Set-Valued Control Functions*

Sukjin Han
School of Economics
University of Bristol
vincent.han@bristol.ac.uk

Hiroaki Kaido
Department of Economics
Boston University
hkaido@bu.edu

January 30, 2025

Abstract

The *control function* approach allows the researcher to identify various causal effects of interest. While powerful, it requires a strong invertibility assumption in the selection process, which limits its applicability. This paper expands the scope of the nonparametric control function approach by allowing the control function to be *set-valued* and derive sharp bounds on structural parameters. The proposed generalization accommodates a wide range of selection processes involving discrete endogenous variables, random coefficients, treatment selections with interference, and dynamic treatment selections. The framework also applies to partially observed or identified controls that are directly motivated from economic models.

Keywords: Control Function, Control Variable, Partial Identification, Spillover Effects, Dynamic Treatment Effects.

*We appreciate feedback from Marinho Bertanha, Andrew Chesher, Ben Deaner, Áureo de Paula, Dongwoo Kim, Pat Kline, Francesca Molinari, Kenichi Nagasawa, Adam Rosen, Andrei Zeleneev, and participants in seminars at Johns Hopkins, Stanford, Berkeley, Yale, UCSC, and UC Davis, UCL, Warwick and workshops at Manchester and Seoul National, as well as at the North American Winter Meeting 2025, the North American Summer Meeting 2024, and the Asian Meeting 2024 of the Econometric Society. We thank Shuowen Chen, Qifan Han, and Zhanyuan Tian for their excellent research assistance. We gratefully acknowledge financial support from NSF grant SES-2018498.

1 Introduction

Endogeneity is the main challenge in conducting causal inference with observational data. The control function (CF) approach has been a valuable tool in addressing endogeneity and recovering various causal parameters. Although this approach originated in parametric models (e.g., Dhrymes, 1970; Heckman, 1979), it has been proven to be a powerful tool for identification and estimation in nonparametric models that allow causal effect heterogeneity. The CF approach constructs control variables V, which define a *latent type* conditional on which endogenous explanatory variables D can be viewed as unconfounded. In observational settings, such V is typically constructed by inverting treatment selection processes so that it is written as a function of observables—thus a *control function*. Many empirical studies build on this insight to construct and utilize control variables (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Kline and Walters, 2016; Card et al., 2019; Abdulkadiroğlu et al., 2020; Bishop et al., 2022). While powerful, this approach relies on the *invertibility* of selection models. For example, in nonparametric triangular models, invertibility requires D to be continuously distributed and the selection equation for D to be strictly monotone in a scalar unobservable variable. This type of restriction is viewed as the most important limitation of the CF approach (Blundell and Powell, 2003). More generally, whether selection models are involved or not, empirical researchers encounter situations where control variables are only partially observed or identified; see below.

This paper expands the scope of the CF approach by allowing the control function to be set-valued. That is, one only needs to know the set of values V takes for each value of observable variables. Formally, a *set-valued control function* $\boldsymbol{V}$ is a random closed set, constructed from observable variables, that contains the true control variable V inside it. When selection processes are involved, this approach allows us to drop invertibility. This adaptation accommodates a wide range of selection processes. Observational data are often generated through complex selection processes. For example, D can be a binary variable generated by a generalized Roy model (Eisenhauer et al., 2015). One can allow for richer heterogeneity by considering a selection model with binary D that violates the local average treatment effect (LATE) monotonicity (Imbens and Angrist, 1994) or, analogously, a model with continuous D with vector unobservables (e.g., selection with random coefficients). Other examples are the cases where D is determined through interaction of multiple agents (Tamer, 2003; Ciliberto et al., 2021; Balat and Han, 2023) (e.g., due to violation of the stable-unit treatment value assumption (SUTVA) in forming outcomes); where D and outcomes are dynamically determined over time (Han, 2021, 2023); and where D results from censoring (Manski and Tamer, 2002a) or as corner solutions. Such processes typically violate the invertibility assumption, as the mapping from observables to V is only a correspondence. We show that the CF approach can still be used with these selection processes to partially identify structural (i.e., causal) parameters, such as average and quantile structural functions for outcomes. By allowing control functions to be set-valued, we can also incorporate a wider range of examples that use controls without relying on selection models. Bertanha et al. (2024) control for a set of true preferences using strategic reports, while Auerbach (2022) recovers a control variable from a friendship network. In other examples, controls are simply interval data (e.g., wealth, debt, biometric measures, psychological traits). Our framework can be applied to such scenarios, enabling researchers to conduct sensitivity analyses.

To our knowledge, sharp identifying restrictions under the control function assumption were unknown without requiring the full observability of the control variable. Our innovation is to construct a random set that contains outcome values consistent with the control function assumption by combining a set-valued control function $\boldsymbol{V}$ with an augmented outcome equation. This is a crucial step to formulate the model's incomplete prediction. Utilizing tools from the theory of random sets (Molchanov, 2017), such as the *containment functional* and *Aumann expectation* of the prediction set, we establish restrictions that lead to the sharp identified set for the structural parameters. Assuming conditional full independence (Assumption 1), the identifying restrictions result in inequality constraints on the conditional choice probabilities. If we assume a weaker conditional mean independence (Assumption 3), these restrictions become conditional moment inequality restrictions. Inference methods based on such restrictions have been extensively studied (see e.g. Canay and Shaikh, 2017; Molinari, 2020). Hence, practitioners can apply existing methods to our identifying restrictions. To demonstrate this point, we illustrate our approach by studying individuals' HIV preventive behavior under an informational provision intervention studied by Thornton (2008). This empirical illustration involves an ordered outcome, an endogenous treatment, multiple instruments, and various observed controls. We examine the effectiveness of the intervention by constructing confidence intervals for policy-relevant parameters such as the counterfactual switching probability of individual choices.

This paper contributes to the vast literature on identification and estimation in nonparametric models with endogenous explanatory variables. In linear models, the two-stage least squared (TSLS) estimator can have two different interpretations: the instrumental variable (IV) approach and the CF approach (Blundell and Powell, 2003). The noparametric version of the IV approach is considered in Newey and Powell (2003); Hall and Horowitz (2005); Chernozhukov and Hansen (2005); Darolles et al. (2011); D'Haultfoeuille and Février (2015); Torgovitsky (2015); Vuong and Xu (2017); Chen and Christensen (2018). Typically, this approach assumes invertibility in the outcome equation and thus relies on a scalar unobservable, so that the IV assumption can be utilized. The CF approach is generalized to nonparametric models by Newey et al. (1999); Chesher (2003); Das et al. (2003); Blundell and Powell (2004); Imbens and Newey (2009); D'Haultfœuille et al. (2021); Newey and Stouli (2021); Nagasawa (2024), following the adaptation to nonlinear parametric models in Newey (1987); Rivers and Vuong (1988); Smith and Blundell (1986); Blundell and Smith (1989). The nonparametric CF literature typically assumes a model for endogenous explanatory variables and its invertibility in a scalar unobservable. Then, this approach generates control variables and combines it with the CF assumption (non-nested to the IV assumption) to identify structural parameters. Although this approach restricts selection behavior and is not applicable to discrete treatments, its advantage is the freedom from restricting heterogeneity directly relevant in generating causal effects. Another important strand of the causal inference literature concerns a binary or discrete treatment with a monotonicity assumption (Imbens and Angrist, 1994; Abadie et al., 2002) or equivalently (Vytlacil, 2002) a threshold-crossing model (Heckman and Vytlacil, 2005).

Chesher and Rosen (2017) propose a generalized IV (GIV) framework that allows for partial identification of structural parameters in a range of complete and incomplete models (see also Chesher and Smolinski (2012); Chesher and Rosen (2013)) under stochastic restrictions governing the relationship between the latent variables and instrumental variables. Chesher and Rosen (2017) applies their analysis to single-equation IV models, extending the IV approach to partial identification. This paper proposes the CF approach to partial identification, filling the gap in the literature. Sharing the aspect of the CF literature above, we allow for arbitrary causal effect heterogeneity (e.g., multi-dimensional outcome unobservables). Overcoming the aspect of the CF literature, we accommodate discrete treatments along with heterogeneity and complexity in treatment selection (e.g., multi-dimensional selection unobservables, incompleteness of selection models).

An incomplete model can often be formulated using one of the following structures: (a) Given covariates $X \in \mathcal{X}$, latent variables $U \in \mathcal{U}$, and parameters $\theta \in \Theta$, the model predicts a set $\boldsymbol{Y}(X, U; \theta) \subseteq \mathcal{Y}$ of values for the outcome $Y \in \mathcal{Y}$; or (b) Given covariates $X \in \mathcal{X}$ and observed outcome $Y \in \mathcal{Y}$ and parameters $\theta \in \Theta$, the model predicts a set $\boldsymbol{U}(X, Y; \theta) \subseteq \mathcal{U}$ for the latent variable $U \in \mathcal{U}$. One can choose how to formulate an incomplete model depending on the objects of interest, assumptions, and observability of variables. For example, many discrete choice models that specify a parametric family for F can be formulated using (a) (Jovanovic, 1989; Galichon and Henry, 2011). Beresteanu et al. (2011) study an extended setting, allowing for solution concepts that involve randomization, e.g., correlated equilibria. Chesher and Rosen (2017) employ (b) to define a set of latent variables whose selection satisfies specific stochastic restrictions. A similar approach is used in more recent work by

Chesher et al. (2023a) for IV Tobit models and Chesher et al. (2023b) in the context of panel data models. Our approach to defining a random set closely relates to the previous work but differs in the following respects. The first step of our approach uses (b) to construct the set-valued CF as a set of unobserved control variables. The second step plugs the set-valued control function into an augmented outcome equation to define a model in structure (a). This hybrid approach allows us to extend the control function approach naturally to the current setting.

Chesher (2005) also considers partial identification without requiring invertibility in selection processes. He assumes that discrete endogenous variables are generated from ordered structure and focuses on local parameters. This paper in contrast focuses on global parameters, while encompassing a range of selection processes including ordered selection. Shaikh and Vytlacil (2011); Jun et al. (2011); Mourifié (2015); Mogstad et al. (2018); Machado et al. (2019); Han and Yang (2024) consider partial identification in nonparametric models without requiring invertibility in selection processes; they consider either a binary treatment generated from threshold-crossing models (equivalently, under the LATE monotonicity) or a discrete treatment with similar restrictions. These models are nested within the class of models we consider, but our distinct features include the generality in selection processes and the use of the CF approach.

2 Setup

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ be the outcome of interest generated according to the following *outcome equation*:

$$Y = \mu(D, X, U), \tag{2.1}$$

where $D \in \mathcal{D} \subseteq \mathbb{R}^{d_D}$ is a vector of endogenous treatment variables, $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ is a vector of covariates, and $U \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$ is a vector of latent variables. All random variables are defined on a complete probability space $(\Omega, \mathfrak{F}, P)$. The *structural function* μ determines the value of the *potential outcome* $Y(d) = \mu(d, X, U)$ that would realize when the endogenous variable is set to $d \in \mathcal{D}$. Many policy-relevant parameters are features of the potential outcome, and hence functionals of μ. Examples are the *average structural function* and the *distributional structural function*: $\text{ASF}(d) \equiv E[\mu(d, X, U)] = E[Y(d)]$ and $\text{DSF}(d) \equiv F_{\mu(d,X,U)} = F_{Y(d)}$, respectively. Other examples are the *policy-relevant structural function* and the *mediated structural function*, defined later.

A vector of control variables $V \in \mathcal{V}$ ($\subseteq \mathbb{R}^{d_V}$ for example) is such that, the assignment of

D becomes independent of U, once we condition on V and the observable covariates X:

$$D \perp U | X, V. \tag{2.2}$$

Such variables allow the researcher to identify various causal parameters without additional parametric assumptions on μ or the distribution of unobservables.

For this approach to work, one needs to express V as a function of observable variables. Suppose D is generated from a selection process and Z are the vector of instrumental variables. A commonly used specification for the selection process is the additive model $D = \Pi(Z) + V$, in which one may express $V = D - \Pi(Z)$ (e.g., Newey et al., 1999). Imbens and Newey (2009) consider a nonseparable system, in which a single endogenous variable is modeled as $D = h(Z, \tilde{V})$ of a vector of instrumental variables Z and a continuously distributed *scalar* latent variable $\tilde{V}$, where h is strictly monotonic in $\tilde{V}$. They show that, under the independence of $(U, \tilde{V})$ and Z, one may use the conditional cumulative distribution function $V = F_{D|Z}(D|Z)$ as a control variable. The key assumption is the invertibility of h in the latent variable, which ensures that there is a one-to-one relationship between $\tilde{V}$ and V.[1]

When D is binary, there are other approaches employed in the literature to use control functions without invertibility. These approaches maintain a scalar unobservable V in the selection and additive separability in the outcome equation and either (i) impose parametric assumptions, such as a parametric distribution of unobservables (Heckman, 1979; Dal Bó et al., 2021); or (ii) introduce the marginal treatment effects (MTE) (Heckman and Vytlacil, 2005) as a control function, where a parametric restriction is imposed on the MTE function to deal with discrete "forcing" variables (Brinch et al., 2017); see Kline and Walters (2019) for related discussions.

The previous approaches rely on either invertibility in the selection or certain parametric restrictions. The parametric restrictions are subject to misspecification and they are often combined with scalar V. The invertibility requirement restricts the form of the selection equation and the dimension of V. When V is continuously distributed, the invertibility also requires D to be continuous, which limits the scope of the control function assumption. Moreover, having vector V is important in allowing for rich heterogeneity in the selection process. For example, a multidimensional V can capture individuals' heterogeneous responses to a determinant of the treatment take-up decision by allowing for both compliers and defiers. We, therefore, aim to remove these restrictions.

Another line of research builds control variables V from observables that are not directly related to selection processes. For instance, Auerbach (2022) utilizes friendship networks to

[1]This approach can be extended to a class of simultaneous equations models that meet a certain separability condition (Blundell et al., 2013; Blundell and Matzkin, 2014).

infer a control for latent student abilities. Our framework can also be employed to construct a set-valued control that relies on less stringent assumptions, thereby enabling the researcher to perform a sensitivity analysis.

2.1 Motivating Examples

A *set-valued control function* $\boldsymbol{V}$ is a random closed set that contains the true control V almost surely and is a function of observable variables. We state this as a formal assumption in the next section (Assumption 2), together with a precise measurability requirement.

Before proceeding, we introduce motivating examples. The examples share the following features. First, they involve control variables, conditional on which the treatment decisions can be viewed as random. Second, they do not allow the researcher to uniquely recover the control variables. Nonetheless, it is possible to construct a set-valued control function. Finally, the above features are related to the fact that the control variable V may be interpreted as structural unobservables in these examples.

We start with examples involving selection processes (Examples 1-2). When the control variable appears as a component in a selection process, the above features can be summarized in the following generalized selection equation:

$$D = \pi(Z, X, V). \tag{2.3}$$

Note that $(D, X, Z) \mapsto V$ is, in general, a correspondence because either $\pi(Z, X, \cdot)$ is not necessarily strictly monotonic or V is not scalar. Therefore, the selection process (2.3) only restricts V to the following set almost surely: $\{v : D = \pi(Z, X, v)\} \subseteq \mathbb{R}^{d_V}$. Motivated by this, we define the set-valued control function as follows in the following two examples:

$$\boldsymbol{V}(D, Z, X; \pi) = \text{cl}\{v : D = \pi(Z, X, v)\} \subseteq \mathbb{R}^{d_V}. \tag{2.4}$$

We define $\boldsymbol{V}$ as a closed set in order to utilize the theory of random sets. In Examples 1-2, we illustrate specific forms of (2.3) and (2.4).

Example 1 (Binary and Censored Treatment Decisions): Let D be a binary treatment that is determined by the selection equation

$$D = 1\{\pi(Z, X) \geq V\}, \tag{2.5}$$

where we normalize $V|X$ to the uniform distribution without loss of generality. The selection equation can be motivated by the generalized Roy model (Eisenhauer et al., 2015). Suppose

$Y = DY(1) + (1-D)Y(0)$ where $Y(d)$ follows

$$Y(d) = \mu(d, X) + U_d \quad \text{for } d = 0, 1. \tag{2.6}$$

We allow the unobservables U_d to be treatment-specific. This makes $U = (U_1, U_0)$ a vector. Let $C = \mu_c(Z, X) + U_c$ be the cost of choosing one alternative over the other, where Z is a vector of variables that shifts the cost but not the outcome.[2] The treatment decision is based on the net surplus S from the treatment:

$$D = 1\{S \geq 0\} = 1\{Y(1) - Y(0) - C \geq 0\}. \tag{2.7}$$

We may write the surplus as $S = \pi(Z, X) - V$, where $\pi(Z, X) = \mu(1, X) - \mu(0, X) - \mu_c(Z, X)$ is the observable part of the surplus, and $V = (U_c - U_1 + U_0)$ is the unobserved part of the surplus. Then, we can express the treatment decision as (2.5). Clearly, V depends on (U_0, U_1).

Suppose we are interested in the causal effect of D on Y. Suppose Z is independent of U given (X, V). Then, (X, V) are valid control variables because D's remaining variation is independent of U conditional on them. What prevents us from applying the existing approach is that we cannot recover V by inverting (2.5) because D is binary. Nonetheless, the model restricts V to the following set almost surely:

$$\boldsymbol{V}(D, Z, X; \pi) = \begin{cases} [0, \pi(Z, X)] & \text{if } D = 1 \\ [\pi(Z, X), 1] & \text{if } D = 0, \end{cases} \tag{2.8}$$

which is a set-valued analog of the control function we may condition on.

The previous specification satisfies the local average treatment effect (LATE) monotonicity, eliminating either compliers or defiers (Imbens and Angrist, 1994; Vytlacil, 2002). Next, we consider a selection model that allows richer compliance types, and thus increased heterogeneity in the population. Suppose the value of the instrument is set to z. Let the potential treatment be

$$D(z) = 1\{\pi(z, X) \geq V_z\} \quad \text{for } z \in \mathcal{Z}. \tag{2.9}$$

The observed treatment is $D = \sum_{z \in \mathcal{Z}} D(z) 1\{Z = z\}$. Suppose Z is binary below. Given

[2] The generalized Roy model above nests the classical Roy model where C is degenerate (Heckman and Honoré, 1990) and the extended Roy model where U_c is degenerate (Heckman and Vytlacil, 2007).

(2.9), both compliers and defiers can have nonzero shares:

$$\{D(0) = 0, D(1) = 1\} = \{V_0 > \pi(0, X), V_1 \le \pi(1, X)\},$$
$$\{D(0) = 1, D(1) = 0\} = \{V_0 \le \pi(0, X), V_1 > \pi(1, X)\}.$$

The observed treatment D satisfies

$$D = 1\{D(0) + (D(1) - D(0))Z \ge 0\} \equiv 1\{\tilde{\pi}(Z, X) + (V_1 - V_0)Z + V_0 \ge 0\}, \tag{2.10}$$

where $\tilde{\pi}(Z, X) = \pi(0, X) + Z(\pi(1, X) - \pi(0, X))$.[3] One may view the last expression as a *random-coefficient* model, in which the individuals respond heterogeneously to interventions to Z (Gautier and Hoderlein, 2011; Kline and Walters, 2019). Our framework allows us to proceed with mild assumptions, which may not be sufficient for point identification as in the previous work. Suppose the outcome Y is generated according to (2.1) and Z is independent of U conditional on (X, V_0, V_1). Then, (X, V_0, V_1) are valid control variables. By (2.10), $V = (V_0, V_1)$ belongs to the following set almost surely:

$$\boldsymbol{V}(D, Z, X; \pi) = \begin{cases} \{(v_0, v_1) : \tilde{\pi}(Z, X) + (1 - Z)v_0 + Zv_1 \ge 0\} & \text{if } D = 1 \\ \{(v_0, v_1) : \tilde{\pi}(Z, X) + (1 - Z)v_0 + Zv_1 \le 0\} & \text{if } D = 0. \end{cases} \tag{2.11}$$

Similar to the binary case is D being a censored decision as a corner solution in an agent's optimization problem. In this case, the generalized selection equation is not invertible and a corresponding set-valued control function can be constructed; see Section B.3 in the Appendix for details. □

Example 2 (Treatment Responses as Vectors): The next example involves a *vector* of treatments generated by either (i) strategic decisions of multiple individuals (Balat and Han, 2023) or (ii) a single agent's dynamic decisions over multiple periods (Han, 2021, 2023; Han and Lee, 2023). Let D be the vector of binary decisions across individuals or the vector of binary treatments and previous outcomes over periods. We are interested in the effect of the entire profile D on an outcome Y. To that end, suppose we have a vector of (individual- or time- specific) IVs, Z, and let $\pi(\cdot)$ be the generalized selection function for D. Note that $\pi(\cdot)$ is not invertible in the corresponding unobservables due to the discreteness of D, as in Example 1. Moreover, in case (i), decisions across individuals can be generated from multiple equilibria and, in case (ii), decisions across time can involve dynamic endogeneity. These aspect further complicates the CF approach. However, we can construct appropriate

[3]Note that $D = ZD(1) + (1 - Z)D(0) = 1\{Z(\pi(1, X) - V_1) + (1 - Z)(\pi(0, X) - V_0) \ge 0\}$.

control variables V and corresponding $\boldsymbol{V}(D, Z; \pi)$ in multi-dimensional spaces. We detail these examples in Sections 5.2-5.3.

The examples above constructed $\boldsymbol{V}$ from selection processes. The next example concerns control variables that are not necessarily generated from selection models. Set-valued control functions $\boldsymbol{V}(Z, X)$ in such settings are functions of variables other observables (Z, X), where Z is not necessarily an instrument. Often, it contains information on the true control V, and some components of Z may be excluded from the outcome equation.

Example 3 (Set-Valued Controls Without Selection): We provide three examples using the notation of our paper. First, control variables V may simply be partially observed via interval or censored measurement. In many administrative data, information such as wealth, debt, biometric measures, and psychological traits is observed as an interval due to limitations in data collection or privacy concerns.[4] In this case, $\boldsymbol{V}$ would be directly obtained from observed intervals and can be expressed as $\boldsymbol{V}(Z, X) = [X_L, X_U]$; see Manski and Tamer (2002b) for a related setting.

In the context of school matching mechanisms, Bertanha et al. (2024) estimate the causal effects of school assignment using students' local preferences as control variables ($V \in \mathcal{V}$ where $\mathcal{V}$ is the set of preference relations) to enable regression-discontinuity comparisons. The key feature of their setup is that, under capacity constraints, students have incentives to misreport their preferences. Based on students' reported partial order of preferences, they recover local preference sets $\boldsymbol{V}(Z, X)$ (with reported preferences Z and test scores X) that contain the true preference V a.s., and subsequently characterize the bounds on the effects of school assignment.

In a social network setting, Auerbach (2022) considers a partial linear model with a nonparametric function ($\lambda(\cdot)$) of social characteristics as an unknown control variable (V), which is seldom identified to be used as a control function. Instead, he proposes to use the link function $\pi(\cdot)$ in a nonparametric link formation model that is identified from the distribution of social links. Then, under the assumption that individuals with similar link functions have similar values of control $\lambda(V)$ (Assumption 3 therein), he identifies the slope parameters. However, one may want to relax this identifying assumption and allow individuals with similar link functions to have values of $\lambda(V)$ with discrepancy bounded by a sensitivity margin. This can be achieved by constructing a set-valued control $\boldsymbol{V} = \{(v, v') : \|\pi_v - \pi_{v'}\|_{L^2} \leq \delta\}$ that contains the latent characteristics of a pair of individuals whose link functions are within a certain distance (δ). Then, we can recover a set of controls ($\lambda(\boldsymbol{V})$) and partially identify the

[4] For example, wealth in the Health and Retirement Survey (HRS) and income in the Current Population Survey (CPS) are measured as intervals.

linear parameters. □

3 Model Prediction

As a preparation for the identification analysis, we formulate the model prediction. We show that our limited knowledge of the unobserved control variable V can be formalized as an *incomplete model.*

Throughout, $V : \Omega \to \mathcal{V}$ is a random element taking values in a Polish space $\mathcal{V}$. First, we assume that V and observable covariates X form control variables.

ASSUMPTION 1: $U|D, X, V \sim U|X, V$.

By Assumption 1, the treatment decision is independent of U once we condition on (X, V). Next, we introduce *random closed sets* and their *measurable selections*, which are commonly used in the partial identification literature (see Molchanov and Molinari, 2018).

DEFINITION 1 (Random Closed Set): *A map* $\boldsymbol{X}$ *from a probability space* $(\Omega, \mathfrak{F}, P)$ *to the family* $\mathcal{F}(\mathbb{E})$ *of closed subsets of a Polish space* $\mathbb{E}$ *is called a* random closed set *if*

$$\boldsymbol{X}^{-}(K) \equiv \{\omega \in \Omega : \boldsymbol{X}(\omega) \cap K \neq \emptyset\} \tag{3.1}$$

is in $\mathfrak{F}$ *for each compact set* $K \subseteq \mathbb{E}$.

DEFINITION 2 (Measurable Selections): *For any random set* $\boldsymbol{X}$*, a measurable selection of* $\boldsymbol{X}$ *is a random element* X *with values in* $\mathbb{E}$ *such that* $X(\omega) \in \boldsymbol{X}(\omega)$ *almost surely. We denote by* $\mathrm{Sel}(\boldsymbol{X})$ *the set of all selections from* $\boldsymbol{X}$*.*

We assume one can construct a set-valued control function in the following sense.

ASSUMPTION 2: *(i) There is a random closed set* $\boldsymbol{V} : \Omega \to \mathcal{F}(\mathcal{V})$ *such that* $V \in \boldsymbol{V}$ *with probability 1; (ii)* $\boldsymbol{V}$ *is a measurable function of observable variables and a parameter* π*.*

The set-valued control function $\boldsymbol{V}$ is a random closed-set constructed from the observables.[5] A leading case would be $\boldsymbol{V}$ generated by a selection equation $D = \pi(Z, X, V)$ where Z is a vector of instrumental variables excluded from μ (e.g., Examples 1-2).[6] However, $\boldsymbol{V}$ can also be generated from other sources with corresponding observables and π (e.g., Example 3). Assumption 2 is agnostic about the genesis of a set-valued control function. The set can

[5] A singleton-valued control function in the literature is a special case of Assumption 2.

[6] In each example of Section 2.1, observe that we use closed intervals or sets to construct random set $\boldsymbol{V}$ that is closed.

depend on an unknown parameter π, which can be infinite-dimensional. In some applications, π can be point identified from additional restrictions. We incorporate such restrictions into our identification framework below. We write $\boldsymbol{V}(D, X, Z; \pi)$ whenever it is useful to show its dependence on (D, X, Z) and π.

Let us discuss Assumptions 1-2 further. In the conventional CF approach, we use the control variable V for two main purposes. First, we use V to adjust for the effects of confounding factors as outlined in Assumption 1. Second, we condition on the subpopulation for which this assumption holds. We may use these properties simultaneously if V is observable or can be recovered from other observable variables. However, in the current scenario, the second property is not available. Therefore, we use $\boldsymbol{V}$ (recovered from other observables ensured by Assumption 2) to condition on a "coarser" subpopulation. Failing to condition on V can result in a loss of identifying power. Nevertheless, our framework enables the researcher to use all information available under the stated assumptions and establish sharp bounds on the parameters of interest.

We represent U as $U = Q(\eta; D, X, V)$ for a measurable function $Q : [0,1]^{d_U} \times \mathcal{D} \times \mathcal{X} \times \mathcal{V} \to \mathcal{U} \subseteq \mathbb{R}^{d_U}$ determined by the conditional distribution of U given (D, X, V) and a random vector $\eta \in \mathbb{R}^{d_U}$, which is independent of (D, X, V) and is uniformly distributed over $[0,1]^{d_U}$. This representation holds generally. To see this, consider an example in which $U = (U_0, U_1)$ is two dimensional as in Example 1. Let $(\eta_0, \eta_1) \sim U[0,1]^2$. One can represent $(U_0, U_1) \sim F_{U|D,X,V}$ sequentially by letting

$$U_0 = Q_0(\eta; D, X, V) \equiv F^{-1}_{U_0|D,X,V}(\eta_0|D, X, V), \tag{3.2}$$

$$U_1 = Q_1(\eta; D, X, V) \equiv F^{-1}_{U_1|U_0,D,X,V}(\eta_1|U_0, D, X, V), \tag{3.3}$$

where for any cumulative distribution function F, $F^{-1}(c) \equiv \inf\{u : F(u) > c\}$, which is the quantile function when F is a continuous distribution.[7]

By Assumption 1, we may drop D from the right-hand side of (3.2)-(3.3)

$$\begin{aligned} U_0 &= Q_0(\eta; X, V) \equiv F^{-1}_{U_0|X,V}(\eta_0|X, V), \\ U_1 &= Q_1(\eta; X, V) \equiv F^{-1}_{U_1|U_0,X,V}(\eta_1|U_0, X, V). \end{aligned}$$

This argument can be generalized to settings with any finite d_U. In general, under Assumption

[7]This sequential transformation is known as the Knothe-Rosenblatt transform (see, e.g., Villani, 2008; Carlier et al., 2010; Joe, 2014).

1, we may drop D from Q's argument and represent U by

$$U = Q(\eta; X, V). \tag{3.4}$$

The map $Q : [0,1]^{d_U} \times \mathcal{X} \times \mathcal{V} \to \mathcal{U} \subseteq \mathbb{R}^{d_U}$ is determined by the conditional distribution of $U|X, V$, which we denote as

$$F \equiv F_{U|X,V}.$$

In (3.4), we may view η as the remaining source of randomness in the potential outcome after controlling for (X, V).

Let $\theta \equiv (\mu, F, \pi)$ collect the structural parameters. Consider the model's prediction given θ. For the moment, suppose we may condition on (X, V). The observed outcome is determined by

$$Y = \mu(D, X, U) = \mu(D, X, Q(\eta; X, V)). \tag{3.5}$$

One can view the right-hand side of (3.5) as an outcome equation augmented by an *adjustment term*, $Q(\eta; X, V)$, which involves the control variable V and a "clean" error term η that is independent of D.[8]

Using the fact that V is a measurable selection of $\boldsymbol{V}$, we define the following random closed set:[9]

$$\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F) \equiv \mathrm{cl}\left\{y \in \mathcal{Y} : y = \mu(D, X, Q(\eta; X, V)), V \in \mathrm{Sel}(\boldsymbol{V})\right\}. \tag{3.6}$$

This set collects all outcome values (and their closure) compatible with the model structure for some unknown control variable V taking values in the set-valued control function $\boldsymbol{V}$. This formulation allows us to capture (i) the role of V as a control variable entering the augmented outcome equation through Q and (ii) model incompleteness due the coarse information provided by $\boldsymbol{V}$. To our knowledge, summarizing the model prediction by a random set in (3.6) is new.

Representing the model's prediction in this way has several advantages. First, $\boldsymbol{Y}$ collects all outcome values given all *observable exogenous variables* $(D, X, \boldsymbol{V})$ and latent variables η. It represents the prediction of an *incomplete model* in the sense of Jovanovic (1989).[10]

[8] This is analogous to an additive model, in which the error term can be decomposed into a control function and an error term that is independent of the treatment.

[9] Lemma 1 in the appendix establishes $\boldsymbol{Y}$ is a well-defined random closed set.

[10] The general formulation of Jovanovic (1989) is characterized by observed endogenous variables y, latent variables η, and a *structure* (ν, ϕ), where ν is the distribution of η, and ϕ is a relation such that $(y, \eta) \in \phi$.

Systematic ways to obtain sharp identifying restrictions in such models have been developed in the partial identification literature, and we apply them in the next section. Second, $\boldsymbol{Y}$ builds on an augmented outcome equation, which often helps derive closed-form bounds; e.g., see the discussion of the next paragraph. Finally, the framework can accommodate both continuous and discrete outcomes. We provide further details in Section 5.

Assumption 1 plays an important role in obtaining identifying restrictions for structural parameters via $\boldsymbol{Y}$. Each measurable selection of $\boldsymbol{Y}$ is represented by the *augmented outcome equation* $Y = \mu(D, X, Q(\eta; X, V))$, which involves two functions with different features: the structural function μ takes D as its argument, while the adjustment term Q excludes D but accounts for (X, V). This separation is possible due to Assumption 1. Since Q does not depend on D, it facilitates recovering structural parameters from the equation $\mu(d, x, Q(\eta, x, v))$. For example, we may represent the potential outcome using the augmented outcome equation: $Y(d) = \mu(d, X, Q(\eta; X, V))$. This allows us to express structural quantities such as the *average conditional response* $E[Y(d)|X = x, V = v]$ by integrating out η

$$E[Y(d)|X = x, V = v] = \int_{[0,1]^{d_U}} \mu(d, x, Q(\eta; x, v))d\eta. \tag{3.7}$$

After characterizing the sharp identification region for θ, we use this property to obtain bounds on various structural functions of interest (see Section 4.2).

4 Identification

Let P_0 be the joint distribution of the observable variables (Y, D, X, Z). Let $\Theta \equiv \mathsf{M} \times \mathsf{F} \times \mathsf{\Pi}$ be the parameter space for $\theta = (\mu, F, \pi)$, which embodies a priori restrictions on the parameter. As discussed earlier, some models provide additional restrictions on π.[11] We let $\mathsf{\Pi}_r(P_0) \subset \mathsf{\Pi}$ be the set of selection parameters satisfying them.

We define the sharp identification region for θ as follows.

Definition 3 (Sharp Identification Region under Full Independence): *The* sharp identification region $\Theta_I(P_0) \subset \mathsf{M} \times \mathsf{F} \times \mathsf{\Pi}_r(P_0)$ *is a subset of* Θ *such that each of its elements* $\theta = (\mu, F, \pi)$ *satisfies the following statements: (i) For any* $Y \sim P_0(\cdot|D, X, Z)$*, one can represent the outcome as* $Y = \mu(D, X, U)$ *for some* U *whose conditional law* F *satisfies Assumption 1 for some* $V : \Omega \to \mathcal{V}$*. (ii) The control variable* V *is a measurable selection of a*

The observable exogenous variables are allowed to shift ϕ. In our setting, the graph of $\boldsymbol{Y}$ corresponds to ϕ. The representation of U in (3.4) allows us to incorporate the structural parameter (μ, F) into the model's incomplete prediction (ϕ in Jovanovic (1989)), whereas the remaining randomness is captured by $\eta \sim U[0,1]^{d_U}$.

[11] Consider Example 1. Under an additional independence assumption $Z \perp V|X$, $\Pi_r(P_0) = \{\pi \in \Pi : \pi(z, x) = P_0(D = 1|Z = z, X = x)\}$.

set-valued control function $\boldsymbol{V}$ satisfying Assumption 2.

The main result (Theorem 1) of this section characterizes $\Theta_I(P_0)$ through inequality restrictions on θ. For this, we introduce the *containment functional* $\mathbb{C}_\theta$ of the random set $\boldsymbol{Y}$. Note that the distribution of η satisfies $F_\eta(\eta|D = d, X = x, Z = z) = F_\eta(\eta|D = d, X = x, V = v) = \eta$. For any closed set $A \subset \mathcal{Y}$ and $(d, x, z) \in \mathcal{D} \times \mathcal{X} \times \mathcal{Z}$, let

$$\mathbb{C}_\theta(A|D = d, X = x, Z = z) \equiv \int_{[0,1]^{d_U}} 1\big\{\boldsymbol{Y}(\eta, d, x, \boldsymbol{V}(d, x, z); \mu, F) \subseteq A\big\} d\eta. \tag{4.1}$$

This functional uniquely determines the distribution of $\boldsymbol{Y}$ (Molchanov, 2017). Recall that η is a random vector distributed uniformly over $[0, 1]^{d_U}$. As such, it is straightforward to compute the right-hand side of (4.1) analytically or by simulation (see Section 5). We use the property of $\mathbb{C}_\theta$ that it characterizes the distribution of *all* measurable selections of $\boldsymbol{Y}$. That is, for any $Y \sim P(\cdot|D, X, Z)$[12],

$$\begin{aligned} &Y \in \boldsymbol{Y}(\eta, d, x, \boldsymbol{V}(d, x, z); \mu, F),\ a.s. \\ &\qquad\Leftrightarrow\ P(Y \in A|d, x, z) \geq \mathbb{C}_\theta(A|d, x, z),\ \forall A \in \mathcal{F}(\mathcal{Y}),\ (d, x, z) - a.s. \end{aligned} \tag{4.2}$$

The inequality restriction in (4.2) is known as *Artstein's inequality* (see, e.g., Molchanov and Molinari, 2018, Theorem 2.13), which relates the distribution of Y with the distribution of $\boldsymbol{Y}$. Artstein's inequality is the central device to derive sharp identifying restrictions in incomplete models. The left-hand side $P_0(Y \in A|D, X, Z)$ of the inequality can be recovered from a large sample of the observable variables (Y, D, X, Z). The right-hand side $\mathbb{C}_\theta(A|D, X, Z)$ can be computed from model primitives. We demonstrate how to do so through examples (see Sections 5.2 and B.2).

The following theorem characterizes the sharp identification region using $\mathbb{C}_\theta$.

Theorem 1: *Suppose Assumptions 1-2 hold. Then, the sharp identification region for the structural parameter $\theta = (\mu, F, \pi)$ is*

$$\Theta_I(P_0) = \{\theta \in \Theta : P_0(Y \in A|D, X, Z) \geq \mathbb{C}_\theta(A|D, X, Z),\ a.s.\ \forall A \in \mathcal{F}(\mathcal{Y}),\ \pi \in \Pi_r(P_0)\}. \tag{4.3}$$

Artstein's inequality is introduced to the identification literature in econometrics by Galichon and Henry (2011) and has been extensively used. As in other work, we use this result to convert the model's set-valued prediction into a system of inequality restrictions that do

[12] This equivalence holds up to an ordered coupling (Molchanov and Molinari, 2018, Chapter 2).

not involve the unobserved control variable V, making the resulting restrictions amenable to estimation. Practitioners can use (4.2) to make inference for the elements of $\Theta_I(P_0)$ or their functions. For example, one may use inference methods for conditional moment inequalities (Andrews and Shi, 2013; Chernozhukov et al., 2013) or likelihood-based inference methods (Chen et al., 2018; Kaido and Molinari, 2024). We provide an empirical illustration utilizing a likelihood-based inference method in Section 6.

Remark 1: For a given (d, x, z), the number of the inequalities in (4.2) is finite as long as $\mathcal{Y}$ is a finite set. Furthermore, it often suffices to impose a subset of inequalities to characterize $\Theta_I(P_0)$. Such a class $\mathcal{A} \subseteq \mathcal{F}(\mathcal{Y})$ is called the *core determining class* (CDC) (Galichon and Henry, 2011). The smallest core determining class only depends on the graph representation of $\boldsymbol{Y}(\cdot, D, X, Z; \mu, F)$ and does not depend on P_0 (Luo and Wang, 2017; Ponomarev, 2022).[13]

If Y is a continuous variable, (4.2) involves infinitely many inequalities.[14] However, it is still possible to obtain a sharp characterization with a finite number of inequalities in a commonly used empirical specification under a weaker conditional mean independence assumption; see Section 4.1.

4.1 Conditional Mean Restrictions

So far, we worked with the control function assumption in the form of conditional independence assumption (Assumption 1). A weaker conditional mean independence assumption is also considered in the literature (see e.g. Newey et al., 1999; Pinkse, 2000). This section explores identifying restrictions that can be obtained from this assumption.

We focus on a scalar outcome Y that is continuously distributed. Let $U \equiv (U_d, d \in \mathcal{D})$ and $U_D = \sum_{d\in\mathcal{D}} U_d 1\{D = d\}$.[15] Consider the following additive model:

$$Y = \mu(D, X) + U_D, \tag{4.4}$$

It nests the linear model $\mu(d, x) = \alpha d + x'\beta$ with scalar unobservable U as a special case.[16] This way, Y is a function of vector U, making this model a special case of (2.1).

Suppose the following assumption holds.

[13] Ponomarev (2022) provides an algorithm based on the connectedness of suitable subgraphs to determine the smallest core determining class. See also Chesher and Rosen (2017); Bontemps and Kumar (2020).

[14] While we do not pursue this here, an alternative approach to deal with continuous variables would be to characterize the sharp identification region through an optimal transport problem instead of Artstein's inequalities (see Li and Henry, 2024).

[15] For continuous D, we may define U_D by $U_D = \int_{\mathcal{D}} U_s d\delta_D(s)$, where δ_D is a Dirac measure at D.

[16] A similar argument can be applied to a nonadditive model $Y = \mu(D, X, U)$ for which U is a scalar and μ is invertible with respect to U. We focus on the additive model only for notational simplicity.

Assumption 3: *For each* $d \in \mathcal{D}$, $E[|U_d|] < \infty$, *and* $E[U_d|D,X,V] = E[U_d|X,V]$, *a.s.*

This is a mean independence analog of Assumption 1. For each $d \in \mathcal{D}$, let $\lambda_d(X,V) \equiv E[U_d|X,V]$ and $\eta_d \equiv U_d - E[U_d|X,V]$. Under Assumption 3, we may write

$$E[Y|D=d, X=x, V=v] = \mu(d,x) + \lambda_d(x,v) \tag{4.5}$$

and $Y = \mu(D,X) + \lambda_d(X,V) + \eta_d$. We note that λ_d is a known function of F and plays the role of the adjustment term similarly to Q. Finally, we assume that U is continuously distributed.

Assumption 4: $U|D,X,V$ *has a strictly positive density with respect to Lebesgue measure on* $\mathbb{R}^{d_U}$ *almost surely.*

This assumption helps us derive tractable identifying restrictions. It can be dropped when $\boldsymbol{Y}$ is interval-valued almost surely. We now define the sharp identification region as follows.

Definition 4 (Sharp Identification Region under Mean Independence): *The* sharp identification region under mean independence $\Theta_I(P_0) \subset \mathsf{M} \times \mathsf{F} \times \Pi_r(P_0)$ *is a set such that each* $\theta = (\mu, F, \pi) \in \Theta_I(P_0)$ *satisfies the following statement: (i) For any* Y *whose conditional mean is* $E_{P_0}[Y|D,X,Z]$, *one can represent the outcome as in* (4.4), *where* U*'s conditional law* F *satisfies Assumptions 3 and 4 for some* $V : \Omega \to \mathcal{V}$. *(ii) The control variable* V *is a measurable selection of a set-valued control function* $\boldsymbol{V}$ *satisfying Assumption 2.*

Let $\eta \equiv (\eta_d, d \in \mathcal{D})$. Define

$$\boldsymbol{Y}(\eta, D, X, Z; \mu, F) \equiv \mathrm{cl}\big\{y \in \mathcal{Y} : y = \mu(D,X) + \lambda_D(X,V) + \eta_D, V \in \mathrm{Sel}(\boldsymbol{V})\big\}. \tag{4.6}$$

Since $Y \in \mathrm{Sel}\,\boldsymbol{Y}$ by Assumption 2 and by (4.6), the observed conditional mean $E_{P_0}[Y|D,X,Z]$ belongs to the set of the conditional mean of measurable selections of $\boldsymbol{Y}(\eta, D, X, Z; \mu, F)$ for some $\theta \in \Theta_I(P_0)$. To use this observation, we introduce the *conditional Aumann expectation* of a random set.

A random closed set $\boldsymbol{X}$ is said to be *integrable* if $\boldsymbol{X}$ has at least one integrable selection. We define the Aumann (or selection) expectation of an integrable random closed set following Beresteanu et al. (2011) and Molinari (2020). For this, we let $\mathrm{Sel}^1(\boldsymbol{X})$ denote the set of integrable selections of $\boldsymbol{X}$.

Definition 5: *The* Aumann expectation *of an integrable random closed set* $\boldsymbol{X}$ *is given* *by*

$$\mathbb{E}[\boldsymbol{X}] \equiv \text{cl}\left\{E[X] :, X \in \text{Sel}^1(\boldsymbol{X})\right\}. \tag{4.7}$$

For each sub σ-algebra $\mathfrak{B} \subset \mathfrak{F}$, the conditional Aumann expectation of X given $\mathfrak{B}$ is the $\mathfrak{B}$-measurable random closed set $\boldsymbol{R} \equiv \mathbb{E}(\boldsymbol{X}|\mathfrak{B})$ such that the family of $\mathfrak{B}$-measurable integrable selections of $\boldsymbol{R}$, denoted $\text{Sel}^1_{\mathfrak{B}}(\boldsymbol{R})$, satisfies

$$\text{Sel}^1_{\mathfrak{B}}(\boldsymbol{R}) \equiv \text{cl}\left\{E[X|\mathfrak{B}] :, X \in \text{Sel}^1(\boldsymbol{X})\right\}. \tag{4.8}$$

where the closure in the right-hand side is taken in L^1.

Using the Aumann expectation, the model's prediction is summarized by $E_{P_0}[Y|D,X,Z] \in \mathbb{E}[\boldsymbol{Y}(\eta,D,X,Z;\mu,F)|D,X,Z]$. This is equivalent to

$$bE_{P_0}[Y|D,X,Z] \le s(b, \mathbb{E}[\boldsymbol{Y}(\eta,D,X,Z;\mu,F)|D,X,Z]),\ b \in \{-1,1\}, \tag{4.9}$$

where $s(b,K) = \sup_{k\in K} bk$ is the *support function* of K. As pointed out in the literature, directly working with the conditional Aumann expectation operator can be computationally demanding. We therefore use Assumption 4 to ensure the convexification property (Molinari, 2020, Theorem A.2.) of $\mathbb{E}[\boldsymbol{Y}(\eta,D,X,Z;\mu,F)|D,X,Z]$. The convexification property allows us to interchange the expectation and support function operations and derive a tractable restriction in the following theorem.

THEOREM 2: *Suppose Assumptions 2-4 hold. Suppose $E_{P_0}[|Y|] < \infty$. Then, the sharp identification region is*

$$\Theta_I(P_0) = \{\theta \in \Theta : \mu(d,x) + \lambda_L(d,x,z) \le E_{P_0}[Y|D=d,X=x,Z=z] \\ \le \mu(d,x) + \lambda_U(d,x,z),\ \pi \in \Pi_r(P_0)\}, \tag{4.10}$$

where

$$\lambda_L(d,x,z) = \inf_{v\in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x,v), \quad \lambda_U(d,x,z) = \sup_{v\in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x,v). \tag{4.11}$$

4.2 Causal and Counterfactual Objects

Based on Theorems 1 or 2, one can construct bounds on functionals of θ. Let $W \equiv (X, V)$ and let F_W be its distribution. Given $\varphi : \mathbb{R} \to \mathbb{R}$, let

$$\kappa(d) \equiv E[\varphi(Y(d))] = \int\int \varphi(\mu(d, x, Q(\eta; w)))d\eta dF_W(w). \tag{4.12}$$

The average and distributional structural functions are special cases of κ.

For the average structural function $\mathrm{ASF}(d) \equiv E[\mu(d, X, U)] = E[Y(d)]$ considered by Blundell and Powell (2003), we may set $\varphi(Y(d)) = Y(d)$, which yields

$$\mathrm{ASF}(d) = \int\int \mu(d, x, u)dF(u|w)dF_W(w) = \int\int \mu(d, x, Q(\eta; w))d\eta dF_W(w). \tag{4.13}$$

The average treatment effect (ATE) is then $\mathrm{ATE}(d, d') = \mathrm{ASF}(d) - \mathrm{ASF}(d')$. For the distributional structural function (Chernozhukov et al., 2020), we may set $\varphi(Y(d)) = 1\{Y(d) \leq y\}$, which gives

$$\begin{aligned}\mathrm{DSF}(y, d) &= \int\int 1\{\mu(d, x, u) \leq y\}dF_U(u|w)dF_W(w)\\ &= \int\int 1\{\mu(d, x, Q(\eta; w)) \leq y\}d\eta dF_W(w).\end{aligned} \tag{4.14}$$

The *quantile structural function* (QSF), the τ-th quantile of $Y(d)$, can be obtained using $\mathrm{QSF}(d) = \mathrm{DSF}^{-1}(\tau, d)$ (Imbens and Newey, 2002).

The following proposition characterizes the identification region for κ.

THEOREM 3: *Suppose the conditions of Theorem 1 or 2 hold. Suppose φ is bounded, and the underlying probability space is non-atomic. Then, the sharp identification region for κ is*

$$\mathfrak{K}_I(d) = \bigcup_{\theta \in \Theta_I(P_0)} [\underline{\kappa}(d; \theta), \overline{\kappa}(d; \theta)], \tag{4.15}$$

where

$$\overline{\kappa}(d; \theta) = E[\sup_{v \in \boldsymbol{V}(D,X,Z;\pi)} \int \varphi(\mu(d, X, Q(\eta; X, v)))d\eta], \tag{4.16}$$

$$\underline{\kappa}(d; \theta) = E[\inf_{v \in \boldsymbol{V}(D,X,Z;\pi)} \int \varphi(\mu(d, X, Q(\eta; X, v)))d\eta], \tag{4.17}$$

and the expectation above is taken with respect to the distribution of (D, X, Z).

The identification region for κ is expressed as a union of intervals. Practically, one may only be interested in the upper and lower end points of $\mathfrak{K}_I(d)$. They are given by the following corollary.

Corollary 1: *Suppose the conditions of Theorem 3 hold. Then, the upper and lower bounds of $\mathfrak{K}_I(d)$ are*

$$\overline{\kappa}(d) = \sup_{\theta\in\Theta_I(P)} E[\sup_{v\in\boldsymbol{V}(D,X,Z;\pi)} \int \varphi(\mu(d,X,Q(\eta;X,v)))d\eta], \tag{4.18}$$

$$\underline{\kappa}(d) = \inf_{\theta\in\Theta_I(P)} E[\inf_{v\in\boldsymbol{V}(D,X,Z;\pi)} \int \varphi(\mu(d,X,Q(\eta;X,v)))d\eta]. \tag{4.19}$$

If F_W is point identified, the upper and lower bounds are

$$\overline{\kappa}(d) = \sup_{\theta\in\Theta_I(P)} E_{\eta,W}[\varphi(\mu(d,X,Q(\eta;W)))] \tag{4.20}$$

$$\underline{\kappa}(d) = \inf_{\theta\in\Theta_I(P)} E_{\eta,W}[\varphi(\mu(d,X,Q(\eta;W)))]. \tag{4.21}$$

Theorem 3 does not presume point identification of F_W because V is not observable, which leads to general bounds in (4.18)-(4.19). In some examples, F_W is point identified even if V itself is unobserved and is not uniquely recovered.[17] If so, for each $\theta \in \Theta_I(P_0)$, $\overline{\kappa}(d;\theta) = \underline{\kappa}(d;\theta)$. This allows us to simplify the bounds as in (4.20)-(4.21).

In addition to the structural parameter (4.12), one can consider a policy that only changes the selection behavior. Suppose a policy sets Z (e.g., a tuition subsidy) to z, and the treatment selection under this policy is $D(z) = \pi(z,X,V)$. The *policy-relevant structural function (PRSF)* would be

$$\kappa(z) \equiv E[\varphi(Y(D(z))] = \int\int \varphi(\mu(\pi(z,w),x,Q(\eta;w)))d\eta dF_W(w). \tag{4.22}$$

The PRSF is related to the policy-relevant treatment effect (PRTE) and marginal PRTE introduced in Heckman and Vytlacil (2005) and Carneiro et al. (2010).

One can consider another related structural function. Suppose $D = (D_1, D_2)$, and let $Y(d_1, d_2)$ denote the counterfactual outcome given (d_1, d_2) and $D_2(d_1)$ denote the counterfactual treatment of D_2 given d_1. Then the *mediated structural function (MSF)* would be

$$\kappa(d_1, d_1') \equiv E[\varphi(Y(d_1, D_2(d_1')))] = \int\int \varphi(\mu(d_1, \pi_2(d_1', z, w), x, Q(\eta;w)))d\eta dF_{Z,W}(z,w), \tag{4.23}$$

[17] In Example 1, the distribution of V is normalized to $U[0,1]$.

where we allow $d_1 \neq d_1'$. The MSF can be used to define the direct causal effect of one treatment and the indirect causal effect mediated by another treatment. This scenario is relevant in Example 2 on strategic interaction (e.g., a player's decision being mediated by the opponent's decision) and on dynamic treatment effects (e.g., a previous treatment being mediated by the previous outcome; Han and Lee, 2023). One can derive bounds on these objects in a similar manner.

5 Applications of the Identification Results

We illustrate the use of Theorems 1 and 2 through examples.[18] We present examples with various selection processes (Sections 5.1-5.3) and an example with an incomplete control (Section 5.4). Appendix B provides further examples.

5.1 Generalized Roy Model with a Continuous Outcome

We revisit Example 1. Let $U \equiv (U_1, U_0)$, and recall that

$$D = 1\{\pi(Z, X) \geq V\};$$

hence U's conditional mean independence from D holds as long as U is mean independent of the instrument Z. Suppose Y satisfies (4.4). Let $\lambda_d(X, V) \equiv E[U_d|X, V]$ for $d \in \mathcal{D}$, and let

$$\lambda_L(d, x, z) = \inf_{v \in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x, v), \quad \lambda_U(d, x, z) = \sup_{v \in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x, v). \tag{5.1}$$

The model's prediction is

$$\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F) = \mathrm{cl}\{y \in \mathcal{Y} : y = \mu(D, X) + \lambda_D(X, V) + \eta_D, V \in \mathrm{Sel}(\boldsymbol{V})\}. \tag{5.2}$$

By Theorem 2, we obtain the following inequalities:

$$E_{P_0}[Y|D = d, X = x, Z = z] \leq \mu(d, x) + \lambda_U(d, x, z) \tag{5.3}$$

$$E_{P_0}[Y|D = d, X = x, Z = z] \geq \mu(d, x) + \lambda_L(d, x, z). \tag{5.4}$$

Rearranging them and taking their intersections across z give the following result.

[18] In the illustrations, we pair continuous outcome variables with the generalized Roy model and strategic treatment decisions. We pair discrete outcomes with other examples. These choices are arbitrary. Theorems 1 and 2 allow the researcher to combine various outcome variable types, selection models, and other sources of controls.

Corollary 2: *Suppose* $E_{P_0}[|Y|] < \infty$. *Suppose* $U_0, U_1|X, Z$ *have a density with respect to Lebesgue measure, and* $E[U_d|Z, X, V] = E[U_d|X, V], d = 0, 1$. *Then,* $\Theta_I(P_0)$ *is the set of parameter values* $\theta = (\mu, F, \pi)$ *such that, for almost all* (d, x, z)*,*

$$\sup_{z\in\mathcal{Z}}\Big\{E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_U(d, x, z)\Big\} \\ \leq \mu(d, x) \leq \\ \inf_{z\in\mathcal{Z}}\Big\{E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_L(d, x, z)\Big\}. \quad (5.5)$$

The identifying restrictions (5.5) take the form of intersection bounds on μ. For each z, $E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_U(d, x, z)$ defines a lower bound on $\mu(d, x)$. Since z is excluded from μ, we can intersect the lower bounds across all values of z. The upper bound is formed similarly. It is worth noting that (5.5) restricts the parameter vector $\theta = (\mu, F, \pi)$ jointly because λ_L, λ_U are functions of (F, π). Therefore, they are also useful for bounding (F, π). Furthermore, if $Z \perp V|X$, π is point identified as the propensity score $\pi(z, x) = P_0(D = 1|Z = z, X = x)$ only using the model of selection. Hence, in this case, (5.5) gives joint restrictions on (μ, F).

The terms λ_U, λ_L can be seen as *adjustment terms* to account for the effects of V. To see this, suppose $\boldsymbol{V}$ is a singleton $\{V(D, X, Z; \pi)\}$ (e.g., because $D = \pi(Z, X) + V$). Then,

$$\lambda_L(d, x, z) = \lambda_U(d, x, z) = \lambda_d(x, z) = E[U_d|X = x, V = v], \quad (5.6)$$

In this case, (5.3)-(5.4) reduce to

$$E[Y|D = d, X = x, Z = z] = \mu(d, x) + E[U_d|X = x, V = v]. \quad (5.7)$$

Hence, it justifies regressing Y on (D, X) with an additive correction term (Newey et al., 1999). This argument works only when $\boldsymbol{V}$ is singleton-valued. In the general setting with a set-valued control function, one can work with the intersection bounds in (5.5).

5.2 Dynamic Treatment Effects

We consider a model of dynamic treatment decisions with imperfect compliance (Robins, 1997; Han, 2021). In the initial period, binary treatment D_1 and binary outcome Y_1 (e.g., the presence of symptoms) are generated according to

$$D_1 = 1\{\pi_1(Z_1, X) \geq V_1\}, \tag{5.8}$$
$$Y_1 = 1\{\mu_1(D_1, X) \geq U_1\}. \tag{5.9}$$

In the next period, the observed treatment status is determined based on the initial treatment and outcome:

$$D_2 = 1\{\pi_2(Y_1, D_1, Z_2, X) \geq V_2\}. \tag{5.10}$$

Finally, the outcome in period 2 is determined by

$$Y_2 = 1\{\mu_2(Y_1, D_1, D_2, X) \geq U_2\}. \tag{5.11}$$

Throughout, U_t and V_t are normalized to $U[0, 1]$ conditional on $X = x$.

Consider the effect of the initial outcome and treatment history $D = (Y_1, D_1, D_2)$ on Y_2. One may be concerned about endogeneity because U_2 may depend on (U_1, V_1, V_2). For example, U_1 and U_2 may share a time invariant component. Another possibility is that U_2 may be related to (V_1, V_2) through the agent's dynamic treatment take-up decisions.

Below, we let $Y \equiv Y_2$, $U \equiv U_2$ and let $V \equiv (U_1, V_1, V_2)$ be unobserved control variables; also let $Z \equiv (Z_1, Z_2)$, and let $\pi \equiv (\mu_1(\cdot), \pi_1(\cdot), \pi_2(\cdot))$. Inspecting the system of selection equations, one can see that the assignment of $D = (Y_1, D_1, D_2)$ is independent of U_2 conditional on (X, V) as long as the instrumental variables Z are independent of U_2.

For notational simplicity, we rewrite (5.11) as

$$Y = 1\{\mu(D, X) \geq U\}, \tag{5.12}$$

and derive the model prediction $\boldsymbol{Y}$ as follows. First, let $U = Q(\eta|X, V) = F^{-1}(\eta|X, V)$. Then,

$$\begin{aligned} Y &= 1\{\mu(D, X) \geq Q(\eta|X, V)\} \\ &= 1\{F(\mu(D, X)|X, V) \geq \eta\} \\ &= 1\{H(D, X, V) \geq \eta\}, \end{aligned} \tag{5.13}$$

where $H(d, x, v) \equiv F(\mu(d, x)|x, v)$ is the average response conditional on (x, v).

Next, the dynamic selection and outcome equations allow us to restrict V to the following set:

$$\boldsymbol{V}(D,Z,X;\pi)=\boldsymbol{V}_{U_1}(D,X;\mu_1)\times\boldsymbol{V}_1(D,Z_1,X;\pi_1)\times\boldsymbol{V}_2(D,Z_2,X;\pi_2), \tag{5.14}$$

where

$$\boldsymbol{V}_{U_1}(D,X;\mu_1)=\begin{cases}[\mu_1(D_1,X),1] & \text{if } Y_1=0\\ [0,\mu_1(D_1,X)] & \text{if } Y_1=1,\end{cases}\quad \boldsymbol{V}_1(D,Z_1,X;\pi_1)=\begin{cases}[\pi_1(Z_1,X),1] & \text{if } D_1=0\\ [0,\pi_1(Z_1,X)] & \text{if } D_1=1,\end{cases}$$

$$\boldsymbol{V}_2(D,Z_2,X;\pi_2)=\begin{cases}[\pi_2(Y_1,D_1,Z_2,X),1] & \text{if } D_2=0\\ [0,\pi_2(Y_2,D_2,Z_2,X)] & \text{if } D_2=1.\end{cases}$$

By the augmented outcome equation (5.13) and (3.6), we obtain the following model prediction:

$$\boldsymbol{Y}(\eta,D,X,\boldsymbol{V};\mu,F)=\begin{cases}\{0\} & \eta>\sup_{V\in\mathrm{Sel}(\boldsymbol{V})}H(D,X,V)\\ \{0,1\} & \inf_{V\in\mathrm{Sel}(\boldsymbol{V})}H(D,X,V)<\eta\le\sup_{V\in\mathrm{Sel}(\boldsymbol{V})}H(D,X,V)\\ \{1\} & \eta\le\inf_{V\in\mathrm{Sel}(\boldsymbol{V})}H(D,X,V).\end{cases} \tag{5.15}$$

This expression exhibits an *incomplete threshold-crossing structure* shown in Figure 1.[19] If η is below the lower threshold, the model predicts $\boldsymbol{Y}=\{1\}$, whereas $\boldsymbol{Y}=\{0\}$ if η is above the upper threshold. The model predicts $\boldsymbol{Y}=\{0,1\}$ if η is between the two thresholds.

The containment functional of $\boldsymbol{Y}$ in (5.15) satisfies

$$\begin{aligned}\mathbb{C}_\theta(\{1\}|D=d,X=x,Z=z)&=F_\eta(\boldsymbol{Y}(\eta,D,X,\boldsymbol{V};\mu,F)\subseteq\{1\}|D=d,X=x,Z=z)\\ &=\inf_{v\in\boldsymbol{V}(d,x,z;\pi)}H(d,x,v),\\ \mathbb{C}_\theta(\{0\}|D=d,X=x,Z=z)&=F_\eta(\boldsymbol{Y}(\eta,D,X,\boldsymbol{V};\mu,F)\subseteq\{0\}|D=d,X=x,Z=z)\\ &=1-\sup_{v\in\boldsymbol{V}(d,x,z;\pi)}H(d,x,v).\end{aligned}$$

[19] The incomplete threshold-crossing structure also appears in semiparametric binary choice models with interval-valued covariates (Manski and Tamer, 2002a). Manski and Tamer's (2002a) model is considerably different from ours; they consider $Y=1\{W'\theta+\delta X^*+\epsilon>0\}$, where W is exogenous, X^* is observed as an interval (i.e. $X^*\in[X_L,X_U]$), $\delta>0$ and ϵ satisfies a quantile independence condition. See also Molinari (2020) (Section 3.1.1) for an extensive discussion of their model.

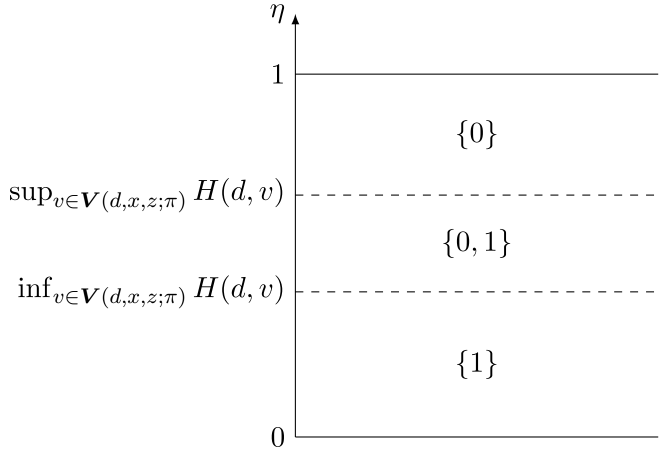


Figure 1: An incomplete threshold-crossing structure.

Note: The figure shows the value of $\boldsymbol{Y}(\eta|d,\boldsymbol{V};\nu,F)$ as a function of η.

Theorem 1 then implies simple identifying restrictions:

$$\inf_{v\in\boldsymbol{V}(d,x,z;\pi)} H(d,x,v) \le P(Y=1|D=d,X=x,Z=z) \le \sup_{v\in\boldsymbol{V}(d,x,z;\pi)} H(d,x,v). \tag{5.16}$$

We may apply this argument sequentially to (5.9)–(5.11). The next step is to take $Y=D_2$ as an outcome, $D=(Y_1,D_1)$ as a treatment, $U=V_2$ as a latent variable in the outcome equation, and $V=(V_1,U_1)$ as control variables, which generates inequalities of the form (5.16). Finally, we may apply the same argument to the outcome and selection equations in period 1. Corollary 4 in the Appendix characterizes the sharp identification region for θ.

5.3 Treatment Responses with Social Interactions

We consider settings where other individuals' treatment status affects one's outcome through spillover or equilibrium effects. Let individuals be indexed by $j=1,\dots,J$. Let $D=(D_1,\dots,D_J)$ be a vector of treatment decisions across individuals and let D_{-j} be the vector D without the element D_j. We are interested in the effect of the entire profile D on some outcome Y. For example, D indicates entries of potential market participants, e.g., airlines, and Y is a market-level outcome, e.g., pollution. Suppose the observed treatments D satisfy

$$D_j = 1\{\pi_j(D_{-j},Z_j,X)\ge V_j\},\ \ j=1,\dots,J, \tag{5.17}$$

where $V_j|X = x$ is normalized to $U[0,1]$.[20]

One way to motivate (5.17) is by relaxing the *Stable Unit Treatment Value Assumption (SUTVA)* (Rubin, 1978) (or equivalently relaxing the *Individualistic Treatment Response (ITR)* by Manski (2013)) and introducing strategic selection. Let $Y_j(d_1, \dots, d_J)$ be the potential outcome of individual j when D is set to $(d_1, \dots, d_J)$. The previous examples assume an individual's outcome only depended on their own treatment (i.e., SUTVA or ITR) that $Y_j(d_1, \dots, d_J) = Y_j(d_j)$. We now allow each individual's outcome to depend on the entire vector of treatments received by the individuals. This generalization is important when treatments are expected to have spillover effects (Graham, 2011; Aronow and Samii, 2017).

For simplicity, consider two individuals. Each faces a binary action d_j. For individual j and $(d_1, d_2) \in \mathcal{D} = \{0,1\}^2$, let

$$Y_j(d_1, d_2) = \mu_j(d_1, d_2, X) + U_{j,d_1,d_2}. \tag{5.18}$$

The observed outcome is generated according to $Y_j = \sum_{(d_1,d_2)\in\mathcal{D}} 1\{D_1 = d_1, D_2 = d_2\} Y_j(d_1, d_2)$. Suppose the individuals are involved in Roy-type decisions:

$$\begin{aligned} D_1 &= 1\{Y_1(1, D_2) - Y_1(0, D_2) \geq \mu_{c1}(Z_j, X) + U_{c1}\}, \\ D_2 &= 1\{Y_2(D_1, 1) - Y_2(D_1, 0) \geq \mu_{c2}(Z_j, X) + U_{c2}\}. \end{aligned}$$

Namely, each individual chooses 1 if the payoff of choosing 1 over 0 weakly exceeds its cost, given the other individual's action. A key difference from the previous examples is the presence of externalities in the selection process.

The selection process is compatible with (5.17) with

$$\begin{aligned} \pi_j(D_{-j}, Z_j, X) &= \mu_j(1, D_{-j}, X) - \mu_j(0, D_{-j}, X) - \mu_{cj}(Z_j, X) \\ V_j &= U_{cj} - U_{j,1,D_{-j}} - U_{j,0,D_{-j}}. \end{aligned}$$

The individuals' social/strategic interaction is captured by the impact of the other individual's treatment status on player j's payoff, which corresponds to $\pi_j(1, z_j, x) - \pi(0, z_j, x)$.

Multiple solutions to the simultaneous equation system (5.17) may exist, which makes the selection process set-valued (Tamer, 2003; Ciliberto et al., 2021; Balat and Han, 2023). For example, suppose the selection process involves strategic substitution, i.e., $\pi_j(1, z_j, x) -$

[20]The joint distribution of $V = (V_1, \dots, V_J)$ is unrestricted.

$\pi(0, z_j, x) \leq 0$ for $j = 1, 2$.[21] The model's prediction is $D \in G(V_1, V_2|Z, X, ; \pi)$ with

$$G(v_1, v_2|z, x; \pi) = \begin{cases} \{(0,0)\} & (v_1, v_2) \in S_{\pi,(0,0)}(z, x) \\ \{(0,1)\} & (v_1, v_2) \in S_{\pi,(0,1)}(z, x) \\ \{(1,0)\} & (v_1, v_2) \in S_{\pi,(1,0)}(z, x) \\ \{(1,1)\} & (v_1, v_2) \in S_{\pi,(1,1)}(z, x) \\ \{(1,0),(0,1)\} & (v_1, v_2) \in S_{\pi,\{(1,0),(0,1)\}}(z, x). \end{cases} \tag{5.19}$$

Figure 2 summarizes the subsets $S_{\pi,(0,0)}(z, x), \dots, S_{\pi,\{(1,0),(0,1)\}}(z, x)$. This model differs from other examples because the selection process itself is incomplete.

We construct a generalized selection equation as in (2.3) by introducing a random variable $V_s : \Omega \to \{0, 1\}$ representing an unknown *selection mechanism*. Without loss of generality, suppose that the treatment status $D = (1, 0)$ ($D = (0, 1)$) is selected when $V_s = 1$ ($V_s = 0$) and multiple values of D are predicted.[22] We do not impose any restrictions on the distribution of V_s, reflecting the researcher's agnosticism against the selection. We then let $V = (V_1, V_2, V_s)$ be a vector of controls.

Suppose that Z is independent of $U = (U_{j,d_1,d_2}, U_{c_j})_{(d_1,d_2)\in\mathcal{D}, j=1,2}$ given the control variables (X, V_1, V_2, V_s). The set-valued control function can be constructed as follows:

$$\boldsymbol{V}(D, Z, X; \pi) = \begin{cases} S_{\pi,(0,0)}(Z, X) \times \{0, 1\} & \text{if } D = (0, 0) \\ \left[S_{\pi,(0,1)}(Z, X) \times \{0, 1\}\right] \cup \left[S_{\pi,\{(1,0),(0,1)\}}(Z, X) \times \{0\}\right] & \text{if } D = (0, 1) \\ \left[S_{\pi,(1,0)}(Z, X) \times \{0, 1\}\right] \cup \left[S_{\pi,\{(1,0),(0,1)\}}(Z, X) \times \{1\}\right] & \text{if } D = (1, 0) \\ S_{\pi,(1,1)}(Z, X) \times \{0, 1\} & \text{if } D = (1, 1). \end{cases} \tag{5.20}$$

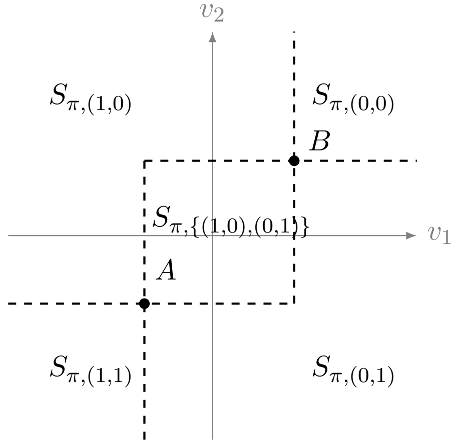
In words, if $D = (0, 0)$ is realized, it implies that $(V_1, V_2) \in S_{\pi,(0,0)}(Z, X)$, regardless of the value of V_s. Therefore, V belongs to $S_{\pi,(0,0)}(Z, X) \times \{0, 1\}$. Similarly, if $D = (0, 1)$ realizes, then either $(0, 1)$ is uniquely predicted due to $(V_1, V_2) \in S_{\pi,(0,1)}(Z, X)$ (and V_s unrestricted) or $(0, 1)$ is selected from the set of treatment statuses due to $(V_1, V_2) \in S_{\pi,\{(1,0),(0,1)\}}(Z, X)$ and $V_s = 0$. The other cases can be analyzed similarly.

We work with the conditional mean-independence assumption. Recall $D = (D_1, D_2)$.

[21] A similar argument can be applied to games of strategic complementarity and even to models with incoherent predictions.

[22] One may represent the selection mechanism by a latent random variable defining a mixture without loss of generality. See Tamer (2010), Ponomareva and Tamer (2011), Molchanov and Molinari (2018, Example 2.6) and Molinari (2020, p.377).

Figure 2: Level sets of $v \mapsto G(v|z;\pi)$ and set-valued CF



Note: $A = (\pi_1(1,z_1,x), \pi_2(1,z_1,x))$; $B = (\pi_1(0,z_1,x), \pi_2(0,z_2,x))$. The subsets are defined as follows.

$$
\begin{aligned}
S_{\pi,(0,0)}(z,x) &= \{v : v_1 > \pi_1(0,z_1,x), v_2 > \pi_2(0,z_2,x)\}\\
S_{\pi,(0,1)}(z,x) &= \{v : \pi_1(1,z_1,x) < v_1 \leq \pi_1(0,z_1,x), v_2 \leq \pi_2(1,z,x)\} \cup \{v : \pi_1(0,z_1,x) < v_1, v_2 \leq \pi_2(0,z_2,x)\}\\
S_{\pi,(1,0)}(z,x) &= \{v : v_1 \leq \pi_1(1,z_1,x), v_2 > \pi_2(1,z_2,0)\} \cup \{v : \pi_1(1,z_1,x) < v_1 \leq \pi_1(0,z_1,x), v_2 > \pi_2(0,z_2,x)\}\\
S_{\pi,(1,1)}(z,x) &= \{v : v_1 \leq \pi_1(1,z_1,x), v_2 \leq \pi_2(1,z_2,x)\}\\
S_{\pi,\{(0,1),(1,0)\}}(z,x) &= \{v : \pi_j(0,z_j,x) < v_j \leq \pi_j(1,z_j,x), j = 1,2\}.
\end{aligned}
$$

Define the model prediction

$$
\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F) = \text{cl}\{y \in \mathcal{Y} : y = \mu(D,X) + \lambda_D(X,V) + \eta_D, V \in \text{Sel}(\boldsymbol{V})\}, \tag{5.21}
$$

where $\lambda_d(x,v)$ is the conditional mean function of $U_d|X,V$.

Let us rewrite the set-valued control function in (5.20) as a union of two random sets.

$$
\boldsymbol{V}(D,X,Z;\pi) = [\tilde{\boldsymbol{V}}_0(D,X,Z;\pi) \times \{0\}] \cup [\tilde{\boldsymbol{V}}_1(D,X,Z;\pi) \times \{1\}], \tag{5.22}
$$

where

$$
\tilde{\boldsymbol{V}}_0(D,X,Z;\pi) \equiv \begin{cases} S_{\pi,(0,1)}(Z) \cup S_{\pi,\{(1,0),(0,1)\}}(Z) & \text{if } D = (0,1) \\ S_{\pi,(d_1,d_2)}(Z) & \text{if } D = (d_1,d_2),\ (d_1,d_2) \neq (0,1), \end{cases} \tag{5.23}
$$

and

$$
\tilde{\boldsymbol{V}}_1(D,X,Z;\pi) \equiv \begin{cases} S_{\pi,(1,0)}(Z) \cup S_{\pi,\{(1,0),(0,1)\}}(Z) & \text{if } D = (1,0) \\ S_{\pi,(d_1,d_2)}(Z) & \text{if } D = (d_1,d_2),\ (d_1,d_2) \neq (1,0). \end{cases} \tag{5.24}
$$

As in the previous examples, the sharp identification region for $\theta = (\mu, \pi, F)$ involves the supremum and infimum of a function f over $v \in \boldsymbol{V}(d, x, z; \pi)$. Eq. (5.22) suggests that the supremum, for example, can be written as

$$\sup_{v \in \boldsymbol{V}(d,x,z;\pi)} f(v_1, v_2, v_s) = \max\{ \sup_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_0(d,x,z)} f(v_1, v_2, 0), \sup_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_1(d,x,z)} f(v_1, v_2, 1)\}. \quad (5.25)$$

We use (4.10) from Theorem 2 and argue as in Example 1 to characterize the sharp identification region.

Corollary 3: *Suppose $E_{P_0}[|Y|] < \infty$. Suppose $U = (U_{00}, U_{10}, U_{01}, U_{11})$ has a strictly positive conditional density given (X, V). Suppose, for each $(d_1, d_2) \in \mathcal{D}$, $E[U_{d_1,d_2}|Z, X, V] = E[U_{d_1,d_2}|X, V]$, a.s. Then, $\Theta_I(P_0)$ is the set of parameter values $\theta = (\mu, \pi, F)$ such that, for almost all (d, x),*

$$\begin{aligned} \sup_{z \in \mathcal{Z}} \Big\{ E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_U(d, x, z) \Big\} \\ \leq \mu(d, x) \leq \\ \inf_{z \in \mathcal{Z}} \Big\{ E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_L(d, x, z) \Big\}, \quad (5.26) \end{aligned}$$

where

$$\lambda_U(d, x, z) = \max\big\{ \sup_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_0(d,x,z;\pi)} \lambda_d(x, v_1, v_2, 0), \sup_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_1(d,x,z;\pi)} \lambda_d(x, v_1, v_2, 1)\big\}, \quad (5.27)$$

$$\lambda_L(d, x, z) = \min\big\{ \inf_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_0(d,x,z;\pi)} \lambda_d(x, v_1, v_2, 0), \inf_{(v_1,v_2) \in \tilde{\boldsymbol{V}}_1(d,x,z;\pi)} \lambda_d(x, v_1, v_2, 1)\big\}. \quad (5.28)$$

Remark 2: One may impose further restrictions on the relationship between U and V_s via a priori restrictions on F. An example is to assume U is independent (or mean independent) of the selection mechanism conditional on other control variables. This assumption is plausible if V_s is viewed as a signal that is only relevant for the treatment decision, e.g., firms' profitability, but irrelevant for the outcome, such as pollution level. Imposing this assumption allows us to exclude v_s from λ_d: $\lambda_d(x, v_1, v_2, 1) = \lambda_d(x, v_1, v_2, 0) = \lambda_d(x, v_1, v_2)$.

Remark 3: With additional assumptions, it is possible to point identify π. Specifically, suppose $(V_1, V_2) \perp (Z_1, Z_2)|X$, and for each j, $Z_j = (Z_{j,k}, Z_{j,-k})$ contains a continuous component Z_{jk} supported on $\mathbb{R}$. Furthermore, *supp*$(Z_j, X|Z_{-j}) =$ *supp*(Z_j, X), *a.s.* Tamer (2003) shows, if one can vary the continuous component to push the choice probabilities toward extreme values, i.e., π_j is such that $\lim_{z_{j,k} \to -\infty} \pi_j(0, z_{j,k}, z_{j,-k}, x) = 0$, and

$\lim_{z_{j,k}\to\infty} \pi_j(1, z_{j,k}, z_{j,-k}, x) = 1$, then, π is point identified. This argument, however, requires a variable with a large support (for each player).

Remark 4: The recent work by Ciliberto et al. (2021) simultaneously estimates the parameters of the incomplete entry model and an outcome model (demand and supply). Their approach can be viewed as a generalization of the simultaneous estimation of a triangular system. In contrast, we use the incomplete entry model to construct a set-valued control function, which is a generalization of the control function approach to a triangular system. Moreover, we are interested in identifying treatment effects, while their focus is recovering market primitives.

5.4 Effects of School Assignment under Capacity Constraints

Following Bertanha et al. (2024), consider a continuum population of students and a set of J schools, $\mathcal{J} := \{1, \ldots, J\}$. Let Q denotes the true preference relation of a student over the set of options $\mathcal{J}^0 := \mathcal{J} \cup \{0\}$ with an outside option 0.[23] For example, if $J = 2$ and $Q = \{1, 2, 0\}$, then 1 is preferred to 2 (i.e., $1Q2$) and so on. Let $\mathbf{S} \equiv (S_1, \ldots, S_J) \subseteq \mathbb{R}^J$ be a vector of placement scores of a student and $Y(d)$ be the potential outcome of a student being assigned to option $d \in \mathcal{J}^0$. For placement scores $\mathbf{S}$ of a student and admission cutoffs $\mathbf{c} \equiv (c_1, \ldots, c_J)$ of schools, the set of feasible options of the student is $B(\boldsymbol{S}) \equiv \{0\} \cup \{j \in \mathcal{J} : S_j \geq c_j\}$. For simplicity, assume that placement scores and cutoffs do not have ties.[24] Finally, for a given preference Q and a school $j \in \mathcal{J}$, define a *local preference* Q_j as the pair $(k, l) \in \mathcal{J}^0 \times \mathcal{J}^0$ where k is the most preferred school in the set of feasible options plus j (i.e., $B(\boldsymbol{S}) \cup \{j\}$) and l is the most preferred school in the set of feasible options minus j (i.e., $B(\boldsymbol{S}) \backslash \{j\}$).

If students' reported preferences P are equal to their true preferences Q, then the local preference $P_j = Q_j$ serves as a control variable, conditional on which one can apply a regression discontinuity design (RDD). Specifically, conditional on the local preference and under the continuity of the distribution of $Y(l)$ $(l \in \{j, k\})$ with respect to S_j, one can identify the effect of assigning school j relative to school k at the cutoff (Kirkeboen et al., 2016; Abdulkadiroğlu et al., 2019):

$$\begin{aligned}
&E\left[Y(j) - Y(k) \mid Q_j = (j,k), S_j = c_j\right] \\
&= \lim_{\epsilon \downarrow 0} E\left[Y \mid P_j = (j,k), S_j = c_j + \epsilon\right] - \lim_{\epsilon \uparrow 0} E\left[Y \mid P_j = (j,k), S_j = c_j + \epsilon\right].
\end{aligned}$$

In reality, school assignment mechanisms face various constraints (e.g., capacity constraints,

[23]Within this section, we use Q and P to denote preferences, maintaining notational consistency with Bertanha et al. (2024).

[24]Bertanha et al. (2024) consider a more general framework.

application costs), in which case, students have incentives to strategically misreport their preferences (Agarwal and Somaini, 2018; Fack et al., 2019). This case is the main focus of Bertanha et al. (2024). Suppose students submit a weak partial order P of their true preferences Q, that is, P is any selection of up to K schools in Q that are ranked according to Q and preferred to the outside option 0. Also, suppose students are matched to their best feasible options according to P. Under these assumption, following Proposition 2 in Bertanha et al. (2024), we can construct the set-valued control function as follows. Fix a school j with cutoff c_j. Consider a student with scores $\mathbf{S} \equiv (S_j, \mathbf{S}_{-j})$ and reported preferences P, which are the observables. Suppose $(a, b) = P_j$ is the student's reported local preference. Then, the true local preference Q_j of this student belongs to the set-valued control function $\mathbf{Q}_j$ defined as

$$\mathbf{Q}_j(\boldsymbol{S}, P) \equiv \begin{cases} \{(a,b)\}, & \text{if } S_j \geq c_j \text{ and } a = b, \\ \{(a,b)\} \cup \left(\{a\} \times N_j^-\right), & \text{if } S_j \geq c_j \text{ and } a \neq b, \\ \{(a,b)\} \cup \left(\left(N_j^+ \backslash N_j^-\right) \times \{b\}\right) & \text{if } S_j < c_j, \end{cases} \tag{5.29}$$

where $N_j^- \equiv \{B(\boldsymbol{S}) \backslash \{j\}\} \backslash \{P \cup \{0\}\}$ and $N_j^+ \equiv \{B(\boldsymbol{S}) \cup \{j\}\} \backslash \{P \cup \{0\}\}$.[25] Given $\boldsymbol{Q}_j$, the current paper's procedure can be applied to construct *sharp* bounds on the effect of school assignment, $E\left[Y(j) - Y(k) \mid Q_j = (j,k), S_j = c_j\right]$, by proceeding as follows.

Suppose $Y(j) = \mu(j, Q_j, S_j) + U$ (and similarly for $Y(k)$) and let $\lambda(Q_j, S_j) \equiv E[U|Q_j, S_j]$.[26] Then, the parameter of interest becomes $\mu(j, (j,k), c_j) - \mu(k, (j,k), c_j)$. Assume that $\mu(j, Q_j, \cdot)$, $\mu(k, Q_j, \cdot)$ and $\lambda(Q_j, \cdot)$ are continuous, which implies the usual continuity of $E[Y(l)|Q_j, \cdot]$ for $l \in \{j, k\}$. Even if we do not observe Q_j, we observe $\boldsymbol{Q}_j(\boldsymbol{S}, P)$. Assume $Q_j \in \boldsymbol{Q}_j$ with probability 1 conditional on S_j.[27] Note that, conditional on $Q_j = (j,k)$ and $S_j \geq c_j$, $Y = \mu(j, Q_j, S_j) + \lambda(Q_j, S_j) + \eta$ where $\eta \equiv U - \lambda(Q_j, S_j)$ satisfies $E[\eta|Q_j, S_j] = 0$. Therefore, conditional on $S_j \geq c_j$ and $P_j = (j,k)$, we can write

$$\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda) \equiv \text{cl}\left\{y \in \mathcal{Y} : y = \mu(j, Q_j, S_j) + \lambda(Q_j, S_j) + \eta, Q_j \in \text{Sel}(\boldsymbol{Q}_j)\right\},$$

because according to the definition (5.29), when $S_j \geq c_j$ and $P_j = (j,k)$, school j is the school that is truthfully preferred (i.e., j is point-identified as the first element of Q_j). Similarly,

[25] See the proof of a more general version of this result in Bertanha et al. (2024).

[26] Following Section 4.1, we can allow for a more general specification $Y(j) = \mu(j, Q_j, S_j) + U_j$ where U_j replaces U. With the vector unobservables, (U_j, U_k), one can define $\lambda_l(Q_j, S_j) \equiv E[U_l|Q_j, S_j]$ for $l \in \{j, k\}$. Then, the RDD approach will require assuming $\lambda_j(Q_j, c_j) = \lambda_k(Q_j, c_j)$, while our partial identification approach do not rely on such an assumption.

[27] This corresponds to Assumption 6(i) of Bertanha et al. (2024).

conditional on $S_j < c_j$ and $P_j = (j,k)$,

$$\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda) \equiv \text{cl}\left\{y \in \mathcal{Y} : y = \mu(k, Q_j, S_j) + \lambda(Q_j, S_j) + \eta, Q_j \in \text{Sel}(\boldsymbol{Q}_j)\right\},$$

because, when $S_j < c_j$ and $P_j = (j,k)$, school k is truthfully disliked (i.e., k is point-identified as the second element of Q_j). Now, using the Aumann expectation, $E[Y|\boldsymbol{S}, P] \in \mathbb{E}[\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda)|\boldsymbol{S}, P]$, which is equivalent to

$$bE[Y|\boldsymbol{S}, P] \leq s(b, \mathbb{E}[\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda)|\boldsymbol{S}, P]), \quad b \in \{-1, 1\}.$$

Then, by the convexification theorem,

$$s(b, \mathbb{E}[\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda)|\boldsymbol{S}, P]) = \mathbb{E}[s(b, \boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda))|\boldsymbol{S}, P], \quad b \in \{-1, 1\}.$$

Suppose $P_j = (j,k)$. For $b = 1$ and $c \geq c_j$,

$$\begin{aligned}
&\mathbb{E}[s(b, \boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda))|S_j = c, S_{-j} = s_{-j}, P = p] \\
&\quad = E[\sup_{Y \in \text{Sel}(\boldsymbol{Y}(\eta, \boldsymbol{S}, P; \mu, \lambda))} Y|S_j = c, S_{-j} = s_{-j}, P = p] = \sup_{q \in \boldsymbol{Q}_j(c, s_{-j}, p)} \{\mu(j, q, c) + \lambda(q, c)\}.
\end{aligned}$$

Then, by combining the results above and taking the limit (allowing that lim and sup are interchangeable) and under continuity, we have

$$\lim_{c \downarrow c_j} E[Y|S_j = c, S_{-j} = s_{-j}, P = p] \leq \sup_{q \in \boldsymbol{Q}_j(c_j, s_{-j}, p)} \{\mu(j, q, c_j) + \lambda(q, c_j)\},$$

where we use an assumption that $\lim_{c \downarrow c_j} \boldsymbol{Q}_j(c, s_{-j}, p)$ is a well-defined set.[28] A similar argument can be made with $b = -1$ and $c \geq c_j$, as well as $b = 1, -1$ and $c < c_j$. Consequently, there are in total four inequalities constructed for a given conditioning value. Note that $\lambda(q, c_j)$ is common in all four inequalities. When $\boldsymbol{Q}_j$ is a singleton, $\lambda(q, c_j)$ would cancel out when taking the difference between the equalities to calculate the treatment effect. Now, the researcher can further assume separability, such as $\mu(j, Q_j, S_j) = \mu_1(j) + \mu_2(Q_j, S_j)$, and derive intersection bounds on $\mu_1(j)$. Such a separability assumption would become more plausible once additional covariates (X, S_{-j}) are incorporated: $\mu(j, Q_j, \boldsymbol{S}, X) = \mu_1(j, \boldsymbol{S}, X) + \mu_2(Q_j, \boldsymbol{S}, X)$.

Remark 5: The bounds resulting from our approach are sharp by construction. These bounds may not coincide with the bounds derived in Bertanha et al. (2024), as our bounding

[28]This may relate to Assumption 7 in Bertanha et al. (2024).

strategy is different from theirs, which builds on Horowitz and Manski (1995). Also, the two approaches rely on different sets of assumptions.

6 Empirical Illustration

We apply the identification framework to a model of health preventive actions. We use data from Thornton (2008) who studied the impacts of learning HIV status and receiving voluntary counseling about safe sexual practices on HIV preventive behavior, as measured by condom purchases. The study was based on the Malawi Diffusion and Ideational Change Project (MDICP), which provided randomized monetary incentives (vouchers) to individuals. These incentives encouraged individuals to visit voluntary counseling and testing (VCT) centers to learn about their HIV status and redeem the vouchers. During a follow-up survey conducted at the respondents' homes, they were offered to purchase condoms at a discounted price. Our primary goal here is to illustrate how our framework can be used to conduct inference for policy-relevant objects in an empirical example that involves self-selection and a discrete endogenous variable. Specifically, we examine the average health preventive response of individuals when they are informed of their HIV status.

The binary treatment D_i indicates whether an individual i learns their HIV status. The outcome variable Y_i represents the number of condom purchases. We focus on individuals who were sexually active at the time of the survey and purchased either 0, 3, or 6 condoms, which corresponds to 0, 1, or 2 packs of condoms.[29] We construct confidence intervals for average and policy-relevant structural functions and examine how assumptions on the selection process affect them. The instrumental variables are Z_{amt}, the amount of the monetary voucher, and Z_{dist}, the distance to the VCT center, the location of which was randomized. We used the same set of observable controls as in the original study, a dummy variable for the HIV diagnosis ($X_{HIV} = 1$ if positive) and additional controls ($\tilde{X}$): a dummy for male, age, simulated average distance to the VCT center, and a district dummy variable.

Consider a simple ordered choice model of condom purchases:

$$Y = \begin{cases} 0 & \text{if } \mu(D, X) + U \leq c_L \\ 3 & \text{if } c_L < \mu(D, X) + U \leq c_U \\ 6 & \text{if } \mu(D, X) + U > c_U. \end{cases} \tag{6.1}$$

The latent variable U captures each individual's latent preference toward HIV prevention.

[29] This subsample accounts for 94% of the sample of sexually active individuals. Those outside this sample either purchased condoms at a premium (not by packs) or purchased more than 2 packs, which was rare.

We specify U's conditional distribution as $U = Q(\eta; X, V) = g(V) + Q(\eta)$ as in (B.9), where Q is a quantile function of some distribution.[30] Our specification of the selection process is

$$D = 1\{\pi(Z, X) \geq V\}. \tag{6.2}$$

As before, we normalize the distribution of $V|Z, X$ to $U[0,1]$. This selection process yields the set-valued control function in (2.8), which allows us to define the model prediction $\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F)$ and derive its containment functional. An inspection of $\boldsymbol{Y}$ shows it suffices to obtain the containment functional $\mathbb{C}_\theta$ (and its conjugate $\mathbb{C}^*_\theta$) for $A = \{0\}$ and $\{6\}$ (see the proof of Proposition 1). Let $F \equiv Q^{-1}$, and let

$$\begin{aligned}
\mathbb{C}_\theta(\{0\}|d, x, z) &= F(c_L - \mu(d, x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)) \\
\mathbb{C}^*_\theta(\{0\}|d, x, z) &= F(c_L - \mu(d, x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v)) \\
\mathbb{C}_\theta(\{6\}|d, x, z) &= 1 - F(c_U - \mu(d, x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v)) \\
\mathbb{C}^*_\theta(\{6\}|d, x, z) &= 1 - F(c_U - \mu(d, x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)).
\end{aligned}$$

The following proposition characterizes the sharp identification region.

PROPOSITION 1: *Suppose Assumptions 1-2 hold. Then, $\theta = (\mu, c_L, c_U, F, \pi)$ is in the sharp identification region $\Theta_I(P_0)$ if and only if*

$$\mathbb{C}_\theta(\{0\}|D, X, Z) \leq P_0(Y = 0|D, X, Z) \leq \mathbb{C}^*_\theta(\{0\}|D, X, Z) \tag{6.3}$$

$$\mathbb{C}_\theta(\{6\}|D, X, Z) \leq P_0(Y = 6|D, X, Z) \leq \mathbb{C}^*_\theta(\{6\}|D, X, Z), \;\; a.s., \tag{6.4}$$

and $\pi \in \Pi_r(P_0)$.

We conduct inference for structural objects based on (6.3)-(6.4). Specifically, we examine two structural functions: the average structural function, denoted as $\mathrm{ASF}(d, x_{HIV})$, and the counterfactual switching probability described below. The ASF represents the expected number of condom purchases when the treatment and HIV status are set to (d, x_{HIV}). This function can be expressed as a function of θ (see Appendix D for details).

For each object, we construct a confidence interval using the universal inference method by Kaido and Zhang (2024). For any $\varphi : \Theta \to \mathbb{R}$, this confidence interval covers each value of $\varphi(\theta)$ in its sharp identification region (see $\mathfrak{K}_I(d)$ in Theorem 3) with at least a prescribed confidence level in finite samples. It is suitable for our setting because it can accommodate

[30]This corresponds to $U|V$ being in a location family with location parameter $g(V)$.

both discrete and continuous covariates and nonparametric components in θ. We provide details of the implementation in Appendix D.

Table 1 reports confidence intervals for the ASF, where we specify (μ, g, Q) by $\mu(d,x) = \mu_1 d + \mu_{int}(d \times x_{HIV}) + \tilde{x}'\beta$, $g(v) = \rho\Phi^{-1}(v)$, and $Q(\eta) = \sqrt{1-\rho^2}\Phi^{-1}(\eta)$.[31] We find that there may be differences in how people respond to receiving an HIV-positive diagnosis compared to an HIV-negative diagnosis. People who receive an HIV-negative diagnosis typically purchase an average of 3 or fewer condoms (1 pack). On the other hand, those who receive an HIV-positive diagnosis may purchase more than 3 condoms, as suggested by the wider confidence interval. However, the confidence interval for the HIV-positive group is wide and includes zero, possibly due to the small sample size of this group. Similarly, the confidence interval for the HIV-negative group is shorter but still includes zero, making it difficult to determine the average treatment effect for this group.

The HIV-negative individuals form a large subpopulation that can benefit from policy-induced preventive actions. We now examine whether informing them of their HIV status has any sizeable effects. We conduct this analysis by sequentially adding assumptions regarding the treatment response and selection processes.

First, we consider monotone treatment selection (MTS) (Manski and Pepper, 2000): For each $d = 0, 1$, $E[Y(d)|D=1] \geq E[Y(d)|D=0]$. One may argue those who choose to learn their HIV status are more likely to buy condoms (regardless of the learning of HIV status), because they are proactively health-conscious. Given the specification above, this assumption corresponds to $-1 < \rho \leq 0$. Panel B of Table 1 shows that this assumption tightens the confidence interval for the ASF at $d = 1$ (and ATE) of the HIV-negative group but has no effects otherwise. Next, we consider monotone treatment response (MTR) (Manski, 1990, 1997) for the HIV-negative group: $Y(1)|X_{HIV} = 0 \geq Y(0)|X_{HIV} = 0$. This assumption presumes that, keeping everything else fixed, learning a negative HIV diagnosis encourages individuals to protect themselves from future infections, restricting the effect's sign *a priori*. This version of MTR corresponds to $\mu_1 \geq 0$. Panel C of Table 1 shows that the information provision increases condom purchases by at most 2 units. By combining the MTS and MTR (for HIV-negative), we may obtain an even tighter confidence interval as shown in Panel D. While the ATE is positive under these assumptions, the magnitude of the effect in this case is at most 1.176. These results suggest learning an HIV-negative diagnosis has a limited impact on condom purchases. [32]

To understand the individual decisions behind the average effects, we also create confi-

[31] This corresponds to a parametric model, in which $U|\tilde{V} = \tilde{v} \sim N(\rho\tilde{v}, 1-\rho^2)$, where $\tilde{V} = \Phi(V)$.

[32] Although we do not pursue this, one may preemptively assume positive effect for overall population, which may be supported by previous RCTs. This corresponds to MTR for all x. Under this assumption, one may be interested in the magnitude of the effects.

		HIV+	HIV−
Panel A: (Baseline)			
	$d = 1$	$[0.030, 5.578]$	$[0.090, 2.864]$
	$d = 0$	$[0.030, 5.216]$	$[0.151, 3.256]$
	ATE	$[-4.794, 4.492]$	$[-2.563, 2.141]$
Panel B: (MTS)			
	$d = 1$	$[0.030, 5.307]$	$[0.090, 1.779]$
	$d = 0$	$[0.030, 5.276]$	$[0.181, 3.286]$
	ATE	$[-4.854, 3.889]$	$[-2.563, 1.176]$
Panel C: (MTR for HIV−)			
	$d = 1$	$[0.060, 5.397]$	$[0.271, 2.683]$
	$d = 0$	$[0.030, 4.070]$	$[0.151, 1.568]$
	ATE	$[-2.985, 4.372]$	$[0.030, 2.020]$
Panel D: (MTS & MTR for HIV−)			
	$d = 1$	$[0.030, 5.276]$	$[0.241, 1.719]$
	$d = 0$	$[0.030, 4.281]$	$[0.181, 1.719]$
	ATE	$[-3.045, 3.950]$	$[0.030, 1.176]$

Table 1: 95% Confidence Intervals for ASFs and ATE (Model 1)

		HIV+	HIV−
Panel A: (Baseline)			
	Switch Pr.	$[0, 0.829]$	$[0, 0.508]$
Panel B: (MTS)			
	Switch Pr.	$[0, 0.724]$	$[0, 0.291]$
Panel C: (MTR for HIV−)			
	Switch Pr.	$[0, 0.824]$	$[0, 0.492]$
Panel D: (MTS & MTR for HIV−)			
	Switch Pr.	$[0, 0.729]$	$[0, 0.286]$

Table 2: 95% Confidence Intervals for the Switching Probabilities (Model 1)

dence intervals for the counterfactual probability of changing decisions. Specifically, we focus on the probability of switching from not buying any condoms to buying some condoms due to the intervention. Let $Y(d, x_{HIV})$ be the counterfactual outcome given the values (d, x_{HIV}) of treatment and HIV status. The share of switchers can be represented as follows:

$$
\begin{aligned}
&P(Y(0, x_{HIV}) = 0, Y(1, x_{HIV}) > 0) \\
&\quad = \int [F(c_L - \mu(0, x_{HIV}, \tilde{x}) - g(v)) - F(c_L - \mu(1, x_{HIV}, \tilde{x}) - g(v))]_+ dF_{(\tilde{X},V)}(\tilde{x}, v). \qquad (6.5)
\end{aligned}
$$

Table 2 shows that, under the baseline assumption, at most 50% of the individuals who received HIV-negative diagnosis would change their behavior in response to the treatment. With an average treatment effect of about 2, most of them would only purchase no more than a single pack. However, under the MTS assumption, this upper bound on the switching probability is reduced to 29%, which explains why the ATE was lower under this additional assumption. They suggest that the intervention encourages at most 30-50% of the individuals to switch their behavior, but the quantity of the purchased condoms would be limited.

The study by Thornton (2008) used a linear IV model with interactive terms and observed a mild disparity in the impacts of learning an HIV-positive diagnosis compared to an HIV-negative diagnosis. It also found that the two-stage least squares (2SLS) estimate for the HIV-negative group was small. The original interpretation of the 2SLS estimand as a local average treatment effect (LATE) should be approached with caution due to the presence of multiple continuous instruments and continuous covariates. In such settings, the 2SLS estimand can be interpreted as a positively weighted average of LATEs only under restrictive assumptions (Mogstad et al., 2021; Blandhol et al., 2022; Słoczyński, 2022). In contrast, we target causal estimands that can be expressed as a function of the underlying parameter. We find that, for the HIV-negative group which is the majority of the population, the treatment effect is limited and can be negative unless the MTR assumption holds. This conclusion holds with or without the MTS assumption. In addition, the analysis of the switching probability reveals a limited fraction of individuals would switch their purchasing behavior in response to the information provision, even though implementing such a policy is costly. These results suggest providing information on HIV status does not seem to be an effective way to induce health-preventive behavior broadly.

7 Concluding remarks

Observational data are often generated through complex decision processes. Allowing control functions to be set-valued, this paper expands the scope of the control function approach. The proposed framework accommodates, for example, selection processes that involve rich heterogeneity, dynamic optimizing behavior, or social interaction. Our identifying restrictions are inequalities on the conditional choice probabilities. One can conduct inference using moment-based methods or likelihood-based inference methods. Practitioners can use the results of this paper for various purposes. First, they can evaluate social programs nonparametrically, while taking into account potentially complex treatment selection processes. Second, the bounds in our main identification results can easily be combined with a range of shape restrictions and parametric assumptions, allowing practitioners to conduct a sensitivity analysis to assess the additional identifying power of specific assumptions. The tools from random set theory enable us to guarantee sharpness of bounds one obtains in such a sensitivity analysis, without needing to prove sharpness case after case.

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.

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2019): "Breaking Ties: Regression Discontinuity Design Meets Market Design," *Econometrica*.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do Parents Value School Effectiveness?" *American Economic Review*, 110, 1502–39.

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83, 2411–2451.

Agarwal, N. and P. Somaini (2018): "Demand Analysis Using Strategic Reports: An Application To A School Choice Mechanism," *Econometrica*, 86, 391–444.

Andrews, D. W. K. and X. Shi (2013): "Inference based on conditional moment inequalities." *Econometrica*, 81, 609–666.

Aronow, P. M. and C. Samii (2017): "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 11, 1912–1947.

Auerbach, E. (2022): "Identification and estimation of a partially linear regression model using network data," *Econometrica*, 90, 347–365.

Balat, J. F. and S. Han (2023): "Multiple treatments with strategic substitutes," *Journal of Econometrics*, 234, 732–757.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821.

Bertanha, M., M. Luflade, and I. Mourifié (2024): "Causal Effects in Matching Mechanisms with Strategically Reported Preferences," *National Bureau of Economic Research*.

Bishop, K. C., J. D. Ketcham, and N. V. Kuminoff (2022): "Hazed and Confused: The Effect of Air Pollution on Dementia," *The Review of Economic Studies*, 90, 2188–2214.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When is TSLS Actually LATE?" Working Paper 29709, National Bureau of Economic Research.

Blundell, R., D. Kristensen, and R. L. Matzkin (2013): "Control Functions and Simultaneous Equations Methods," *American Economic Review*, 103, 563–569.

Blundell, R. and R. L. Matzkin (2014): "Control functions in nonseparable simultaneous equations models," *Quantitative Economics*, 5, 271–295.

Blundell, R. and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric Society Monographs*, 36, 312–357.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, 71, 655–679.

Blundell, R. W. and R. J. Smith (1989): "Estimation in a class of simultaneous equation limited dependent variable models," *The Review of Economic Studies*, 56, 37–57.

Bontemps, C. and R. Kumar (2020): "A geometric approach to inference in set-identified entry games," *Journal of Econometrics*, 218, 373–389.

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

Canay, I. A. and A. M. Shaikh (2017): "Practical and theoretical advances in inference for partially identified models," *Advances in Economics and Econometrics*, 2, 271–306.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2019): "Are Referees and Editors in Economics Gender Neutral?*," *The Quarterly Journal of Economics*, 135, 269–327.

Carlier, G., A. Galichon, and F. Santambrogio (2010): "From Knothe's Transport to Brenier's Map and a Continuation Method for Optimal Transport," *SIAM Journal on Mathematical Analysis*, 41, 2554–2576.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 78, 377–394.

Chen, X. and T. M. Christensen (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, 9, 39–84.

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2020): "Semiparametric estimation of structural functions in nonseparable triangular models," *Quantitative Economics*, 11, 503–533.

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

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models1," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: estimation and inference." *Econometrica*, 81, 667–737.

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

——— (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73, 1525–1550.

Chesher, A., D. Kim, and A. M. Rosen (2023a): "IV methods for Tobit models," *Journal of Econometrics*, 235, 1700–1724.

Chesher, A. and A. M. Rosen (2013): "What Do Instrumental Variable Models Deliver with Discrete Dependent Variables?" *American Economic Review*, 103, 557–62.

——— (2017): "Generalized instrumental variable models," *Econometrica*, 85, 959–989.

Chesher, A., A. M. Rosen, and Y. Zhang (2023b): "Identification analysis in models with unrestricted latent variables: Fixed effects and initial conditions," *Cemmap Working Paper*.

Chesher, A. and K. Smolinski (2012): "IV models of ordered choice," *Journal of Econometrics*, 166, 33–48.

Ciliberto, F., C. Murry, and E. Tamer (2021): "Market structure and competition in airline markets," *Journal of Political Economy*, 129, 2995–3038.

Dal Bó, E., F. Finan, N. Y. Li, and L. Schechter (2021): "Information technology and government decentralization: Experimental evidence from paraguay," *Econometrica*, 89, 677–701.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric instrumental regression," *Econometrica*, 79, 1541–1565.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric estimation of sample selection models," *The Review of Economic Studies*, 70, 33–58.

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

D'Haultfœuille, X., S. Hoderlein, and Y. Sasaki (2021): "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*.

Dhrymes, P. J. (1970): *Econometrics: Statistical foundations and applications*, New York: Springer-Verlag.

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.

Fack, G., J. Grenet, and Y. He (2019): "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions," *The American Economic Review*, 109, 1486–1529.

Galichon, A. and M. Henry (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, rdr008.

Gautier, E. and S. Hoderlein (2011): "A triangular treatment effect model with random coefficients in the selection equation," .

Graham, B. S. (2011): "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," .

Hall, P. and J. L. Horowitz (2005): "Nonparametric methods for inference in the presence of instrumental variables," *The Annals of Statistics*, 33, 2904–2929.

Han, S. (2021): "Identification in nonparametric models for dynamic treatment effects," *Journal of Econometrics*, 225, 132–147.

——— (2023): "Optimal dynamic treatment regimes and partial welfare ordering," *Journal of the American Statistical Association*, 1–11.

Han, S. and S. Lee (2023): "Semiparametric Models for Dynamic Treatment Effects and Mediation Analyses with Observational Data," *U of Bristol and Sogang U.*

Han, S. and S. Yang (2024): "A computational approach to identification of treatment effects for policy evaluation," *Journal of Econometrics*, 240, 105680.

Heckman, J. J. (1979): “Sample selection bias as a specification error,” *Econometrica*, 153–161.

Heckman, J. J. and B. E. Honoré (1990): “The Empirical Content of the Roy Model,” *Econometrica*, 58, 1121–1149.

Heckman, J. J. and E. Vytlacil (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669–738.

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.

Horowitz, J. L. and C. F. Manski (1995): “Identification and robustness with contaminated and corrupted data,” *Econometrica*, 281–302.

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

Imbens, G. W. and W. K. Newey (2002): “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” Working Paper 285, National Bureau of Economic Research.

——— (2009): “Identification and estimation of triangular simultaneous equations models without additivity,” *Econometrica*, 77, 1481–1512.

Joe, H. (2014): *Dependence Modeling with Copulas*, CRC press.

Jovanovic, B. (1989): “Observable implications of models with multiple equilibria,” *Econometrica*, 1431–1437.

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

Kaido, H. and F. Molinari (2024): “Information Based Inference In Models with Set-Valued Predictions and Misspecification,” ArXiv: 2401.11046.

Kaido, H. and Y. Zhang (2024): “Universal Inference for Incomplete Discrete Choice Models,” Working Paper, Boston University.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): “Field of study, earnings, and self-selection,” *The Quarterly Journal of Economics*, 131, 1057–1111.

Kline, P. and C. R. Walters (2016): "Evaluating Public Programs with Close Substitutes: The Case of Head Start*," *The Quarterly Journal of Economics*, 131, 1795–1848.

——— (2019): "On Heckits, LATE, and Numerical Equivalence," *Econometrica*, 87, 677–696.

Levinsohn, J. and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341.

Li, L. and M. Henry (2024): "Finite Sample Inference in Incomplete Models," *arXiv:2204.00473*.

Luo, Y. and H. Wang (2017): "Core Determining Class: Construction, Approximation and Inference," Working Paper, University of Florida.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212, 522–555.

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

——— (1997): "Monotone treatment response," *Econometrica*, 1311–1334.

——— (2013): "Identification of treatment response with social interactions," *The Econometrics Journal*, 16, S1–S23.

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

Manski, C. F. and E. Tamer (2002a): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

——— (2002b): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–98.

Molchanov, I. (2017): *Theory of Random Sets*, vol. 19, Springer London, 2 ed.

Molchanov, I. and F. Molinari (2018): *Random sets in econometrics*, vol. 60, Cambridge University Press.

Molinari, F. (2020): "Chapter 5 - Microeconometrics with partial identification," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 355–486.

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81.

Nagasawa, K. (2024): "Treatment effect estimation with noisy conditioning variables," *arXiv preprint arXiv:1811.00667*.

Newey, W. K. (1987): "Efficient estimation of limited dependent variable models with endogenous explanatory variables," *Journal of Econometrics*, 36, 231–250.

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

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

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

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Petrin, A. and K. Train (2010): "A Control Function Approach to Endogeneity in Consumer Choice Models," *Journal of Marketing Research*, 47, 3–13.

Pinkse, J. (2000): "Nonparametric two-step regression estimation when regressors and error are dependent," *Canadian Journal of Statistics*, 28, 289–300.

Ponomarev, K. (2022): "Essays in Econometrics," Ph.D. thesis, UCLA.

Ponomareva, M. and E. Tamer (2011): "Misspecification in moment inequality models: back to moment equalities?" *The Econometrics Journal*, 14, 186–203.

Rivers, D. and Q. H. Vuong (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39, 347–366.

Robins, J. M. (1997): "Latent Variable Modeling and Applications to Causality," *Lecture Notes in Statistics*, 69–117.

Rubin, D. B. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

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

Smith, R. J. and R. W. Blundell (1986): "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply," *Econometrica*, 679–685.

Sosa-Rubí, S. G., O. Galárraga, and J. E. Harris (2009): "Heterogeneous impact of the "Seguro Popular" program on the utilization of obstetrical services in Mexico, 2001–2006: A multinomial probit model with a discrete endogenous variable," *Journal of Health Economics*, 28, 20–34.

Słoczyński, T. (2022): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *arXiv:2011.06695.*

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165.

——— (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

Thornton, R. L. (2008): "The Demand for, and Impact of, Learning HIV Status," *American Economic Review*, 98, 1829–63.

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

Villani, C. (2008): *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg.

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

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

A Comparison with the IV Approach

Chesher and Rosen (2017) applies their method to a single-equation IV model and employ the IV assumption. They characterize identified sets for structural parameters, applying random set theory to the level set of unobservables. In this section, we compare our approach with theirs. The main propose of the comparison is to illustrate that the two frameworks are non-nested and complementary.

We summarize the characterization of the identified set in Chesher and Rosen (2017) with notation close to ours. Let Y be a vector of endogenous variables, Z be a vector of exogenous variables (e.g., IVs), and U be a vector of structural unobservables. Then define a random closed set of U as

$$\boldsymbol{U}(Y,Z;h) \equiv \{u : h(Y,Z,u) = 0\},$$

where h is a structural function, the features of which are of interest. Assume $(h, F_{U|Z}) \in \mathcal{M}$ where $\mathcal{M}$ incorporates identifying assumptions. Chesher and Rosen (2017) use the following Artstein's inequality: For $F_{U|Z}(\cdot|z) \in \mathsf{F}_{U|Z}$ being the distribution of one of the measurable selections of $\boldsymbol{U}(Y,Z;h)$,

$$F_{U|Z}(B|z) \geq \mathbb{C}_h(B|z)$$

holds for all closed sets $B \in \mathcal{F}(\mathcal{U})$ where $\mathbb{C}_h(B|Z) \equiv P[\boldsymbol{U}(Y,Z;\mu) \subseteq B|Z]$. Then, the identified set can be characterized as

$$\{(h, F_{U|Z}) \in \mathcal{M} : F_{U|Z}(B|Z) \geq \mathbb{C}_h(B|Z),\ a.s.\ \forall B \in \mathcal{F}(\mathcal{U})\}.$$

They also provide characterization using the Aumman expectation of $\boldsymbol{U}(Y,Z;h)$.

On the other hand, we characterize the identified set as

$$\{(\mu, F_{U|V}) : P_0(A|D,Z) \geq \mathbb{C}_{\mu,F}(A|D,Z),\ a.s.\ \forall A \in \mathcal{F}(\mathcal{Y}),\ \pi \in \Pi_r(P_0)\},$$

where $\mathbb{C}_{\mu,F}(A|D,Z) \equiv F_\eta(\boldsymbol{Y}(\eta, D, \boldsymbol{V}; \mu, F_{U|V}) \subseteq A)$ and X is suppressed. We also provide characterization using the Aumman expectation of $\boldsymbol{Y}$.

The two approaches share similar features in that only *sets* of unobservables can be recovered from observed data, while stochastic restrictions are imposed on true unobservables. However, the two differ in several ways. First, the CF assumption is imposed on $F_{U|D,V}$ whereas the IV assumption is imposed on $F_{U|Z}$. Note that the two stochastic assumptions are not nested. Second, the CF assumption is imposed in generating $\boldsymbol{Y}(\eta, D, \boldsymbol{V}; \mu, F_{U|V})$, while the IV assumption is imposed when applying Artstein's inequality using $\boldsymbol{U}(Y, Z; h)$. Therefore, even if we were to use the IV approach by treating our (U, V) as their U and μ as part of h, it is unknown how the CF assumption can be utilized in their framework. Third, due to this difference, the containment functional is compared to the observed distribution to construct the identified set in our setting, while it is compared to the unobserved distribution in theirs. This may have implications on implementation in practice.

In sum, the two frameworks offer complementary tools to practitioners for robust causal analyses. Practitioners can select the most appropriate approach based on the specific model at hand and their belief on the stochastic nature of their problem.

B Details on Dynamic Treatment Effects and Additional Examples

B.1 Dynamic Treatment Effects

We complete the characterization of the sharp identification region for the dynamic treatment effect example in Section 5.2. Define the following objects

$$\begin{aligned} H_{Y_2}(y_1, d_1, d_2, x, v; \theta) &\equiv F_{U_2|X,U_1,V_1,V_2}(\mu_2(y_1, d_1, d_2, x)|x, u_1, v_1, v_2) \\ H_{D_2}(y_1, d_1, z_2, x, \tilde{v}; \theta) &\equiv F_{V_2|X,U_1,V_1}(\pi_2(y_1, d_1, z_2, x)|x, u_1, v_1) \\ H_{Y_1}(d_1, x, v_1; \theta) &\equiv F_{U_1|X,V_1}(\mu_1(d_1, x)|x, v_1), \end{aligned}$$

where $\tilde{v} = (u_1, v_1)$. Recall $\boldsymbol{V}$ was defined as in (5.14). Also define

$$\begin{aligned} \boldsymbol{V}_{D_2}(Y_1, D_1, Z, X; \theta) &= \boldsymbol{V}_{U_1}(D, X; \mu_1) \times \boldsymbol{V}_1(D, Z_1, X; \pi_1) \\ \boldsymbol{V}_{Y_1}(D_1, Z_1, X; \theta) &= \boldsymbol{V}_1(D, Z_1, X; \pi_1). \end{aligned}$$

Then, we can use (4.2) from Theorem 1 and characterize the sharp identification region as follows.

Corollary 4: *Suppose (i)* $U_1 \perp Z_1 | X, V_1$ *(ii)* $V_2 \perp Z_1 | X, U_1, V_1$ *and (iii)* $U_2 \perp (Z_1, Z_2) | X, V_2, U_1, V_1$. *Then,* $\Theta_I(P_0)$ *is the set of parameter values* $\theta = (\mu_1, \mu_2, \pi_1, \pi_2, F)$ *such that, for almost all*

(d, x, z),

$$\inf_{v \in \boldsymbol{V}(d,x,z;\theta)} H_{Y_2}(y_1, d_1, d_2, x, v; \theta) \\ \leq P_0(Y_2 = 1 | Y_1 = y_1, D_1 = d_1, D_2 = d_2, X = x, Z = z) \\ \leq \sup_{v \in \boldsymbol{V}(d,x,z;\theta)} H_{Y_2}(y_1, d_1, d_2, x, v; \theta), \quad \text{(B.1)}$$

$$\inf_{\tilde{v} \in \boldsymbol{V}_{D_2}(y_1,d_1,x,z;\theta)} H_{D_2}(y_1, d_1, x, z_2, \tilde{v}; \theta) \\ \leq P_0(D_2 = 1 | Y_1 = y_1, D_1 = d_1, X = x, Z = z) \\ \leq \sup_{\tilde{v} \in \boldsymbol{V}_{D_2}(y_1,d_1,x,z;\theta)} H_{D_2}(y_1, d_1, x, z_2, \tilde{v}; \theta), \quad \text{(B.2)}$$

and

$$\inf_{v_1 \in \boldsymbol{V}_{Y_1}(d_1,x,z_1;\theta)} H_{Y_1}(d_1, x, v_1; \theta) \\ \leq P_0(Y_1 = 1 | D_1 = d_1, X = x, Z_1 = z_1) \\ \sup_{v_1 \in \boldsymbol{V}_{Y_1}(d_1,x,z_1;\theta)} H_{Y_1}(d_1, x, v_1; \theta). \quad \text{(B.3)}$$

In the corollary, the conditional independence assumptions (i), (ii), (iii) for the IVs are useful in constructing informative bounds on μ_1, π_2, and μ_2, respectively. As in Example 1, π_1 can be point identified from the selection equation in period 1 if $V_1 \perp Z_1 | X$.

B.2 Multinomial Choice with a Generalized Selection Model

Suppose an individual chooses an option Y out of mutually exclusive alternatives $\mathcal{Y} = \{1, \ldots, J\}$ by maximizing her utility:[33]

$$Y \in \arg\max_{j \in \mathcal{Y}} \mu_j(D, X) + U_j. \quad \text{(B.4)}$$

The individual's utility from alternative j depends on whether she is enrolled in a certain program ($D = 1$) or not ($D = 0$).[34] For example, Sosa-Rubí et al. (2009) analyze the choice of pregnant women in Mexico who choose sites for their obstetric care. The treatment of

[33] Similar to Section 5.1, one can allow further heterogeneity by replacing U_j with $U_{j,D}$ in this model.
[34] It is also possible to let μ be a function of individual-specific unobservables (e.g., random coefficients) and treat them as part of U. For simplicity, we do not pursue this extension here.

interest is enrollment in a public health insurance program that provides access to health services for vulnerable populations.

Suppose there is a binary instrument (e.g., eligibility), and $D = D(1)Z + D(0)(1 - Z)$ with

$$D(z) = 1\{\pi(z, X) \geq V_z\}. \tag{B.5}$$

This is the flexible selection model in Example 1. Allowing for flexibility is relevant in this context, as the insurance program may not be mandatory for the eligible or exclusive against the non-eligible. The set-valued control function is as in (2.11).

Let $V = (V_0, V_1)$ and let $U_k = Q_k(\eta; X, V), k = 1, \ldots, J$. This model's prediction is

$$\begin{aligned} &\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F) \\ &\quad = \Big\{ j \in \mathcal{Y} : \mu_j(D, X) \geq \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} \Big(\max_{k \neq j} [\mu_k(D, X) + Q_k(\eta; X, V)] - Q_j(\eta; X, V) \Big) \Big\}. \end{aligned} \tag{B.6}$$

Each element of $\boldsymbol{Y}$ is the maximizer of the utility index $\mu_k(D, X) + Q_k(\eta; X, V), k = 1, \ldots, J$ for some $V \in \mathrm{Sel}(\boldsymbol{V})$. When $\boldsymbol{V}$ is singleton-valued,

$$\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F) = \arg\max_{j \in \mathcal{Y}} \mu_j(D, X) + Q_j(\eta; X, V). \tag{B.7}$$

The model prediction in (B.7) nests Petrin and Train's (2010) specification, which assumes the additive separability of Q_j between functions of V and η:[35]

$$Q_j(\eta; X, V) = g(V; \lambda) + Q_j(\eta_j). \tag{B.8}$$

Let $\theta \equiv (\mu_1, \ldots, \mu_J, \pi, F)$. Also, let $A \subseteq \mathcal{Y}$. One can show the containment functional is

$$\begin{aligned} &\mathbb{C}_\theta(A | D = d, X = x, Z = z) \\ &= \sum_{\{j_1, \ldots, j_m\} \subset A} F_\eta \Big(\mu_{j_\ell}(d, x) \geq \inf_{v \in \boldsymbol{V}(d, x, z; \pi)} \Big(\max_{k \neq j_\ell} [\mu_k(d, x) + Q_k(\eta; x, v)] - Q_{j_\ell}(\eta; x, v) \Big), \ell = 1, \ldots, m \Big). \end{aligned} \tag{B.9}$$

The containment functional can be computed by simulating $\eta \sim U[0, 1]^J$. The following corollary characterizes $\Theta_I(P_0)$, applying Theorem 1.

[35] In their notation, $g(V; \lambda)$ is $CF(\mu_n; \lambda)$, and $Q_j(\eta_j)$ is $\tilde{\varepsilon}_{nj}$. Their specification only allows V to shift the location of the conditional distribution of U. They show that this specification holds for several parametric models of $U|V$.

COROLLARY 5: *Suppose $U \perp Z|X, V$. Then, $\Theta_I(P_0)$ is the set of parameter values $\theta = (\mu_1, \ldots, \mu_J, \pi, F)$ such that, for almost all (d, x, z),*

$$P_0(Y \in A|D = d, X = x, Z = z) \geq \sum_{B \subseteq A} F_\eta\Big(\Big\{\mu_{j_\ell}(d,x) \geq \inf_{v \in \boldsymbol{V}(d,x,z;\pi)} \Big(\max_{k \neq j_\ell}[\mu_k(d,x) + Q_k(\eta; x, v)] - Q_{j_\ell}(\eta; x, v)\Big)\Big\} \cap \Big\{\mu_{j_m}(d,x) < \inf_{v \in \boldsymbol{V}(d,x,z;\pi)} \Big(\max_{k \neq j_m}[\mu_k(d,x) + Q_k(\eta; x, v)] - Q_{j_m}(\eta; x, v)\Big)\Big\}, j_\ell \in B, j_m \in A \backslash B\Big),$$
$$A \subseteq \{1, \ldots, J\}. \quad \text{(B.10)}$$

As in the previous example, (B.10) jointly restricts μ, F (through Q), and π (via $\boldsymbol{V}$). Suppose further that $V \perp Z|X$. Then, π is point identified as $\pi(z, x) = P_0(D = 1|Z = z, X = x)$, which also ensures $\tilde{\pi}$ in (2.11) is point identified.

B.3 Decisions as Corner Solutions or Censored Decisions

Another important class of selection models are models with censored, missing, or interval data. Consider censored D and suppose

$$D = \max\{D^*, 0\}, \quad \text{(B.11)}$$
$$D^* = \pi^*(Z, X) + V. \quad \text{(B.12)}$$

This model is nested in (2.3) as $D = \pi(Z, X, V) \equiv \max\{\pi^*(Z, X) + V, 0\}$. If D results as a corner solution of an economic agent's optimization, we may be interested in $Y = \mu(D, X) + U$. Examples of such D are the hours of training or the amount of subsidy affecting certain outcomes. On the other hand, if D results from data censoring, it is reasonable to consider $Y = \mu(D^*, X) + U$. The latter D can be viewed as a special case of interval regressor (Manski and Tamer, 2002a), namely, $[D^l, D^u]$ that satisfy

$$D^* \in [D^l, D^u], \text{ a.s.} \quad \text{(B.13)}$$

This latter case is related to interval data mentioned in Example 3, although we considered interval variable D rather than interval-observed V here.

Consider (B.11)–(B.12). Then, we have a set-valued CF as

$$\boldsymbol{V}(d, x, z; \pi^*) = \begin{cases} [-\pi^*(z, x), \infty) & \text{if } d > 0, \\ (-\infty, -\pi^*(z, x)] & \text{if } d = 0. \end{cases}$$

Assume $E[U|Z,X,V] = E[U|X,V]$, which implies $E[U|D,X,V] = E[U|X,V]$. Then, it holds that $E[Y|D,X,V] = \mu(D,X) + \lambda(X,V)$ where $\lambda(x,v) \equiv E[U|X=x,V=v]$. Accordingly, define

$$\boldsymbol{Y}(\eta, D, X, Z; \mu, F, \pi^*) \equiv \mathrm{cl}\big\{y \in \mathcal{Y} : y = \mu(D,X) + \lambda(X,V) + \eta, V \in \mathrm{Sel}(\boldsymbol{V}(D,X,Z;\pi^*))\big\}.$$

Then, by Theorem 2,

$$\mu(d,x) + \lambda_l(d,x,z) \leq E[Y|D=d, X=x, Z=z] \leq \mu(d,x) + \lambda_u(d,x,z),$$

where

$$\lambda_l(d,x,z) \equiv \inf_{v \in \boldsymbol{V}(d,x,z;\pi^*)} \lambda(x,v), \qquad \lambda_u(d,x,z) \equiv \sup_{v \in \boldsymbol{V}(d,x,z;\pi^*)} \lambda(x,v).$$

Next, consider (B.12)–(B.13). This example involves two set-valued control functions:

$$\begin{aligned}\boldsymbol{V}(D^l, D^u, X, Z; \pi^*) &\equiv [D^l - \pi^*(Z,X), D^u - \pi^*(Z,X)], \\ \boldsymbol{D}^*(D^l, D^u) &\equiv [D^l, D^u].\end{aligned}$$

Note that (B.13) implies $V \in \boldsymbol{V}$ and $D \in \boldsymbol{D}^*$ a.s. Assume that $E[U|D^l, D^u, X, V] = E[U|X,V]$ and $\mu(d^*, x)$ is weakly increasing in d^*.[36] Then, define

$$\boldsymbol{Y}(\eta, D^l, D^u, X, Z; \mu, F, \pi^*) \equiv \mathrm{cl}\big\{y \in \mathcal{Y} : y \in \mu(D^*, X) + \lambda(X,V) + \eta, D^* \in \mathrm{Sel}(\boldsymbol{D}^*), V \in \mathrm{Sel}(\boldsymbol{V})\big\}.$$

Then, by slightly modifying Theorem 2 and its proof, one can show that

$$\mu(d^l, x) + \lambda_l(d^l, d^u, x, z) \leq E[Y|D^l = d^l, D^u = d^u, X = x, Z = z] \leq \mu(d^u, x) + \lambda_u(d^l, d^u, x, z),$$

where

$$\lambda_l(d^l, d^u, x, z) \equiv \inf_{v \in \boldsymbol{V}(d^l, d^u, x, z)} \lambda(x,v), \qquad \lambda_u(d^l, d^u, x, z) \equiv \sup_{v \in \boldsymbol{V}(d^l, d^u, x, z)} \lambda(x,v).$$

Manski and Tamer (2002a) focus on characterizing bounds on the conditional mean $E[Y|D^*, X]$, which is related to our parameter of interest, μ. The main difference is that we assume a selection process and utilize control variables under different identifying assumptions.

[36] Analogous assumptions appear in Manski and Tamer (2002a), although they do not consider control variables or a treatment selection process.

C Proofs

C.1 Proofs of Theorems 1, 2, and 3

Proof of Theorem 1. By Assumptions 1, one may represent the outcome as $Y = \mu(D, X, U) = \mu(D, X, Q(\eta; X, V))$. By Assumption 2, V is a measurable selection of $\boldsymbol{V}$, and therefore Y is a measurable selection of $\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F)$. Therefore, the model's prediction is summarized by

$$Y \in \boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F), \ a.s. \tag{C.1}$$

By Assumption 2 (ii), $\boldsymbol{V}$ is a function of (D, X, Z). Hence, one may condition on $(D, X, \boldsymbol{V})$ by conditioning on (D, X, Z). By Artstein's inequality (see Molinari, 2020, Theorem A.1.), the distribution $P_0(A|D, X, Z)$ is the conditional law of a measurable selection of $\boldsymbol{Y}$ if and only if

$$P_0(A|D, X, Z) \geq \mathbb{C}_\theta(A|D, X, Z), \ \forall A \in \mathcal{F}(\mathbb{R}^{d_Y}). \tag{C.2}$$

This ensures the representation of the sharp identification region by the inequalities above. $\square$

Proof of Theorem 2. Let $\mathfrak{B} \equiv \sigma(D, X, Z)$ be the σ-algebra generated by (D, X, Z). By Assumptions 2 and 3, we may represent the model's set-valued prediction by $\boldsymbol{Y}$ in (4.6), the random set of outcomes $Y = \mu(D, X) + \lambda_D(X, V) + \eta_D$, where $\eta = (\eta_d, d \in \mathcal{D})$ is conditionally mean independent of D. $\boldsymbol{Y}$ is integrable because its measurable selection Y is assumed to be integrable. Because of $Y \in \mathrm{Sel}^1(\boldsymbol{Y})$, the model's prediction on the conditional mean is summarized by

$$E_{P_0}[Y|\mathfrak{B}] \in \mathbb{E}[\boldsymbol{Y}(\eta, D, X, Z; \mu, F)|\mathfrak{B}], \ a.s., \tag{C.3}$$

where the right-hand side is the conditional Aumann expectation of $\boldsymbol{Y}$. Let $b \in \{-1, 1\}$. Then, (C.3) is equivalent to

$$bE_{P_0}[Y|\mathfrak{B}] \leq s(b, \mathbb{E}[\boldsymbol{Y}(\eta, D, X, Z; \mu, F)|\mathfrak{B}]), \quad \text{for all } b \in \{-1, 1\}, \tag{C.4}$$

where $s(b, K) \equiv \sup_{x \in K} bx$ is the support function of K.

Now, we use the convexification property (Molinari, 2020, Theorem A.2.) of the Aumann expectation of random closed sets to exchange the support function and expectation in (C.4). Technically, this property relies on the probability space used to define $\boldsymbol{Y}$ (Molchanov, 2017, Sec. 2.1.2). Hence, we proceed as follows. Let $\Omega = \mathbb{R}^{d_U} \times \mathbb{R}^{d_D} \times \mathbb{R}^{d_X} \times \mathbb{R}^{d_Z}$ be the sample space, and let $\mathfrak{F} = \mathfrak{F}_{\mathbb{R}^{d_U}} \otimes \mathfrak{F}_{\mathbb{R}^{d_D}} \otimes \mathfrak{F}_{\mathbb{R}^{d_X}} \otimes \mathfrak{F}_{\mathbb{R}^{d_Z}}$ be the product σ-algebra, where $\mathfrak{F}_E$ is the Borel σ-algebra over E. Let $\mathbb{F}$ be a probability measure on $(\Omega, \mathfrak{F})$. Measurable maps (η, D, X, Z) are defined on this space. Consider a measurable rectangle $A = A_\eta \times A_{D,X,Z}$, where $A_\eta \subset \mathbb{R}^{d_U}$ and $A_{D,X,Z} \subset \mathbb{R}^{d_D} \times \mathbb{R}^{d_X} \times \mathbb{R}^{d_Z}$. Then, $\mathbb{F}(A|\mathfrak{B}) = F_\eta(A_\eta)$. By Assumption 4 and the construction of η, F_η is atomless. Since any $A \in \mathfrak{F}$ can be approximated by a countable union of measurable rectangles, conclude that $\mathbb{F}$ is atomless over $\mathfrak{B}$.[37]

Therefore, we can apply the convexification theorem (Molinari, 2020, Theorem A.2.), which yields that $\mathbb{E}[\boldsymbol{Y}(\eta, D, X, Z; \mu, F)|\mathfrak{B}]$ is convex and

$$s(b, \mathbb{E}[\boldsymbol{Y}(\eta, D, X, Z; \mu, F)|\mathfrak{B}]) = E[s(b, \boldsymbol{Y}(\eta, D, X, Z; \mu, F))|\mathfrak{B}],\ b \in \{1, -1\}. \tag{C.5}$$

For $b = 1$,

$$\begin{aligned} E[s(b, \boldsymbol{Y}(\eta, D, X, Z; \mu, F))|\mathfrak{B}] &= E[\sup_{Y \in \mathrm{Sel}(\boldsymbol{Y}(\eta, D, X, Z; \mu, F))} Y|\mathfrak{B}] \\ &= \mu(d, x) + E[\sup_{V \in \mathrm{Sel}(\boldsymbol{V}(D, X, Z; \pi))} \lambda_d(X, V)|\mathfrak{B}] \\ &= \mu(d, x) + \sup_{v \in \boldsymbol{V}(d, x, z; \pi)} \lambda_d(x, v). \end{aligned} \tag{C.6}$$

By (C.4)-(C.6), $E_{P_0}[Y|D = d, X = x, Z = z] \leq \mu(d, x) + \sup_{v \in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x, v)$. For $b = -1$, the argument is similar. $\square$

Proof of Theorem 3. Let $\theta \in \Theta_I(P_0)$. Let $\boldsymbol{W} \equiv \{X\} \times \boldsymbol{V}$. Define

$$\mathfrak{K}_I(d; \theta) \equiv \{\kappa(d) \in \mathbb{R} : \kappa(d) = \int \varphi(\mu(d, x, Q(\eta; w))d\eta dF_W(w),\ W \in \mathrm{Sel}(\boldsymbol{W})\}. \tag{C.7}$$

This set collects the values of $\kappa(d)$ compatible with θ for some measurable selection W of $\boldsymbol{W}$. The sharp identification region for $\kappa(d)$ is

$$\mathfrak{K}_I(d) = \bigcup_{\theta \in \Theta_I(P_0)} \mathfrak{K}_I(d; \theta). \tag{C.8}$$

Hence, for the conclusion of the theorem, it suffices to show $\mathfrak{K}_I(d; \theta) = [\underline{\kappa}(d; \theta), \overline{\kappa}(d; \theta)]$.

For this, we represent $\mathfrak{K}_I(d; \theta)$ as the Aumann expectation of a random set and apply the convexification theorem. Define

$$\boldsymbol{K}(d; \theta) \equiv \Big\{r \in \mathbb{R} : r = \int \varphi(\mu(d, x, Q(\eta; W))d\eta,\ W \in \mathrm{Sel}(\boldsymbol{W})\Big\}. \tag{C.9}$$

[37] An event $A' \in \mathfrak{B}$ is called a $\mathfrak{B}$-atom if $\mathbb{F}(0 < \mathbb{F}(A|\mathfrak{B}) < \mathbb{F}(A'|\mathfrak{B})) = 0$ for all $A \subset A'$ such that $A \in \mathfrak{F}$.

Then, by construction, $\mathfrak{K}_I(d;\theta)$ is the Aumann expectation of $\boldsymbol{K}(d;\theta)$. Under the assumption that the underlying probability space is non-atomic, we may apply the convexification theorem (Molinari, 2020, Theorem A.2.). It ensures $\mathfrak{K}_I(d;\theta) = \mathbb{E}[\boldsymbol{K}(d;\theta)]$ is a convex closed set. Since φ is bounded, $\mathfrak{K}_I(d;\theta)$ is a bounded closed interval. Again, by Theorem A.2. of Molinari (2020), its upper bound is

$$\begin{aligned} s(1, \boldsymbol{K}(d;\theta)) &= s(1, \mathbb{E}[\boldsymbol{K}(d;\theta)]) = E[s(1, \boldsymbol{K}(d;\theta))] \\ &= E[\sup_{w\in \boldsymbol{W}} \int \varphi(\mu(d,x,Q(\eta;W))d\eta] = E[\sup_{v\in \boldsymbol{V}} \int \varphi(\mu(d,x,Q(\eta;X,v))] = \overline{\kappa}(d;\theta), \end{aligned} \tag{C.10}$$

where we used $\boldsymbol{W} = \{X\} \times \boldsymbol{V}$. The argument for the lower bound is similar and is omitted. $\square$

C.2 Lemmas

Lemma 1: *Suppose μ is a measurable function. Then, $\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F)$ is a random closed set.*

Proof. $\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F)$ being closed is immediate from the definition. We show its measurability below. Write $\boldsymbol{Y}(\eta(\omega), D(\omega), X(\omega), \boldsymbol{V}(\omega); \mu, F)$ as $\boldsymbol{Y}(\omega)$ for short. Since $\boldsymbol{V}$ is a random closed set, there is a sequence $\{V_n\}$ such that $\boldsymbol{V} = \mathrm{cl}(\{V_n, n \geq 1\})$ by Theorem 1.3.3 in Molchanov (2017). Let $\upsilon_n(\omega) = \mu(D(\omega), X(\omega), Q(\eta(\omega), V_n(\omega)))$ and note that $\boldsymbol{Y} = \mathrm{cl}(\{\upsilon_n, n \geq 1\})$ by Lemma 2. Then, for any $x \in \mathcal{Y}$, the distance function

$$\rho(x, \boldsymbol{Y}(\omega)) = \inf\{\|x - y\|, y \in \boldsymbol{Y}(\omega)\} = \inf\{\|x - \upsilon_n(\omega)\|, n \geq 1\} \tag{C.11}$$

is a random variable in $[0, \infty]$. Again, by Theorem 1.3.3 in Molchanov (2017), the conclusion follows. $\square$

Consider a random closed set $\boldsymbol{X}$ that is nonempty almost surely. A countable family of selections $\xi_n \in \mathrm{Sel}(\boldsymbol{X}), n \geq 1$ is called the *Castaing representation* of $\boldsymbol{X}$ if $\boldsymbol{X} = \mathrm{cl}(\{\xi_n, n \geq 1\})$. Such representation exists for any random closed set (Molchanov, 2017).

Lemma 2: *Let $\boldsymbol{X}$ be a random closed set, and let $\{\xi_n, n \geq 1\}$ be its Castaing representation. For each $\omega \in \Omega$, let $\boldsymbol{Y}(\omega) \equiv cl\{y \in \mathcal{Y} : y = f(\omega, \xi(\omega)), \xi \in \mathrm{Sel}(\boldsymbol{X})\}$ for a measurable map $f : \Omega \times \mathcal{X} \to \mathcal{Y}$. Then $\boldsymbol{Y}$ is a random closed set with a Castaing representation $\{\upsilon_n\}$ with $\upsilon_n(\omega) = f(\omega, \xi_n(\omega))$ for $n \geq 1$.*

Proof. Le $\{y_n, n \geq 1\}$ be an enumeration of a countable dense set in $\mathcal{Y}$. For each $n, k \geq 1$ and $\omega \in \Omega$, let $C_{k,n}(\omega) = \{x \in \mathcal{X} : f(\omega, x) \cap B_{2^{-k}}(y_n) \neq \emptyset\}$. Let $\boldsymbol{X}_{k,n} \equiv \boldsymbol{X} \cap C_{k,n}$ if the intersection is nonempty and let $\boldsymbol{X}_{k,n} = \boldsymbol{X}$ otherwise. Note that $\boldsymbol{X}_{k,n}$ itself is a random closed set. For each k, n, there is $m \in \mathbb{N}$ such that ξ_m is a measurable selection of $\boldsymbol{X}_{k,n}$. For each ω with $y \in \boldsymbol{Y}(\omega)$, we have $y \in B_{2^{-k}}(y_n)$ for some k, n, and

$$\|y - \upsilon_{k,n}\| \leq \|y - y_n\| + \|y_n - \upsilon_{k,n}\| \leq 2^{-k+1}, \tag{C.12}$$

where $\upsilon_{k,n} = f(\omega, \xi_m)$. Therefore, the conclusion follows. □

C.3 Proofs of Corollaries

Proof of Corollary 2. We show Assumptions 2-4 and invoke Theorem 2. First, define

$$\boldsymbol{V}(D, Z, X; \pi) = \begin{cases} [0, \pi(Z, X)] & \text{if } D = 1 \\ [\pi(Z, X), 1] & \text{if } D = 0. \end{cases} \tag{C.13}$$

Then Assumption 2 (i) holds by the selection equation (2.7). Assumption 2 (ii) holds because $\boldsymbol{V}$ is a function of (D, Z, X) and π. Assumptions 3-4 hold by hypothesis. By Theorem 2, each $\theta = (\mu, F, \pi)$ in the sharp identified set satisfies

$$\mu(d, x) + \lambda_L(d, x, z) \leq E_{P_0}[Y|D = d, X = x, Z = z] \leq \mu(d, x) + \lambda_U(d, x, z).$$

Rearranging them yields

$$E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_U(d, x, z) \leq \mu(d, x) \leq E_{P_0}[Y|D = d, X = x, Z = z] - \lambda_L(d, x, z).$$

Note that μ does not depend on z. Taking the supremum of the lower bounds and taking the infimum of the upper bounds over $z \in \mathcal{Z}$ yields the desired result. □

Proof of Corollary 3. The main argument is essentially the same as the proof of Corollary 2. Hence, we omit it. Here, we derive (5.27). By Theorem 2,

$$\lambda_U(d, x, z) = \sup_{v \in \boldsymbol{V}(d,x,z;\pi)} \lambda_d(x, v), \tag{C.14}$$

where $v = (v_1, v_2, v_s)$. By (5.25), the identifying restrictions in (5.27) follow. One can show (5.28) by a similar argument.

□

Proof of Corollary 4. We first note that, conditional on (X, U_1, V_1, V_2), the endogenous variables (Y_1, D_1, D_2) are a function of the instruments determined by the following triangular system:

$$\begin{aligned} D_2 &= 1\{\pi_2(Y_1, D_1, Z_2, X) \geq V_2\} \\ Y_1 &= 1\{\mu_1(D_1, X) \geq U_1\} \\ D_1 &= 1\{\pi_1(Z_1, X) \geq V_1\}. \end{aligned}$$

Therefore, Assumption 1 follows from $U_2 \perp (Z_1, Z_2)|X, U_1, V_1, V_2$, which in turn is implied by $(U_1, U_2, V_1, V_2) \perp (Z_1, Z_2)|X$. The sets $\boldsymbol{V}_{U_1}, \boldsymbol{V}_1, \boldsymbol{V}_2$ are defined by inverting the triangular system above with respect to (U_1, V_1, V_2), which ensures Assumption 2 (i). Assumption 2 (ii) also holds because these sets are functions of (D, Z, X) and π.

By Theorem 1, $\theta \in \Theta_I(P_0)$ if and only if

$$P_0(Y = 1|D = d, X = x, Z = z) \geq \mathbb{C}_\theta(\{1\}|D = d, X = x, Z = z) \tag{C.15}$$

$$P_0(Y = 0|D = d, X = x, Z = z) \geq \mathbb{C}_\theta(\{0\}|D = d, X = x, Z = z). \tag{C.16}$$

By (5.15) (and as argued in the text),

$$\mathbb{C}_\theta(\{1\}|D = d, X = x, Z = z) = \inf_{v \in \boldsymbol{V}(d,x,z;\pi)} H(d, x, v) \tag{C.17}$$

$$\mathbb{C}_\theta(\{0\}|D = d, X = x, Z = z) = 1 - \sup_{v \in \boldsymbol{V}(d,x,z;\pi)} H(d, x, v) \tag{C.18}$$

The identifying restriction (B.1) follows from (C.15)-(C.18) and noting that $P_0(Y = 0|D = d, X = x, Z = z) = 1 - P_0(Y = 1|D = d, X = x, Z = z)$.

The identifying restrictions (B.2)-(B.3) follow from applying the same argument sequentially. For example, letting $Y \equiv D_2$, $D \equiv (Y_1, D_1)$, $U \equiv V_2$, and $V \equiv (U_1, V_1)$ and applying the argument above yields (B.2). $\square$

Proof of Corollary 5. We show Assumptions 1-2 and invoke Theorem 1. Assumption 1 holds because D is a function of (Z, X, V) in (2.10), and $U \perp Z|X, V$. Define

$$\boldsymbol{V}(D, Z, X; \pi) = \begin{cases} \big\{v \in \mathbb{R}^2 : \tilde{\pi}(Z, X) + (1 - Z)v_0 + Zv_1 \geq 0\big\} & \text{if } D = 1 \\ \big\{v \in \mathbb{R}^2 : \tilde{\pi}(Z, X) + (1 - Z)v_0 + Zv_1 \leq 0\big\} & \text{if } D = 0, \end{cases} \tag{C.19}$$

where $\tilde{\pi}(z, x) = \pi(0, x) + z(\pi(1, x) - \pi(0, x))$. Assumption 2 (i) holds by (2.9) and (2.10). Also, Assumption 2 (ii) holds because $\boldsymbol{V}$ is a function of (D, Z, X) and π. By Theorem 1, θ is in the sharp identified set if and only if $P_0(A|D, X, Z) \geq \mathbb{C}_\theta(A|D, X, Z)$ holds.

For the main result, it remains to show (B.9). Let $A \subset \{1, \ldots J\}$ and write the model's prediction as $\boldsymbol{Y}$ for short. Then,

$$
\begin{aligned}
&\{\boldsymbol{Y} \subseteq A\} \\
&= \bigcup_{B \subseteq A} \{\boldsymbol{Y} = B\} \\
&= \bigcup_{B \subseteq A} \left(\bigcap_{j_\ell \in B} \left\{ \mu_{j_\ell}(D, X) \geq \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} \left(\max_{k \neq j_\ell} [\mu_k(D, X) + Q_k(\eta; X, V)] - Q_{j_\ell}(\eta; X, V) \right) \right\} \right) \\
&\qquad \cap \left(\bigcap_{j_m \in A \setminus B} \left\{ \mu_{j_n}(D, X) < \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} \left(\max_{k \neq j_m} [\mu_k(D, X) + Q_k(\eta; X, V)] - Q_{j_m}(\eta; X, V) \right) \right\} \right).
\end{aligned}
$$

Conditioning on (D, X, Z) and evaluating the probability on the right-hand side by F_η yields (B.9).

$\square$

C.4 Proof of Propositions

Proof of Proposition 1. As before, we write $\boldsymbol{Y}$ for $\boldsymbol{Y}(\eta, D, X, \boldsymbol{V}; \mu, F)$. Let $\varepsilon = Q(\eta)$. Then,

$$
\begin{aligned}
0 \in \boldsymbol{Y} \quad & \text{if } \mu(D, X) + \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) + \varepsilon \leq c_L \\
3 \in \boldsymbol{Y} \quad & \text{if } c_L < \mu(D, X) + \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) + \varepsilon \cap \mu(D, X) + \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) + \varepsilon \leq c_U \\
6 \in \boldsymbol{Y} \quad & \text{if } c_U < \mu(D, X) + \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) + \varepsilon,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
0 \in \boldsymbol{Y} \quad & \text{if } \varepsilon \leq c_L - \mu(D, X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\
3 \in \boldsymbol{Y} \quad & \text{if } c_L - \mu(D, X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \cap \varepsilon \leq c_U - \mu(D, X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\
6 \in \boldsymbol{Y} \quad & \text{if } c_U - \mu(D, X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon.
\end{aligned}
$$

Hence, if $c_L - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < c_U - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V)$,

$$\boldsymbol{Y} = \begin{cases} \{6\} & c_U - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \\ \{3,6\} & c_U - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_U - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{3\} & c_L - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_U - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{0,3\} & c_L - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_L - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{0\} & \varepsilon \le c_L - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V). \end{cases}$$

If $c_U - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \le c_L - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V)$,

$$\boldsymbol{Y} = \begin{cases} \{6\} & c_U - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \\ \{3,6\} & c_L - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_U - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{0,3,6\} & c_U - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_L - \mu(D,X) - \inf_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{0,3\} & c_L - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) < \varepsilon \le c_U - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V) \\ \{0\} & \varepsilon \le c_L - \mu(D,X) - \sup_{V \in \mathrm{Sel}(\boldsymbol{V})} g(V). \end{cases}$$

Let $F(\cdot) = Q^{-1}(\cdot)$. Then,

$$\mathbb{C}_\theta(\{0\}|d,x,z) = F(c_L - \mu(d,x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)) \tag{C.20}$$

$$\begin{aligned}\mathbb{C}_\theta(\{3\}|d,x,z) = [&F(c_U - \mu(d,x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)) \\ &- F(c_L - \mu(d,x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v))] \vee 0\end{aligned} \tag{C.21}$$

$$\mathbb{C}_\theta(\{6\}|d,x,z) = 1 - F(c_U - \mu(d,x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v)) \tag{C.22}$$

$$\mathbb{C}_\theta(\{0,3\}|d,x,z) = F(c_U - \mu(d,x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)) \tag{C.23}$$

$$\mathbb{C}_\theta(\{3,6\}|d,x,z) = 1 - F(c_L - \mu(d,x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v)) \tag{C.24}$$

$$\mathbb{C}_\theta(\{0,6\}|d,x,z) = \mathbb{C}_\theta(\{0\}|x) + \mathbb{C}_\theta(\{6\}|x). \tag{C.25}$$

Note that

$$\mathbb{C}^*_\theta(\{0\}|x) = 1 - \mathbb{C}_\theta(\{3,6\}|x) = F(c_L - \mu(d,x) - \inf_{v \in \boldsymbol{V}(d,x,z)} g(v)) \tag{C.26}$$

$$\mathbb{C}^*_\theta(\{6\}|x) = 1 - \mathbb{C}_\theta(\{0,3\}|x) = 1 - F(c_U - \mu(d,x) - \sup_{v \in \boldsymbol{V}(d,x,z)} g(v)). \tag{C.27}$$

Assumptions 1-2 ensure that we can invoke Theorem 1. By Theorem 1, θ is in the sharp identification region if and only if $P_0(A|d,x,z) \geq \mathbb{C}_\theta(A|d,x,z), A \in \mathcal{C}$. The restriction $P_0(\{0,6\}|d,x,z) \geq \mathbb{C}_\theta(\{0,6\}|d,x,z)$ is redundant because (C.25) shows $\mathbb{C}_\theta(\{0,6\}|d,x,z)$ is the sum of $\mathbb{C}_\theta(\{0\}|x)$ and $\mathbb{C}_\theta(\{6\}|x)$. Note that, for any $A \in \mathcal{C}$, $P_0(A|d,x,z) \geq \mathbb{C}_\theta(A|d,x,z)$ is equivalent to $P_0(A^c|d,x,z) \leq \mathbb{C}^*_\theta(A^c|d,x,z)$. Hence, $P_0(\{3\}|d,x,z) \geq \mathbb{C}_\theta(\{3\}|d,x,z)$ is also redundant. This leaves us with the following restrictions

$$\begin{aligned}
P_0(\{0\}|d,x,z) &\geq \mathbb{C}_\theta(\{0\}|d,x,z)\\
P_0(\{6\}|d,x,z) &\geq \mathbb{C}_\theta(\{6\}|d,x,z)\\
P_0(\{0,3\}|d,x,z) &\geq \mathbb{C}_\theta(\{0,3\}|d,x,z)\\
P_0(\{3,6\}|d,x,z) &\geq \mathbb{C}_\theta(\{3,6\}|d,x,z).
\end{aligned}$$

Again, for any $A \in \mathcal{C}$, $P_0(A|d,x,z) \geq \mathbb{C}_\theta(A|d,x,z)$ is equivalent to $P_0(A^c|d,x,z) \leq \mathbb{C}^*_\theta(A^c|d,x,z)$. Hence, the restrictions above are equivalent to

$$\begin{aligned}
P_0(\{0\}|d,x,z) &\geq \mathbb{C}_\theta(\{0\}|d,x,z)\\
P_0(\{6\}|d,x,z) &\geq \mathbb{C}_\theta(\{6\}|d,x,z)\\
P_0(\{6\}|d,x,z) &\leq \mathbb{C}^*_\theta(\{6\}|d,x,z)\\
P_0(\{0\}|d,x,z) &\leq \mathbb{C}^*_\theta(\{0\}|d,x,z).
\end{aligned}$$

This completes the proof. □

D Empirical Application

D.1 Target Objects

Let $X \equiv (X_{HIV}, \tilde{X})$, and let β be defined similarly. We show that the ASF can be expressed as functions of the structural parameter $\theta = (\mu, c_L, c_U, g, \pi, F, F_{\tilde{X},V})$. We write $\mu(d,x) = \mu(d, x_{HIV}, \tilde{x})$.

Average Structural Function:

$$\begin{aligned}
\varphi(\theta) = \mathrm{ASF}(d, x_{HIV}) = \int 3 \times [F(c_U - \mu(d, x_{HIV}, \tilde{x}) - g(v)) - F(c_L - \mu(d, x_{HIV}, \tilde{x}) - g(v))]\\
+ 6 \times [1 - F(c_U - \mu(d, x_{HIV}, \tilde{x}) - g(v))] dF_{\tilde{X},V}(\tilde{x}, v). \quad \text{(D.1)}
\end{aligned}$$

Similarly, the average treatment effect $\mathrm{ASF}(1, x_{HIV}) - \mathrm{ASF}(0, x_{HIV})$ can be expressed as a function of θ.

D.2 Confidence Intervals

We outline how we construct confidence intervals using Kaido and Zhang (2024). With a slight abuse of notation, we write all observable variables (D, X, Z) except the outcome as X in order to keep the notation below consistent with the one used in their paper.

Let $\varphi : \Theta \to \mathbb{R}$, and let $\varphi^* \in \mathbb{R}$. Their procedure inverts a test for the following hypothesis:

$$H_0 : \varphi(\theta) = \varphi^*, \quad v.s. \quad H_1 : \varphi(\theta) \neq \varphi^*. \tag{D.2}$$

A confidence interval is obtained by comparing a cross-fit likelihood-ratio (LR) statistic to a fixed critical value.

Their test statistic is constructed as follows. First, we split the sample $i = 1, \ldots, n$ into two subsamples $\mathcal{S}_0$ and $\mathcal{S}_1$ and compute

$$T_n(\varphi^*) = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} = \frac{\prod_{i\in\mathcal{S}_0} q_{\hat{\theta}_1}(Y_i|X_i)}{\sup_{\theta\in\{\theta':\varphi(\theta')=\varphi^*\}} \prod_{i\in\mathcal{S}_0} q_\theta(Y_i|X_i)}. \tag{D.3}$$

This is a split-sample likelihood ratio statistic, where a likelihood function (see below) is evaluated at an unrestricted estimator $\hat{\theta}_1$ and a restricted estimator $\hat{\theta}_0$.

The unrestricted estimator $\hat{\theta}_1$ is constructed from sample $\mathcal{S}_1$. We use the minimizer of a sample criterion function $\hat{\mathsf{Q}}_1(\theta) = \sup_{A\in\mathcal{C}} \sum_{i\in\mathcal{S}_1} \{\mathbb{C}_\theta(A|X_i) - \hat{P}_1(A|X_i)\}_+^2$ (see, e.g., Chernozhukov et al., 2007; Chernozhukov et al., 2013) as $\hat{\theta}_1$, where $\hat{P}_1$ is the empirical (conditional) distribution of Y_i.[38]

The restricted estimator $\hat{\theta}_0$ is constructed from $\mathcal{S}_0$,

$$\hat{\theta}_0 \in \operatorname*{arg\,max}_{\theta\in\{\theta':\varphi(\theta')=\varphi^*\}} \prod_{i\in\mathcal{S}_0} q_\theta(Y_i|X_i) \tag{D.4}$$

Here, the function $\theta \mapsto q_\theta$ is the least-favorable-pair (LFP) based density. While we refer to Kaido and Zhang (2024) for details, we note that this density q_θ can be calculated from the model primitives and represents the worst-case distribution for controlling the test's size when testing a parameter value θ satisfying the restriction $\varphi(\theta) = \varphi^*$ against an unrestricted parameter value $\hat{\theta}_1$ (treated as a fixed parameter value by conditioning on $\mathcal{S}_1$). In this empirical exercise, q_θ is available in closed form.[39]

[38] Other choices of $\hat{\theta}_1$ are also possible as long as they are calculated from $\mathcal{S}_1$.

[39] See Supplementary Material (Lemma D.1) for its expression and derivation.

Define the cross-fit LR statistic by

$$S_n(\varphi^*) = \frac{T_n(\varphi^*) + T_n^{\text{swap}}(\varphi^*)}{2}. \tag{D.5}$$

$T_n^{\text{swap}}(\varphi^*)$ is defined similarly to $T_n(\varphi^*)$ while swapping the roles of $\mathcal{S}_0$ and $\mathcal{S}_1$. Recall that $\varphi(\theta) \in \mathbb{R}$ is the target object. We define a confidence interval by

$$CI_n = \big\{\varphi^* \in \mathbb{R} : S_n(\varphi^*) \leq \frac{1}{\alpha}\big\}. \tag{D.6}$$

Let $\mathcal{P}_{\theta,x} = \{Q \in \mathcal{M}(\Sigma_Y, \mathcal{X}) : Q(A|x) \geq \mathbb{C}_\theta(A|x),\ A \in \mathcal{C}\}$ be the set of conditional probabilities of Y satisfying Artstein's inequality at θ. Let $\mathcal{S} = \mathcal{Y} \times \mathcal{X}$, and let $\Delta(\mathcal{S}^n)$ be the space of probability measures on the product space $\mathcal{S}^n$. The set of joint distributions of the observable variables compatible with θ is

$$\mathcal{P}_\theta^n = \Big\{P^n \in \Delta(\mathcal{S}^n) : P^n = \bigotimes_{i=1}^n P_i, P_i(\cdot|x) \in \mathcal{P}_{\theta,x},\ \forall i,\ x \in \mathcal{X}\Big\}. \tag{D.7}$$

Define the sharp identification for $\varphi(\theta)$ by

$$\mathcal{H}_{P^n}[\varphi] = \{\varphi^* : \varphi(\theta) = \varphi^*, P^n \in \mathcal{P}_\theta^n,\ \text{for some } \theta \in \Theta\}. \tag{D.8}$$

The confidence interval in (D.6) has the following universal coverage property (Kaido and Zhang, 2024, Corollary 1):

$$\inf_{P^n \in \mathcal{P}_\theta^n, \theta \in \Theta} \inf_{\varphi \in \mathcal{H}_{P^n}[\varphi]} P^n(\varphi^* \in CI_n) \geq 1 - \alpha, \tag{D.9}$$

That is, it covers each element of the sharp identification region of $\varphi(\theta)$ with probability at least $1 - \alpha$ uniformly over a family of data generating processes compatible with the model.

In our application, we construct a grid of $K = 200$ equally spaced points over the range of $\varphi(\theta)$. For each φ^* in this grid, we compute $S_n(\varphi^*)$ and compare it to $1/\alpha$, where we use $\alpha = 0.05$. Each confidence interval in Table 1 consists of the smallest and largest values of of φ^* that are not rejected by the test.