Chesher (2003) showed identification of the quantile effect of continuous treatments on continuous outcomes with continuous instruments assuming strict monotonicity of the outcome and treatment selection equations with respect to scalar unobservables. Under similar assumptions, D'Haultfoeuille and Février (2015) and Torgovitsky (2015) found that quantile effects can be identified with discrete instruments. In this class of models, Torgovitsky (2017) proposes a two-step minimum distance estimator for the finite dimensional parameter of the outcome function. Again, these restrictions are different and not nested with CI. Moreover, in the case of continuous treatments, our identification strategy is constructive, yielding straightforward plug-in estimators. There are also partial identification solutions that impose less structure on the treatment assignment and potential outcomes (e.g., Manski (1990) and Balke and Pearl (1997) for earlier references, and Chesher and Rosen (2020) for a more recent survey).

The copula is a powerful tool that has been previously employed in econometrics for identification and estimation. For example, Chen et al. (2006) used a parametric copula to achieve efficient estimation in a class of multivariate distributions.[4] In a semiparametric triangular model with binary dependent variables, Han and Vytlacil (2017) introduced a class of single-parameter copulas to model the dependence structure between the unobservables

[4]In time series, Chen and Fan (2006b,a), Beare (2010), Chen et al. (2021, 2022), and Fan et al. (2023), among others, use copulas to model temporal dependence.

and established a condition on the copula under which the parameters are identified. They showed that many well-known copulas including the Gaussian copula satisfy the condition. When we restrict our attention to a binary treatment and binary outcome, the current paper's framework is relevant to Han and Vytlacil (2017). However, while they assumed a parametric copula for the dependence structure, we assume CI. Assuming a Gaussian copula in Han and Vytlacil (2017) can be viewed as an extreme special case of CI. Han and Lee (2019) developed sieve estimation and inference methods based on Han and Vytlacil (2017), and Han and Lee (2023) extended them to semiparametric models for dynamic treatment effects. Mourifié and Wan (2021) use copula as a channel to impose assumptions in characterizing identified sets in the framework of marginal treatment effects.

Arellano and Bonhomme (2017) and Chernozhukov et al. (2020a) studied IV identification of selection models using assumptions on the copula between the latent outcome and selection unobservable. Arellano and Bonhomme (2017) assumed a real analytical copula and required continuous instruments. Chernozhukov et al. (2020a) used the local Gaussian representation and copula exclusion, which is a special case of CI, with a binary instrument. Therefore, our framework with binary treatment is related to their setup. However, even in the case of binary treatment, the current setting differs from Chernozhukov et al. (2020a) in several dimensions. First, our setting requires two-way sample selection due to the switching of treatment status. Second, we introduce a general selection model that does not follow the typical threshold-crossing structure, which is important to allow for rich selection patterns. Third, because of these features, the identification analysis involves local representation and copula invariance that are specific to treatment status and the value of the IV. More importantly, the use of local Gaussian representation and CI for ordered and continuous treatments is completely new to this paper. Moreover, while relying on the same CI assumption, the identification strategies and proof techniques in these two cases are distinct.

Finally, in the difference-in-differences setup, Athey and Imbens (2006) showed that average and quantile treatment effects on the treated (and the untreated) can be identified when the unobservable determinant of the untreated potential outcome is independent of time within groups. Ghanem et al. (2023) provide general identification results under a time invariance assumption on the copula between the potential outcomes and the group indicator and show that their assumption is equivalent to the assumptions in Athey and Imbens (2006) with continuous outcomes. Their time invariance in the copula can be viewed as a version of CI where the IV is a time indicator.

1.2 Organization of the Paper

Section 2.1 introduces the key variables and parameters of interest, Section 2.2 states the local Gaussian representation, and Section 2.3 posits the main identifying assumptions that will be used throughout the analyses. We devote Sections 3.1–3.3 to the identification analyses with binary, ordered discrete, and continuous treatments, respectively. Section 4 discusses copula invariance in further detail and Section 5 discusses estimation and inference. Section 6 provides an empirical illustration. Section 7 presents concluding remarks. The Appendix contains the proofs and additional results.

1.3 Notation

For scalar random variables X and Y and possibly multivariate random variable Z, $F_{X,Y|Z}$ denotes the joint distribution of X and Y conditional on Z, $F_{X|Z}$ denotes the (marginal) distribution of X conditional on Z, and F_Z denotes the marginal (joint) distribution of Z. We use calligraphic letters to denote support sets of random variables. For example, $\mathcal{Z}$ denotes the support of Z. The symbol $\perp\!\!\!\perp$ denotes (stochastic) independence; for example, $X \perp\!\!\!\perp Y$ means that X is independent of Y. The interior of the set $\mathcal{D}$ is denoted as $\text{int}(\mathcal{D})$.

2 Setup and Assumptions

We consider three classes of models depending on the type of treatment variable: binary, discrete ordered, and continuous. Before investigating identification, we introduce the setup, parameters of interest and identifying assumptions that are common to all classes of models.

2.1 Preliminaries

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the scalar outcome and $D \in \mathcal{D} \subseteq \mathbb{R}$ denote the scalar treatment. We consider binary, discrete ordered, and continuous treatments with $\mathcal{D} = \{0, 1\}$, $\mathcal{D} = \{1, ..., K\}$, and $\mathcal{D}$ equal to an uncountable set, respectively. The outcome is not restricted, it can be continuous, discrete or mixed continuous-discrete. Let $Z \in \{0, 1\}$ be the binary IV. We focus on a binary instrument as the most challenging case; the analysis readily extends to multi-valued discrete or continuous Z. Let Y_d denote the potential outcome given $d \in \mathcal{D}$ and D_z the potential treatment given $z \in \{0, 1\}$. They are related to the observed outcome and treatment through $Y = Y_D$ and $D = D_Z$.[5] Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, for some positive integer d_x, be a vector of covariates. All the identification analysis in Section 3 is conditional on X, but

[5]When D is continuous, we require that Y_d is suitably measurable (Hirano and Imbens, 2004).

we keep the dependence implicit to lighten the notation. We make it explicit in Appendix D and when we discuss estimation and inference in Section 5.

We consider a general treatment assignment equation:

$$D_z = h(z, V_z), \quad z \in \{0, 1\}, \tag{2.1}$$

where $v \mapsto h(z, v)$ is weakly increasing, and we normalize $V_z \sim U[0, 1]$. We provide examples of the function h for each type of treatment below. By allowing for a different unobservable V_z at each value of z, we essentially permit D to be a function of the *vector* of unobservables (V_0, V_1). Even this general version of a treatment assignment model may not be necessary for our analyses (see Remark 3.4) but simplifies the exposition.

We are interested in identifying the distribution of Y_d, F_{Y_d}, for $d \in \mathcal{D}$, and functionals of F_{Y_d}, such as QSFs and ASFs. Thus, by using appropriate operators:

$$QSF_\tau(d) \equiv Q_{Y_d}(\tau) = \mathcal{Q}_\tau(F_{Y_d}), \quad ASF(d) \equiv E[Y_d] = \mathcal{E}(F_{Y_d}),$$

where $\mathcal{Q}_\tau(F) \equiv \inf\{y \in \mathcal{Y} : F(y) \geq \tau\}$ and $\mathcal{E}(F) \equiv \int_{\mathcal{Y}}[1 - F(y)]dy$. QTE and ATE can be expressed as $\{QSF_\tau(d) - QSF_\tau(d')\}/(d - d')$ and $\{ASF(d) - ASF(d')\}/(d - d')$, $d, d' \in \mathcal{D}$ with $d' \neq d$, and for continuous treatments we can also consider $\partial QSF_\tau(d)/\partial d$ and $\partial ASF_\tau(d)/\partial d$, $d \in \mathcal{D}$. When the treatment is binary, we may also be interested in the distribution of Y_d in subpopulations such as the treated, $F_{Y_d|D}(\cdot \mid 1)$, and untreated, $F_{Y_d|D}(\cdot \mid 0)$, for $d \in \{0, 1\}$, and functionals of these distributions. More generally, we may be interested in these objects for subpopulations defined by values of the covariates in X.

2.2 Local Gaussian Representation

Treatment endogeneity can be captured by the joint distribution of the potential outcome and unobservable of the treatment assignment equation (2.1). We use a conditional version of the local Gaussian representation (LGR) to represent such a joint distribution. This representation is the basis of our identification and estimation strategies. Throughout the paper, let $C(u_1, u_2; \rho)$ denote the Gaussian copula with correlation coefficient ρ, that is

$$C(u_1, u_2; \rho) \equiv \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho),$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the standard bivariate Gaussian distribution with parameter ρ and Φ is the standard univariate Gaussian distribution.

The following lemma shows that the conditional distribution of any bivariate random

variable has a local Gaussian representation (Anjos and Kolev, 2005; Kolev et al., 2006; Chernozhukov et al., 2020a).

Lemma 2.1 (LGR). *For any random variables Y, V and Z, the joint distribution of Y and V conditional on Z admits the representation:*

$$F_{Y,V|Z}(y,v \mid z) = C(F_{Y|Z}(y \mid z), F_{V|Z}(v \mid z); \rho_{Y,V;Z}(y,v;z)), \text{ for all } (y,v,z),$$

where $\rho_{Y,V;Z}(y,v;z)$ is the unique solution in ρ to

$$F_{Y,V|Z}(y,v \mid z) = C(F_{Y|Z}(y \mid z), F_{V|Z}(v \mid z); \rho).$$

In the lemma, the LGR is fully nonparametric and is not imposing Gaussianity on the joint or marginal distribution. The distribution is equated to the Gaussian copula (with the nonparametric marginal distributions) by adjusting the value of the correlation coefficient ρ for each evaluation point (y,v,z). Note that the solution $\rho_{Y,V;Z}(y,v;z)$ depends on both the dependence structure and marginals. Lemma 2.1 can be equivalently stated as the LGR of a copula instead of a distribution; see Section B.1.

2.3 Assumptions

We maintain the following assumptions:

Assumption EX (Independence). *For $d \in \mathcal{D}$ and $z \in \{0,1\}$, $Z \perp\!\!\!\perp Y_d$ and $Z \perp\!\!\!\perp V_z$.*

Assumption REL (Relevance). *(i) $Z \in \{0,1\}$; (ii) $0 < \Pr[Z=1] < 1$; and (iii) for $\mathcal{D} = \{0,1\}$, $\Pr[D=1 \mid Z=1] \neq \Pr[D=1 \mid Z=0]$ and $0 < \Pr[D=1 \mid Z=z] < 1$, $z \in \{0,1\}$; for $\mathcal{D} = \{1,...,K\}$, $F_{D|Z}(d \mid 1) \neq F_{D|Z}(d \mid 0)$, $d \in \mathcal{D} \setminus \{K\}$, and $\Pr[D=d \mid Z=z] > 0$, $(z,d) \in \{0,1\} \times \mathcal{D}$; and for uncountable $\mathcal{D}$, $F_{D|Z}(d \mid 1) \neq F_{D|Z}(d \mid 0)$ and $0 < F_{D|Z}(d \mid z) < 1$, $(z,d) \in \{0,1\} \times \text{int}(\mathcal{D})$.*

EX is a standard exogeneity condition in IV strategies. It is weaker than $Z \perp\!\!\!\perp (\{Y_d\}_{d\in\mathcal{D}}, V_0, V_1)$ or $Z \perp\!\!\!\perp (Y_d, V_z)$ for $(d,z) \in \mathcal{D} \times \{0,1\}$. Also, in EX, a standard exclusion restriction is implicit in the notation: $Y_d = Y_{d,z}$ almost surely, where $Y_{d,z}$ is the potential outcome given (d,z). REL(ii)–(iii) are the usual IV relevance and non-degeneracy conditions. REL(iii) for $\mathcal{D} = \{0,1\}$ can be formulated a special case of REL(iii) for $\mathcal{D} = \{1,...,K\}$ with $K = 2$, but we state it separately for clarity.[6]

[6]For $K > 2$, a weaker but less interpretable condition for REL(iii) is $\Pr[D=d \mid Z=z] > 0$ for $(z,d) \in \{0,1\} \times \mathcal{D}$, $F_{D|Z}(1 \mid 1) \neq F_{D|Z}(1 \mid 0)$, and either $F_{D|Z}(d \mid 1) \neq F_{D|Z}(d \mid 0)$ or $F_{D|Z}(d-1 \mid 1) \neq F_{D|Z}(d-1 \mid 0)$ for $d \in \{2,...,K\}$. Note that when $d = K$ the last condition requires that $F_{D|Z}(K-1 \mid 1) \neq F_{D|Z}(K-1 \mid 0)$ because $F_{D|Z}(K \mid 1) = F_{D|Z}(K \mid 0) = 1$.

We make the following assumption about the local dependence parameter of the LGR of (Y_d, V_z) conditional on Z:

Assumption CI (Copula Invariance)**.** *For $d \in \mathcal{D}$, $\rho_{Y_d,V_z;Z}(y,v;z)$ is a constant function of (v,z), that is*

$$\rho_{Y_d,V_z;Z}(y,v;z) = \rho_{Y_d}(y), \quad (y,v,z) \in \mathcal{Y} \times \mathcal{V} \times \{0,1\},$$

and $\rho_{Y_d}(y) \in (-1,1)$.

CI is a high-level condition that imposes a shape restriction on the dependence between Y_d and V_z. This condition (together with the other assumptions we maintain) is sufficient for identification in all the cases that we consider, but it is not necessary. We provide weaker conditions for each case in the following section. Section 4 provides more interpretable conditions for CI and compares CI with alternative identifying assumptions that have been used in the literature, such as rank invariance and rank similarity.

Figure 1 shows examples of joint distributions of (Y_d, V_z) that satisfy CI. Panel (a) corresponds to a bivariate Gaussian distribution (i.e., a Gaussian copula and univariate Gaussian marginals); Panel (b) corresponds to a non-Gaussian copula (i.e., a Gaussian copula with varying correlation) and standard univariate Gaussian marginals; and Panel (c) corresponds to a Gaussian copula coupled with non-Gaussian marginals. The upper figures display contour plots of the joint distribution and the lower figures display the corresponding local dependence functions. These examples showcase that CI is compatible with very diverse shapes for the joint distribution including multimodality and asymmetry. Combinations of these features and more complex shapes are possible by mixing and expanding these examples (e.g., a non-Gaussian copula with non-Gaussian marginals).[7] Moreover, even if Gaussian copula does not have tail dependence, CI allows for that because the local dependence parameter can change with the level of Y_d.

3 Identification Analysis

3.1 Binary Treatment

We start by considering the identification of the causal effects of a binary treatment $D \in \mathcal{D} = \{0,1\}$. To reflect this, we consider a treatment selection equation

$$D_z = h(z, V_z) = 1\{V_z \leq \pi(z)\}, \quad z \in \{0,1\}, \tag{3.1}$$

[7] We refer to Liu et al. (2009) and Chernozhukov et al. (2020a) for more examples.

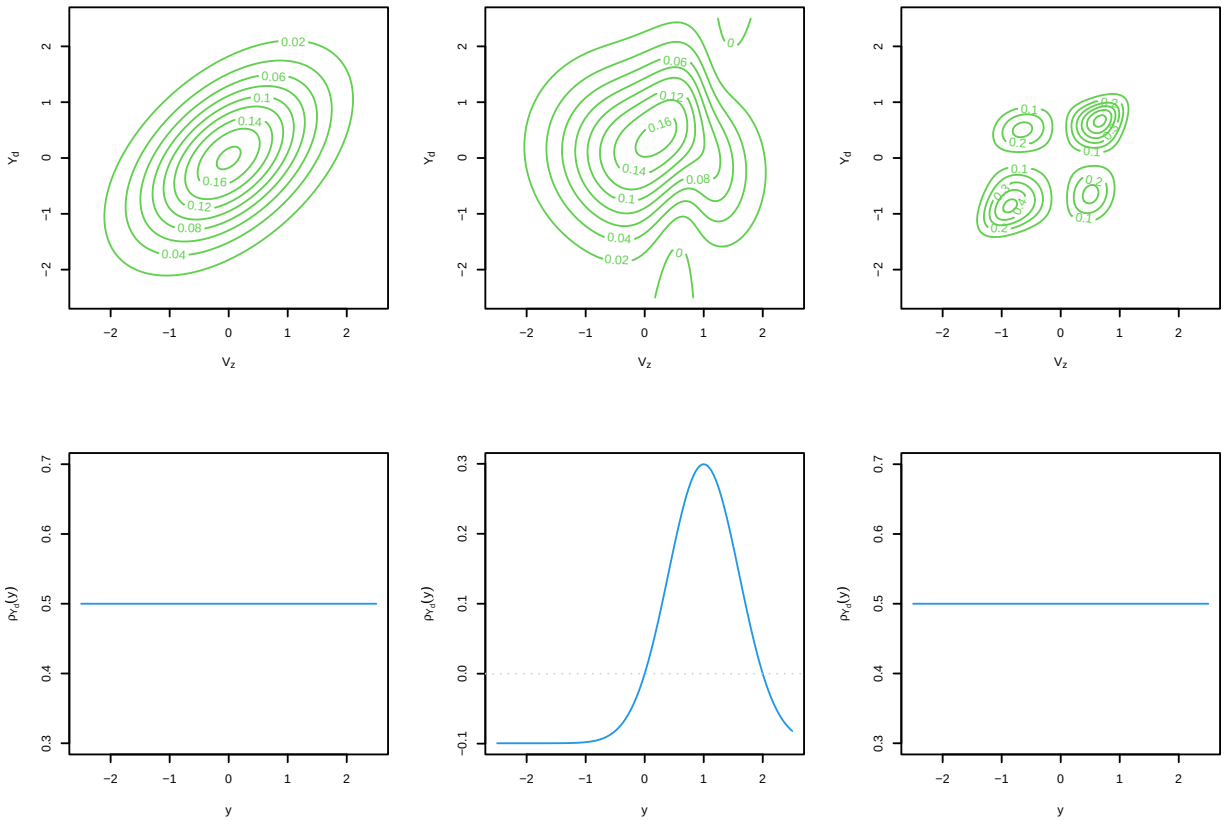


Figure 1: Examples of Distributions that Satisfy Copula Invariance

Notes: Panel (a): Bivariate Gaussian distribution (i.e., Gaussian copula and univariate Gaussian marginals) with correlation 0.5; Panel (b): Gaussian copula with local correlation $\rho_{Y_d}(y) = \phi(5(y-1)/3) - \phi(5/3)$ and univariate Gaussian marginals; Panel (c): Gaussian copula with correlation 0.5 and non-Gaussian marginals $\Phi(a(\text{sign}(u) + 1(u > 0)/2) + b)$ for both Y_d and V_z, where a and b are calibrated such that the distribution has zero mean and unit variance. The upper figures display contour plots of the joint distribution and the lower figures display the corresponding local dependence functions.

with propensity score

$$\Pr[D = 1 \mid Z = z] = \Pr[D_z = 1 \mid Z = z] = \Pr[V_z \leq \pi(z)] = \pi(z),$$

by EX and the normalization $V_z \sim U[0, 1]$. Note that V_1 and V_0 are two distinct unobservables that do not restrict the behavior of D_0 and D_1. The LATE monotonicity assumption of Imbens and Angrist (1994) imposes either $D_1 \geq D_0$ or $D_0 \geq D_1$, almost surely, which corresponds to $V_1 = V_0$, almost surely (Vytlacil, 2002).

For the identification analysis, consider

$$\begin{aligned}\Pr[Y \leq y, D = 1 \mid Z = z] &= \Pr[Y_1 \leq y, D_z = 1 \mid Z = z] \\ &= C(F_{Y_1|Z}(y|z), \pi(z); \rho_{Y_1,V_z;Z}(y, \pi(z); z)) \\ &= C(F_{Y_1}(y), \pi(z); \rho_{Y_1,V_z;Z}(y, \pi(z); z)), \quad (y, z) \in \mathcal{Y} \times \{0, 1\}, \quad (3.2)\end{aligned}$$

where the second equality uses equation (3.1) and Lemma 2.1, and the last equality follows from EX. For each $y \in \mathcal{Y}$, this is a system of two equations in three unknowns, $F_{Y_1}(y)$, $\rho_{Y_1,V_z;Z}(y, \pi(0); 0)$ and $\rho_{Y_1,V_z;Z}(y, \pi(1); 1)$. The number of unknowns can be reduced to two by the condition

$$\rho_{Y_1,V_1;Z}(y, \pi(1); 1) = \rho_{Y_1,V_0;Z}(y, \pi(0); 0) \equiv \rho_{Y_1}(y), \quad y \in \mathcal{Y}, \tag{3.3}$$

which is implied by CI. Equation (3.3) only requires that copula invariance holds at two points, $(\pi(1), 1)$ and $(\pi(0), 0)$. The following theorem shows that the nonlinear system of equations (3.2) has a unique solution under (3.3). This result follows from a global univalence theorem of Gale and Nikaido (1965), because the Jacobian of the system of equations is a P-matrix under REL(iii), the map defined by the system is differentiable (as the copula is differentiable) and the parameter space $(0, 1) \times (-1, 1)$ for $(F_{Y_1}(y), \rho_{Y_1}(y))$ is open and rectangular. A similar argument shows that the distribution of Y_0 and the local dependence parameter of the LGR of Y_0 and V_z are identified.

Theorem 3.1 (Identification for Binary Treatment)**.** *Suppose $D_z \in \{0, 1\}$ satisfies (3.1) for $z \in \{0, 1\}$. Under EX, REL, and CI, the functions $y \mapsto F_{Y_d}(y)$ and $y \mapsto \rho_{Y_d}(y)$ are identified on $y \in \mathcal{Y}$, for $d \in \{0, 1\}$.*

The proof of Theorem 3.1 is contained in Appendix G. It is interesting to compare Theorem 3.1 to Chernozhukov and Hansen (2005) who also establish identification of F_{Y_d} when D and Z are binary. Compared to them, we do not impose rank similarity and thus, as we show in Appendix A.1, allow for richer effect heterogeneity. As a trade-off, we restrict the form of endogeneity by imposing CI. For more detailed comparison, see Section 4.2 and Appendix A.1.

Remark 3.1 (Identification of $F_{Y_1|D}$ and $F_{Y_0|D}$)**.** *We focus on identification for the treated, $D = 1$; identification for the untreated, $D = 0$, follows by a similar argument. The distribution of Y_1 is trivially identified from $F_{Y_1|D}(y \mid 1) = F_{Y|D}(y \mid 1)$. Identification of the distribution of Y_0 follows from*

$$F_{Y_0|D}(y \mid 1) = \frac{F_{Y_0}(y) - (1 - \pi)F_{Y|D}(y \mid 0)}{\pi},$$

where $\pi \equiv \Pr[D=1]$ *and* $F_{Y_0}(y)$ *is identified by Theorem 3.1.*

Remark 3.2 (One-sided non-compliance)**.** *Under one-sided non-compliance or random intention to treat,* $D=0$ *whenever* $Z=0$ *(i.e.,* $\Pr[D=0 \mid Z=0]=1$*), REL(iii) is violated. In this case* $F_{Y_1}(y)$ *is no longer identified because one of the equations in (3.2) becomes uninformative as* $\pi(0)=0$*. We can still identify* $F_{Y_d|D}(y \mid 1)$ *for* $d=0,1$ *using the same analysis of Remark 3.1, because* $F_{Y_0}(y)=F_{Y|Z}(y \mid 0)$*.*

Remark 3.3 (Overidentification with nonbinary instruments)**.** *It is clear from the identification analysis that, when the IV takes more than two values (or equivalently, when there exist multiple binary IVs), the resulting system of equations produces overidentifying restrictions. These restrictions can be used to test the model specification, and CI in particular. Multi-valued IVs can also be used to make the model more flexible by allowing the local dependence parameter to partially vary with the IV as CI only needs to be satisfied for two values (see Appendix C). Although we shall not repeat it, this remark also applies to the subsequent cases of discrete ordered and continuous* D*.*

Remark 3.4 (Alternative selection equation)**.** *The treatment selection equation (3.1) is not necessary for our analysis. We can alternatively consider* $D_z = 1\{D_z^* \leq 0\}$*, where* D_1^* *and* D_0^* *are two distinct random variables. This model nests (3.1) as a special case with* $V_z \equiv D_z^* + \pi(z)$*. The alternative LGR using* D_z^* *still requires a similar version of CI because*

$$F_{Y|D,Z}(y \mid D=1, Z=z)\pi(z) = C(F_{Y_1}(y), F_{D_z^*|Z}(0 \mid z); \rho_{Y_1}(y)),$$

by EX and the CI $\rho_{Y_1,D_1^*;Z}(y,0;1) = \rho_{Y_1,D_0^*;Z}(y,0;0)$*. A similar discussion applies to the subsequent cases of discrete ordered and continuous* D*.*

3.2 Ordered Discrete Treatment

We consider identification of the causal effect of a multi-valued ordered treatment $D \in \mathcal{D} = \{1,\ldots,K\}$ using a binary instrument $Z \in \{0,1\}$. We assume a threshold-crossing model for the treatment selection equation, which can be viewed as a natural extension of model (3.1) from two to multiple treatment levels,

$$D_z = h(z,V_z) = \sum_{d\in\mathcal{D}} d\ 1\{\pi_{d-1}(z) < V_z \leq \pi_d(z)\} = \begin{cases} 1, & \pi_0(z) < V_z \leq \pi_1(z) \\ 2, & \pi_1(z) < V_z \leq \pi_2(z) \\ \vdots & \vdots \\ K, & \pi_{K-1}(z) < V_z \leq \pi_K(z) \end{cases}, \quad (3.4)$$

where $\pi_0(z) = 0$ and $\pi_K(z) = 1$. Equation (3.4) generalizes the model in Section 7.2 of Heckman and Vytlacil (2007) by allowing for two unobservables (V_0, V_1) and a different impact of the instrument on the different cutoffs; see Remark 3.5. It is not fully general, however, as it imposes that the unobservable is the same for all the treatment levels.[8]

Under the normalization $V_z \sim U[0,1]$ and EX, the threshold functions $\pi_d(z)$ are identified by the distribution of the observed treatment conditional on the instrument, $\pi_d(z) = F_{D|Z}(d \mid z)$ for $d \in \mathcal{D}$. For the identification analysis, consider

$$\begin{aligned}\Pr[Y \leq y, D = d \mid Z = z] &= \Pr[Y_d \leq y, \pi_{d-1}(z) < V_z \leq \pi_d(z) \mid Z = z] \\ &= C(F_{Y_d}(y), \pi_d(z); \rho_{Y_d,V_z;Z}(y, \pi_d(z); z)) \\ &\quad - C(F_{Y_d}(y), \pi_{d-1}(z); \rho_{Y_d,V_z;Z}(y, \pi_{d-1}(z); z)), \quad (y,d,z) \in \mathcal{Y} \times \mathcal{D} \times \{0,1\}, \quad (3.5)\end{aligned}$$

where the first equality follows from (3.4) and the second equality from EX and Lemma 2.1. For each $d \in \mathcal{D}$ and $y \in \mathcal{Y}$, (3.5) is a system of two equations in five unknowns: $F_{Y_d}(y)$, $\rho_{Y_d,V_0;Z}(y, \pi_{d-1}(0); 0)$, $\rho_{Y_d,V_0;Z}(y, \pi_d(0); 0)$, $\rho_{Y_d,V_1;Z}(y, \pi_{d-1}(1); 1)$, and $\rho_{Y_d,V_1;Z}(y, \pi_d(1); 1)$. REL(iii) guarantees that the two equations of the system are not redundant.

For $d \in \{1, K\}$, one of the terms on the right hand side drops out because either $\pi_{d-1}(z) = 0$ or $\pi_d(z) = 1$, yielding a system of two equations on three unknowns. Consequently, the distribution of the potential outcome and local dependence parameter can be identified by combining REL(iii) with the condition

$$\rho_{Y_d,V_1;Z}(y, \pi_{d'}(1); 1) = \rho_{Y_d,V_0;Z}(y, \pi_{d'}(0); 0), \quad y \in \mathcal{Y}, \quad d \in \mathcal{D}, \quad d' \in \{d-1, d\}, \qquad (3.6)$$

which is analogous to condition (3.3) from the binary treatment case.

For $d \in \mathcal{D} \backslash \{1, K\}$, condition (3.6) reduces the number of unknowns to three but is not sufficient to identify the unknowns. We impose additionally copula invariance between consecutive treatment levels

$$\rho_{Y_d,V_z;Z}(y, \pi_d(z); z) = \rho_{Y_d,V_z;Z}(y, \pi_{d-1}(z); z) \equiv \rho_{Y_d}(y), \quad (y,d,z) \in \mathcal{Y} \times \mathcal{D} \times \{0,1\}.$$

This condition is also implied by CI and reduces the number of unknowns to two: $F_{Y_d}(y)$ and $\rho_{Y_d}(y)$. The Jacobian of the resulting system of equations, however, does not satisfy the conditions to apply the global univalence results of Gale and Nikaido (1965) even under REL(iii). We show uniqueness of solution using an alternative global univalence result of Ambrosetti and Prodi (1995). To apply this result we impose the following sufficient

[8] Cunha et al. (2007) showed that ordered choice models with multivariate unobservables are point identified under strong assumptions.

condition on the distribution of the treatment conditional on the instrument:

Assumption U_{OC} (Uniformity in Ordered Choice)**.** *Either* $F_{D|Z}(d \mid 0) > F_{D|Z}(d \mid 1)$ *for all* $d \in \mathcal{D}\backslash\{K\}$ *or* $F_{D|Z}(d \mid 0) < F_{D|Z}(d \mid 1)$ *for all* $d \in \mathcal{D}\backslash\{K\}$.

U_{OC} does not necessarily follow from REL(iii) and imposes stochastic dominance between $F_{D|Z}(d \mid 0)$ and $F_{D|Z}(d \mid 1)$. For example, the ordered choice model considered by Heckman and Vytlacil (2007) satisfies U_{OC}. Like REL(iii), U_{OC} can be directly tested from the data. It is interesting to see what type of compliance behavior with respect to D_0 and D_1 is ruled out by this sufficient condition. To explore this, define the compliers and defiers of order $j \in \mathcal{D}\backslash\{K\}$ as

$$C_j \equiv \bigcup_{d=1}^{K-j} \{D_0 = d, D_1 = d + j\}, \quad B_j \equiv \bigcup_{d=1}^{K-j} \{D_1 = d, D_0 = d + j\}.$$

Assumption EG (Exchangeability)**.** V_0 *and* V_1 *are exchangeable, i.e.,* $C(v_0, v_1) = C(v_1, v_0)$.

EG states that the distribution for (V_0, V_1) is symmetric; most known copulas are symmetric. It holds trivially if $V_0 = V_1$, almost surely. Under EG, we can interpret Assumption U_{OC} in terms of compliance behavior:

Lemma 3.1 (Compliance Shares)**.** *Under EG,* $F_{D|Z}(d \mid 0) > F_{D|Z}(d \mid 1)$ *(resp.* $<$*) for all* $d \in \mathcal{D}\backslash\{K\}$ *implies that the share of all complier groups is larger (resp. smaller) than the share of all defier groups, that is,* $\Pr[\bigcup_{j=1}^{K-1} C_j] > \Pr[\bigcup_{j=1}^{K-1} B_j]$ *(resp.* $<$*).*

The condition about the share of compliers and defiers is reminiscent of a similar assumption used in De Chaisemartin (2017) in the case of binary treatment. Another simple interpretation of Lemma 3.1 can be made under the restriction $V_0 = V_1$, almost surely. In this special case, we can easily see that there is no defier groups (i.e., $\Pr[D_1 < D_0] = 0$) if and only if $F_{D|Z}(d \mid 0) > F_{D|Z}(d \mid 1)$ for all $k \in \mathcal{D}\backslash\{K\}$. In general, when V_z is not restricted, Assumption EG alone does not eliminate compliers or defiers.

We summarize the identification result in the following theorem:

Theorem 3.2 (Identification for Ordered Treatment)**.** *Suppose* D_z, $z \in \{0, 1\}$, *satisfies* (3.4)*. Under EX, REL, CI, and* U_{OC}*, the functions* $y \mapsto F_{Y_d}(y)$ *and* $y \mapsto \rho_{Y_d}(y)$ *are identified on* $y \in \mathcal{Y}$*, for* $d \in \mathcal{D}$.

In the proof of Theorem 3.2, contained in Appendix G, we proceed as follows to show that (3.5) has a unique solution. First, we show that the function that defines the system is proper (due to properties of copula) and its Jacobian has full-rank (by U_{OC}). Then, we

apply Corollary 1.4 of Ambrosetti and Prodi (1995) to show that the system has a set of solutions whose cardinality is invariant over the parameter space. Since the system has a unique solution when $\rho_{Y_d}(y) = 0$ (locally no endogeneity), we can then conclude the solution is unique everywhere in the parameter space. A similar proof strategy was previously used in De Paula et al. (2019) in the different setting of panel models with peer effects.

Remark 3.5 (Comparison with Heckman and Vytlacil (2007)). *Heckman and Vytlacil (2007, Section 7.2) consider an ordered choice model, where the instrument is restricted to shift all cutoffs by the same amount. Suppose that*

$$D_z = \begin{cases} 1, & -\infty < \mu(z) + \tilde{V} \leq \pi_1 \\ 2, & \pi_1 < \mu(z) + \tilde{V} \leq \pi_2 \\ \vdots & \vdots \\ K, & \pi_{K-1} < \mu(z) + \tilde{V} < \infty \end{cases}. \tag{3.7}$$

where $\tilde{V} \mid Z \sim N(0,1)$. *This model is a special case of the model we consider in this section if we impose* $\pi_d(z) = \pi_d - \mu(z)$, $d \in \{1, \ldots, K-1\}$, $V_0 = V_1$, *almost surely, and normalize* $V_z \sim N(0,1)$.

3.3 Continuous Treatment

Suppose $D \in \mathcal{D} \subseteq \mathbb{R}$ is an uncountable set and $d \mapsto F_{D|Z}(d \mid z)$ is strictly increasing on $\mathcal{D}$, for $z \in \{0,1\}$. Assume the treatment selection equation,

$$D_z = h(z, V_z) = F^{-1}_{D|Z}(V_z \mid z), \quad z \in \{0,1\}, \tag{3.8}$$

where $V_z \sim U(0,1)$.

For the identification analysis, consider

$$F_{Y|D,Z}(y \mid d, z) = F_{Y_d|D_z,Z}(y \mid d, z) = F_{Y_d|V_z,Z}(y \mid F_{D|Z}(d \mid z), z), \tag{3.9}$$

where the second equality holds from equation (3.8) and a change of variable. By the properties of the conditional distribution, Lemma 2.1, and EX,

$$\begin{aligned} F_{Y_d|V_z,Z}(y \mid v, z) = \frac{(\partial/\partial v) F_{Y_d,V_z|Z}(y, v \mid z)}{(\partial/\partial v) F_{V_z|Z}(v \mid z)} &= C_2(F_{Y_d}(y), v; \rho_{Y_d,V_z;Z}(y, v; z)) \\ &\quad + C_\rho(F_{Y_d}(y), v; \rho_{Y_d,V_z;Z}(y, v; z))(\partial/\partial v)\rho_{Y_d,V_z;Z}(y, v; z), \end{aligned} \tag{3.10}$$

where C_2 and C_ρ are the derivatives of $C(\cdot, v; \rho)$ with respect to v and ρ, respectively. Assume that

$$\rho_{Y_d,V_z;Z}(y, F_{D|Z}(d \mid 1); 1) = \rho_{Y_d,V_z;Z}(y, F_{D|Z}(d \mid 0); 0) \equiv \rho_{Y_d}(y), \quad y \in \mathcal{Y}, \tag{3.11}$$

and

$$(\partial/\partial v)\rho_{Y_d,V_z;Z}(y, F_{D|Z}(d \mid z); z) = 0, \quad z \in \{0, 1\}, \tag{3.12}$$

where the differentiability of $v \mapsto \rho_{Y_d,V_z;Z}(y, v; z)$ follows by the differentiability of $C(\cdot, v; \cdot)$ with respect to v and the implicit function theorem; see Section B.1 for related discussions. Note that (3.11) and (3.12) are implied by CI. Then, by the properties of Gaussian copula, combining (3.9) and (3.10) yields

$$\Phi^{-1}\left(F_{Y|D,Z}(y \mid d, z)\right) = a_{d,y} + b_{d,y}\Phi^{-1}(F_{D|Z}(d \mid z)), \quad z \in \{0, 1\}, \tag{3.13}$$

where $a_{d,y} \equiv \Phi^{-1}(F_{Y_d}(y))/\sqrt{1 - \rho_{Y_d}(y)^2}$ and $b_{d,y} \equiv -\rho_{Y_d}(y)/\sqrt{1 - \rho_{Y_d}(y)^2}$. Equation (3.13) yields a linear system of two equations on two unknowns, $a_{d,y}$ and $b_{d,y}$, which has solution

$$\begin{aligned} a_{d,y} &= \frac{\Phi^{-1}(F_{Y|D,Z}(y \mid d, 0))\Phi^{-1}(F_{D|Z}(d \mid 1)) - \Phi^{-1}(F_{Y|D,Z}(y \mid d, 1))\Phi^{-1}(F_{D|Z}(d \mid 0))}{\Phi^{-1}(F_{D|Z}(d \mid 1)) - \Phi^{-1}(F_{D|Z}(d \mid 0))}, \\ b_{d,y} &= \frac{\Phi^{-1}(F_{Y|D,Z}(y \mid d, 1)) - \Phi^{-1}(F_{Y|D,Z}(y \mid d, 0))}{\Phi^{-1}(F_{D|Z}(d \mid 1)) - \Phi^{-1}(F_{D|Z}(d \mid 0))}, \end{aligned} \tag{3.14}$$

under REL.

This discussion implies the following identification result.

Theorem 3.3 (Identification for Continuous Treatment)**.** *Suppose D_z, $z \in \{0, 1\}$, satisfies (3.8). Under EX, REL, and CI, the functions $y \mapsto F_{Y_d}(y)$ and $y \mapsto \rho_{Y_d}(y)$ are identified on $y \in \mathcal{Y}$, for $d \in \mathcal{D}$, by*

$$F_{Y_d}(y) = \Phi\left(\frac{a_{d,y}}{\sqrt{1 + b_{d,y}^2}}\right), \quad \rho_{Y_d}(y) = \frac{-b_{d,y}}{\sqrt{1 + b_{d,y}^2}},$$

where $a_{d,y}$ and $b_{d,y}$ are defined in (3.14).

It is interesting to compare this identification result to Imbens and Newey (2009) and Torgovitsky (2010, 2015)[9] who also provide identification results for nonseparable models

[9]Portions of Torgovitsky (2010) were published in Torgovitsky (2015). Since the role of the conditional copula invariance assumption is only discussed in Torgovitsky (2010), we focus on comparing our method and assumptions to this paper.

with continuous D. Unlike Imbens and Newey (2009), our approach does not require an instrument with large support nor rank invariance in the treatment selection equation (i.e., $V_1 = V_0$ almost surely), but instead imposes CI. Unlike Torgovitsky (2010), our approach does not require rank invariance in the outcome and treatment equation, but again imposes CI. See Appendix A for more detailed comparisons.

Remark 3.6 (Censored Treatment)**.** *Suppose D is a continuous treatment censored at zero (i.e., $D = \max\{D^*, 0\}$ where D^* is continuous). Examples include worked hours or amount of subsidy. Then, by REL(iii) for uncountable $\mathcal{D}$, one can apply the analysis of this section to identify $F_{Y_d}(y)$ for $d > 0$ from $F_{Y|D,Z}(y \mid d, z)$ for $d > 0$ and $z \in \{0, 1\}$ and the analysis of Section 3.1 to identify $F_{Y_0}(y)$ from $\Pr[Y \leq y, D = 0 \mid Z = z]$ for $z \in \{0, 1\}$. In particular, assuming the treatment selection equation*

$$D_z = h(z, V_z) = \max\{F^{-1}_{D^*|Z}(V_z \mid z), 0\},$$

where $V_z \sim U(0, 1)$, we can identify $F_{Y_0}(y)$ and $\rho_{Y_0}(y)$ from

$$\Pr(Y \leq y, D = 0 \mid Z = z) = C\left(F_{Y_0|Z}(y \mid z), F_{D|Z}(0 \mid z); \rho_{Y_0,V_z;Z}(y, F_{D|Z}(0 \mid z); z)\right),$$

following the same analysis as in Section 3.1; and we can identify $F_{Y_d}(y)$ and $\rho_{Y_d}(y)$, $d > 0$, from

$$F_{Y|D,Z}(y \mid d, z) = F_{Y_d|V_z,Z}(y \mid F_{D|Z}(d \mid z), z),$$

following the same analysis as in Section 3.3.

4 Discussions on Copula Invariance

To further understand CI, we provide sets of simple equivalent conditions (Sections 4.1), compare it with rank similarity (Section 4.2), and discuss its flexibility in the context of Roy models (Section 4.3).

4.1 Equivalent Conditions

4.1.1 Joint Independence

Here we provide equivalent conditions to EX and CI that highlight the trade-offs between copula invariance and instrument independence assumptions.

Assumption EX′ (Joint Independence)**.** *For $d \in \mathcal{D}$ and $z \in \{0, 1\}$, $Z \perp\!\!\!\perp (Y_d, V_z)$.*

Assumption CI′ (Unconditional CI)**.** *For $d \in \mathcal{D}$,*

$$\rho_{Y_d,V_z}(y,v) = \rho_{Y_d}(y), \quad (y,v,z) \in \mathcal{Y} \times \mathcal{V} \times \{0,1\}.$$

Proposition 4.1. *EX′ and CI′ are equivalent to EX and CI.*

Note that $\rho_{Y_d,V_z;Z}(\cdot,\cdot;z) = \rho_{Y_d,V_z}(\cdot,\cdot)$ by EX′ and $\rho_{Y_d,V_z}(y,\cdot) = \rho_{Y_d}(y)$ by CI′.[10] Conversely, EX and CI imply EX′. Then, EX′ together with CI imply CI′. When D is binary, a sufficient condition for EX′ is $(Y_0, Y_1, V_z) \perp\!\!\!\perp Z$, which is imposed in Imbens and Angrist (1994) and Vytlacil (2002) with $V_0 = V_1$ almost surely, although it is sufficient for the LATE result to have $(Y_d, V) \perp\!\!\!\perp Z$ for $d \in \{0,1\}$.

Remark 4.1 (CI′)**.** *We might wonder if CI′ implies rank invariance in selection, $V_0 = V_1$, almost surely. The following example shows that this is not the case. Let*

$$\begin{pmatrix} Y_d \\ V_0 \\ V_1 \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{Y_d,V_0} & \rho_{Y_d,V_1} \\ \rho_{Y_d,V_0} & 1 & \rho_{V_0,V_1} \\ \rho_{Y_d,V_1} & \rho_{V_0,V_1} & 1 \end{bmatrix} \right).$$

Under CI′, $\rho_{Y_d,V_0} = \rho_{Y_d,V_1}$, so that (Y_d, V_0) and (Y_d, V_1) have the same distribution. Moreover, the matrix

$$\begin{bmatrix} 1 & \rho_{Y_d,V_0} & \rho_{Y_d,V_0} \\ \rho_{Y_d,V_0} & 1 & \rho_{V_0,V_1} \\ \rho_{Y_d,V_0} & \rho_{V_0,V_1} & 1 \end{bmatrix}$$

can be positive definite for $|\rho_{V_0,V_1}| \neq 1$. In other words, the condition $\rho_{Y_d,V_0} = \rho_{Y_d,V_1}$ does not imply that $V_1 = V_0$. We refer to Appendix A.2 for a more detailed comparison to the LATE framework, where we relate CI to a rank similarity condition between V_0 and V_1.

4.1.2 Local Single Index

We provide an equivalent condition to CI′.

Assumption SI (Local Single Index)**.** *For $d \in \mathcal{D}$ and $z \in \{0,1\}$,*

$$F_{Y_d|V_z}(y \mid v) = \Phi\left(a_{d,y} + b_{d,y}\Phi^{-1}(v)\right), \quad (y,v) \in \mathcal{Y} \times \mathcal{V},$$

where $a_{d,y} = \Phi^{-1}(F_{Y_d}(y))/\sqrt{1-\rho_{Y_d}(y)^2}$ and $b_{d,y} = -\rho_{Y_d}(y)/\sqrt{1-\rho_{Y_d}(y)^2}$.

[10] In fact, not only EX′ implies $\rho_{Y_d,V_z;Z}(\cdot,\cdot;z) = \rho_{Y_d,V_z}(\cdot,\cdot)$, but the converse is also true. See Remark B.2 below.

Proposition 4.2. *CI′ is equivalent to SI.*

Note that CI′ implies that, for $d \in \mathcal{D}$ and $z \in \{0, 1\}$,

$$F_{Y_d,V_z}(y, v) = C(F_{Y_d}(y), v; \rho_{Y_d}(y)), \quad (y, v) \in \mathcal{Y} \times \mathcal{V}.$$

By the properties of the conditional distribution and Gaussian copula,

$$F_{Y_d|V_z}(y \mid v) = \frac{(\partial/\partial v)F_{Y_d,V_z}(y, v)}{(\partial/\partial v)F_{V_z}(v)} = \Phi\left(a_{d,y} + b_{d,y}\Phi^{-1}(v)\right), \quad (y, v) \in \mathcal{Y} \times \mathcal{V},$$

where $a_{d,y} = \Phi^{-1}(F_{Y_d}(y))/\sqrt{1 - \rho_{Y_d}(y)^2}$ and $b_{d,y} = -\rho_{Y_d}(y)/\sqrt{1 - \rho_{Y_d}(y)^2}$.[11] It can be shown that the converse is also true.

SI is a single index restriction on the local relationship between the potential outcome Y_d and the unobservable of the treatment assignment V_z. This restriction does *not* require Gaussianity as the correlation coefficient is allowed to be a function of y but implies, for example, that the sign of $(\partial/\partial v)F_{Y_d|V_z}(y \mid v)$ does not depend on the value of v, although it can change with the value of y. This is a stochastic monotonicity restriction in the dependence between Y_d and V_z. Chernozhukov et al. (2020a) showed the equivalence between a special case of CI and a single index restriction for the selection model under the rank invariance $V_0 = V_1$, almost surely. Here, we extend the equivalence to a larger class of models that include non-binary treatments without imposing rank invariance. SI allows for rich forms of dependence between Y_d and V_z, as illustrated in Figure 1.

SI can also be viewed as a means of extrapolation. For example, Brinch et al. (2017) and Kowalski (2023) impose linearity in the marginal treatment effect to extrapolate the LATE to other treatment parameters. SI can be viewed as an alternative to their approach.[12] Specifically, Brinch et al. (2017) and Kowalski (2023) impose a linear model for the conditional expectation of Y_d given $V = v$, $\mathrm{E}[Y_d \mid V = v] = \alpha_d + \beta_d v$, where $V = V_0 = V_1$ under the LATE assumptions. By contrast, Proposition 4.2 shows that we impose a flexible model for the entire conditional distribution, $F_{Y_d|V}(y \mid v) = \Phi\left(a_{d,y} + b_{d,y}\Phi^{-1}(v)\right)$. When Y_d is binary, their approach corresponds to using linear probability model, whereas SI corresponds to a Probit model.

[11] This is reminiscent of the derivation in Section 3.3, although SI applies to both discrete and continuous D.

[12] Other approaches exist for extrapolating from the LATE to externally valid global treatment effects. Examples include approaches based on structural models (e.g., Heckman et al., 2003), covariates (e.g., Angrist and Fernández-Val, 2013), rank similarity (e.g., Wüthrich, 2020), and the smoothness of the marginal treatment effect function (e.g., Mogstad et al., 2018; Han and Yang, 2023).

4.2 Rank Similarity

The instrumental variables quantile regression (IVQR) model of Chernozhukov and Hansen (2005) provides an alternative set of conditions under which the QSF_τ is point identified, provided that the outcome is continuous. Here, we discuss how their identifying restriction is related to our results and CI, focusing on the case where D is binary.

The IVQR model is based on the Skorohod representation for continuous variables: $Y_d = Q_{Y_d}(U_d)$, $U_d \sim U[0,1]$, for $d \in \{0,1\}$. Chernozhukov and Hansen (2005) consider a general selection mechanism, $D = \delta(Z,V)$, where V can be vector-valued and the instrument is assumed to satisfy $U_d \perp\!\!\!\perp Z$ (which is implied by Assumption EX). The key assumption of the IVQR model is rank similarity (RS), $U_1 \overset{d}{=} U_0 \mid Z,V$. RS weakens the classical rank invariance (RI) assumption, which requires $U_1 = U_0$ almost surely.

The IVQR model yields the following conditional moment restriction (Chernozhukov and Hansen, 2005, Theorem 1),

$$\tau = \Pr[Y_1 \leq Q_{Y_1}(\tau), D_z = 1 \mid Z = z] + \Pr[Y_0 \leq Q_{Y_0}(\tau), D_z = 0 \mid Z = z], \quad z \in \{0,1\}. \tag{4.1}$$

Under the selection model (2.1) and $Y_0 \perp\!\!\!\perp Z$, (4.1) can be rewritten as

$$\Pr[Y_1 \leq Q_{Y_1}(\tau), V_z \leq \pi(z) \mid Z = z] = \Pr[Y_0 \leq Q_{Y_0}(\tau), V_z \leq \pi(z) \mid Z = z], \quad z \in \{0,1\}.$$

Using Lemma 2.1, we can further rewrite it as

$$C(\tau, \pi(z); \rho_{Y_1,V_z;Z}(Q_{Y_1}(\tau), \pi(z); z)) = C(\tau, \pi(z); \rho_{Y_0,V_z;Z}(Q_{Y_0}(\tau), \pi(z); z)), \quad z \in \{0,1\}.$$

This shows that the IVQR model also relies on a version of copula invariance,

$$\rho_{Y_1,V_z;Z}(Q_{Y_1}(\tau), \pi(z); z) = \rho_{Y_0,V_z;Z}(Q_{Y_0}(\tau), \pi(z); z), \quad z \in \{0,1\}. \tag{4.2}$$

The IVQR model imposes restrictions across potential outcomes for each instrument level, whereas CI imposes restrictions across instrument levels for each potential outcome. The two conditions are non-nested. We show in Appendix A.1 that the copula invariance assumption implied by the IVQR model and RS restricts treatment effect heterogeneity.

This discussion shows that both the IVQR model and our approach rely on non-nested and complementary copula invariance assumptions to identify causal effects for the overall population using instruments. Relative to the IVQR model, the proposed identification approach has three main advantages. First, it does not rely on continuity of the outcome to achieve point identification, and it naturally accommodates discrete and mixed discrete-

continuous outcomes. Second, it imposes less restrictions on the relationship across potential outcomes and thus effect heterogeneity (see Appendix A.1). Finally, it relies on a weaker relevance condition (see Remark 4.2).

Remark 4.2 (Weaker Relevance Conditions)**.** *The moment restriction* (4.1) *is not sufficient for point identification. To establish point identification, Chernozhukov and Hansen (2005) require the Jacobian of* (4.1) *to be of full rank and continuous. Vuong and Xu (2017) provide weaker conditions, requiring piecewise strict monotonicity of*

$$y \mapsto (-1)^d(\Pr[Y \leq y, D = d|Z = 0] - \Pr[Y \leq y, D = d|Z = 1])$$

for some d (in addition to REL). Both of these conditions impose assumptions on how the distribution of $(Y, D) \mid Z = z$ changes with z. These conditions can be restrictive in applications because they correspond to full support conditions for certain subpopulations (see, e.g., Wüthrich, 2020, Section 3.5). By contrast, under CI, we only require the weaker (and very standard) relevance condition REL, which only restricts how the distribution of $D \mid Z = z$ changes with z.

4.3 Roy Models

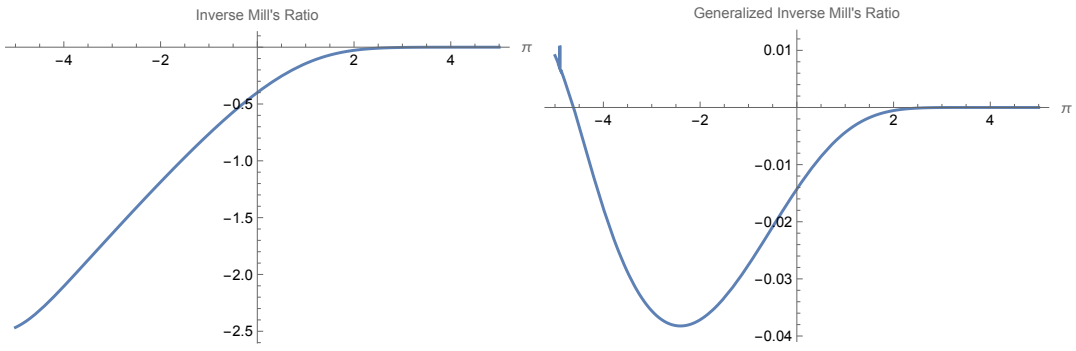
Figure 1 in Section 2.3 demonstrates the flexibility of CI in modeling the joint distribution of (Y_d, D_z). Our analyses show how point identification can be achieved under much weaker modeling restrictions than Gaussianity, which can be appealing to empirical researchers. To illustrate this further, consider a generalized Roy model for two sectors indexed by binary D and binary cost shock Z. Let the potential log wage in sector d be $Y_d = \mathrm{E}[Y_d] + U_d$ for $d \in \{0, 1\}$, and suppose that individuals choose sector $D = 1$ if the benefit $Y_1 - Y_0$ exceeds the cost $\tilde{\pi}(z) + \tilde{V}_z$. This implies that $D_z = 1\{V_z \leq \pi(z)\}$, where $V_z \equiv U_0 - U_1 + \tilde{V}_z$ and $\pi(z) \equiv \mathrm{E}[Y_1] - \mathrm{E}[Y_0] - \tilde{\pi}(z)$.[13] Suppose $Z \perp\!\!\!\perp (Y_d, V_z)$. Then, $\mathrm{E}[Y \mid D = 1, Z = z] = \mathrm{E}[Y_1] + \mathrm{E}[U_1 \mid V_z \leq \pi(z)]$, where $\mathrm{E}[U_1 \mid V_z \leq \pi(z)]$ is the control function that captures endogenous section into sector $d = 1$. We can obtain a similar expression for $d = 0$.

Under joint Gaussianity of (Y_d, V_z), one can show that $\mathrm{E}[U_1 \mid V_z \leq \pi(z)]$ is equal to the inverse Mill's ratio (Heckman, 1979). This function has a specific shape, namely, strictly monotone with respect to the propensity score $\pi(z)$, which is depicted in Figure 2(a). Relatedly, Heckman and Honore (1990) show that Gaussianity in Roy models implies that (i) the Roy economy results in less dispersed log wages than when there is no endogenous se-

[13] Note that this model is more flexible than the standard generalized Roy model as the cost unobservable, $\tilde{V}_z$, is z-specific, which is the aspect we can allow for.

lection of sectors (Theorem 2); (ii) the Roy economy exhibits a right-skewed distribution of aggregate log wages (Theorem 3).

On the other hand, under CI for the joint distribution of (Y_d, V_z), one can generate general non-monotonic selection patterns in the control function, as we illustrate in Figure 2(b). In this sense, in CI Roy models, economic implications can be more nuanced and data-driven, unlike in Gaussian Roy models where implications are largely model-driven. Moreover, the analysis in Section 3.1 implies that the parameters in the sector wage equations can be identified even with a binary cost shock.[14]



(a) (b)

Figure 2: Examples of Control Functions that Satisfy Copula Invariance

Notes: Panel (a): Bivariate Gaussian distribution (corresponding to Figure 1(a)); Panel (b): Non-Gaussian copula and univariate Gaussian marginals (corresponding to Figure 1(b)). Both panels depict $\pi \mapsto \mathrm{E}[U_1 \mid V_z \leq \pi]$ for $U_1 = Y_1 - \mathrm{E}[Y_1]$.

5 Estimation and Inference

An attractive feature of the identification results in Sections 3.1–3.3 is that they are constructive and lead to tractable estimators. Here, we propose semiparametric estimators of the potential outcome distributions for the three types of treatment based on distribution regression (DR). We also show how to construct estimators of treatment effect parameters such as unconditional QTE using the plug-in rule. We focus on target parameters that yield $\sqrt{n}$-consistent and asymptotically normal estimators. We do not consider parameters like $\partial QSF_\tau(d)/\partial d$, because estimating such parameters would involve non-parametric methods, which require choosing tuning parameters and exhibit slower convergence rates.

[14] Without any distributional assumptions or assumptions on the dependence structure, one would require IVs with large support to identify the wage parameters (Eisenhauer et al., 2015).

In this section, we make the role of the covariates X explicit; see Appendix D for a more detailed discussion of identification with covariates. We assume we have access to a random sample of size n from (Y, D, Z, X), $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$, for estimation. Let $B(X_i)$, $B(Z_i, X_i)$, and $B(D_i, Z_i, X_i)$ denote vectors of transformations of X_i, (Z_i, X_i), and (D_i, Z_i, X_i), respectively. Define the indicators $I_i(y) \equiv 1\{Y_i \leq y\}$ and $J_i(d) \equiv 1\{D_i \leq d\}$. Let $\bar{\mathcal{D}}$ and $\bar{\mathcal{Y}}$ be two finite grids covering $\mathcal{D}$ and $\mathcal{Y}$.[15]

5.1 Binary Treatments

We consider separate DR models for the conditional potential outcome distributions,

$$F_{Y_d|X}(y|x) = \Phi(B(x)'\beta_d(y)), \quad d \in \{0, 1\}, \tag{5.1}$$

where $y \mapsto \beta_d(y)$ is a function-valued unknown parameter, and a Probit model for the propensity score,

$$\pi(z, x) = \Pr[D = 1 \mid Z = z, X = x] = \Phi(B(z, x)'\pi). \tag{5.2}$$

We model the local dependence parameter as

$$\rho_{Y_d;X}(y; x) = \rho(B(x)'\gamma_d(y)), \quad d \in \{0, 1\}, \tag{5.3}$$

where $\rho(u) = \tanh(u) \in [-1, 1]$, the Fisher transformation, and $y \mapsto \gamma_d(y)$ is a function-valued unknown parameter.[16] Together, (5.1), (5.2), and (5.3) imply the bivariate DR model

$$\begin{aligned}
\Pr[Y \leq y, D = 1 \mid Z = z] &= \Phi_2(B(x)'\beta_1(y), B(z, x)'\pi; \rho(B(x)'\gamma_1(y))), \\
\Pr[Y \leq y, D = 0 \mid Z = z] &= \Phi_2(B(x)'\beta_0(y), -B(z, x)'\pi; -\rho(B(x)'\gamma_0(y))),
\end{aligned}$$

where we have used the symmetry properties of the bivariate Gaussian distribution to simplify the previous expressions.

We propose a computationally tractable two-step maximum likelihood estimator, building on Chernozhukov et al. (2020a).

Algorithm 5.1 (Estimation of Binary Treatment Model)**.** *We compute the estimator in two stages:*

[15] We can set $\bar{\mathcal{Y}} = \mathcal{Y}$ when $\mathcal{Y}$ is finite. We only use $\bar{\mathcal{D}}$ when D is continuous.

[16] To simplify the notation, we use the same vector of transformations in models (5.1) and (5.3). This is not essential, and one can use different specifications in both models.

1. *Treatment equation: estimate π using a Probit regression*

$$\widehat{\pi} \in \arg\max_{c} \sum_{i=1}^{n} \left[D_i \log \Phi(B(Z_i, X_i)'c) + (1 - D_i) \log(1 - \Phi(B(Z_i, X_i)'c)) \right].$$

2. *Outcome equation: for $y \in \bar{\mathcal{Y}}$ and $d \in \{0, 1\}$, $\widehat{F}_{Y_d|X}(y|x) = \Phi(B(x)'\widehat{\beta}_d(y))$ and $\widehat{\rho}_{Y_d;X}(y; x) = \rho(B(x)'\widehat{\gamma}_d(y))$, where*

$$\begin{aligned}
(\widehat{\beta}_1(y), \widehat{\gamma}_1(y)) \in \arg\max_{b,g} \sum_{i=1}^{n} D_i \big[& I_i(y) \log \Phi_2(B(X_i)'b, B(Z_i, X_i)'\widehat{\pi}, \rho(B(X_i)'g)) \\
& + (1 - I_i(y)) \log \Phi_2(-B(X_i)'b, B(Z_i, X_i)'\widehat{\pi}, \rho(B(X_i)'g)) \big], \\
(\widehat{\beta}_0(y), \widehat{\gamma}_0(y)) \in \arg\max_{b,g} \sum_{i=1}^{n} (1 - D_i) \big[& I_i(y) \log \Phi_2(B(X_i)'b, -B(Z_i, X_i)'\widehat{\pi}, -\rho(B(X_i)'g)) \\
& + (1 - I_i(y)) \log \Phi_2(-B(X_i)'b, -B(Z_i, X_i)'\widehat{\pi}, -\rho(B(X_i)'g)) \big].
\end{aligned}$$

Rearrange the estimates $y \mapsto \widehat{F}_{Y_d|X}(y \mid x)$ on $\bar{\mathcal{Y}}$ if needed.

Remark 5.1 (Computation)**.** *The first stage of Algorithm 5.1 is a conventional Probit regression, and estimation can proceed using existing software. The second stage is computationally more expensive since it involves a nonlinear smooth optimization problem. This optimization problem can be solved using standard algorithms such as Newton-Raphson.*

5.2 Ordered Discrete Treatments

As for binary treatments, we model all the components using flexible generalized linear and DR models:

$$\begin{aligned}
& F_{Y_d|X}(y \mid x) = \Phi(B(x)'\beta_d(y)), \quad \rho_{Y_d;X}(y; x) = \rho(B(x)'\gamma_d(y)), \\
& \pi_d(z, x) = F_{D|Z,X}(d \mid z, x) = \Phi(B(z, x)'\pi(d)),
\end{aligned}$$

where $\rho(u) = \tanh(u)$.

Algorithm 5.2 (Estimation of Ordered Treatment Model)**.** *We compute the estimator in two stages:*

1. *Treatment equation: set $\widehat{\pi}_0(z, x) = 0$ and $\widehat{\pi}_K(z, x) = 1$ for all (z, x). For $d \in$*

$\{1, \ldots, K-1\}$, $\widehat{\pi}_d(z,x) = \Phi(B(z,x)'\widehat{\pi}(d))$, *where*

$$\widehat{\pi}(d) \in \arg\max_p \sum_{i=1}^n \left[J_i(d) \log \Phi(B(Z_i, X_i)'p) + (1 - J_i(d)) \log \Phi(-B(Z_i, X_i)'p) \right].$$

Rearrange the estimates $d \mapsto \widehat{\pi}_d(z,x)$ *on* $\mathcal{D}$ *if needed. This rearrangement is important to avoid having logarithms of negative numbers in the second stage.*[17]

2. *Outcome equation: for* $y \in \bar{\mathcal{Y}}$ *and* $d \in \mathcal{D}$, $\widehat{F}_{Y_d|X}(y \mid x) = \Phi(B(x)'\widehat{\beta}_d(y))$ *and* $\widehat{\rho}_{Y_d;X}(y;x) = \rho(B(x)'\widehat{\gamma}_d(y))$, *where*

$$(\widehat{\beta}_d(y), \widehat{\gamma}_d(y)) \in \arg\max_{b,g} \sum_{i=1}^n 1\{D_i = d\} \left[I_i(y) \log g_{d,i}(b,g) + (1 - I_i(y)) \log \bar{g}_{d,i}(b,g) \right],$$

where

$$g_{d,i}(b,g) \equiv \Phi_2(B(X_i)'b, \Phi^{-1}\left(\widehat{\pi}_d(Z_i, X_i)\right), \rho(B(X_i)'g)) \\ - \Phi_2(B(X_i)'b, \Phi^{-1}\left(\widehat{\pi}_{d-1}(Z_i, X_i)\right), \rho(B(X_i)'g)),$$

and

$$\bar{g}_{d,i}(b,g) \equiv \widehat{\pi}_d(Z_i, X_i) - \widehat{\pi}_{d-1}(Z_i, X_i) - g_{d,i}(b,g).$$

Rearrange the estimates $y \mapsto \widehat{F}_{Y_d|X}(y \mid x)$ *on* $\bar{\mathcal{Y}}$ *if needed.*

Remark 5.2 (Computation)**.** *The first stage is a sequence of Probit regressions that can be solved using standard software, as in Algorithm 5.1. The second stage is a nonlinear smooth optimization problem that can be solved using standard algorithms such as Newton-Raphson.*

5.3 Continuous Treatments

We construct plug-in estimators based on the closed-form solutions in Section 3.3. We consider DR models for $F_{Y|D,Z,X}$ and $F_{D|Z,X}$,

$$F_{Y|D,Z,X}(y \mid d, z, x) = \Phi(B(d, z, x)'\beta(y)), \tag{5.4}$$

$$F_{D|Z,X}(d \mid z, x) = \Phi(B(z, x)'\pi(d)), \tag{5.5}$$

where $y \mapsto \beta(y)$ and $d \mapsto \pi(d)$ are unknown function-valued parameters.

[17] In the second stage, $g_{d,i}(b,g) > 0$ and $\bar{g}_{d,i}(b,g) > 0$ a.s. if $\widehat{\pi}_d(Z_i, X_i) > \widehat{\pi}_{d-1}(Z_i, X_i)$ a.s.

Algorithm 5.3 (Estimation of Continuous Treatment Model)**.** *We compute the estimator in two stages:*

1. *Observable conditional distributions: for* $y \in \bar{\mathcal{Y}}$ *and* $d \in \bar{\mathcal{D}}$, $\widehat{F}_{Y|D,Z,X}(y|d,z,x) = \Phi(B(d,z,x)'\widehat{\beta}(y))$ *and* $\widehat{F}_{D|Z,X}(d|z,x) = \Phi(B(z,x)'\widehat{\pi}(d))$*, where*

$$\widehat{\beta}(y) \in \arg\max_b \sum_{i=1}^n \left[I_i(y) \log \Phi(B(D_i,Z_i,X_i)'b) + (1-I_i(y)) \log(1-\Phi(B(D_i,Z_i,X_i)'b)) \right],$$

$$\widehat{\pi}(d) \in \arg\max_p \sum_{i=1}^n \left[J_i(d) \log \Phi(B(Z_i,X_i)'p) + (1-J_i(d)) \log(1-\Phi(B(Z_i,X_i)'p)) \right].$$

2. *Potential outcome distributions: for* $y \in \bar{\mathcal{Y}}$ *and* $d \in \bar{\mathcal{D}}$, $\widehat{F}_{Y_d|X}(y|x) = \Phi\left(\widehat{a}_{d,y;x}/\sqrt{1+\widehat{b}^2_{d,y;x}}\right)$ *and* $\widehat{\rho}_{Y_d;X}(y;x) = -\widehat{b}_{d,y;x}/\sqrt{1+\widehat{b}^2_{d,y;x}}$*, where*

$$\widehat{a}_{d,y;x} = \frac{(B(d,0,x)'\widehat{\beta}(y))(B(1,x)'\widehat{\pi}(d)) - (B(d,1,x)'\widehat{\beta}(y))(B(0,x)'\widehat{\pi}(d))}{B(1,x)'\widehat{\pi}(d) - B(0,x)'\widehat{\pi}(d)},$$

$$\widehat{b}_{d,y;x} = \frac{B(d,1,x)'\widehat{\beta}(y) - B(d,0,x)'\widehat{\beta}(y)}{B(1,x)'\widehat{\pi}(d) - B(0,x)'\widehat{\pi}(d)}.$$

Rearrange the estimates $y \mapsto \widehat{F}_{Y_d|X}(y \mid x)$ *on* $\bar{\mathcal{Y}}$ *if needed.*

5.4 Marginal Distributions and Functionals

In the presence of covariates, the marginal distributions of the potential outcomes are identified by

$$F_{Y_d}(y) = \int F_{Y_d|X}(y \mid x) dF_X(x), \quad d \in \mathcal{D},$$

where F_X is the distribution of X. We can use this expression to construct estimators of F_{Y_d} and functionals of interest, such as the QSF and QTE, by plugging in the estimators obtained above. To provide a unified approach to all the treatment cases, we set $\bar{\mathcal{D}} = \mathcal{D}$ when D is binary or discrete ordered.

Algorithm 5.4 (Estimation of F_{Y_d}, QSF and QTE)**.** *Estimation proceeds in two steps.*

1. *Unconditional distribution: for* $y \in \bar{\mathcal{Y}}$ *and* $d \in \bar{\mathcal{D}}$*,*

$$\widehat{F}_{Y_d}(y) = \frac{1}{n} \sum_{i=1}^n \widehat{F}_{Y_d|X}(y \mid X_i).$$

For $y \in \mathcal{Y} \setminus \bar{\mathcal{Y}}$ and $d \in \bar{\mathcal{D}}$, $\widehat{F}_{Y_d}(y) = \max\{\widehat{F}_{Y_d}(\bar{y}) : \bar{y} < y, \bar{y} \in \bar{\mathcal{Y}}\}$.[18]

2. *Unconditional QSF and QTE: $\widehat{QSF}_\tau(d) = \mathcal{Q}_\tau(\widehat{F}_{Y_d})$ and $\widehat{QTE}_\tau(d, d') = \{\widehat{QSF}_\tau(d) - \widehat{QSF}_\tau(d')\}/(d - d')$.*

5.5 Asymptotic Theory and Inference

The target parameters in Section 5.4 are function-valued. Inference on these parameters can be performed using resampling methods. To provide a unified framework, we denote the functional parameters by $u \mapsto \delta_u, u \in \mathcal{U}$, where $\mathcal{U} \subset \tilde{\mathcal{Y}} \times \tilde{\mathcal{D}} \times \mathcal{T}$, where $\mathcal{T} \subset (c, 1-c)$ for $c > 0$, $\tilde{\mathcal{Y}}$ is a compact subset of $\mathcal{Y}$ when $\mathcal{Y}$ is uncountable and equal to $\mathcal{Y}$ otherwise, and $\tilde{\mathcal{D}}$ is defined analogously. For example, if we are interested in $\tau \mapsto QSF_\tau(d)$ on $[.05, .95]$, then $u = \tau$, $\delta_u = QSF_u(d)$ and $\mathcal{U} = [.05, .95]$ for fixed d. In practice, we approximate $\mathcal{U}$ using a fine grid $\bar{\mathcal{U}}$. We denote the estimator of δ_u obtained from Algorithms 5.1, 5.2, 5.3, and 5.4 as $\widehat{\delta}_u$.

We show in Appendix F that $\sqrt{n}(\widehat{\delta}_u - \delta_u)$ converges in distribution to a mean-zero Gaussian process Z_δ and the bootstrap is valid for a wide range of parameters of interest, including F_{Y_d} and its functionals. These results imply that the bootstrap algorithm below is theoretically valid. When the outcomes are discrete or mixed discrete-continuous, the estimators of the QSF or QTE are not asymptotically Gaussian in general. In this case, one can use the inference methods proposed by Chernozhukov et al. (2020b).

We focus on constructing pointwise and uniform confidence bands, $CB^{pt}_{(1-\alpha)}(\delta_{\bar{u}})$ and $CB_{(1-\alpha)}(\delta_u)$ respectively, satisfying

$$\lim_{n\to\infty} \Pr[\delta_{\bar{u}} \in CB^{pt}_{(1-\alpha)}(\delta_{\bar{u}})] = 1 - \alpha, \qquad \lim_{n\to\infty} \Pr[\delta_u \in CB_{(1-\alpha)}(\delta_u) \text{ for all } u \in \mathcal{U}] = 1 - \alpha.$$

Uniform confidence bands can be used to test a variety of hypotheses of interest, such as the hypotheses of no effect or constant effects when applied to QTE, or stochastic dominance.

The following algorithm provides a generic bootstrap construction of $CB_{(1-\alpha)}(\delta_u)$. A similar algorithm can be used for $CB^{pt}_{(1-\alpha)}(\delta_u)$, which we omit.

Algorithm 5.5 (Uniform Confidence Bands for Functional Parameters[19])**.**

1. *For $u \in \bar{\mathcal{U}}$, obtain B bootstrap draws of the estimator $\widehat{\delta}_u$, $\{\widehat{\delta}^{(b)}_u : 1 \le b \le B\}$.*

2. *For $u \in \bar{\mathcal{U}}$, compute the robust standard error,*

$$SE(\widehat{\delta}_u) = (\widehat{Q}_{0.75}(\widehat{\delta}_u) - \widehat{Q}_{0.25}(\widehat{\delta}_u))/(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)),$$

[18] In practice, one can also use linear extrapolation when D is continuous.

[19] See, for example, Chernozhukov et al. (2013, 2020a,b) for similar algorithms.

where $\widehat{Q}_\tau(\widehat{\delta}_u)$ *is the* τ*-quantile of* $\{\widehat{\delta}_u^{(b)} : 1 \leq b \leq B\}$.

3. *Compute the critical value as*

$$cv(1-\alpha) = (1-\alpha)\text{-quantile of } \left\{ \max_{u \in \bar{\mathcal{U}}} \frac{|\widehat{\delta}_u^{(b)} - \widehat{\delta}_u|}{SE(\widehat{\delta}_u)} : 1 \leq b \leq B \right\}.$$

4. *Compute the* $(1-\alpha)$ *uniform confidence band as*

$$CB_{(1-\alpha)}(\delta_u) = [\widehat{\delta}_u \pm cv(1-\alpha)SE(\widehat{\delta}_u)], \quad u \in \bar{\mathcal{U}}.$$

The estimation algorithms for binary and ordered treatments (Algorithms 5.1 and 5.2) involve nonlinear optimization problems. Therefore, we recommend using the multiplier bootstrap in Step 1 of Algorithm 5.5, which avoids re-estimating the parameters in Algorithms 5.1 and 5.2 in each of the B bootstrap iterations.

The estimation approach for continuous treatments in Algorithm 5.3 does not involve solving a nonlinear optimization problem in the second step and is computationally less expensive than Algorithms 5.1 and 5.2. Therefore, the standard empirical bootstrap is a natural alternative to the multiplier bootstrap in Step 1, provided that the sample size is not too large, as for example in Section 6.

6 Empirical Illustration

We illustrate our methods by estimating the distributional effects of sleep on well-being.[20] We reanalyze the data from Bessone et al. (2021b), who studied the effects of randomized interventions to increase sleep time of low-income adults in India.[21] According to an expert survey conducted by Bessone et al. (2021b), increasing sleep could have large benefits in this empirical context, which is characterized by low baseline levels of sleep per night and sleep efficiency.

Bessone et al. (2021b) considered two main treatments (see their Section III for details): (i) *devices + encouragement* (information, encouragements and sleep trackers, various devices for improving sleep environment) and (ii) *devices + incentives* (same as (i) and payments for each minute of sleep increase). In addition, they cross-randomized a *nap treatment* (the opportunity to nap at work).

[20] The empirical results were obtained using the statistical software `R` (R Core Team, 2024).
[21] We downloaded the data from the Harvard Dataverse replication package (Bessone et al., 2021a).

The outcome of interest (Y) is an overall index of individual well-being. The treatment (D) is sleep time per night (in hours). We use an indicator for whether an individual received either of the main treatments as an instrument (Z) for sleep. The vector of covariates (X) includes controls for gender, three age indicators, and the baseline well-being index, as in Bessone et al. (2021b, Table A.XVII). Following Dong and Lee (2023), we restrict the sample to individuals who did not receive the nap treatment, so that the total sample size is $n = 226$.

The treatment D takes on many values in this application and is treated as continuous. We therefore use the estimators for continuous treatments described in Algorithm 5.3. We choose fully saturated (in Z_i) specifications for $B(d, z, x)$ and $B(z, x)$. The resulting DR models can be written as

$$F_{Y|D,Z,X}(y \mid d, z, x) = \Phi((d, x')\beta_z(y)), \quad F_{D|Z,X}(d \mid z, x) = \Phi(x'\pi_z(d)), \quad z \in \{0, 1\}.$$

Our flexible semiparametric estimators allow us to analyze the full distributional impact of sleep on well-being. Our results thus complement the empirical analyses of the average effect of sleep on well-being in Bessone et al. (2021b) using two-stage least squares (2SLS) and in Dong and Lee (2023) using 2SLS and semiparametric doubly robust methods.

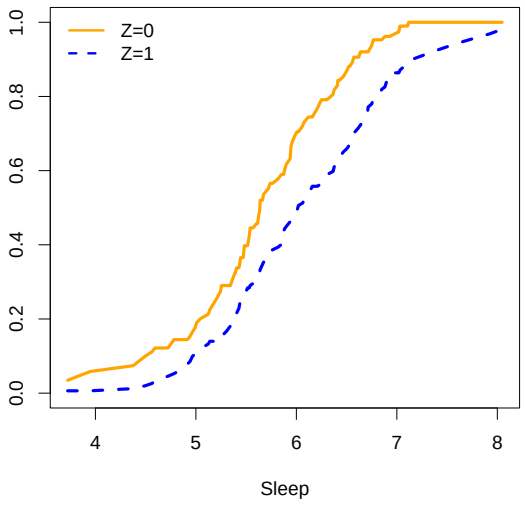
We begin by analyzing the (distributional) first-stage relationship between D and Z to shed light on the plausibility of REL. Figure 3 plots $\widehat{F}_{D|Z}(\cdot \mid 1) = n^{-1}\sum_{i=1}^{n} \widehat{F}_{D|Z,X}(\cdot \mid 1, X_i)$ and $\widehat{F}_{D|Z}(\cdot \mid 0) = n^{-1}\sum_{i=1}^{n} \widehat{F}_{D|Z,X}(\cdot \mid 0, X_i)$. It shows that the instrument induces a shift in the distribution of D. The distribution of sleep under the experimental treatments first order stochastically dominates the distribution of sleep without treatments.

In addition to REL, our method relies on EX and CI. The experimental random assignment renders the independence assumptions in EX plausible.[22] CI allows the local dependence between potential well-being and the unobservable determinants of sleep to depend on the level of well-being, but not on the level of the unobservable determinants of sleep and the instrument.

Figure 4 plots estimates of the QTE, $\widehat{QTE}_{\tau}(d, d')$, including 90% pointwise and uniform confidence intervals (CIs) computed using empirical bootstrap. We set $d = \widehat{Q}_D(0.75)$ and $d' = \widehat{Q}_D(0.25)$, where $\widehat{Q}_D(\tau)$ is the τ-quantile of the empirical distribution of sleep. Figure 4(a) suggests interesting effect heterogeneity across the distribution of well-being. The QTEs are the largest and pointwise significant in the tails and smaller and insignificant around the median. The uniform CI includes zero at all quantiles. This finding is not very surprising given the relatively small sample size. In Figure 4(b), we "zoom-in" on the lower tail of the well-being distribution, which is of particular policy interest. The corresponding uniform CI

[22] Note that random assignment does not automatically imply the (implicit) exclusion restriction, which requires that the instrument has no direct effect on the well-being.

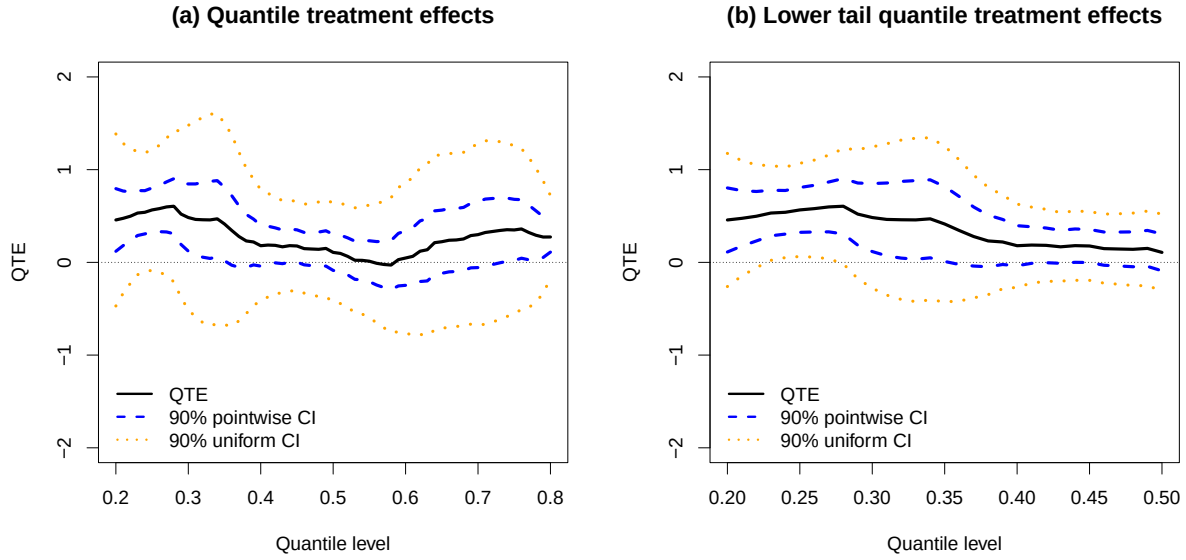
Figure 3: Distributional First Stage



Notes: All specifications control for gender, three age indicators, and the baseline well-being index.

do not include the zero QTE line, so that we can reject the null hypothesis that sleep has no impact on well-being at the lower tail.

Figure 4: Quantile Treatment Effects



Notes: Pointwise and uniform CIs for the QTE are computed using the empirical bootstrap with 5,000 repetitions. All specifications control for gender, three age indicators, and the baseline well-being index.

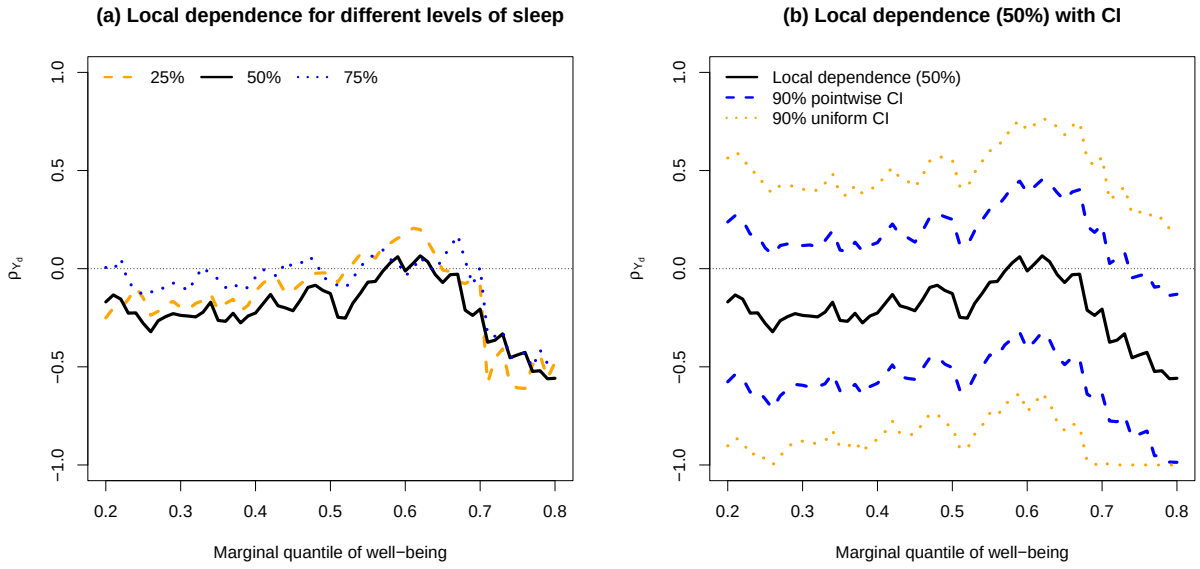
An interesting feature of our method is that it allows for estimating the local dependence parameter $\rho_{Y_d;X}(y;x)$, which can be interpreted as a local measure of endogeneity or self-

selection. Figure 5(a) plots the average local dependence parameter,

$$\tau \mapsto \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}_{Y_d;X}(\widehat{Q}_Y(\tau); X_i), \quad d \in \left\{\widehat{Q}_D(0.25), \widehat{Q}_D(0.50), \widehat{Q}_D(0.75)\right\},$$

where $\widehat{Q}_Y(\tau)$ is the τ-quantile of the empirical distribution of well-being. The average local dependence is negative at most quantiles of well-being, and there is interesting heterogeneity both across the distribution of well-being and across the three different values of sleep. This negative selection means that the unobserved propensity to sleep and the potential well-being by level of sleep are negatively correlated. In other words, individuals with relatively poor underlying health conditions sleep more hours. Figure 5(b) shows that the average local dependence for $d = \widehat{Q}_D(0.50)$ is pointwise significant at the upper tail, although the uniform CI includes zero at all quantile levels considered.

Figure 5: Local Dependence



Notes: Pointwise and uniform CIs for the average local dependence are computed using the empirical bootstrap with 5,000 repetitions. All specifications control for gender, three age indicators, and the baseline well-being index.

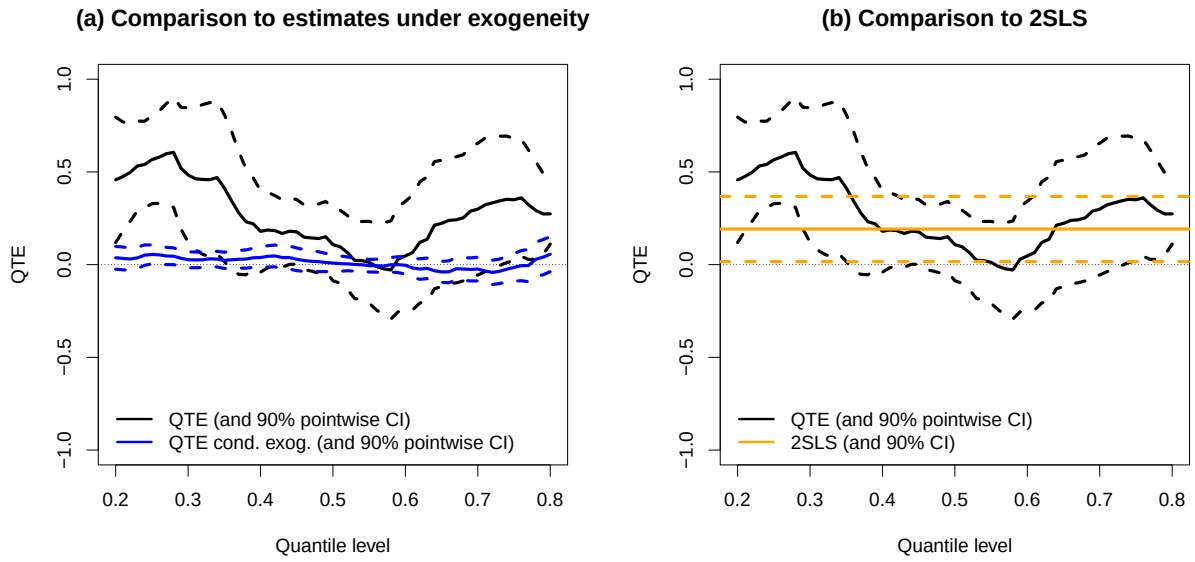
The negative local dependence suggests that estimators that ignore the endogeneity of sleep will underestimate the effect of sleep on well-being. Figure 6(a) demonstrates this phenomenon by comparing the QTE estimates obtained using our IV method to the corresponding estimates under conditional exogeneity, $Y_d \perp\!\!\!\perp D \mid X$ for $d \in \mathcal{D}$.[23] The QTE

[23]Under conditional exogeneity, F_{Y_d} is identified as $F_{Y_d}(y) = \int F_{Y|D,X}(y|d,x)dF_X(x)$. For estimation, we consider the DR model $F_{Y|D,X}(y|d,x) = \Phi(d\gamma(y) + x'\beta(y))$. The DR estimator is $\widehat{F}_{Y|D,X}(y|d,x) =$

estimates under conditional exogeneity are small and insignificant at most quantiles, and thus miss the positive well-being effects of sleep in the tails of the well-being distribution. This finding demonstrates the importance of using IV methods in this empirical context.

Figure 6(b) further showcases the value-added that our method can bring to standard empirical analyses focusing on average effects. It compares the QTE estimates obtained from our method to standard 2SLS estimates using the same set of covariates. While the 2SLS analysis suggests that sleep has moderate average effects on well-being, our method uncovers interesting patterns of heterogeneity in the effect of sleep on well-being.

Figure 6: Comparison to Estimates under Conditional Exogeneity and 2SLS



Notes: Pointwise CIs for the QTE are computed using the empirical bootstrap with 5,000 repetitions. All specifications control for gender, three age indicators, and the baseline well-being index.

7 Concluding Remarks

In identifying causal effect of endogenous treatments, researchers inevitably face modeling trade-offs. This paper proposes a new direction to deal with these trade-offs by imposing assumptions on the local dependence between potential outcomes and unobservables determining treatment assignment. In doing so, we make minimal assumptions on the equations determining potential outcomes and treatments, thereby allowing for rich heterogeneity in these equations. The proposed framework applies to binary, discrete ordered, and continuous

$\Phi(d\widehat{\gamma}(y)+x'\widehat{\beta}(y))$, where $\widehat{\gamma}(y)$ and $\widehat{\beta}(y)$ are the coefficients obtained from a Probit regression of $1\{Y_i \leq y\}$ on D_i and X_i. The final estimator of $F_{Y_d}(y)$ under conditional exogeneity is $\widehat{F}_{Y_d}(y) = \frac{1}{n}\sum_{i=1}^{n}\Phi(d\widehat{\gamma}(y)+X_i'\widehat{\beta}(y))$.

treatments, where point identification is achieved even with instruments that have limited variation (e.g., a binary instrument). In all three cases, the identification analysis leads to practical estimation and inference procedures that might appeal to practitioners.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117. 1.1

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843. 3

Ambrosetti, A. and G. Prodi (1995): *A primer of nonlinear analysis*, 34, Cambridge University Press. 3.2, 3.2, G.2

Angrist, J. D. and I. Fernández-Val (2013): *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*, Cambridge University Press, vol. 3 of *Econometric Society Monographs*, 401–434. 12

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527. 1

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 1

Anjos, U. U. d. and N. Kolev (2005): "Representation of bivariate copulas via local measure of dependence," . 2.2, B.2, B.2

Arellano, M. and S. Bonhomme (2017): "Quantile selection models with an application to understanding changes in wage inequality," *Econometrica*, 85, 1–28. 1.1

Athey, S. and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 74, 431–497. 1.1

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 1.1

Beare, B. K. (2010): "Copulas and Temporal Dependence," *Econometrica*, 78, 395–410. 4

Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021a): "Replication Data for: 'The Economic Consequences of Increasing Sleep among the Urban Poor'," . 21

——— (2021b): "The Economic Consequences of Increasing Sleep Among the Urban Poor*," *The Quarterly Journal of Economics*, 136, 1887–1941. 1, 1, 6

Black, S. E. and P. J. Devereux (2011): "Recent developments in intergenerational mobility," *Handbook of labor economics*, 4, 1487–1541. 1

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669. 1.1

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125, 985–1039. 4.1.2

Carneiro, P. and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149, 191–208. 1.1

Chay, K. and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113, 376–424. 1

Chen, X. and Y. Fan (2006a): "Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification," *Journal of Econometrics*, 135, 125–154. 4

——— (2006b): "Estimation of copula-based semiparametric time series models," *Journal of Econometrics*, 130, 307–335. 4

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association*, 101, 1228–1240. 1.1

Chen, X., Z. Huang, and Y. Yi (2021): "Efficient estimation of multivariate semi-nonparametric GARCH filtered copula models," *Journal of Econometrics*, 222, 484–501. 4

Chen, X. and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152, 46–60. 3

——— (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80, 277–321. 3

——— (2015): "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models," *Econometrica*, 83, 1013–1079. 3

Chen, X., Z. Xiao, and B. Wang (2022): "Copula-based time series with filtered non-stationarity," *Journal of Econometrics*, 228, 127–155. 4

Chernozhukov, V., I. Fernández-Val, and S. Luo (2020a): "Distribution regression with sample selection, with an application to wage decompositions in the UK," *arXiv preprint arXiv:1811.11603*. 1.1, 2.2, 7, 4.1.2, 5.1, 19, F

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268. 19, F, F, G.4

Chernozhukov, V., I. Fernández-Val, B. Melly, and K. Wüthrich (2020b): "Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes," *Journal of the American Statistical Association*, 115, 123–137. 5.5, 19, F

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261. 1, 1.1, 3.1, 4.2, 4.2, A.1

——— (2013): "Quantile models with endogeneity," *Annu. Rev. Econ.*, 5, 57–81. 1.1

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441. 1.1

Chesher, A. and A. M. Rosen (2020): "Generalized instrumental variable models, methods, and applications," in *Handbook of Econometrics*, Elsevier, vol. 7, 1–110. 1.1

Cunha, F., J. J. Heckman, and S. Navarro (2007): "The identification and economic content of ordered choice models with stochastic thresholds," *International Economic Review*, 48, 1273–1309. 8

De Chaisemartin, C. (2017): "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 8, 367–396. 3.2

De Paula, Á., I. Rasul, and P. Souza (2019): "Identifying network ties from panel data: Theory and an application to tax competition," *arXiv preprint arXiv:1910.07452*. 3.2, G.2

D'Haultfoeuille, X. and P. Février (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 83, 1199–1210. 1.1

Dong, Y. and Y.-Y. Lee (2023): "Nonparametric Doubly Robust Identification of Causal Effects of a Continuous Treatment using Discrete Instruments," arXiv:2310.18504. 6

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The generalized Roy model and the cost-benefit analysis of social programs," *Journal of Political Economy*, 123, 413–443. 14

Fan, Y., F. Han, and H. Park (2023): "Estimation and inference in a high-dimensional semiparametric Gaussian copula vector autoregressive model," *Journal of Econometrics*, 237, 105513. 4

Gale, D. and H. Nikaido (1965): "The Jacobian matrix and global univalence of mappings," *Mathematische Annalen*, 159, 81–93. 3.1, 3.2, G.1

Ghanem, D., D. Kédagni, and I. Mourifié (2023): "Evaluating the Impact of Regulatory Policies on Social Welfare in Difference-in-Difference Settings," . 1.1

Hadamard, J. (1906): "Sur les transformations ponctuelles," *Bull. Soc. Math. France*, 34, 71–84. B.2, 29

Han, S. and S. Lee (2019): "Estimation in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Applied Econometrics*, 34, 994–1015. 1.1

——— (2023): "Semiparametric Models for Dynamic Treatment Effects and Mediation Analyses with Observational Data," *University of Bristol and Sogang University*. 1.1

Han, S. and E. J. Vytlacil (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199, 63–73. 1.1, B.2, C, G.1

Han, S. and S. Yang (2023): "A Computational Approach to Identification of Treatment Effects for Policy Evaluation," *arXiv preprint arXiv:2009.13861*. 12

Heckman, J., J. L. Tobias, and E. Vytlacil (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *The Review of Economics and Statistics*, 85, 748–755. 12

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. 4.3

Heckman, J. J. and B. E. Honore (1990): "The empirical content of the Roy model," *Econometrica: Journal of the Econometric Society*, 1121–1149. 4.3

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation1," *Econometrica*, 73, 669–738. 1.1

Heckman, J. J. and E. J. Vytlacil (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," Elsevier, vol. 6 of *Handbook of Econometrics*, 4875–5143. 3.2, 3.2, 3.5, 30

Hirano, K. and G. W. Imbens (2004): "The Propensity Score with Continuous Treatments," in *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman and X.-L. Meng, Wiley, 73–84. 5

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 1.1, 3.1, 4.1.1, A.2

Imbens, G. W. and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512. 1, 1.1, 3.3, A.3

Kolev, N., U. d. Anjos, and B. V. d. M. Mendes (2006): "Copulas: a review and recent developments," *Stochastic models*, 22, 617–660. 2.2

Kowalski, A. E. (2023): "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform," *Review of Economics and Statistics*, 105, 646–664. 4.1.2

Liu, H., J. Lafferty, and L. Wasserman (2009): "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research*, 10. 7

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *The American Economic Review*, 80, 319–323. 1.1

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. 12

Mourifié, I. and Y. Wan (2021): "Layered policy analysis program evaluation using the marginal treatment effect," Tech. rep., cemmap working paper. 1.1

Newey, W. and S. Stouli (2021): "Control variables, discrete instruments, and identification of structural functions," *Journal of Econometrics*, 222, 73–88. 1.1, A.3

Newey, W. K. and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578. 1, 1.1

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603. 1.1

R Core Team (2024): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. 20

Torgovitsky, A. (2010): "Identification and Estimation of Nonparametric Quantile Regressions with Endogeneity," . 3.3, 9, A.4, A.4

——— (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 1.1, 3.3, 9

——— (2017): "Minimum distance from independence estimation of nonseparable instrumental variables models," *Journal of Econometrics*, 199, 35–48. 1.1

van der Vaart, A. W. and J. A. Wellner (1996): *Weak convergence*, Springer. F

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. 1.1, 4.2

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341. 3.1, 4.1.1, A.2

——— (2006): "Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results," *The Review of Economics and Statistics*, 88, 578–581. 1.1

Wüthrich, K. (2020): "A Comparison of Two Quantile Models With Endogeneity," *Journal of Business & Economic Statistics*, 38, 443–456. 12, 4.2

Supplemental Appendix to "Estimating Causal Effects of Discrete and Continuous Treatments with Binary Instruments"

A Comparisons to Previous Studies

Here we compare the proposed CI-based identification approach to the literature and show how it can complement existing methods. To simplify the exposition, we will abstract from covariates.

A.1 Rank Similarity Restricts Effect Heterogeneity

In Section 4.2, we compared our approach to Chernozhukov and Hansen (2005)'s IVQR approach based on RS. Unlike CI, IVQR imposes copula invariance assumptions across potential outcomes. Here we illustrate that these restrictions across potential outcomes restrict the treatment effect heterogeneity in an example.

For ease of notation, we consider a slightly different version of the treatment selection equation (3.1), $D_z = 1\{\tilde{V}_z \leq q(z)\}$, where $\tilde{V}_z \mid Z \sim \mathcal{N}(0,1)$ is an alternative normalization such that $q(z) \equiv \Phi^{-1}(\pi(z))$.[24] Suppose that the LATE assumptions hold, $\tilde{V}_1 = \tilde{V}_0 = \tilde{V}$ and $(Y_1, Y_0, \tilde{V}) \perp\!\!\!\perp Z$, and suppose further that $(Y_1, Y_0, \tilde{V})$ are jointly normal with zero means and unit variances, i.e.,

$$\begin{pmatrix} Y_0 \\ Y_1 \\ \tilde{V} \end{pmatrix} \mid Z = z \sim \mathcal{N}_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{01} & \rho_{0V} \\ \rho_{01} & 1 & \rho_{1V} \\ \rho_{0V} & \rho_{1V} & 1 \end{pmatrix} \right).$$

Here, CI holds by construction, and therefore does not restrict treatment effect heterogeneity since ρ_{01} and hence the relationship between Y_1 and Y_0 is unrestricted. By contrast, RS imposes that $\rho_{0V} = \rho_{1V} = \rho_V$, which restrict treatment effect heterogeneity.

To see the last point, note that the joint distribution of the treatment effect $Y_1 - Y_0$ and treatment propensity $\tilde{V}$ is

$$\begin{pmatrix} Y_1 - Y_0 \\ \tilde{V} \end{pmatrix} \mid Z = z \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2(1-\rho_{01}) & \rho_{1V} - \rho_{0V} \\ \rho_{1V} - \rho_{0V} & 1 \end{pmatrix} \right).$$

[24] This is because of the normalization $V_z \mid Z \sim U[0,1]$ and thus $\tilde{V}_z = \Phi^{-1}(V_z) \mid Z \sim \mathcal{N}(0,1)$.

Under RS, $\rho_{1V} - \rho_{0V} = 0$ and therefore $Y_1 - Y_0 \perp\!\!\!\perp \tilde{V}$, that is, there is no selection on gains.[25]

A.2 LATE Monotonicity in Imbens and Angrist (1994)

The LATE framework of Imbens and Angrist (1994) provides conditions for identifying causal effects for the subpopulation of compliers. The compliers are the individuals who react to the instrument so that $D_1 \geq D_0$ almost surely. Compared to the assumptions in Section 3.1, the LATE framework restricts the selection model by setting $V_1 = V_0 = V$ almost surely (Vytlacil, 2002), and relies a stronger joint independence assumption, $(Y_0, Y_1, V) \perp\!\!\!\perp Z$, but does not impose any copula invariance assumptions.

Note that the specification of D_z in (2.1) allows for rich compliance patterns due to the inclusion of two unobservables: V_1 and V_0. For example with binary D, (3.1) is weaker than LATE monotonicity, and Remark 4.1 shows that CI does not imply $V_0 = V_1$. Consequently, we can generate any compliance patterns from the joint distribution of (V_1, V_0):

$$\begin{aligned}
\Pr[D_1 = 1, D_0 = 1] &= \Pr[V_1 \leq \pi(1), V_0 \leq \pi(0)] \\
\Pr[D_1 = 0, D_0 = 1] &= \Pr[V_1 > \pi(1), V_0 \leq \pi(0)] \\
\Pr[D_1 = 1, D_0 = 0] &= \Pr[V_1 \leq \pi(1), V_0 > \pi(0)] \\
\Pr[D_1 = 0, D_0 = 0] &= \Pr[V_1 > \pi(1), V_0 > \pi(0)]
\end{aligned}$$

In the following, we briefly investigate how CI interacts with the LATE assumptions. Suppose we maintain EX′ for ease of discussion. Consider the following assumptions.

Assumption RI$_S$ (Rank Invariance in Selection)**.** *$V_1 = V_0 = V$ almost surely.*

Assumption RS$_S$ (Joint Rank Similarity in Selection)**.** *For $d \in \mathcal{D}$, (Y_d, V_1) and (Y_d, V_0) are identically distributed such that $\rho_{Y_d,V_0}(y,v) = \rho_{Y_d,V_1}(y,v) \equiv \rho_{Y_d,V}(y,v)$.*

Assumption CI″ (CI in Treatment Propensity)**.** *For $d \in \mathcal{D}$, $\rho_{Y_d,V}(y,v) = \rho_{Y_d}(y)$.*

The next proposition shows that these assumptions are sufficient for CI.

Proposition A.1. *Under EX′, RS$_S$ and CI″ imply CI. RI$_S$ implies RS$_S$.*

The proof is straightforward and omitted. RI$_S$ is equivalent to LATE montonicity when D is binary, and it implies RS$_S$. RS$_S$ and CI″ (or RI$_S$ and CI″) being sufficient for CI shows how CI may interact with the LATE assumptions. Note that Proposition A.1 applies not only to the case of binary D but also to discrete ordered and continuous D.

[25] If Y_1 and Y_0 have non-unit variances σ_1^2 and σ_0^2, then RS becomes $\text{Cov}(Y_1 - Y_0, \tilde{V}) = (\sigma_1 - \sigma_0)\rho_V$, which allows for selection on gains, but the relationship between $Y_1 - Y_0$ and $\tilde{V}$ is restricted. In this case, $\text{Cov}(Y_1 - Y_0, \tilde{V}) = \sigma_1\rho_{1V} - \sigma_0\rho_{0V}$ in general.

A.3 Control Function Approach in Imbens and Newey (2009)

For a continuous treatment, Imbens and Newey (2009) considered identification based on a control function approach. A simple version of their model consists of a structural outcome equation, $Y_d = g(d, \varepsilon)$, and a reduced form treatment assignment equation, $D = \tilde{h}(Z, V)$, where V is scalar and $v \mapsto \tilde{h}(\cdot, v)$ is strictly monotone. The main idea is to use $V = F_{D|Z}(D \mid Z)$ as a control function that satisfies $D \perp\!\!\!\perp \varepsilon \mid V$. The latter holds under the assumption that $(\varepsilon, V) \perp\!\!\!\perp Z$, which is what they maintain.[26] The key to their identification approach is the assumption that the support of V conditional on D equals the support of V, which requires a large support of Z.

While they require a scalar unobservable V and strict monotonicity with respect to V, we allow a vector unobservable (V_0, V_1) and impose strict monotonicity with respect to the unobservables in the equation for the *counterfactual* treatment $D_z = h(z, V_z)$; see (3.8). We also do not require Z to have a large variation and allow for binary Z. On the other hand, we assume CI as a trade-off. To avoid the large support assumption of Imbens and Newey (2009), Newey and Stouli (2021) imposed parametric structure on the conditional distribution of Y_d given D and V for extrapolation. Again this assumption is not nested with CI. While their approach relies on parametric structure in a conditional distribution of Y given D and V, CI restricts the dependence structure of Y_d and V_z.

A.4 Conditional Copula Invariance in Torgovitsky (2010)

For a continuous treatment, Torgovitsky (2010) considered identification based on a conditional copula invariance assumption. He assumes that Y_d and D are continuous and considers the model, $Y = m(D, U)$, where $u \mapsto m(d, u)$ is strictly increasing for every d, which implies rank invariance in potential outcomes (RI). The (possibly binary) instrument Z is assumed to be marginally independent of U, $U \perp\!\!\!\perp Z$, which is implied by EX. Moreover, he imposes a weak local dependence assumption between D and Z.

The key condition of Torgovitsky (2010) is the conditional copula invariance assumption. Let $V \equiv F_{D|Z}(D \mid Z)$ and consider $\Pr[U \leq u, D \leq Q_{D|Z}(v \mid z) \mid Z = z]$. Then, the assumption requires that the copula of $(U, D) \mid Z = 1$ is equal to the copula of $(U, D) \mid Z = 0$ (focusing on binary Z). Under $U \perp\!\!\!\perp Z$, this can be written as

$$\tilde{C}(F_U(u), v; 1) = \tilde{C}(F_U(u), v; 0). \tag{A.1}$$

[26] In contrast, we only need "marginal" independence between Z and Y_d and between Z and V_z.

Using Lemma 2.1, these two copulas have the following LGR:

$$\tilde{C}(F_U(u), v; z) = C(F_U(u), v; \rho_{U,D;Z}(F_U(u), v; z)), \quad z \in \{0, 1\}.$$

Hence, the conditional copula invariance assumption (A.1) can be written as

$$\rho_{U,D;Z}(F_U(u), v; 1) = \rho_{U,D;Z}(F_U(u), v; 0). \tag{A.2}$$

Comparing equation (A.2) to CI, we can see that both copula invariance assumptions restrict the dependence of the joint distribution of (Y_d, D) on Z by requiring the correlation parameter not to depend on $Z = z$. Since Torgovitsky (2010) maintains RI, restricting the copula of (U, D) is sufficient. Our identification strategy does not depend on RI such that we need to impose copula invariance restrictions for both potential outcomes. As a trade-off of not assuming RI, we impose CI, which requires that the local correlation parameter is not a function of v.

Overall, our identification results complement Torgovitsky (2010) by showing that copula invariance assumptions also are useful with binary treatments. While the underlying copula invariance assumptions are related, our identification strategy fundamentally differs from Torgovitsky (2010). It accommodates binary and ordered treatments and does not rely on RI.

B Local Representation in Copula

The LGR in Lemma 2.1 is written in terms of the distribution. Alternatively, one can consider local representations in terms of the copula. This can be helpful for understanding copula invariance assumptions as restrictions on the dependence, purged of the effects from the marginals. Section B.1 discusses this point. Section B.2 shows how copulas other than the Gaussian copula can be used for local representation.

B.1 Local Dependence as Implicit Function and CI

The LGR in Lemma 2.1 can be expressed as

$$\tilde{C}(u_1, u_2 \mid z) = C(u_1, u_2; \rho(u_1, u_2; z)), \tag{B.1}$$

where $\tilde{C}(u_1, u_2 \mid z)$ is the conditional copula of (Y_d, V_z) given $Z = z$, that is the joint distribution of $U_1 = F_{Y_d|Z}(Y_d \mid Z)$ and $U_2 = F_{V_z|Z}(V_z \mid Z)$ conditional on $Z = z$, and C is

the Gaussian copula. Note that $U_2 = V_z$ under EX and the normalization $V_z \sim U(0,1)$.

The parameter $\rho(u_1, u_2; z)$ can be viewed as an implicit function in (B.1). For any $z \neq z'$, consider

$$\begin{aligned}\tilde{C}(u_1, u_2 \mid z) - \tilde{C}(u_1, u_2 \mid z') &= C_\rho(u_1, u_2; \tilde{\rho}) \left\{\rho(u_1, u_2; z) - \rho(u_1, u_2; z')\right\} \\ &= \phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \tilde{\rho}) \left\{\rho(u_1, u_2; z) - \rho(u_1, u_2; z')\right\},\end{aligned}$$

where $\tilde{\rho}$ lies between $\rho(u_1, u_2; z)$ and $\rho(u_1, u_2; z')$, and $\phi_2(\cdot, \cdot; \rho)$ is the density of the standard bivariate Gaussian distribution with parameter ρ such that $\phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho) \neq 0$ for $(u_1, u_2) \in (0,1)^2$.

For simplicity, here we maintain EX′ and CI′, namely, $Z \perp\!\!\!\perp (Y_d, V_z)$ and $\rho_{Y_d, V_1}(y, v) = \rho_{Y_d, V_0}(y, v) = \rho_{Y_d}(y)$. To understand CI′, consider the LGR of the (unconditional) copula of (Y_d, V_z):

$$\tilde{C}(u_1, u_2) = C(u_1, u_2; \rho(u_1, u_2)).$$

CI′ is equivalent to $\rho(u_1, u_2) = \rho(u_1)$. Since $\tilde{C}$ and C are differentiable in u_2 almost everywhere in $(0,1)$ (by the definition of copula), so is ρ with respect to u_2 by the implicit function theorem. Then, for $\tilde{C}(u_1 \mid u_2)$ and $C(u_1 \mid u_2)$ being conditional copulas,

$$\begin{aligned}\tilde{C}(u_1 \mid u_2) &= C(u_1 \mid u_2; \rho(u_1, u_2)) + C_\rho(u_1, u_2; \rho(u_1, u_2)) \frac{\partial \rho(u_1, u_2)}{\partial u_2} \\ &= C(u_1 \mid u_2; \rho(u_1, u_2)) + \phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho(u_1, u_2)) \frac{\partial \rho(u_1, u_2)}{\partial u_2}.\end{aligned}$$

We can interpret $\phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho(u_1, u_2)) \frac{\partial \rho(u_1, u_2)}{\partial u_2}$ as the adjustment term that equates the two conditional copulas.[27] In general, the LGR for the joint distribution does *not* imply the same representation for the conditional distribution. Rewrite the equation to have

$$\frac{\partial \rho(u_1, u_2)}{\partial u_2} = \frac{\tilde{C}(u_1 \mid u_2) - C(u_1 \mid u_2; \rho(u_1, u_2))}{\phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho(u_1, u_2))}$$

for $(u_1, u_2) \in (0,1)^2$, which captures a (normalized) deviation from local Gaussianity. Given this result, CI′ with respect to u_2 is equivalent to $\tilde{C}(u_1 \mid u_2) = C(u_1 \mid u_2; \rho(u_1))$. We thus have the following result:

Proposition B.1. *Under EX′, CI′ holds if and only if* $\tilde{C}(u_1 \mid u_2) = C(u_1 \mid u_2; \rho(u_1))$.

[27] Note that $\phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho(u_1, u_2)) \to 0$ as $u_1 \to 1$ or 0, which is consistent with $\tilde{C}(u_1 \mid u_2)$ and $C(u_1 \mid u_2)$ being CDFs (and similarly in the previous case).

Remark B.1 (Stochastic Monotonicity)**.** $\tilde{C}(u_1 \mid u_2) = C(u_1 \mid u_2; \rho(u_1))$ *implies that* $u_2 \mapsto \tilde{C}(u_1 \mid u_2)$ *is monotonic for each* u_1*. This restricts the dependence between* U_1 *and* U_2*. For example, if* U_1 *and* U_2 *are continuous, then the effect of* U_2 *on the* τ*-quantile of* U_1 *cannot change sign with respect to the value of* U_2*, but can change sign with* τ*. Note that if* U_1 *and* U_2 *are jointly normal, then this* τ*-quantile effect cannot change sign with* τ*. More generally, the stochastic monotonicity condition holds if, for example, the conditional distribution has a monotone likelihood ratio. In this sense, the discussion in this section allows us provide a different perspective on stochastic monotonicity discussed in Section 4.1.2.*

Remark B.2. *Under the LGR (B.1), we can show that the equivalence between statistical independence and a restriction on the dependence parameter as an implicit function:*

$$\begin{aligned}(U_1, U_2) \perp\!\!\!\perp Z &\Leftrightarrow \tilde{C}(u_1, u_2 \mid z) - \tilde{C}(u_1, u_2 \mid z') \neq 0 \quad \text{for any } z \neq z' \text{ and } (u_1, u_2) \in (0,1)^2 \\ &\Leftrightarrow \rho(u_1, u_2; Z) = \rho(u_1, u_2) \quad \text{almost surely, for any } (u_1, u_2) \in (0,1)^2.\end{aligned}$$

B.2 Local Representation with Other Copulas

Gaussianity in the local representation in Lemma 2.1 is convenient to introduce identifying assumptions, interpret these assumptions using joint normality as a benchmark of comparison, and to develop estimators. However, Gaussianity is not be essential for the local representation. To illustrate this, we ask what other single-parameter copulas can be used for local representation. In this section, let $C(u_1, u_2; \rho)$ denote the copula on which the local representation is based.

Consider the case of binary D. In showing the full rank of Jacobian in the identification proof that employs Hadamard (1906)'s global inverse function theorem, the following quantity typically arises once the Jacobian is transformed using elementary operations:

$$\frac{C_\rho(F_{d,y}, \pi(z); \rho)}{C_1(F_{d,y}, \pi(z); \rho)} - \frac{C_\rho(F_{d,y}, \pi(z'); \rho)}{C_1(F_{d,y}, \pi(z'); \rho)},$$

where C_ρ and C_1 are the derivatives with respect to ρ and the first argument, respectively. Therefore, any copula that satisfies

$$\frac{C_\rho(F_{d,y}, \pi(z); \rho)}{C_1(F_{d,y}, \pi(z); \rho)} \neq \frac{C_\rho(F_{d,y}, \pi(z'); \rho)}{C_1(F_{d,y}, \pi(z'); \rho)} \tag{B.2}$$

for $\pi(z) \neq \pi(z')$ will yield the full rank Jacobian. Han and Vytlacil (2017) show that any single-parameter copula that follows the ordering of stochastic increasingness with respect to ρ satisfies (B.2). Therefore, among them, a *comprehensive copula* can be a candidate for the

local representation. Let C_L, C_U and C_I denote the lower and upper Frèchet-Hoeffding copula bounds and independent copula. The following Archimedean copulas are such copulas:

Example 1 (Clayton copula)**.**

$$C(u_1, u_2; \rho) = \max\{u_1^{-\rho} + u_2^{-\rho} - 1, 0\}^{-1/\rho}, \quad \rho \in [-1, \infty) \backslash \{0\}$$

and $C \to C_L$ *when* $\rho \to -1$, $C \to C_U$ *when* $\rho \to \infty$, *and* $C \to C_I$ *when* $\rho \to 0$.

Example 2 (Frank copula)**.**

$$C(u_1, u_2; \rho) = -\frac{1}{\rho} \ln \left(1 + \frac{(e^{-\rho u_1} - 1)(e^{-\rho u_2} - 1)}{e^{-\rho} - 1} \right), \quad \rho \in (-\infty, \infty) \backslash \{0\}$$

and $C \to C_L$ *when* $\rho \to -\infty$, $C \to C_U$ *when* $\rho \to \infty$, *and* $C \to C_I$ *when* $\rho \to 0$.

Next, consider the case of continuous D. Following Proposition 1 of Anjos and Kolev (2005), consider the following local representation for an arbitrary copula $\tilde{C}(u_1, u_2)$

$$\tilde{C}(u_1, u_2) = C(u_1, u_2; \rho(u_1, u_2)) \equiv u_1 u_2 + \rho(u_1, u_2) \sqrt{u_1 u_2 (1 - u_1)(1 - u_2)}. \tag{B.3}$$

Then, the local dependence function satisfies

$$\rho(u_1, u_2) = \frac{\tilde{C}(u_1, u_2) - u_1 u_2}{\sqrt{u_1 u_2 (1 - u_1)(1 - u_2)}},$$

which captures the normalized deviation from the independent copula. In fact, $\rho(u_1, u_2)$ can be interpreted as a local Spearman correlation coefficient that satisfies many interesting properties (Anjos and Kolev, 2005, Sections 3.1 and 3.2). For the identification analysis, by the properties of the conditional distribution, (B.3), EX, and CI applied to the copula in (B.3), that is $\rho_{Y_d, V_z; Z}(y, v; z) = \rho_{Y_d}(y)$,

$$\begin{aligned} F_{Y|D,Z}(y \mid d, z) &= C_2(F_{Y_d}(y), F_{D|Z}(d \mid z); \rho_{Y_d}(y)) \\ &= F_{Y_d}(y) + \frac{\rho_{Y_d}(y)}{2} w(d, z) \sqrt{F_{Y_d}(y)(1 - F_{Y_d}(y))}. \end{aligned} \tag{B.4}$$

where

$$w(d, z) \equiv \frac{1 - 2F_{D|Z}(d \mid z)}{\sqrt{F_{D|Z}(d \mid z)(1 - F_{D|Z}(d \mid z))}}$$

Note that

$$\frac{F_{Y|D,Z}(y \mid d,1) - F_{Y_d}(y)}{w(d,1)} = \frac{F_{Y|D,Z}(y \mid d,0) - F_{Y_d}(y)}{w(d,0)},$$

which implies that

$$F_{Y_d}(y) = \frac{F_{Y|D,Z}(y \mid d,1)w(d,0) - F_{Y|D,Z}(y \mid d,0)w(d,1)}{w(d,0) - w(d,1)},$$

Then, (B.4) gives

$$\rho_{Y_d}(y) = 2\frac{F_{Y|D,Z}(y \mid d,z) - F_{Y_d}(y)}{w(d,z)\sqrt{F_{Y_d}(y)(1-F_{Y_d}(y))}}.$$

C Identification with Multi-Valued IVs

When there are multiple instruments and/or there is an instrument that takes more than two values, Assumption CI only needs to hold for two values of one instrument. Let $\boldsymbol{Z} \equiv (Z_1, ..., Z_K)$ be the vector of binary IVs, that is, $Z_k \in \{0,1\}$ for $k = 1, ..., K$. This vector might arise from having multiple binary instruments or constructing indicators from multi-valued instruments.[28] We focus on the binary treatment case. Define a selection equation

$$D_{\boldsymbol{z}} = 1\{V_{\boldsymbol{z}} \leq \pi(\boldsymbol{z})\}, \tag{C.1}$$

where $\pi(\boldsymbol{z}) \equiv \Pr[D = 1 \mid \boldsymbol{Z} = \boldsymbol{z}]$. We make the following assumptions.

Assumption EX2. *For $d, z_k \in \{0,1\}$ for all k, $\boldsymbol{Z} \perp\!\!\!\perp Y_d$ and $\boldsymbol{Z} \perp\!\!\!\perp V_{\boldsymbol{z}}$, where $\boldsymbol{z} = (z_1, \ldots, z_K)$.*

Assumption CI2 (Partial Copula Invariance)**.** *For $d \in \{0,1\}$, $\rho_{d,y}(0, ..., 0, 1) = \rho_{d,y}(0, ..., 0, 0) \equiv \rho^0_{d,y}$, where $\rho_{d,y}(\boldsymbol{z}) \equiv \rho_{Y_d,V_{\boldsymbol{z}}}(y, \pi(\boldsymbol{z}))$.*

Assumption REL2. *(i) $\boldsymbol{Z} \in \{0,1\}^K$; (ii) $0 < \Pr[\boldsymbol{Z} = \boldsymbol{z}] < 1$ and $0 < \Pr[D = d \mid \boldsymbol{Z} = \boldsymbol{z}] < 1$, for $d \in \mathcal{D}$ and $\boldsymbol{z} \in \{(0, ..., 0, 0), (0, ..., 0, 1)\}$; and (iii) $\Pr[D = d \mid \boldsymbol{Z} = (0, ..., 0, 0)] \neq \Pr[D = d \mid \boldsymbol{Z} = (0, ..., 0, 1)]$ for $d \in \mathcal{D}$.*

EX2 is the analog of EX in the multiple instrument case. CI2 can be justified if, conditional on $Z_k = 0$ for $k = 1, ..., K-1$ (i.e., the status quo), Z_K does not shift the joint distribution of $(Y_d, V_{\boldsymbol{z}})$. In general, with the vector of IVs, there will always be only one

[28] Note that Z_k being binary is not essential and we can have discrete or continuous Z_k with different supports across instruments.

more parameter than the number of identifying equations, which is 2^K. CI2 reduces this additional parameter. For illustration, let $K = 2$, that is, consider two binary IVs, Z_1 and Z_2, in $\{0, 1\}$. Then, the resulting equations for $D = 1$ are

$$\begin{aligned}
F_{Y|D,\boldsymbol{Z}}(y \mid 1, (1,1))\pi(1,1) &= C(F_{Y_1}(y), \pi(1,1); \rho_{1,y}(1,1)), \\
F_{Y|D,\boldsymbol{Z}}(y \mid 1, (1,0))\pi(1,0) &= C(F_{Y_1}(y), \pi(1,0); \rho_{1,y}(1,0)), \\
F_{Y|D,\boldsymbol{Z}}(y \mid 1, (0,1))\pi(0,1) &= C(F_{Y_1}(y), \pi(0,1); \rho^0_{1,y}), \\
F_{Y|D,\boldsymbol{Z}}(y \mid 1, (0,0))\pi(0,0) &= C(F_{Y_1}(y), \pi(0,0); \rho^0_{1,y}),
\end{aligned}$$

where there are four unknowns: $(F_{Y_1}(y), \rho_{1,y}(1,1), \rho_{1,y}(1,0), \rho^0_{1,y})$. Then we can show that the corresponding Jacobian has full rank as long as

$$\frac{C_\rho(F_{Y_1}(y), \pi(0,1); \rho^0_{1,y})}{C_1(F_{Y_1}(y), \pi(0,1); \rho^0_{1,y})} \neq \frac{C_\rho(F_{Y_1}(y), \pi(0,0); \rho^0_{1,y})}{C_1(F_{Y_1}(y), \pi(0,0); \rho^0_{1,y})},$$

which is guaranteed since $\pi(0,1) \neq \pi(0,0)$ by REL2, and Lemma 4.1 in Han and Vytlacil (2017). This identifies $F_{Y_1}(y)$ and $\rho^0_{1,y}$. Then, $\rho_{1,y}(1,1)$ and $\rho_{1,y}(1,0)$ are identified from the first two equations above without additional restrictions. For general K, we can prove that the Jacobian of the corresponding system of equations for $D = 1$ has full rank as long as

$$\frac{C_\rho(F_{Y_1}(y), \pi(0,...,0,1); \rho^0_{1,y})}{C_1(F_{Y_1}(y), \pi(0,...,0,1); \rho^0_{1,y})} \neq \frac{C_\rho(F_{Y_1}(y), \pi(0,...,0,0); \rho^0_{1,y})}{C_1(F_{Y_1}(y), \pi(0,...,0,0); \rho^0_{1,y})},$$

which is guaranteed by $\pi(0,...,0,1) \neq \pi(0,...,0,0)$. Then we identify $2^K \times 1$ vector $(F_{Y_1}(y), \{\rho_{1,y}(\boldsymbol{z}) : \boldsymbol{z}_{-K} \neq \boldsymbol{0}\}, \rho^0_{1,y})$. A desirable aspect is that no matter how large the system is (i.e., how large 2^K is), the proof of full rank always amounts to checking the ratio of copula derivatives between the two groups defined by the last instrument Z_K given $\boldsymbol{Z}_{-K} = (0,...,0)$, the status quo.

This discussion implies the following theorem that gathers the identification result. Let $\mathcal{V}_{\boldsymbol{z}}$ denote the support of $\pi(\boldsymbol{z})$.

Theorem C.1 (Identification Binary Treatment with Multiple Instruments)**.** *Suppose* $D_{\boldsymbol{z}} \in \{0,1\}$ *satisfies* (C.1) *for* $\boldsymbol{z} \in \{0,1\}^K$*. Under EX2, REL2, and CI2, the functions* $y \mapsto F_{Y_d}(y)$ *and* $(y,v) \mapsto \rho_{Y_d,V_{\boldsymbol{z}}}(y,v)$ *are identified on* $y \in \mathcal{Y}$ *and* $(y,v) \in \mathcal{Y} \times \mathcal{V}_{\boldsymbol{z}}$*, respectively, for* $d \in \{0,1\}$*.*

CI2 may become innocuous with large K, because within a finer cell (defined by $\boldsymbol{Z}$), individuals tend to be homogeneous and thus share the same joint distribution of $(Y_1, V_{\boldsymbol{z}})$, justifying the copula invariance. The trade-off is that in this case instruments may be

weak (i.e., $\pi(0,...,0,1) \approx \pi(0,...,0,0)$) for the same reason. Therefore, a large K may not necessarily be preferred. Finally, as discussed in Remark 3.3, if copula invariance holds for more than two values of IVs, we have overidentifying restrictions that can be used to test CI.

D Identification with Covariates

In this section, we repeat the main identification analyses explicitly including covariates. We focus on binary and continuous D; the case with ordered D is analogous. Let $X \in \mathcal{X}$ be a vector of (potentially endogenous) covariates.

Assumption EX3 (Conditional Independence)**.** *For $d \in \mathcal{D}$ and $z \in \{0,1\}$, $Z \perp\!\!\!\perp Y_d \mid X$ and $(Z,X) \perp\!\!\!\perp V_z$.*

Assumption REL3 (Relevance)**.** *(i) $Z \in \{0,1\}$; (ii) $0 < \Pr[Z=1 \mid X] < 1$, almost surely; and (iii) for $\mathcal{D} = \{0,1\}$, $\Pr[D=1 \mid Z=1,X] \neq \Pr[D=1 \mid Z=0,X]$ and $0 < \Pr[D=1 \mid Z=z,X] < 1$ almost surely, for $z \in \{0,1\}$; and, for uncountable $\mathcal{D}$, $F_{D|Z,X}(d \mid 1,X) \neq F_{D|Z,X}(d \mid 0,X)$ and $0 < F_{D|Z,X}(d \mid z,X) < 1$ almost surely, for $(z,d) \in \{0,1\} \times \text{int}(\mathcal{D})$.*

Assumption CI3 (Conditional Copula Invariance)**.** *For $d \in \mathcal{D}$, $\rho_{Y_d,V_z;Z,X}(y,v;z,x)$ is a constant function of (v,z), that is*

$$\rho_{Y_d,V_z;Z,X}(y,v;z,X) = \rho_{Y_d;X}(y;X), \quad (y,v,z) \in \mathcal{Y} \times \mathcal{V} \times \{0,1\},$$

and $\rho_{Y_d;X}(y;X) \in (-1,1)$, almost surely.

Note that copula invariance is allowed to hold conditional on covariates. Therefore, we allow for observed heterogeneity in the dependence structure.

In the following subsections, we show the identifiability of $F_{d,y}(x) \equiv F_{Y_d|X}(y \mid x)$, from which we can construct conditional parameters:

$$QSF_\tau(d;x) \equiv Q_{Y_d|X}(\tau|x) = \mathcal{Q}_\tau(F_{Y_d|X}(\cdot|x)), \quad ASF(d;x) \equiv E[Y_d|X=x] = \mathcal{E}(F_{Y_d|X}(\cdot|x)).$$

Marginal QSF_τ and ASF are also identified from

$$F_{Y_d}(y) = \int F_{Y_d|X}(y \mid x) \mathrm{d}F_X(x),$$

where F_X is the distribution of X.

Remark D.1. *CI3 and CI2 are complementary. Which one to impose depends on the plausibility in given applications. On the one hand, CI3 imposes invariance for every subgroup defined by $X = x$, whereas CI2 imposes invariance for a single subgroup defined by $\boldsymbol{Z}_{-K} = (0, ..., 0)$. On the other hand, CI2 imposes stronger exclusion restrictions.*

D.1 Binary Treatment

Define a selection equation

$$D_z = 1\{V_z \leq \pi(z, X)\}, \quad V_z \sim U(0, 1), \tag{D.1}$$

where $\pi(z, x) \equiv \Pr[D = 1 \mid Z = z, X = x]$. Consider

$$\begin{aligned}\Pr[Y \leq y, D = 1 \mid Z = z, X = x] &= \Pr[Y_1 \leq y, V_z \leq \pi(z, x) \mid Z = z, X = x] \\ &= C(F_{Y_1|X}(y \mid x), \pi(z, x); \rho_{Y_1;X}(y; x)), \quad z \in \{0, 1\},\end{aligned}$$

where the last equation is by EX3 and CI3. Now, let $F_{d,y}(x) \equiv F_{Y_d|X}(y \mid x)$, and $\rho_{d,y}(x) \equiv \rho_{Y_d}(y; x)$. Then, we have the system of two equations

$$\begin{aligned}\Pr[Y \leq y, D = 1 \mid Z = 1, X = x] &= C(F_{1,y}(x), \pi(1, x); \rho_{1,y}(x)), \\ \Pr[Y \leq y, D = 1 \mid Z = 0, X = x] &= C(F_{1,y}(x), \pi(0, x); \rho_{1,y}(x)),\end{aligned}$$

with two unknowns for every $x \in \mathcal{X}$: $(F_{1,y}(x), \rho_{1,y}(x))$. This system has full rank if

$$\frac{C_\rho(F_{1,y}(x), \pi(1, x); \rho_{1,y}(x))}{C_1(F_{1,y}(x), \pi(1, x); \rho_{1,y}(x))} \neq \frac{C_\rho(F_{1,y}(x), \pi(0, x); \rho_{1,y}(x))}{C_1(F_{1,y}(x), \pi(0, x); \rho_{1,y}(x))}$$

for all x, which is guaranteed by REL3.

The following theorem gathers the identification result:

Theorem D.1 (Identification Binary Treatment with Covariates)**.** *Suppose $D_z \in \{0, 1\}$ satisfies (D.1) for $z \in \{0, 1\}$. Under EX3, REL3, and CI3, the functions $(y, x) \mapsto F_{Y_d|X}(y \mid x)$ and $y \mapsto \rho_{Y_d;X}(y; x)$ are identified on $(y, x) \in \mathcal{Y} \times \mathcal{X}$, for $d \in \{0, 1\}$.*

The proof of Theorem D.1 is omitted because it is analogous to the proof of Theorem 3.1.

D.2 Continuous Treatment

Let $F_{d,y}(x) \equiv F_{Y_d|X}(y \mid x)$ and $\pi_d(z,x) \equiv F_{D|Z,X}(d \mid z,x)$. For the generalized selection, assume that $d \mapsto F_{D|Z,X}(d \mid z, X)$ is strictly increasing on $\mathcal{D}$ for $z \in \{0,1\}$, almost surely, and let

$$D_z = h(z, X, V_z) = F^{-1}_{D|Z,X}(V_z \mid z, X), \tag{D.2}$$

where $V_z \sim U(0,1)$, so that $\pi_\cdot(z,x) = h^{-1}(z,x,\cdot)$.

For the identification analysis, consider

$$F_{Y|D,Z,X}(y \mid d,z,X) = F_{Y_d|D_z,Z,X}(y \mid d,z,X) = F_{Y_d|V_z,Z,X}(y \mid \pi_d(z,X),z,X), \tag{D.3}$$

almost surely, where the second equality holds from equation (D.2) and a change of variable. By the properties of the conditional distribution, Lemma 2.1, EX3, properties of the Gaussian copula, and CI3,

$$F_{Y_d|V_z,Z,X}(y \mid v,z,x) = \frac{(\partial/\partial v)F_{Y_d,V_z|Z,X}(y,v \mid z,x)}{(\partial/\partial v)F_{V_z|Z,X}(v \mid z,x)} \equiv \Phi\left(a_{d,y;x} + b_{d,y;x}\Phi^{-1}(\pi_d(z,x))\right)$$

where

$$a_{d,y;x} \equiv \Phi^{-1}(F_{d,y}(x))/\sqrt{1-\rho^2_{d,y}(x)}, \text{ and } b_{d,y;x} \equiv -\rho_{d,y}(x)/\sqrt{1-\rho^2_{d,y}(x)},$$

and $\rho_{d,y}(x) \equiv \rho_{Y_d}(y;x)$.

The argument from here is the same as in the case without covariates and yields:

$$\begin{aligned} a_{d,y;x} &= \frac{\Phi^{-1}(F_{Y|D,Z,X}(y \mid d,0,x))\Phi^{-1}(\pi_d(1,x)) - \Phi^{-1}(F_{Y|D,Z,X}(y \mid d,1,x))\Phi^{-1}(\pi_d(0,x))}{\Phi^{-1}(\pi_d(1,x)) - \Phi^{-1}(\pi_d(0,x))}, \\ b_{d,y;x} &= \frac{\Phi^{-1}(F_{Y|D,Z,X}(y \mid d,1,x)) - \Phi^{-1}(F_{Y|D,Z,X}(y \mid d,0,x))}{\Phi^{-1}(\pi_d(1,x)) - \Phi^{-1}(\pi_d(0,x))}. \end{aligned} \tag{D.4}$$

The following theorem gathers the identification result:

Theorem D.2 (Identification Continuous Treatment with Covariates)**.** *Suppose D_z, $z \in \{0,1\}$, satisfies (D.2). Under EX3, REL3, and CI3, the functions $(y,x) \mapsto F_{Y_d|X}(y \mid x)$ and $(y,x) \mapsto \rho_{Y_d;X}(y;x)$ are identified on $(y,x) \in \mathcal{Y} \times \mathcal{X}$, for $d \in \mathcal{D}$ by*

$$F_{Y_d|X}(y \mid x) = \Phi\left(\frac{a_{d,y;x}}{\sqrt{1+b^2_{d,y;x}}}\right), \quad \rho_{Y_d;X}(y;x) = \frac{-b_{d,y;x}}{\sqrt{1+b^2_{d,y;x}}},$$

where $a_{d,y;x}$ and $b_{d,y;x}$ are defined in (D.4).

Remark D.2 (Marginal Local Dependence Function)**.** *The marginal local dependence function,* $(y, v) \mapsto \varrho_{Y_d,V_z}(y, v)$*, is the correlation function of the LGR of the marginal joint distribution of* (Y_d, V_z)*,*

$$F_{Y_d,V_z}(y, v) = C(F_{Y_d}(y), v; \varrho_{Y_d,V_z}(y, v)).$$

Note that CI conditional on covariates does not imply unconditional CI, that is, $\varrho_{Y_d,V_z}(y, v)$ *might vary with* v *even if* $\rho_{Y_d;X}(y; x)$ *does not. Under conditional EX and conditional CI,* $\varrho_{Y_d,V_z}(y, v)$ *is identified from the nonlinear equation*

$$\int C(F_{Y_d|X}(y \mid x), v; \rho_{Y_d;X}(y; x))dF_x(x) = C(F_{Y_d}(y), v; \varrho_{Y_d,V_z}(y, v)), \tag{D.5}$$

which has a unique solution in $\varrho_{Y_d,V_z}(y, v)$ *because* $\rho \mapsto C(\cdot, \cdot; \rho)$ *is strictly increasing. Note that* $\varrho_{Y_d,V_0} = \varrho_{Y_d,V_1}$ *because the left-hand side does not depend on* z*. The equation (D.5) can be used to construct an analog estimator of* ϱ_{Y_d,V_z} *by plugging-in estimators of* $F_{Y_d|X}$*,* $\rho_{Y_d;X}$*,* F_X *and* F_{Y_d}*.*

E Alternative Identification Strategies

For the case of binary D, we show there can be alternative identification strategies using a version of copula invariance. The analysis can be extended to the ordered treatment case. Here we assume EX and REL and the treatment selection equation $D_z = 1[V_z \leq \pi(z)]$ where $V_z \mid Z = z \sim U[0, 1]$. We consider strategies that use a subpopulation defined by each treatment level separately and strategies that combine the two subpopulations.

E.1 Restrictions Within Treatment Levels

We focus here on the treatment level $d = 1$. A similar analysis follows for $d = 0$. For $y \in \mathcal{Y}$, consider the LGR of the observed probabilities, that is

$$\Pr[Y_1 \leq y, D = 1 \mid Z = z] = C(F_{Y_1}(y), \pi(z); \rho_1(y, \pi(z); z)), \quad z \in \{0, 1\},$$

where $\rho_d(y, \pi(z); z) \equiv \rho_{Y_d,V_z;Z}(y, \pi(z); z)$. The identification problem is that we have two probabilities to identify three parameters: $F_{Y_1}(y)$, $\rho_1(y, \pi(0); 0)$ and $\rho_1(y, \pi(1); 1)$. So far, we have reduced the number of parameters by imposing the condition:

$$\rho_1(y, \pi(0); 0) = \rho_1(y, \pi(1); 1).$$

This restriction was imposed separately for each value of y. However, it is also possible to impose restrictions across values of y. Assume that there exists $y' \in \mathcal{Y}$ be such that $F_{Y_1}(y) \neq F_{Y_1}(y')$ and

$$\rho_1(y, \pi(z); z) = \rho_1(y', \pi(z); z), \quad z \in \{0, 1\}. \tag{E.1}$$

This condition leads to the following system of four equations with four unknowns:

$$\begin{aligned}
\Pr[Y_1 \leq y, D = 1 \mid Z = z] &= C(F_{Y_1}(y), \pi(z); \rho_1(y, \pi(z); z)), \quad z \in \{0, 1\}, \\
\Pr[Y_1 \leq y', D = 1 \mid Z = z] &= C(F_{Y_1}(y'), \pi(z); \rho_1(y, \pi(z); z)), \quad z \in \{0, 1\}.
\end{aligned}$$

Then, it is possible to find conditions under which the solution to this system exists and is unique. This condition is appealing in that it does not imposes restrictions across levels of z.

Let C_1, C_2, and C_ρ denote the partial derivative of the Gaussian copula C with respect to the first and second arguments and ρ. The Jacobian of the system of equations is

$$J(y, y') = \begin{pmatrix}
C_1(y, 1) & 0 & C_\rho(y, 1) & 0 \\
C_1(y, 0) & 0 & 0 & C_\rho(y, 0) \\
0 & C_1(y', 1) & C_\rho(y', 1) & 0 \\
0 & C_1(y', 0) & 0 & C_\rho(y', 0)
\end{pmatrix},$$

where $C_j(k, z) := C_j(F_{Y_1}(k), \pi(z); \rho_1(k, \pi(z); z)) > 0$ for $j \in \{1, \rho\}$, $k \in \{y, y'\}$ and $z \in \{0, 1\}$. By the Laplace expansion, the Jacobian determinant is

$$\det(J(y, y')) = C_1(y, 0)C_1(y', 1)C_\rho(y, 1)C_\rho(y', 0) - C_1(y, 1)C_1(y', 0)C_\rho(y', 1)C_\rho(y, 0),$$

which does not vanish if

$$\frac{C_\rho(y, 1)}{C_1(y, 1)} \frac{C_\rho(y', 0)}{C_1(y', 0)} \neq \frac{C_\rho(y, 0)}{C_1(y, 0)} \frac{C_\rho(y', 1)}{C_1(y', 1)}.$$

Let

$$\lambda(k, z) = \frac{\phi\left(u(k, z)\right)}{\Phi\left(u(k, z)\right)}, \quad u(k, z) := \frac{\Phi^{-1}(\pi(z)) - \rho_1(k, \pi(z); z)\Phi^{-1}(F_{Y_1}(k))}{\sqrt{1 - \rho_1(k, \pi(z); z)^2}}.$$

Then, using that $C_\rho(k, z) = \lambda(k, z)C_1(k, z)$, the previous condition can be expressed as

$$\frac{\lambda(y, 1)}{\lambda(y, 0)} \neq \frac{\lambda(y', 1)}{\lambda(y', 0)} \quad \text{or} \quad \frac{\lambda(y, 1)}{\lambda(y', 1)} \neq \frac{\lambda(y, 0)}{\lambda(y', 0)}, \tag{E.2}$$

that is, the change in the conditional inverse Mills ratio from $z = 0$ to $z = 1$ is different at y and y', or the change in the conditional inverse Mills ration from y to y' is different at $z = 0$ and $z = 1$. For example, if the identification condition holds locally for $y' = y + \mathrm{d}y$, then the condition becomes

$$\frac{\partial \log \lambda(y,1)}{\partial y} \neq \frac{\partial \log \lambda(y,0)}{\partial y}.$$

Theorem E.1. *Suppose $D \in \{0,1\}$ satisfies (3.1). Suppose Assumptions EX and REL holds. Given $y \in \mathcal{Y}$, suppose that there exists $y' \in \mathcal{Y}$ such that (E.1) and (E.2) hold. Then, $F_{Y_d}(y)$ and $\rho_{Y_d,V_z;Z}(y, \pi(z); z)$ are identified for $d \in \{0,1\}$ and $z \in \{0,1\}$.*

The proof of this theorem follows from the arguments appearing before the theorem. In general, let d_y be the number of values of Y that we use to construct the system of equations. Then, we have $2d_y$ equations and $3d_y$ unknowns. Therefore, we need to reduce d_y parameters by whichever combinations of copula invariance (E.1) and the alternative assumptions.

E.2 Restrictions Between Treatment Levels

Alternative to the previous subsection, we can impose restrictions involving parameters for different treatment levels. This strategy is based on the system of equations

$$\begin{aligned}
\Pr[Y \leq y, D = 1 \mid Z = z] &= C(F_{Y_1}(y), \pi(z); \rho_1(y, \pi(z); z)), \quad z \in \{0,1\}, \\
\Pr[Y \leq y, D = 0 \mid Z = z] &= C(F_{Y_0}(y), 1 - \pi(z); -\rho_0(y, \pi(z); z)), \quad z \in \{0,1\},
\end{aligned}$$

where again $\rho_d(y, \pi(z); z) \equiv \rho_{Y_d,V_z;Z}(y, \pi(z); z)$.

Assume that the local dependence function is the same across treatment levels,

$$\rho_0(y, \pi(z); z) = \rho_1(y, \pi(z); z), \quad z \in \{0,1\}. \tag{E.3}$$

This condition is similar to the rank similarity condition in (4.2), but does not require the potential outcomes to be continuous. It leads to the following system of four equations with four unknowns:

$$\begin{aligned}
\Pr[Y \leq y, D = 1 \mid Z = z] &= C(F_{Y_1}(y), \pi(z); \rho_1(y, \pi(z); z)), \quad z \in \{0,1\}, \\
\Pr[Y \leq y, D = 0 \mid Z = z] &= C(F_{Y_0}(y), 1 - \pi(z); -\rho_1(y, \pi(z); z)), \quad z \in \{0,1\}.
\end{aligned}$$

Then, it is possible to find conditions under which the solution to this system exists and is unique. Like (E.1), (E.3) is appealing in that it does not imposes restrictions across levels of z.

Let C_1, C_2, and C_ρ denote the partial derivative of the Gaussian copula C with respect to the first and second arguments and ρ. The Jacobian of the system of equations is

$$J = \begin{pmatrix} C_1(1,1) & 0 & C_\rho(1,1) & 0 \\ C_1(1,0) & 0 & 0 & C_\rho(1,0) \\ 0 & \bar{C}_1(0,1) & -\bar{C}_\rho(0,1) & 0 \\ 0 & \bar{C}_1(0,0) & 0 & -\bar{C}_\rho(0,0) \end{pmatrix},$$

where $C_j(d,z) \equiv C_j(F_{Y_d}(y), \pi(z); \rho_d(y, \pi(z); z)) > 0$ and $\bar{C}_j(d,z) \equiv C_j(F_{Y_d}(y), 1-\pi(z); -\rho_d(y, \pi(z); z)) > 0$ for $j \in \{1, \rho\}$, $d \in \{0,1\}$ and $z \in \{0,1\}$. By the Laplace expansion, the Jacobian determinant is

$$\det(J) = C_1(1,0)\bar{C}_1(0,1)C_\rho(1,1)\bar{C}_\rho(0,0) - C_1(1,1)\bar{C}_1(0,0)C_\rho(1,0)\bar{C}_\rho(0,1),$$

which does not vanish if

$$\frac{C_\rho(1,1)}{C_1(1,1)}\frac{\bar{C}_\rho(0,0)}{\bar{C}_1(0,0)} \neq \frac{C_\rho(1,0)}{C_1(1,0)}\frac{\bar{C}_\rho(0,1)}{\bar{C}_1(0,1)}.$$

Let

$$\lambda(d,z) \equiv \frac{\phi\left(u(d,z)\right)}{\Phi\left(u(d,z)\right)} \quad \text{and} \quad \bar{\lambda}(d,z) \equiv \frac{\phi\left(u(d,z)\right)}{\Phi\left(-u(d,z)\right)},$$

with

$$u(d,z) \equiv \frac{\Phi^{-1}(\pi(z)) - \rho_1(y, \pi(z); z)\Phi^{-1}(F_{Y_d}(y))}{\sqrt{1-\rho_1(y, \pi(z); z)^2}}.$$

Then, using that $C_\rho(1,z) = \lambda(1,z)C_1(1,z)$ and $\bar{C}_\rho(0,z) = \bar{\lambda}(0,z)\bar{C}_1(0,z)$, the previous condition can be expressed as

$$\frac{\lambda(1,1)}{\lambda(1,0)} \neq \frac{\bar{\lambda}(0,1)}{\bar{\lambda}(0,0)} \quad \text{or} \quad \frac{\lambda(1,1)}{\bar{\lambda}(0,1)} \neq \frac{\lambda(1,0)}{\bar{\lambda}(0,0)}, \tag{E.4}$$

that is, the change in the conditional inverse Mills ratio from $z = 0$ to $z = 1$ is different at $d = 1$ and $d = 0$, or the change in the conditional inverse Mills ratio from $d = 1$ to $d = 0$ is different at $z = 0$ and $z = 1$.

Theorem E.2. *Suppose $D \in \{0,1\}$ satisfies (3.1). Suppose Assumptions EX and REL hold. Assume also that (E.3) and (E.4) hold for all $y \in \mathcal{Y}$. Then, $y \mapsto F_{Y_d}(y)$ and $y \mapsto \rho_{Y_d,V_z;Z}(y, \pi(z); z)$ are identified on $\mathcal{Y}$, for $d \in \{0,1\}$ and $z \in \{0,1\}$.*

The proof of this theorem follows from the arguments appearing before the theorem.

F Asymptotic Theory

Here we discuss the asymptotic properties of the estimators $\widehat{F}_{Y_d|X}$ and $\widehat{\rho}_{Y_d;X}$ in Algorithms 5.1, 5.2, and 5.3 and the validity of the bootstrap. The asymptotic properties of the estimators of the target parameters in Algorithm 5.4 and the validity of the bootstrap then follow because all the functionals involved are Hadamard differentiable (e.g., van der Vaart and Wellner, 1996; Chernozhukov et al., 2013). Specifically, the functional delta method implies that $\sqrt{n}(\widehat{\delta}_u - \delta_u)$ converges in distribution to a mean-zero Gaussian process Z_δ and the functional delta method for the bootstrap implies that the bootstrap consistently estimates the limiting law (van der Vaart and Wellner, 1996, Chapter 3.9). Note that Hadamard differentiability of the inverse operator fails for discrete and mixed discrete-continuous outcome variables. In this case, one can perform inference on the QSF and QTE using the method proposed by Chernozhukov et al. (2020b).

To state the formal results, we introduce some additional notation. Let $\tilde{\mathcal{Y}}$ be a compact subset of $\mathcal{Y}$ when $\mathcal{Y}$ is uncountable and let $\tilde{\mathcal{Y}}$ be equal to $\mathcal{Y}$ otherwise. Define $\tilde{\mathcal{D}}$ analogously. Let $\mathcal{AB} \equiv \{(a,b) : a \in \mathcal{A}, b \in \mathcal{B}\}$ for sets $\mathcal{A}$ and $\mathcal{B}$. Denote the set of bounded functions on the set $\mathcal{A}$ by $\ell^\infty(\mathcal{A})$. Finally, let $\rightsquigarrow$ denote weak convergence.

Using arguments similar to those in Chernozhukov et al. (2020a), it is straightforward to show that the estimators $\widehat{F}_{Y_d|X}$ and $\widehat{\rho}_{Y_d;X}$ in Algorithms 5.1 and 5.2 satisfy functional central limit theorems (FCLT),

$$\begin{aligned}
\sqrt{n}\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right) &\rightsquigarrow Z_{F_{Y_d|X}(y|x)} \text{ in } \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}), \\
\sqrt{n}\left(\widehat{\rho}_{Y_d;X}(y;x) - \rho_{Y_d;X}(y;x)\right) &\rightsquigarrow Z_{\rho_{Y_d;X}(y;x)} \text{ in } \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}),
\end{aligned}$$

where $Z_{F_{Y_d|X}(y|x)}$ and $Z_{\rho_{Y_d;X}(y;x)}$ are zero-mean Gaussian processes, and that the bootstrap is valid. We therefore focus on the estimators $\widehat{F}_{Y_d|X}$ and $\widehat{\rho}_{Y_d;X}$ in Algorithm 5.3, which have a different structure.

We start by introducing some additional notation. Define the expected Hessians

$$\begin{aligned}
H(y) &\equiv -E[G(B(D,Z,X)'\beta(y))\phi(B(D,Z,X)'\beta(y))B(D,Z,X)B(D,Z,X)'] \\
H(d) &\equiv -E[G(B(Z,X)'\pi(d))\phi(B(Z,X)'\pi(d))B(Z,X)B(Z,X)']
\end{aligned}$$

and let $G(u) \equiv \phi(u)/[\Phi(u)\Phi(-u)]$. We impose two assumptions. The first assumption is essentially a first-stage assumption ensuring that $a_{d,y;x}$ and $b_{d,y;x}$ are well-defined.

Assumption F.1. $(B(1,x) - B(0,x))'\pi(d) \geq c > 0$ *for* $(d,x) \in \mathcal{DX}$.

The second assumption is a version of Condition DR in Chernozhukov et al. (2013). It

ensures that the first-step coefficient estimators $\widehat{\beta}(y)$ and $\widehat{\pi}(d)$ satisfy FCLTs. We present the theoretical results for the case where the outcome is continuously distributed. Results for discrete outcomes can be obtained similarly.

Assumption F.2. *(i) $\mathcal{D}$ is a compact interval in $\mathbb{R}$ and $\mathcal{X}$ is compact. (ii) The conditional densities $f_{Y|D,Z,X}(y \mid d,z,x)$ and $f_{D|Z,X}(y \mid z,x)$ exist and are uniformly bounded and uniformly continuous. (iii) $E\left[\|B(D,Z,X)\|^2\right] < \infty$ and $E\left[\|B(Z,X)\|^2\right] < \infty$. (iv) The minimum eigenvalues of $H(y)$ and $H(d)$ are bounded away from zero over $\mathcal{Y}$ and $\mathcal{D}$.*

The following theorem establishes a FCLT for the estimators $\widehat{F}_{Y_d|X}(y|x)$ and $\widehat{\rho}_{Y_d;X}(y;x)$ in Algorithm 5.3 and the validity of the bootstrap.

Theorem F.1. *Consider the estimators $\widehat{F}_{Y_d|X}(y|x)$ and $\widehat{\rho}_{Y_d;X}(y;x)$ defined in Algorithm 5.3. Suppose that Assumptions F.1 and F.2 hold. Then,*

$$\begin{aligned}\sqrt{n}\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right) &\rightsquigarrow Z_{F_{Y_d|X}(y|x)} \text{ in } \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}),\\ \sqrt{n}\left(\widehat{\rho}_{Y_d;X}(y;x) - \rho_{Y_d;X}(y;x)\right) &\rightsquigarrow Z_{\rho_{Y_d;X}(y;x)} \text{ in } \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}})\end{aligned}$$

where $Z_{F_{Y_d|X}(y|x)}$ and $Z_{\rho_{Y_d;X}(y;x)}$ are zero-mean Gaussian processes defined in the proof in Appendix G.4, and the bootstrap is consistent for estimating the limiting laws.

G Proofs

G.1 Proof of Theorem 3.1

Note that $\pi(z)$ is identified as a reduced-form parameter. Let $F_{d,y} \equiv F_{Y_d}(y)$ and $\rho_{d,y} \equiv \rho_{Y_d}(y)$ be the structural parameters of interest. Consider the following mapping between the structural and reduced-form parameters:

$$F_{Y|D,Z}(y|1,0)\pi(0) = C(F_{1,y}, \pi(0); \rho_{1,y}), \tag{G.1}$$
$$F_{Y|D,Z}(y|1,1)\pi(1) = C(F_{1,y}, \pi(1); \rho_{1,y}), \tag{G.2}$$

which we can express as $\pi_y = G(\theta_y)$, where $\theta_y \equiv (F_{1,y}, \rho_{1,y})'$, $\pi_y \equiv (F_{Y|D,Z}(y|1,0)\pi(0), F_{Y|D,Z}(y|1,1)\pi(1))'$, and $G : (0,1)\times(-1,1) \to (0,1)^2$. Let C_1, C_2 and C_ρ denote the derivative of copula $C(u_1,u_2;\rho)$ with respect to u_1, u_2 and ρ, respectively. Consider the Jacobian of the system of nonlinear equations (G.1)–(G.2):

$$J = \frac{\partial G}{\partial \theta_y} = \begin{bmatrix} C_1(F_{1,y}, \pi(0); \rho_{1,y}) & C_\rho(F_{1,y}, \pi(0); \rho_{1,y}) \\ C_1(F_{1,y}, \pi(1); \rho_{1,y}) & C_\rho(F_{1,y}, \pi(1); \rho_{1,y}) \end{bmatrix}.$$

The matrix has full rank if and only if

$$\frac{C_\rho(F_{1,y}, \pi(1); \rho_{1,y})}{C_1(F_{1,y}, \pi(1); \rho_{1,y})} \neq \frac{C_\rho(F_{1,y}, \pi(0); \rho_{1,y})}{C_1(F_{1,y}, \pi(0); \rho_{1,y})}, \tag{G.3}$$

which is true by Assumption REL and Lemma 4.1 in Han and Vytlacil (2017) as Gaussian copula satisfies the stochastically increasing ordering condition (Assumption 6 in Han and Vytlacil (2017)). Therefore, the matrix is a weak P-matrix with $\rho \in (-1, 1)$. Moreover, the domain of the mapping G (i.e., $(0,1)\times(-1,1)$) is open and rectangular and G is differentiable as $C(u_1, u_2; \rho)$ is differentiable with respect to (u_1, ρ). Therefore, one can apply (Gale and Nikaido, 1965, Theorem 4w)'s global univalence theorem, which identifies θ_y.[29]

Analogously, we have

$$\begin{aligned} F_{Y|D,Z}(y|D=0, Z=z)(1-\pi(z)) &= \Pr[Y_0 \leq y | Z = z] - \Pr[Y_0 \leq y, V_z \leq \pi(z) | Z = z] \\ &= \Pr[Y_0 \leq y | Z = z] - C(F_{Y_0|Z}(y|z), \pi(z); \rho_{Y_0, V_z; Z}(y, \pi(z); z)) \\ &= \Pr[Y_0 \leq y] - C(F_{Y_0}(y), \pi(z); \rho_{Y_0}(y)) \end{aligned}$$

and

$$F_{y|0,0} \cdot (1 - \pi(0)) = F_{Y_0}(y) - C(F_{Y_0}(y), \pi(0); \rho_{0,y}), \tag{G.4}$$

$$F_{y|0,1} \cdot (1 - \pi(1)) = F_{Y_0}(y) - C(F_{Y_0}(y), \pi(1); \rho_{0,y}), \tag{G.5}$$

and the mapping has a unique solution for $\tilde{\theta}_y \equiv (F_{Y_0}(y), \rho_{0,y})'$ by a similar argument as above.

G.2 Proof of Theorem 3.2

Recall from the text that we additionally impose copula invariance between a pair of levels:

$$\rho_{Y_d, V_z; Z}(y, \pi_d(z); z) = \rho_{Y_d, V_z; Z}(y, \pi_{d-1}(z); z) \equiv \rho_{Y_d}(y) \equiv \rho_{d,y}.$$

[29] In cases where we combine more equations, the principle minors of the resulting Jacobian may be zero. In that case, Hadamard's global inverse function theorem can be applied instead. According to Hadamard's theorem (Hadamard, 1906), the solution of $\pi_y = G(\theta_y)$ is unique if (i) G is proper, (ii) the Jacobian of G vanishes nowhere, and (iii) $G(\Theta_y)$ is simply connected. Condition (i) trivially holds with our definition of G. Since the parameter space $\Theta_y = (0,1) \times (-1,1)$ is simply connected and G is continuous, Condition (iii) holds if the Jacobian of G is positive or negative semi-definite on Θ_y because simple connectedness is preserved under a monotone map. We can show that the Jacobian is semidefinite and has full rank, which prove Conditions (iii) and (ii), respectively, and hence the uniqueness of the solution.

Now, following Ambrosetti and Prodi (1995, Corollary 1.4) and De Paula et al. (2019, proof of Theorem 2), we show that (i) the system has a unique solution when $\rho_{d,y} = 0$, (ii) the function that defines the system is continuous and proper with a range that is a connected set, and (iii) it is locally invertible. Note that (ii) is trivially true with our nonlinear map, which is defined in terms of the copula and which range is a Cartesian product of $(0,1)$'s. Note that (i) is trivially true. Therefore, we are remained to prove (iii) by showing the full rank of the following Jacobian with $F_{d,y} \equiv F_{Y_d}(y)$:

$$J_d = \begin{bmatrix} C_1(F_{d,y}, \pi_d(0); \rho_{d,y}) - C_1(F_{d,y}, \pi_{d-1}(0); \rho_{d,y}) & C_\rho(F_{d,y}, \pi_d(0); \rho_{d,y}) - C_\rho(F_{d,y}, \pi_{d-1}(0); \rho_{d,y}) \\ C_1(F_{d,y}, \pi_d(1); \rho_{d,y}) - C_1(F_{d,y}, \pi_{d-1}(1); \rho_{d,y}) & C_\rho(F_{d,y}, \pi_d(1); \rho_{d,y}) - C_\rho(F_{d,y}, \pi_{d-1}(1); \rho_{d,y}) \end{bmatrix}.$$

This Jacobian has full rank if and only if

$$\frac{C_\rho(F_{d,y}, \pi_d(1); \rho_{d,y}) - C_\rho(F_{d,y}, \pi_{d-1}(1); \rho_{d,y})}{C_1(F_{d,y}, \pi_d(1); \rho_{d,y}) - C_1(F_{d,y}, \pi_{d-1}(1); \rho_{d,y})} \neq \frac{C_\rho(F_{d,y}, \pi_d(0); \rho_{d,y}) - C_\rho(F_{d,y}, \pi_{d-1}(0); \rho_{d,y})}{C_1(F_{d,y}, \pi_d(0); \rho_{d,y}) - C_1(F_{d,y}, \pi_{d-1}(0); \rho_{d,y})}. \tag{G.6}$$

Showing this is more involved than showing (G.3) with the binary treatment, because the equality can arise due to two points on the indifference curve. Nonetheless, the full-rank condition (G.6) can be expressed as $\lambda(0) \neq \lambda(1)$, where

$$\lambda(z) \equiv \frac{\phi(r_d(z)) - \phi(r_{d-1}(z))}{\Phi(r_d(z)) - \Phi(r_{d-1}(z))}, \quad r_\ell(z) \equiv \frac{\Phi^{-1}(\pi_\ell(z)) - \rho_{d,y} F_{d,y}}{\sqrt{1 - \rho_{d,y}^2}}.$$

To interpret this condition, we note that it can be related to the mean of truncated Gaussian random variable: for $A \sim N(\mu, \sigma^2)$,

$$E[A \mid l < A < u] = \mu - \left(\frac{\phi\left(\frac{u-\mu}{\sigma}\right) - \phi\left(\frac{l-\mu}{\sigma}\right)}{\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)} \right) \sigma.$$

Therefore, the full-rank condition $\lambda(0) \neq \lambda(1)$ can be equivalently expressed as

$$E[A \mid \pi_{d-1}(0) < \Phi(A) < \pi_d(0)] \neq E[A \mid \pi_{d-1}(1) < \Phi(A) < \pi_d(1)]$$

with $\mu = \rho_{d,y} F_{d,y}$ and $\sigma^2 = 1 - \rho_{d,y}^2$. For example, this holds when threshold functions are such that $\pi_{d-1}(0) < \pi_{d-1}(1)$ and $\pi_d(0) < \pi_d(1)$. By transitivity, Assumption U_{OC} guarantees this.[30]

[30] It is interesting to discuss Heckman and Vytlacil (2007)'s model in comparison to ours using the notation introduced here. The cutoffs of the two models are related as $\pi_\ell(z) = \pi_\ell - \mu(z)$. Under this model, we have

G.3 Proof of Lemma 3.1

Before presenting a formal proof, it is helpful to consider an illustrative example with $K = 4$. In this case we have three complier groups and three defier groups:

$$
\begin{aligned}
C_1 &\equiv \{D_0 = 1, D_1 = 2\} \cup \{D_0 = 2, D_1 = 3\} \cup \{D_0 = 3, D_1 = 4\}, \\
C_2 &\equiv \{D_0 = 1, D_1 = 3\} \cup \{D_0 = 2, D_1 = 4\}, \\
C_3 &\equiv \{D_0 = 1, D_1 = 4\}, \\
B_1 &\equiv \{D_1 = 1, D_0 = 2\} \cup \{D_1 = 2, D_0 = 3\} \cup \{D_1 = 3, D_0 = 4\}, \\
B_2 &\equiv \{D_1 = 1, D_0 = 3\} \cup \{D_1 = 2, D_0 = 4\}, \\
B_3 &\equiv \{D_1 = 1, D_0 = 4\}.
\end{aligned}
$$

Note that the union of C_1, C_2 and C_3 is identical to the multiple-counting union of

$$
\begin{aligned}
C_1 &\equiv \{D_0 = 1, D_1 = 2\} \cup \{D_0 = 2, D_1 = 3\} \cup \{D_0 = 3, D_1 = 4\}, \\
C_2 &\equiv \{D_0 = 1, D_1 = 3\} \cup \{D_0 = 2, D_1 = 4\}, \\
C_3 &\equiv \{D_0 = 1, D_1 = 4\}, \\
C_2 &\equiv \qquad\qquad \{D_0 = 1, D_1 = 3\} \cup \{D_0 = 2, D_1 = 4\}, \\
C_3 &\equiv \qquad\qquad \{D_0 = 1, D_1 = 4\} \cup \{D_0 = 1, D_1 = 4\}.
\end{aligned}
$$

$r_\ell(z) = \frac{\pi_\ell - \mu(z) - \rho_{d,y} F_{d,y}}{\sqrt{1-\rho_{d,y}^2}}$, which is particularly easy-to-interpret because $r_d(z) - r_{d-1}(z) = \frac{\pi_d - \pi_{d-1}}{\sqrt{1-\rho_{d,y}^2}}$. On the other hand, in the general model with the normalization $V_z \mid Z \sim N(0,1)$, we have $r_d(z) - r_{d-1}(z) = \frac{\pi_d(z) - \pi_{d-1}(z)}{\sqrt{1-\rho_{d,y}^2}}$. Note that in the simplified model, the full-rank condition holds by construction. Specifically, since $r_d(z) - r_{d-1}(z)$ does not depend on z, we can write $r_d(z) = r_{d-1}(z) + c$ for $c > 0$. Therefore, we can rewrite $\lambda(z)$ as $\frac{\phi(r_{d-1}(z)+c) - \phi(r_{d-1}(z))}{\Phi(r_{d-1}(z)+c) - \Phi(r_{d-1}(z))}$. This function is monotonically decreasing in $r_{d-1}(z)$. Thus, as long as $\mu(1) \neq \mu(0)$ so that $r_{d-1}(1) \neq r_{d-1}(0)$, the full rank condition holds.

Then, by taking the union in each column of above expression, we have

$$\begin{aligned}
\Pr\left[\bigcup_{j=1}^{3} C_j\right] &= \Pr[\{0 < V_0 \leq \pi_1(0), \pi_1(1) < V_1 \leq 1\} \\
&\qquad \cup \{0 < V_0 \leq \pi_2(0), \pi_2(1) < V_1 \leq 1\} \\
&\qquad \cup \{0 < V_0 \leq \pi_3(0), \pi_3(1) < V_1 \leq 1\}] \\
&= \Pr[\{0 < V_1 \leq \pi_1(0), \pi_1(1) < V_0 \leq 1\} \\
&\qquad \cup \{0 < V_1 \leq \pi_2(0), \pi_2(1) < V_0 \leq 1\} \\
&\qquad \cup \{0 < V_1 \leq \pi_3(0), \pi_3(1) < V_0 \leq 1\}] \\
&< \Pr[\{0 < V_1 \leq \pi_1(1), \pi_1(0) < V_0 \leq 1\} \\
&\qquad \cup \{0 < V_1 \leq \pi_2(1), \pi_2(0) < V_0 \leq 1\} \\
&\qquad \cup \{0 < V_1 \leq \pi_3(1), \pi_3(0) < V_0 \leq 1\}] \\
&= \Pr\left[\bigcup_{j=1}^{3} B_j\right],
\end{aligned}$$

where the second equality is by Assumption EG and the inequality is by $\pi_d(1) > \pi_d(0)$ for all $d = 1, 2, 3$.

Now, the following is the formal proof of the lemma. Let $\pi_0(z) = 0$ and $\pi_K(z) = 1$ for all z. Then,

$$\begin{aligned}
\Pr\left[\bigcup_{j=1}^{K-1} C_j\right] &= \Pr\left[\bigcup_{j=1}^{K-1} \bigcup_{s+j+1=K}^{s=0} \{\pi_s(0) < V_0 \leq \pi_{s+1}(0), \pi_{s+j}(1) < V_1 \leq \pi_{s+j+1}(1)\}\right] \\
&= \Pr\left[\bigcup_{j=1}^{K-1} \{0 < V_0 \leq \pi_j(0), \pi_j(1) < V_1 \leq 1\}\right] \\
&= \Pr\left[\bigcup_{j=1}^{K-1} \{0 < V_1 \leq \pi_j(0), \pi_j(1) < V_0 \leq 1\}\right] \\
&< \Pr\left[\bigcup_{j=1}^{K-1} \{0 < V_1 \leq \pi_j(1), \pi_j(0) < V_0 \leq 1\}\right] \\
&= \Pr\left[\bigcup_{j=1}^{K-1} B_j\right],
\end{aligned}$$

where the second equality is from the derivation similar to the case of $K = 4$, the third equality is by Assumption EG, and the inequality is by $\pi_d(1) > \pi_d(0)$ for all $d \in \mathcal{D}\backslash\{K\}$. The proof of the opposite direction of inequality is symmetric.

G.4 Proof of Theorem F.1

The proof proceeds in two steps.

Step 1: FCLT for $(\widehat{\beta}(y)', \widehat{\pi}(d)')'$ and bootstrap validity. In this step, we establish FCLTs for $\widehat{\beta}(y)$ and $\widehat{\pi}(d)$ and the validity of the bootstrap, building on Chernozhukov et al. (2013).

Under Assumption F.2, Corollary 5.3 in Chernozhukov et al. (2013) implies that

$$\sqrt{n}(\widehat{\beta}(y) - \beta(y)) = H^{-1}(y)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\Phi(B(D_i, Z_i, X_i)'\beta(y)) - I_i(y)}{\Phi(B(D_i, Z_i, X_i)'\beta(y))\Phi(-B(D_i, Z_i, X_i)'\beta(y))}\phi(B(D_i, Z_i, X_i)'\beta(y))B(D_i, Z_i, X_i) + o_P(1),$$

$$\sqrt{n}(\widehat{\pi}(d) - \pi(d)) = H^{-1}(d)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\Phi(B(Z_i, X_i)'\pi(d)) - J_i(d)}{\Phi(B(Z_i, X_i)'\pi(d))\Phi(-B(Z_i, X_i)'\pi(d))}\phi(B(Z_i, X_i)'\pi(d))B(Z_i, X_i) + o_P(1),$$

and

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\beta}(y)\\ \widehat{\pi}(d)\end{pmatrix} - \begin{pmatrix}\beta(y)\\ \pi(d)\end{pmatrix}\right) \rightsquigarrow \begin{pmatrix}Z_\beta(y)\\ Z_\pi(d)\end{pmatrix} \quad \text{in} \quad \ell^\infty(\tilde{\mathcal{D}}\tilde{\mathcal{Y}})^{\dim(B(D,Z,X))+\dim(B(Z,X))},$$

where $Z_\beta(y)$ and $Z_\pi(d)$ are mean-zero Gaussian processes, and that the bootstrap is valid.

Step 2: FCLT for $\widehat{\rho}_{Y_d;X}(y;x)$ and $\widehat{F}_{Y_d|X}(y|x)$ and bootstrap validity. In this step, we build on Step 1 to establish a FCLT for $\widehat{\rho}_{Y_d;X}(y;x)$ and $\widehat{F}_{Y_d|X}(y|x)$ and the validity of the bootstrap. Because all maps involved are Hadamard differentiable, the result follows from the functional delta method and the functional delta method for the bootstrap.

Under Assumption F.1, by Step 1 and the functional delta method,

$$\sqrt{n}\left(\begin{pmatrix}\hat{a}_{d,y;x}\\ \hat{b}_{d,y;x}\end{pmatrix} - \begin{pmatrix}a_{d,y;x}\\ b_{d,y;x}\end{pmatrix}\right) \rightsquigarrow \begin{pmatrix}Z_{a_{d,y;x}}\\ Z_{b_{d,y;x}}\end{pmatrix} \quad \text{in} \quad \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}})^2,$$

where

$$\begin{aligned}Z_{a_{d,y;x}} = {} & \frac{B(1,x)'\pi(d)B(d,0,x)' - B(0,x)'\pi(d)B(d,1,x)'}{(B(1,x) - B(0,x))'\pi(d)}Z_\beta(y)\\ & + \left(\frac{(B(d,0,x)'\beta(y)B(1,x)' - B(d,1,x)'\beta(y)B(0,x)')\,(B(1,x) - B(0,x))'\pi(d)}{((B(1,x) - B(0,x))'\pi(d))^2}\right.\\ & \left. - \frac{(B(d,0,x)'\beta(y)B(1,x)'\pi - B(d,1,x)'\beta(y)B(0,x)'\pi(d))\,(B(1,x) - B(0,x))'}{((B(1,x) - B(0,x))'\pi(d))^2}\right)Z_\pi(d)\end{aligned}$$

and

$$Z_{b_{d,y;x}} = \frac{(B(d,1,x) - B(d,0,x))'}{(B(1,x) - B(0,x))'\pi(d)} Z_\beta(y) + \frac{(B(d,1,x) - B(d,0,x))'\beta(y)(B(0,x) - B(1,x))'}{((B(1,x) - B(0,x))'\pi(d))^2} Z_\pi(d).$$

Then, by the functional delta method, we have

$$\sqrt{n}(\widehat{\mu}_{d,y;x} - \mu_{d,y;x}) \rightsquigarrow Z_{\mu_{d,y;x}} \quad \text{in} \quad \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}),$$

where

$$Z_{\mu_{d,y;x}} = \frac{1}{\sqrt{1 + b_{d,y;x}^2}} Z_{a_{d,y;x}} - \frac{a_{d,y;x} b_{d,y;x}}{(1 + b_{d,y;x}^2)^{3/2}} Z_{b_{d,y;x}}.$$

Moreover, we have that

$$\sqrt{n}\left(\widehat{\rho}_{Y_d;X}(y;x) - \rho_{Y_d;X}(y;x)\right) \rightsquigarrow Z_{\rho_{Y_d;X}(y;x)} \text{ in } \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}),$$

where

$$Z_{\rho_{Y_d;X}(y;x)} = -\frac{1}{(1 + b_{d,y;x}^2)^{3/2}} Z_{b_{d,y;x}}.$$

Finally, another application of the functional delta method yields

$$\sqrt{n}\left(\widehat{F}_{Y_d|X}(y|x) - F_{Y_d|X}(y|x)\right) \rightsquigarrow \phi(\mu_{d,y;x}) Z_{\mu_{d,y;x}} \equiv Z_{F_{Y_d|X}(y|x)} \quad \text{in} \quad \ell^\infty(\tilde{\mathcal{D}}\mathcal{X}\tilde{\mathcal{Y}}).$$

The validity of the bootstrap follows from the functional delta method for the bootstrap.