cemmap

centre for microdata methods and practice

Partial Independence in Nonseparable Models

Matthew A. Masten
Alexandre Poirier

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP26/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Partial Independence in Nonseparable Models*

Matthew A. Masten† Alexandre Poirier‡

June 17, 2016

Abstract

We analyze identification of nonseparable models under three kinds of exogeneity assumptions weaker than full statistical independence. The first is based on quantile independence. Selection on unobservables drives deviations from full independence. We show that such deviations based on quantile independence require non-monotonic and oscillatory propensity scores. Our second and third approaches are based on a distance-from-independence metric, using either a conditional cdf or a propensity score. Under all three approaches we obtain simple analytical characterizations of identified sets for various parameters of interest. We do this in three models: the exogenous regressor model of Matzkin (2003), the instrumental variable model of Chernozhukov and Hansen (2005), and the binary choice model with nonparametric latent utility of Matzkin (1992).

JEL classification: C14; C21; C25; C26; C51

Keywords: Nonparametric Identification, Partial Identification, Sensitivity Analysis, Nonseparable Models, Selection on Unobservables, Instrumental Variables, Binary Choice

*This paper was presented at the 2015 Midwest Econometrics Group, Duke, the MIT econometrics lunch, the University of Chicago, the University of Virginia, and UC Davis. We thank audiences at those seminars, as well as Federico Bugni, Ivan Canay, Joachim Freyberger, Joe Hotz, Shakeeb Khan, Chuck Manski, Suyong Song, and Alex Torgovitsky, for helpful conversations and comments.

†Department of Economics, Duke University, `matt.masten@duke.edu`

‡Department of Economics, University of Iowa, `alexandre-poirier@uiowa.edu`

1 Introduction

Exogeneity is a critical assumption in much empirical work. In all models, exogeneity refers to assumptions on the statistical dependence between an observable term and an unobservable term. There are many such assumptions, however, including zero correlation, median independence, and full statistical independence. The choice of a formal definition of exogeneity is not innocuous. Different definitions have different substantive interpretations and different implications for identification. In this paper, we study three different classes of exogeneity assumptions we call *partial independence*. In the first half we study their substantive interpretation. In the second half we study their implications for identification in nonseparable models.

Our first class, called $\mathcal{T}$-*independence*, specifies that quantile independence holds at all quantiles $\tau \in \mathcal{T}$ where $\mathcal{T}$ is a possibly strict subset of $(0, 1)$. This class builds on a large literature beginning with Manski (1975) and Koenker and Bassett (1978) which has studied identification and estimation under quantile independence. This approach has faced a longstanding question, however. Manski (1988) describes it as follows:

> "Quantile independence restrictions sometimes make researchers uncomfortable. The assertion that a given quantile of u does not vary with x may lead one to ask: Why this quantile but not others? In the absence of a persuasive answer, one may feel compelled to adopt an extreme position: Either no quantiles of u are independent of x or all are. On reflection, though, one may feel that the former position understates the available information and the latter overstates it." (page 733)

Put differently, this concern is about how one can substantively interpret and judge the credibility of a given set of quantile independence conditions. To this end, our first main contribution provides a treatment assignment characterization of quantile independence conditions. Specifically, we consider the relationship between an unobserved variable U and an observed discrete treatment variable X. Supposing X is binary for simplicity, the dependence structure is fully characterized by the propensity score

$$p(u) = \mathbb{P}(X = 1 \mid U = u).$$

Constant propensity scores correspond to full statistical independence while nonconstant propensity scores represent deviations from full independence. In this sense, any exogeneity assumption weaker than full independence allows for certain kinds of selection on unobservables. Our first main theorem characterizes the set of propensity scores consistent with a set of quantile independence conditions. That is, we fully describe the kind of selection on unobservables that quantile independence allows for. We then describe several properties of this set. Most notably, nonconstant propensity scores which are consistent with a single quantile independence condition must be non-monotonic. Furthermore, if multiple quantile independence conditions hold, then nonconstant propensity scores must also oscillate up and down. These results apply anytime one makes a quantile independence assumption, including in binary choice and quantile regression models.

A first implication of these results is that prior knowledge on the shape of the propensity score may help one determine which quantile independence conditions hold. We use several stylized Roy models to illustrate the relationship between selection on unobservables and quantile independence conditions. A second implication, however, is that empirical researchers may find the non-monotonic and oscillatory properties of the set of propensity scores consistent with quantile independence undesirable. This suggests that one should consider alternative approaches to weakening full independence. This leads us to our second and third definitions of partial independence.

Our second definition, called *c-independence*, specifies that the propensity score is at most c away from the unconditional probability of being treated, in the sup-norm distance. That is, we constrain the distance between the conditional distribution of $X \mid U$ and the unconditional distribution of X. Our third definition, called *d-independence*, is similar, but bounds the distance between the conditional cdf of $U \mid X$ and the unconditional cdf of U by a constant d. Unlike $\mathcal{T}$-independence, both of these approaches allow for monotonic propensity scores, and do not require nonconstant propensity scores to oscillate. c-independence is simple to interpret since it is an assumption directly on the propensity score. We explain how $\mathcal{T}$- and c-independence are qualitatively very different kinds of deviations from independence. Both imply d-independence, however, making d-independence the qualitatively weakest assumption of the three.

After studying these three partial independence concepts on their own, in the second half of the paper we analyze their identifying power in nonparametric, nonseparable models. In our second main contribution, we first derive sharp sets of cdfs $F_{U|X}$ that are consistent with each of the partial independence concepts. These results do not depend on the specific choice of econometric model. We then show how to use these results in three popular nonseparable models: the exogenous regressor model of Matzkin (2003), the instrumental variable model of Chernozhukov and Hansen (2005), and the binary choice model with nonparametric latent utility of Matzkin (1992). Each of these papers imposes full statistical independence to obtain point identification of the entire model structure.[1] By relaxing full independence to partial independence, we typically obtain partial identification.

In our third main contribution, we obtain simple, analytical characterizations of identified sets under partial independence for various parameters of interest, in all three models. For brevity we focus our analysis primarily on Matzkin's (2003) model. In this model we study identification of the quantile structural function, the average structural function, average treatment effects, the distribution of treatment effects, and spread parameters. For the nonseparable instrumental variables model we study identification of the quantile structural function. For the binary response model we study identification of the average structural function.

Our identification results can be used in two ways. First, if an empirical researcher's prior knowledge about the shape of the propensity score can be formalized via one of our three partial

[1] Full independence, sometimes conditional on covariates, is a common assumption in the nonseparable model literature. For example, see Brown (1983), Roehrig (1988), Matzkin (2008, 2012, 2015), Altonji and Matzkin (2005), Athey and Imbens (2006), Florens, Heckman, Meghir, and Vytlacil (2008), and Imbens and Newey (2009), among many others.

independence concepts, then our identified sets describe the set of parameters consistent with the data and that researcher's maintained assumptions on the shape of the propensity score. Second, an empirical researcher might find full statistical independence credible a priori, but nonetheless want to check the sensitivity of their conclusions to this assumption. In this case our results can be used to perform a sensitivity analysis for the full independence assumption.

The rest of this paper is organized as follows. Below we review the related literature. The first half of the paper, sections 2 and 3, discusses our three concepts of partial independence. Section 2 gives the formal definitions while section 3 gives our main results on how to interpret these concepts. The second half of the paper, sections 4–7, use these concepts to do identification analysis. In section 4 we discuss the implications of partial independence assumptions on the conditional cdf of the unobservable term given the observable term. We then apply these results to each of the three nonseparable models mentioned above in sections 5, 6, and 7. Section 8 concludes.

Related literature

In the rest of this section, we review the related literature. We begin with the literature in statistics which examines deviations from full independence. We give additional details on this literature on page 47 in appendix A. Rosenbaum and Rubin (1983) do a sensitivity analysis in a fully parametric logit model for binary outcomes, binary treatments, and with deviations from independence driven by a binary omitted variable. Like most papers in this literature, their model is point identified for any known choice of the sensitivity parameters. Lin, Psaty, and Kronmal (1998) replace their binary U assumption with a normally distributed U, and also allow for duration outcomes, among other parametric modifications. Imbens (2003) extends Rosenbaum and Rubin (1983) to continuous outcomes by assuming homoskedastic normally distributed outcomes with a linear conditional mean function. Ichino, Mealli, and Nannicini (2008) avoid specifying a parametric model by fixing the distribution of $U \mid X, Y$ a priori, which requires picking four parameters under the assumption that Y and X are binary as well. We consider binary and continuous outcomes Y. Their approach requires setting multiple different sensitivity parameters, while our recommendations only use one parameter. Their model is point identified for any choice of the sensitivity parameters.

Robins, Rotnitzky, and Scharfstein (2000) extend Rosenbaum and Rubin (1983) by relaxing the parametric assumptions on the outcomes. They introduce a slight subtlety which we have not yet mentioned. The Rosenbaum and Rubin (1983) analysis ultimately relies on a conditional independence assumption holding, given the unobserved covariate. Deviations from independence are modeled via the distribution of treatment assignment given this unobserved covariate. Robins et al. (2000), however, do not model deviations from independence as arising from an unobserved covariate. They directly model the relationship between treatment assignment and potential outcomes. This is quite similar to our approach, since under rank invariance our U variable is simply a transformation of the potential outcomes. Consequently, both our analysis and that of Robins et al. (2000) do not ultimately rely on a full independence assumption. Robins et al. (2000), however, use a parametric propensity score which delivers point identification for each fixed sensitivity parameter. They also do not consider instrumental variables models. The Robins et al. (2000) approach has been developed and used in several other papers, including Rotnitzky, Robins, and Scharfstein (1998), Rotnitzky, Scharfstein, Su, and Robins (2001), and Scharfstein, Manski, and Anthony (2004).

Kline and Santos (2013) analyze deviations from the missing-at-random assumption using a sup-norm distance metric equivalent to our d-independence. Also in the context of missing data analysis, Manski (2016) proposes c-independence. In section 4.2.1, however, he notes that the identified set under c-independence "typically has no simple explicit form." While he considered missing data problems, in the models we study we are able to get simple explicit expressions for identified sets under c-independence, as well as under $\mathcal{T}$- and d-independence.

Several papers study the sensitivity of assumptions used for inference rather than assumptions used for identification. Rosenbaum (1995, 2002) proposes a sensitivity analysis within the context of randomization inference for testing the sharp null hypotheses of no unit level treatment effects for the units in our dataset. Small and Rosenbaum (2008) extend this analysis to an instrumental variables setting. This approach begins with a quite different framework and has different goals than ours. See chapter 5 of Imbens and Rubin (2015) for a general discussion of randomization inference. Chamberlain and Leamer (1976) and Leamer (1978, 1982, 1985) considered how changes in one's prior affects the corresponding posterior in Bayesian analyses. See section 2.2.2 of Leamer (1983) for additional discussion.

A large literature initiated by Manski has studied identification problems under various assumptions which typically do not point identify the parameters. Here we focus on the papers which study exogeneity assumptions. Manski and Pepper (2000) begin with a baseline assumption of mean independence and relax this to a monotonicity constraint in the conditioning variable. They derive the identified set for the average potential outcome under this assumption. They repeat this analysis in the instrumental variables case as well. Manski and Pepper (2009) build on this nonparametric analysis by considering parametric models of treatment response, among other issues. Nevo and Rosen (2012) study parametric additively separable instrumental variables models. They relax the assumption that the instrument is uncorrelated with the unobservable by allowing for correlation in the same direction as that between treatment and the unobservable, and of less magnitude than the latter correlation. They derive the analytical identified set in the linear IV model. Our motivation is similar to that of these papers: to study departures from exogeneity assumptions. However, we begin from a baseline of full independence rather than mean independence or zero correlation. Our identification results build on specific nonseparable models, rather than general potential outcome models or linear parametric models. Finally, we do not consider monotonicity restrictions or constraints on relative magnitudes and signs. Incorporating such constraints into our analysis is a useful direction for future work.

Conley, Hansen, and Rossi (2012) consider the linear IV model. They study deviations from the exclusion restriction, while maintaining a classical zero correlation assumption as their exogeneity condition. In our IV analysis, we impose exclusion throughout and focus on deviations from the full independence exogeneity condition. Hotz, Mullin, and Sanders (1997) assume there are two populations: one where mean independence holds and another where it doesn't, and our observed data are a mixture of the two. For a fixed bound on the mixing probability, they derive the identified set for ATT. By varying the bounds on the mixing probability, they provide one way of examining sensitivity to mean independence. Altonji, Elder, and Taber (2005, 2008) estimate bivariate probit selection models, but holding the correlation between the two unobservables fixed. They present a range of estimates for different values of this correlation, noting that independence holds if and only if the correlation is zero.

One of the earliest papers to study identification of a nonseparable model under weak exogeneity assumptions was Manski (1975), who analyzed the binary response model under median independence, rather than full independence. See Manski (1988) for more discussion. A large literature on semiparametric identification and estimation has subsequently developed (see Powell 1994). This literature often focuses on identification under a single quantile independence condition, while we consider identification under a set of quantile independence conditions, as well as under c- and d-independence. We also consider other models, not just the binary response model. Our baseline analysis of the binary response model follows Matzkin (1992), who studies identification of nonparametric binary threshold crossing models under full independence.

More recently, Chesher (2003) studied identification of nonseparable models under quantile independence rather than full independence. This work is an important exception to the usual practice of assuming full independence when analyzing identification of nonseparable models. He has extended this approach in several directions in Chesher (2005, 2007b,a, 2010). In Chesher (2005), he motivated this approach as follows:

> "Nonparametric identification of entire nonseparable structural functions seems to require the unpalatable restriction that latent variates and covariates are *independently* distributed as in the analysis of Roehrig (1988), Imbens and Newey [2009], and Matzkin (2003). One might be prepared to believe in a degree of local independence even in situations in which global independence seems untenable." (page 1541)

Our results are motivated by the same sentiment. We build on Chesher's work in several ways. First, under $\mathcal{T}$-independence his work shows that the structural function is point identified at some points in its domain. We note that this implies partial identification of the structural function at all points in its domain where it is not point identified. Hence we derive identified sets for the entire structural function. This is an important step for obtaining identified sets for functionals like ATE. We also derive identified sets under c- and d-independence, which both allow for the possibility that no quantile independence conditions hold at all.

Torgovitsky (2015b) provides a general framework for computing identified sets. His work can also be used to compute identified sets under partial independence in some of the models we consider. For example, his approach allows for computation of the identified set for ATE in the nonparametric binary response model under median independence, as well as under $\mathcal{T}$-independence for finite $\mathcal{T}$. Chesher and Rosen (2015) also provide a general approach to characterizing identified sets, and also consider models subject to quantile independence constraints. Rather than providing a general framework, our identification analysis is tailored to several specific models. This allows us to provide simple analytical characterizations of the identified sets. One benefit of an analytical characterization is that it transparently shows how the identified set depends on features of the data. It may also lead to simpler estimation and inference methods. We furthermore consider identification under c- and d-independence, which is not considered in these two papers.

2 Partial independence

In this section we first define the three concepts of partial independence we study in this paper. We then discuss alternative approaches to defining partial independence.

Defining partial independence

In this paper we focus on models which have an observable covariate of interest X and a scalar unobservable U. In all three models we consider, U enters a nonparametric function nonseparably. As usual, this implies that the scale of U is not separately identified from this nonparametric function. Hence we require a normalization. In all of our analysis, we use the following normalization assumption. See page 28 for further discussion.

Assumption A1 (Normalization). U is uniformly distributed on $[0, 1]$.

Exogeneity assumptions place restrictions on the joint distribution of (X, U). As discussed in the introduction, we define and analyze concepts of *partial independence* which lay between full independence and arbitrary dependence. There are many such ways to define such an intermediate assumption. We focus on three specific approaches.

The first approach is based on the well known result that statistical independence between U and X is equivalent to

$$F_{U|X}(\tau \mid x) = F_U(\tau) \tag{1}$$

for *all* $(\tau, x) \in \text{supp}(U, X)$, where supp denotes the support of the random variables. Say U is τ-cdf independent of X if (1) holds for all $x \in \text{supp}(X)$ at τ. If these cdfs are strictly increasing, then this is equivalent to the quantile independence conditions

$$Q_{U|X}(\tau \mid x) = Q_U(\tau)$$

for all $\tau \in (0, 1)$ and $x \in \text{supp}(X)$. Existing research focuses on two extreme assumptions: a single quantile independence condition holds, such as $Q_{U|X}(0.5 \mid x) = Q_U(0.5)$ for all $x \in \text{supp}(X)$, or all quantile independence conditions hold (statistical independence). Hence a natural middle ground is to suppose that *some*, but not all, of the possible quantile independence conditions hold. As we'll see later, it is more natural to work with the cdf independence condition (1). This motivates the following definition.

Definition 1. Let $\mathcal{T}$ be a subset of $(0,1)$. Say U is *$\mathcal{T}$-independent* of X if for all $\tau \in \mathcal{T}$, the cdf independence condition (1) holds for all $x \in \text{supp}(X)$.

A large literature starting with Manski (1975) and Koenker and Bassett (1978) has considered the implications of a single quantile independence assumption on identification and estimation in various models. More recently, for example, Chesher (2003) refers to such a single quantile independence condition as 'local independence'. Here we simply generalize this existing approach by also considering multiple quantile independence conditions, rather than just one. We only consider subsets $\mathcal{T}$ of $(0,1)$ since $\text{supp}(U) = [0,1]$ under the normalization A1.

Our second and third concepts use the common notion of distance-from-independence, whereby we measure failures of (1) to hold in terms of a metric on cdfs or propensity scores. In particular, we focus on the sup-norm distance. While our other two concepts allow for X to be discrete or continuous, for the second concept we focus only on discrete X's. For simplicity we consider the binary X case in our exposition, although many of our results can be extended to the discrete X case. Recall the definition of the propensity score

$$p(u) = \mathbb{P}(X = 1 \mid U = u).$$

Definition 2. Let c be a scalar between 0 and 1. Say U is *c-independent* of X if

$$\sup_{u \in [0,1]} |p(u) - \mathbb{P}(X = 1)| \leq c.$$

Using $p(u)$ rather than $1 - p(u)$ in this definition is innocuous; see page 81 in the appendix. Rosenbaum and Rubin (1983), Robins et al. (2000), and Imbens (2003) used parametric propensity scores to model departures from independence similar to our nonparametric concept of c-independence. Only c's between zero and $\overline{c} = \max\{\mathbb{P}(X = 1), \mathbb{P}(X = 0)\}$ are considered since that is the logical range of the absolute value of a difference between the propensity score and $\mathbb{P}(X = 1)$. With $c = 0$, c-independence is equivalent to full independence, while $c = \overline{c}$ imposes no constraints on the relationship between U and X whatsoever. Values of c strictly between zero and $\overline{c}$ lead to intermediate cases.

Our third and final concept is similar to c-independence, but reversing the order of X and U in the conditioning.

Definition 3. For each $x \in \text{supp}(X)$, let $d(x)$ be a scalar between 0 and 1. Say U is *d-independent* of X if

$$\sup_{u \in [0,1]} |F_{U|X}(u \mid x) - F_U(u)| \leq d(x)$$

for all $x \in \text{supp}(X)$.

This approach to measuring distance from independence was used by Manski (1983) for estimation. A slightly different definition is based on comparing $F_{U|X}(\cdot \mid 1)$ with $F_{U|X}(\cdot \mid 0)$; see page 81 in the appendix. Kline and Santos (2013) used this latter approach for partial identification analysis in missing data models. Similarly to c-independence, with $d(x) = 0$, d-independence is equivalent to full independence, while $d(x) = 1$ imposes no constraints on the relationship between U and X whatsoever. Values of $d(x)$ strictly between zero and one lead to intermediate cases.

Alternative definitions

There are several potentially useful alternative approaches to defining partial independence. The first generalizes mean independence. It is known that if U is bounded, then

$$\mathbb{E}(U^k \mid X) = \mathbb{E}(U^k) \tag{2}$$

for all $k = 1, 2, \ldots$ is equivalent to $X \perp\!\!\!\perp U$. (This is not true if U is unbounded; see de Paula 2008.) Hence an alternative approach is to assume that equation (2) holds for some, but not all $k \in \mathbb{N}$. This assumption is not generally nested with any of the three concepts of partial independence discussed above. In particular, identified sets based on (2) and those based on $\mathcal{T}$-, c-, or d-independence are not necessarily the same.

One reason to prefer the cdf based approaches, however, is that the mean based version is not invariant to monotone transforms, whereas the cdf based versions work well with monotone transforms. Specifically, for a smooth, nonlinear, strictly increasing transform f,

$$\mathbb{E}[U \mid X] = \mathbb{E}[U] \qquad \not\Leftrightarrow \qquad \mathbb{E}[f(U) \mid X] = \mathbb{E}[f(U)]$$

while

$$F_{U|X}[u \mid x] = F_U[u] \qquad \Leftrightarrow \qquad F_{f(U)|X}[f(u) \mid x] = F_{f(U)}[f(u)].$$

This invariance importantly ensures that A1 truly is a normalization. Moreover, such invariance is a well studied principle in other contexts, such as testing and decision making. For example, see chapter 6 of Berger (1985). A second reason to prefer the cdf and propensity score based approaches is practical: As will be clear from our analysis below, they are far more tractable than the approach based on means. This is because it is not straightforward to characterize the set of conditional distributions of U given X that satisfy equation (2) for some $k \in \mathbb{N}$. Whereas, as we study in section 4, it is possible to characterize the set of conditional distributions of U given X that satisfy the $\mathcal{T}$-, c-, and d-independence conditions.

A second alternative approach is based on mixing conditions. There are many possible ways of using mixing conditions to characterize statistical independence of X and U (for example, see Doukhan 1994). Each of these ways can also lead to various concepts of partial independence. As with the moments based approach, however, these mixing conditions are not as tractable as our cdf and propensity score based approaches.

A third approach would replace the sup-norm distance (i.e., a Kolmogorov-Smirnov type metric) in the definitions of c- and d-independence with an L_2-norm distance (i.e., a Cramér-von-Mises type metric). That is, measure distances between functions using average square differences in values rather than the largest difference in values. This approach is also not as tractable as that based on the sup-norm metric.

3 Interpreting partial independence

In this section we discuss four topics regarding the interpretation of partial independence. We begin by discussing two ways to choose $\mathcal{T}$, c, and d. First, as with any other modeling assumption, analysis under partial independence can be used to do sensitivity analysis (section 3.1). Second, in our first main contribution we characterize the set of propensity scores consistent with $\mathcal{T}$-independence. Hence the choice of $\mathcal{T}$ can be interpreted as a choice of a certain set of propensity scores (section 3.2). We then discuss the set of propensity scores consistent with d-independence, and compare and contrast our three definitions of partial independence (section 3.3). Finally, we illustrate the relationship between selection on unobservables and partial independence using several stylized Roy models (section 3.4). There we discuss the kinds of structural assumptions one would have to make to obtain certain partial independence restrictions.

3.1 Sensitivity analysis

Many researchers may find the full independence assumption $X \perp\!\!\!\perp U$ to be credible a priori. Full independence, however, is not innocuous—it often has identifying power. Our results can thus be used to do a sensitivity analysis: just how strongly do one's conclusions depend on assuming full statistical independence? If weakening full independence to partial independence yields a substantively small identified set, then we know that the full identifying power of statistical independence is not necessary to obtain the same substantive conclusion. Hence our confidence in our conclusions will likely be stronger. But if weakening full statistical independence leads to a substantively large identified set, then we know that full statistical independence is critical in coming to a tight conclusion. Hence the conclusion of the study will hinge on how strongly one believes in the full statistical independence assumption a priori.

To perform such a sensitivity analysis for c- and d-independence is simple. Our identification results provide an identified set for each value of $c \in [0, 1]$ and $d(x) \in [0, 1]$ and one can examine how these identified sets vary as c or $d(x)$ varies from zero to one. Performing a sensitivity analysis using $\mathcal{T}$-independence is trickier, since this requires picking the set $\mathcal{T}$. There are many possible ways to do this, and any choice of $\mathcal{T}$ that is a strict subset of $(0, 1)$ will, strictly speaking, relax the full statistical independence assumption. Formally, we say that the $\mathcal{T}$-independence assumption is weaker than the $\mathcal{T}'$-independence assumption if $\mathcal{T} \subseteq \mathcal{T}'$. The set of all possible choices one could consider, the power set of $\mathcal{T}$, is not totally ordered under set inclusion. There are many totally ordered subsets, however.

In the absence of any persuasive argument in favor of a particular one of these subsets, researchers can still make an ad hoc choice to perform sensitivity analysis. For example, one could pick a sequence of points $\{\tau_j\}_{j=1}^{\infty}$ that is dense in $(0, 1)$. Next, one could obtain the identified set under the $\mathcal{T}_J$-independence assumption, where $\mathcal{T}_J = \{\tau_1, \ldots, \tau_J\}$, a grid of J points. By increasing J our assumptions become 'closer' to full statistical independence, while by decreasing J our assumptions become 'farther' from full statistical independence. Hence the goal of the sensitivity analysis is to see how the size of the identified set varies with J. A reasonable choice of the sequence $\{\tau_j\}_{j=1}^{\infty}$ would be an equidistributed sequence, as defined in chapter 9 of Judd (1998), who notes that such sequences "[formalize] the idea of a uniformly distributed sequence". A well known example are Halton draws.

A second choice of a sequence of totally ordered subsets of the power set of $(0,1)$ is $\mathcal{T}_\tau = (0, \tau]$ for $\tau \in (0,1)$. For this choice, $\mathcal{T}$-independence means that U is t-cdf independent of X for all $t \leq \tau$, but not necessarily for any larger values. As τ gets closer to 1, our assumptions become 'closer' to full statistical independence, while as τ gets closer to 0 our assumptions become 'farther' from full statistical independence. This particular choice of $\mathcal{T}$ has a simple substantive interpretation. For example, suppose X is binary. Then there are two unknown distributions: $U \mid X = 0$ and $U \mid X = 1$. These distributions have support in $[0,1]$ under the $U \sim \text{Unif}[0,1]$ normalization. But beyond that, they are unrestricted. $\mathcal{T}_\tau$-independence says that these two distributions look identical in the left tail. Formally, $\mathcal{T}_\tau$-independence implies that $X \perp\!\!\!\perp U \mid \{U \leq \tau\}$. This is a special case of the following result, which shows that $\mathcal{T}$-independence yields full independence for a subpopulation defined by $\mathcal{T}$.

Proposition 1. Suppose $\mathbb{P}(U \in \mathcal{T}) > 0$. Then $\mathcal{T}$-independence implies $X \perp\!\!\!\perp U \mid \{U \in \mathcal{T}\}$.

The proof of this result, along with all others in the paper, is in the appendix. Consider again the choice $\mathcal{T}_\tau = (0, \tau]$. Suppose, for example, that X is an indicator of participation in a job training program and U is ability. Then $\mathcal{T}_\tau$-independence implies that all low ability people randomly choose whether to participate, while high ability people may select in to the program on the basis of their value of U. The cutoff between 'low' and 'high' ability is determined by τ. In practice, researchers can begin with the usual point estimates under full statistical independence. They can then compute our identified sets for a sequence of τ's and see, for example, how the length of the set changes as τ decreases from one to zero. A similar analysis can also be done under the choice $\mathcal{T} = [\tau, 1)$.

A third choice of a sequence of totally ordered subsets of $(0,1)$ is $\mathcal{T} = [\varepsilon, 1 - \varepsilon]$ for a decreasing sequence of ε's in $(0, 0.5)$ with $\varepsilon \to 0$. This choice is similar to the previous suggestions $(0, \tau]$ and $[\tau, 1)$. With this choice, proposition 1 implies that there is random assignment of treatment in the middle of the distribution of unobservables, but potentially nonrandom assignment in the tails.

Finally, when considering this difficult question of picking $\mathcal{T}$, we find it helpful to remember Manski's (1988) sentiment: *Any* choice of $\mathcal{T}$, however seemingly ad hoc, is still weaker than full statistical independence and hence can give researchers some idea of the identifying power of statistical independence assumptions.

3.2 $\mathcal{T}$-independence and treatment assignment rules

In the previous subsection we gave several suggested choices of $\mathcal{T}$. We showed that our second and third examples, $\mathcal{T}_\tau = (0, \tau]$ and $\mathcal{T} = [\varepsilon_J, 1 - \varepsilon_J]$, have an interpretation in terms of treatment assignment. In this subsection, we expand on this point. In particular, in our first main contribution we characterize the deviations from full independence allowed by $\mathcal{T}$-independence as a restriction on the propensity score. We study d-independence in the next subsection.

Throughout this section, we focus on the binary treatment case, $X \in \{0, 1\}$. Full statistical independence, $X \perp\!\!\!\perp U$, is equivalent to the propensity score being constant:

$$p(u) = \mathbb{P}(X = 1)$$

for all $u \in [0, 1]$. Consequently, any nonconstant propensity score represents a deviation from full independence. $\mathcal{T}$-independence restricts the form of these deviations. The following theorem characterizes the set of propensity scores consistent with $\mathcal{T}$-independence.

Theorem 1 (Average value characterization)**.** Suppose A1 holds and that X is binary with $\mathbb{P}(X = 1) \in (0, 1)$. Then U is $\mathcal{T}$-independent of X if and only if

$$\frac{1}{t_2 - t_1} \int_{[t_1, t_2]} p(u) \, du = \mathbb{P}(X = 1) \tag{3}$$

for all $t_1, t_2 \in \mathcal{T} \cup \{0, 1\}$ with $t_1 < t_2$.

With minor extra notation, this theorem extends to any discretely distributed X. We assume $\mathbb{P}(X = 1) \in (0, 1)$ since $\mathbb{P}(X = 1) \in \{0, 1\}$ implies $X \perp\!\!\!\perp U$. Theorem 1 says that $\mathcal{T}$-independence holds if and only if for every interval with endpoints in $\mathcal{T} \cup \{0, 1\}$ the average value of the propensity score over that interval equals the overall average of the propensity score, since

$$\int_0^1 p(u) \, du = \mathbb{P}(X = 1).$$

This overall average is just the unconditional probability of being treated. When $\mathcal{T}$ is finite, theorem 1 simplifies, as shown in corollary 8 in the appendix. The following lemma gives us a different view of the average value characterization of $\mathcal{T}$-independence.

Lemma 1. Suppose A1 holds and that X is binary. For $t_1, t_2 \in [0, 1]$ with $t_1 < t_2$ we have

$$\mathbb{P}(X = 1 \mid U \in [t_1, t_2]) = \frac{1}{t_2 - t_1} \int_{[t_1, t_2]} p(u) \, du.$$

Hence the average value condition (3) is

$$\mathbb{P}(X = 1 \mid U \in [t_1, t_2]) = \mathbb{P}(X = 1).$$

Thus $\mathcal{T}$-independence implies that there are many (potentially overlapping) subpopulations for which the probability of being treated within that subpopulation is the same as in the entire population. Thus partial independence implies that we have random assignment *across* subpopulations defined by each interval $[t_1, t_2]$, but it allows for nonrandom assignment *within* subpopulations, in the sense that the conditional propensity score

$$\mathbb{P}(X = 1 \mid U = u, U \in [t_1, t_2])$$

may be a nontrivial function of u.

To illustrate theorem 1, suppose $\mathcal{T} = \{0.5\}$ and $\mathbb{P}(X = 1) = 0.5$. Here we have just a single nontrivial cdf independence condition which is analogous to the classical median independence assumption. Figure 1 plots five different propensity scores which are consistent with $\mathcal{T}$-independence under this choice of $\mathcal{T}$. This figure illustrates several features of such propensity scores: The value of $p(u)$ may vary over the entire range $[0, 1]$. p does not need to be symmetric about $\tau_1 = 0.5$, nor does it need to be continuous. Finally, as suggested by the pictures, p must actually be nonmonotonic; we show this in corollary 1 next.

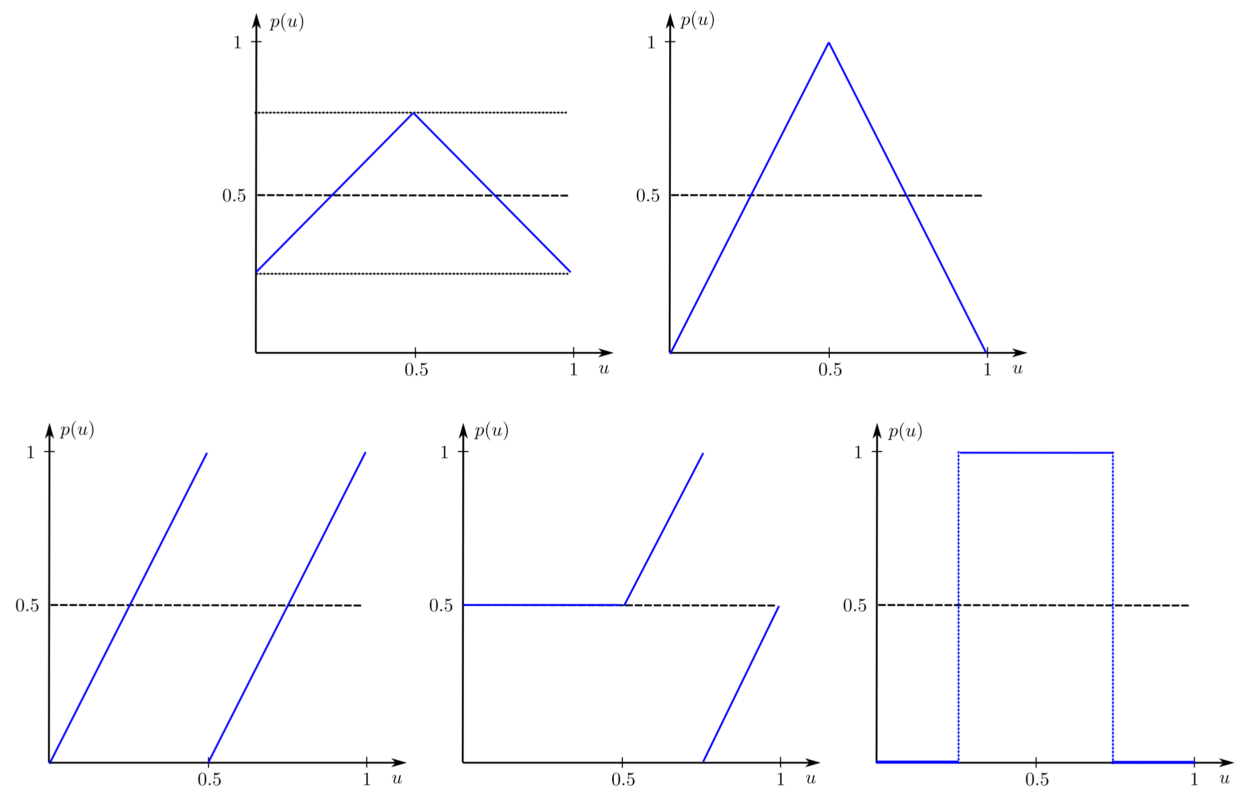


Figure 1: Various propensity scores consistent with $\mathcal{T} = \{0.5\}$-independence, when $\mathbb{P}(X = 1) = 0.5$.

Corollary 1. Suppose the propensity score p is weakly monotonic and not constant. Then U is not τ-cdf independent of X for all $\tau \in (0, 1)$.

Hence monotonic propensity scores imply that no nontrivial cdf independence conditions can hold. For example, none of the propensity scores in figure 3 are consistent with any nontrivial cdf independence condition. Put differently, corollary 1 shows that a nonconstant propensity score must be non-monotonic if it is to satisfy a τ-cdf independence condition. This result can be extended as follows. Say that a function f *changes direction at least K times* if there exists a partition of its domain into K intervals such that f is not monotonic on each interval.

Corollary 2. Suppose U is $\mathcal{T}$-independent of X. Partition $[0,1]$ by the sets $\mathcal{U}_k = [t_{k-1}, t_k)$ for $k = 1, \ldots, K-1$ with $t_0 = 0$, $t_K = 1$, and $\mathcal{U}_K = [t_{K-1}, t_K]$ and such that for each k there is a $\tau_k \in \mathcal{T}$ with $\tau_k \in \mathcal{U}_k$. Suppose p is not constant over each set $\mathcal{U}_k$, $k = 1, \ldots, K$. Then p changes direction at least K times.

This result essentially says that such propensity scores must oscillate up and down at least K times. For example, as in figure 1, suppose we continue to have $\mathbb{P}(X = 1) = 0.5$ but we add a few more τ's to $\mathcal{T}$. Figure 2 shows several propensity scores consistent with $\mathcal{T}$-independence for larger choices of $\mathcal{T}$. These sawtooth and triangle propensity scores illustrate the oscillation required by corollary 2.

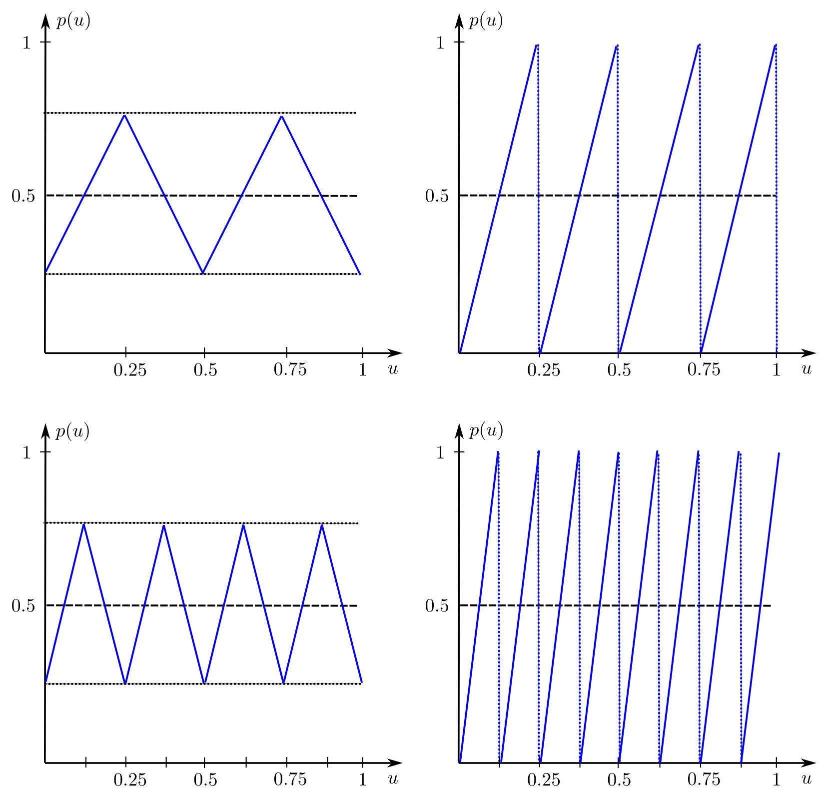


Figure 2: Some propensity scores consistent with $\mathcal{T}$-independence when $\mathbb{P}(X = 1) = 0.5$. Top: $\mathcal{T} = \{0.25, 0.5, 0.75\}$. Bottom: $\mathcal{T} = \{0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875\}$.

One final feature we document is that as long as there is some interval which is not in $\mathcal{T}$ then there is a propensity score which takes the most extreme values possible, 0 and 1.

Corollary 3. Suppose $(0,1) \setminus \mathcal{T}$ contains a non-degenerate interval. Then there exists a propensity score which is consistent with $\mathcal{T}$-independence and for which the sets

$$\{u \in [0,1] : p(u) = 0\} \qquad \text{and} \qquad \{u \in [0,1] : p(u) = 1\}$$

have positive measure.

Figure 6 on page 25 illustrates this point. For any finite $\mathcal{T}$, no matter how large, we can always find such propensity scores that are consistent with $\mathcal{T}$-independence.

Our results in this section are not specific to any given model; they only concern the stochastic relationship between U and X. Consequently, our characterization of the deviations from full independence allowed by $\mathcal{T}$-independence applies to any model based on quantile/cdf independence at a specific τ. This includes semiparametric binary response models based on median independence (Manski 1975), quantile regression models (Koenker and Bassett 1978), and work on identification of nonseparable models under 'local independence' (Chesher 2003). Hence, in those models, for discrete X, the kinds of joint distributions of (U, X) which are allowed and do not correspond to full independence must have non-monotonic propensity scores, and always allow for propensity scores which take extreme values.

Non-monotonic and oscillatory propensity scores may not be the kind of deviations from full independence empirical researchers wish to allow for. A perhaps more natural approach is c-independence, which places restrictions directly on the propensity score. For example, figure 3 illustrates several different propensity scores satisfying c-independence. All of these are monotonic propensity scores and hence are ruled out by $\mathcal{T}$-independence.

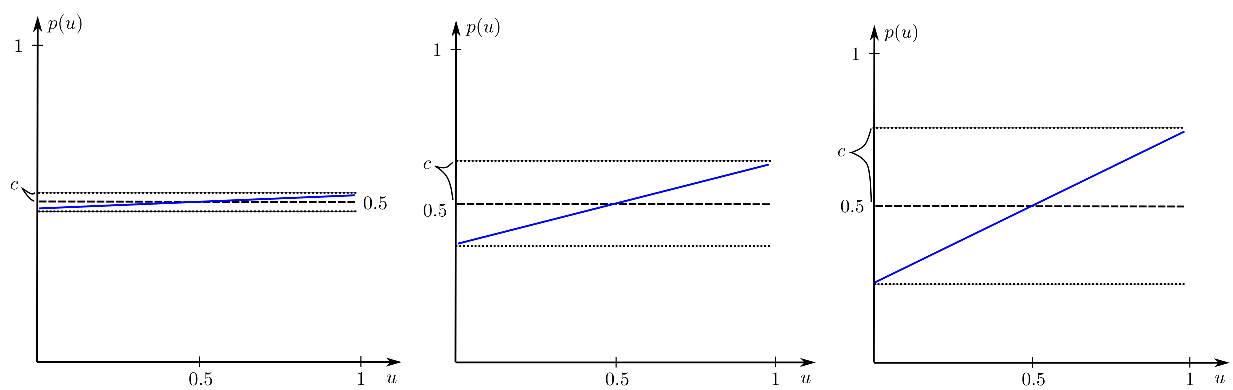


Figure 3: Some propensity scores satisfying c-independence with $\mathbb{P}(X = 1) = 0.5$, for various values of c.

The first plot in figure 3 illustrates a key difference between $\mathcal{T}$- and c-independence: For any $c \in (0, 1]$ and any nonempty $\mathcal{T} \subseteq (0, 1)$, c-independence does not imply $\mathcal{T}$-independence. In particular, for any arbitrarily small positive c, there always exists a monotonic propensity score which satisfies c-independence, but due to its monotonicity does not satisfy τ-cdf independence for any $\tau \in (0, 1)$. Conversely, because of corollary 3, for any subset $\mathcal{T} \subseteq (0, 1)$ which is missing an interval and any $c \in [0, 1)$, $\mathcal{T}$-independence does not imply c-independence. In particular, no matter how 'close' $\mathcal{T}$ is to $(0,1)$, we can always find a propensity score whose sup-norm distance from $\mathbb{P}(X = 1)$ is as large as possible. For these reasons, these two concepts are qualitatively different.

There are, however, propensity scores which are consistent with both $\mathcal{T}$ and c-independence for any arbitrarily small c and any $\mathcal{T}$ arbitrarily 'close' to $(0,1)$ containing an interval. These are just the propensity scores which oscillate but remain close to $\mathbb{P}(X = 1)$ in magnitude. In practice empirical researchers may wish to impose both $\mathcal{T}$- and c-independence. For brevity we omit this case from our identification analysis below.

3.3 d-independence

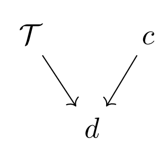
In the previous subsection we characterized the kinds of propensity scores consistent with $\mathcal{T}$-independence. Since these propensity scores have perhaps undesirable properties, we discussed placing restrictions directly on the propensity score via c-independence. d-independence is a similar concept, although the restrictions are placed on the distribution of $U \mid X$ rather than the distribution of $X \mid U$. In this subsection we study the properties of d-independence.

While we do not explicitly give a characterization of the set of propensity scores consistent with d-independence, our theorem 4 in section 4 does so implicitly by characterizing the set of cdfs $F_{U|X}$ consistent with d-independence, which can then be translated to a set of propensity scores. In this section, we instead briefly discuss the relationships between d-independence and the other two concepts of partial independence. The following proposition shows that c-independence implies d-independence.

Proposition 2. Suppose X is binary. Let $c \in [0, \bar{c}]$ for $\bar{c} = \max\{\mathbb{P}(X = 1), \mathbb{P}(X = 0)\}$. Suppose c-independence holds. Then d-independence holds with $d(1) = c/\mathbb{P}(X = 1)$ and $d(0) = c/\mathbb{P}(X = 0)$.

The converse, however, does not hold. For any $d(x) \in (0,1]$ and $c \in [0, \bar{c})$, d-independence does not imply c-independence. This can be seen by constructing a counterexample, such as that in figure 8 on page 27. Put differently, regardless of how small $d(x)$ is, there always exists a propensity score which takes on the extreme values of 0 and 1. From this same example we also see that monotonic propensity scores are allowed. Hence for any $d(x) \in (0,1)$ and any nonempty $\mathcal{T} \subseteq (0,1)$, d-independence does not imply $\mathcal{T}$-independence. The converse, however, does hold: $\mathcal{T}$-independence implies d-independence for a nontrivial d. And as $\mathcal{T}$ gets 'closer' to $(0,1)$, the implied d converges to zero. Although we omit a formal statement of this result, it can be seen by examining figure 6 in the next section. This figure shows the cdfs of $F_{U|X}$ which are farthest from the diagonal cdf $F_U(u) = u$. The corresponding sup-norm distance can thus be read off from these two cdfs.

The diagram at right summarizes the qualitative relationships between the three partial independence concepts. Any missing arrows indicate that there is not an implication in that direction. $\mathcal{T}$- and c-independence are generally unrelated.



Both imply d-independence, but not conversely. Hence d-independence is the qualitatively weakest partial independence concept we consider.

3.4 Partial independence in Roy models of treatment choice

A key message of theorem 1 and the subsequent results above is that failure of full independence can be thought of as a kind of selection on unobservables. Hence failure of full independence is not a purely statistical issue. In this final subsection, we illustrate this point in the context of several stylized Roy models of treatment choice. In particular, we show what kinds of assumptions on structural parameters are necessary to deliver partial independence.

Suppose there is a population of agents $\mathcal{I}$. Each person $i \in \mathcal{I}$ has a number $U_i \in [0,1]$ associated with them. There is a binary treatment. There is a function m such that $m(X_i, U_i)$ is the outcome person i experiences if they choose treatment X_i. Each person i knows m and their value U_i. They choose X_i to maximize $m(X_i, U_i)$:

$$X_i = 1 \qquad \text{if and only if} \qquad m(1, U_i) \geq m(0, U_i).$$

Assume for simplicity that either ties happen with zero probability or people always choose treatment when ties occur. Let

$$\mathcal{U} = \{u \in [0,1] : m(1,u) \geq m(0,u)\}.$$

This is the set of values of the unobservable that lead a person to choose $X = 1$. Hence

$$X = \mathbb{1}(U \in \mathcal{U}).$$

In this model, treatment is deterministically determined by one's value of U. Despite this deterministic selection, we will apply theorem 1 to show that some non-trivial quantile independence conditions may hold. Specifically, we will show that the location of the set $\mathcal{U}$ completely determines what quantile independence conditions arise. This set $\mathcal{U}$, in turn, depends solely on the structure of the unit level causal effects $m(1,u) - m(0,u)$.

First we consider an example with perfect complementarity between U and the treatment X. Suppose

$$m(1,u) = u \qquad \text{and} \qquad m(0,u) = 1 - u.$$

Then $\mathcal{U} = [1/2, 1]$. This setup is illustrated in figure 4. For example, if X indicates the decision to attend college, U is unobserved ability, and $m(X,U)$ is one's wage, this model says that all students in the top half of the ability distribution attend college, while none of the students with below median ability attend college. Thus the propensity score in this model is

$$p(u) = \begin{cases} 0 & \text{if } u < 1/2 \\ 1 & \text{if } u \geq 1/2. \end{cases}$$

This propensity score is monotonic and hence *no* nontrivial quantile independence conditions hold.

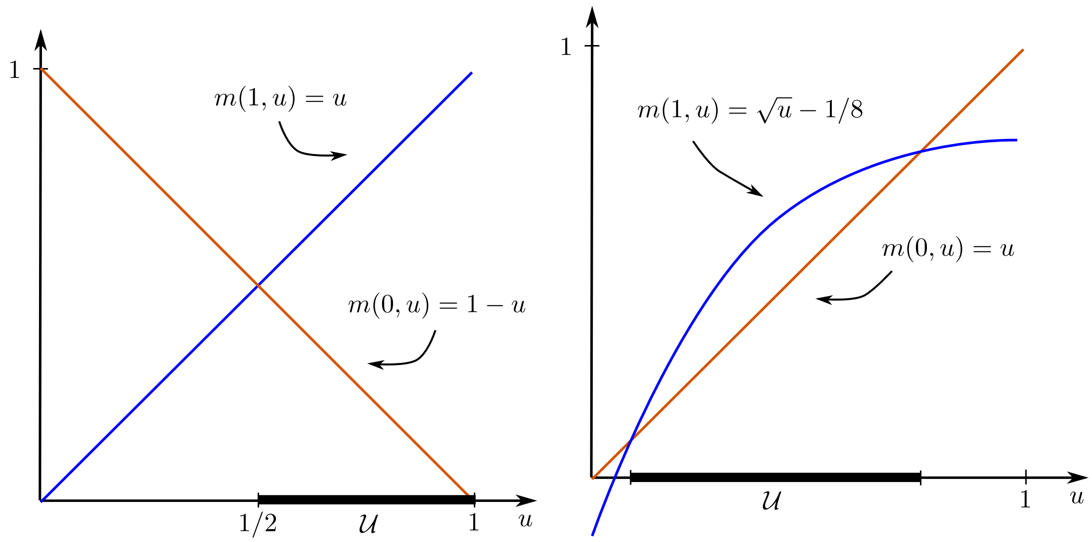


Figure 4: Obtaining partial independence from structural Roy models: On the left, with perfect complementarity between U and treatment X, no nontrivial quantile independence conditions hold. On the right, with a cost of treatment and concave returns, a single nontrivial quantile independence condition holds.

Next suppose

$$m(1,u) = \sqrt{u} - 1/8 \qquad \text{and} \qquad m(0,u) = u.$$

In this example, there are diminishing returns to ability under treatment, but linear returns without treatment. Moreover, there is a fixed cost of treatment. Here $\mathcal{U} \approx [0.021, 0.728]$. This setup is also illustrated in figure 4. Think about the college attendance decision again. In this model, only 'middle' ability people find it worthwhile to attend college. The low and high ability people are both better off not attending college. In this model, the propensity score is approximately

$$p(u) = \begin{cases} 0 & \text{if } u < 0.021 \\ 1 & \text{if } 0.021 \leq u \leq 0.728 \\ 0 & \text{if } u > 0.728. \end{cases} \tag{4}$$

This propensity score is non-monotonic and changes direction exactly once. Hence it is possible that exactly one nontrivial quantile independence restriction holds. By applying theorem 1 we can numerically compute which quantile the independence restriction holds at. Specifically, we need to find the τ which makes the proportion of treated people the same above and below the cutoff:

$$\frac{\tau - 0.021}{\tau} = \mathbb{P}(X = 1 \mid U \leq \tau) = \mathbb{P}(X = 1 \mid U \geq \tau) = \frac{0.728 - \tau}{1 - \tau}.$$

Solving for τ yields $\tau = 0.073$.

In these examples we assumed m was known. This is clearly not a useful starting place for practice since identifying and estimating m is often the main goal of an empirical analysis. Firstly, one goal of these examples is to simply illustrate the connection between selection models and partial independence, as an application of theorem 1. Beyond this, however, the examples and our earlier results suggest a qualitative framework for thinking about how to choose τ: It depends on one's prior beliefs about the shape of the propensity score. In fact, notice that in these examples knowledge of m and the decision rule imply knowledge of the propensity score. But we only required knowledge of the propensity score to compute τ; we did not need to know m precisely. And many different m functions can lead to the same propensity score. Hence any m which leads to the propensity score (4) will yield $\mathcal{T} = \{\tau\}$-independence with $\tau = 0.073$.

We can generalize this point even further. In the above examples, treatment was deterministic given U. Hence the propensity score only takes the values 0 and 1. Therefore only a single propensity score can be consistent with a certain τ-cdf independence condition. If we allow for all propensity scores, not just those that take the values 0 and 1, then we obtain a large class of propensity scores that are consistent with τ-cdf independence. This class is described by theorem 1.

To obtain such propensity scores in the Roy model, it is necessary to allow for additional variables to affect people's treatment decisions. To this end, we consider the *generalized* Roy model (Heckman and Vytlacil 2007). In this model person i incurs the cost V_{1i} of choosing $X_i = 1$ and likewise for V_{0i}. These costs are unobserved to us, the analysts. Each person i chooses treatment to maximize outcomes minus costs:

$$X_i = 1 \qquad \text{if and only if} \qquad m(1, U_i) - V_{1i} \geq m(0, U_i) - V_{0i}.$$

Hence the decision rule is

$$X = \mathbb{1}[V \leq m(1, U) - m(0, U)]$$

where $V \equiv V_1 - V_0$. In the college attendance example, V_1 represents the utility or disutility of physically attending college. Alternatively, one could think of V_1 and V_0 as errors in people's expectations about their own outcomes.

The propensity score is

$$\begin{aligned} p(u) &= \mathbb{P}(X = 1 \mid U = u) \\ &= \mathbb{P}(V \leq m(1, u) - m(0, u) \mid U = u) \\ &= F_{V|U}[m(1, u) - m(0, u) \mid u]. \end{aligned} \tag{5}$$

The shape of the propensity score, and hence the set $\mathcal{T}$, depends on the shapes of both

1. the heterogeneous treatment effects, $m(1, u) - m(0, u)$, and

2. the conditional cdf of $V \mid U$.

Consequently, in order for any nontrivial cdf independence conditions to hold, we must have non-monotonicities in either $F_{V|U}$ (as a function of u) or in the treatment effects $m(1,u) - m(0,u)$. The precise location of the τ's in $\mathcal{T}$, in turn, depend on the location of the direction changes in the propensity score p, which depend on the two pieces above. Thus if one wants to allow for deviations from full independence via $\mathcal{T}$-independence then the choice of τ's in $\mathcal{T}$ depends on prior beliefs about the shapes of the two objects above. As mentioned earlier, the shape constraints imposed by $\mathcal{T}$-independence do not pin down a unique propensity score and also do not require precise parametric knowledge of m. In section 5 we will see the implications of $\mathcal{T}$-independence on identification of m.

Next we consider the implications of c-independence in the generalized Roy model. c-independence holds if the propensity score (5) is not farther than c from $\mathbb{P}(X = 1)$. For example, suppose we observe that $\mathbb{P}(X = 1) = 0.5$. Suppose $V \perp\!\!\!\perp U$ with $V \sim \mathcal{N}(0, 1)$. Then

$$p(u) = \Phi[m(1,u) - m(0,u)].$$

First notice that if treatment effects are homogeneous then full independence holds. If we instead allow for heterogeneous treatment effects, but restrict their magnitude to be between $[-1, 1]$ then $p(u)$ is approximately between $[0.16, 0.84]$ and hence c-independence holds with $c = 0.34$. Our magnitude assumption on the range of treatment effects here is relevant only with respect to the scale of V. This follows because shrinking the variance of V down to zero brings us back to the case where the propensity score is either zero or one, and hence when c-independence does not hold. So, in this example, a substantively restrictive c-independence assumption is akin to a statement that treatment effect variation does not swamp the unobservables V which also drive treatment choice.

In this subsection we have illustrated the relationship between selection on unobservables and partial independence. In a generalized Roy model of treatment choice, we outlined a qualitative approach for selecting $\mathcal{T}$ or c based on prior beliefs about the shape of the two components which make up the propensity score. If one finds non-monotonicities in these components implausible, then this suggests that exogeneity based on c- or d-independence may be more plausible than $\mathcal{T}$-independence.

4 Partial identification of cdfs under partial independence

As we show in sections 5, 6, and 7, our identification results can be split into two steps. The first step is to characterize sets of cdfs $F_{U|X}$ consistent with each of the three partial independence assumptions and the normalization A1. We present these results, which are our second main contribution, in this section. These results do not depend on the specific choice of econometric model.

Specifically, let $x \in \text{supp}(X)$. We first characterize the set of all cdfs $F_{U|X}(\cdot \mid x)$ which are consistent with $\mathcal{T}$-independence and A1, where $\mathcal{T}$ is a union of intervals. We denote this set by $\mathcal{F}^{\mathcal{T}}_{U|X}(x)$. We repeat this analysis for c- and d-independence to obtain the sets $\mathcal{F}^{c}_{U|X}(x)$ and $\mathcal{F}^{d}_{U|X}(x)$. For distinct $x, x' \in \text{supp}(X)$ we also characterize the set of all cdfs $(F_{U|X}(\cdot \mid x), F_{U|X}(\cdot \mid x'))$ that are consistent with $\mathcal{T}$-partial independence and A1. We denote this set by $\mathcal{F}^{\mathcal{T}}_{U|X}(x, x')$. We repeat this analysis for c- and d-independence. For $\mathcal{T}$-independence we consider the discretely supported X case while for c- and d-independence we focus on the binary X case. Our derivations in this section are variations on the decomposition of mixtures problem. See Cross and Manski (2002), Manski (2007) chapter 5, and Molinari and Peski (2006) for more discussion.

4.1 Bounds under $\mathcal{T}$-independence

We begin with $\mathcal{T}$-independence. Throughout this subsection we only consider discretely supported X. We consider continuously supported X on page 82 in the appendix.

Assumption A2 (Discrete support). $\text{supp}(X) = \{x_1, \ldots, x_K\}$ for some positive integer K.

Next, we restrict our attention to sets $\mathcal{T}$ which can be written as a union of closed intervals.

Assumption A3 ($\mathcal{T}$-independence). U is $\mathcal{T}$-independent of X, where $\mathcal{T}$ is a union of disjoint closed intervals:

$$\mathcal{T} = \bigcup_{j=1}^{J} [a_j, b_j] \tag{6}$$

where $b_j < a_{j+1}$ for $j = 1, \ldots, J-1$.

Throughout our discussion, we let $a_0 = b_0 = 0$ and $a_{J+1} = b_{J+1} = 1$ to simplify some expressions. The specification of $\mathcal{T}$ in equation (6) includes most cases of interest. With $J = 0$ we obtain $\mathcal{T} = \emptyset$. In this case the distribution of $U \mid X = x$ is completely unrestricted. With $J = 1$ and $[a_1, b_1] = [0, 1]$ we obtain full independence. In this case $\mathcal{F}^{\mathcal{T}}_{U|X}(x)$ is a singleton: $\{F_U\}$. Finally, with $a_j = b_j$ for $j = 1, \ldots, J$ the set $\mathcal{T}$ is a grid of points.

Let $p_k = \mathbb{P}(X = x_k)$ denote the probability masses. Without loss of generality, assume $p_k > 0$ for all $k = 1, \ldots, K$. Let

$$\begin{aligned} h_k(u) &= F_{U|X}(u \mid x_k) \\ &= \mathbb{P}(U \leq u \mid X = x_k) \end{aligned}$$

denote the conditional cdf of U given $X = x_k$.

For a given set $\mathcal{T}$ written as in equation (6), define the following functions:

$$\overline{h}_k^J(u) = \begin{cases} \dfrac{u}{p_k} & \text{if } b_0 \leq u \leq p_k a_1 \\ a_1 & \text{if } p_k a_1 \leq u \leq a_1 \\ u & \text{if } a_1 \leq u \leq b_1 \\ \dfrac{u - b_1}{p_k} + b_1 & \text{if } b_1 \leq u \leq (1-p_k)b_1 + p_k a_2 \\ a_2 & \text{if } (1-p_k)b_1 + p_k a_2 \leq u \leq a_2 \\ u & \text{if } a_2 \leq u \leq b_2 \\ \quad\vdots & \\ 1 & \text{if } (1-p_k)b_J + p_k \tau_{J+1} \leq u \leq a_{J+1} \end{cases} \tag{7}$$

and

$$\underline{h}_k^J(u) = \begin{cases} b_0 & \text{if } b_0 \leq u \leq p_k b_0 + (1-p_k)a_1 \\ \dfrac{u - a_1}{p_k} + a_1 & \text{if } p_k b_0 + (1-p_k)a_1 \leq u \leq a_1 \\ u & \text{if } a_1 \leq u \leq b_1 \\ b_1 & \text{if } b_1 \leq u \leq p_k b_1 + (1-p_k)a_2 \\ \dfrac{u - a_2}{p_k} + a_2 & \text{if } p_k b_1 + (1-p_k)a_2 \leq u \leq a_2 \\ u & \text{if } a_2 \leq u \leq b_2 \\ \quad\vdots & \\ \dfrac{u - a_{J+1}}{p_k} + a_{J+1} & \text{if } p_k b_J + (1-p_k)a_{J+1} \leq u \leq a_{J+1}. \end{cases} \tag{8}$$

These are stepwise linear cdfs on $[0,1]$: they are non-decreasing, right-continuous, and are equal to 0 and 1 when evaluated at 0 and 1, respectively. The following theorem shows that $\mathcal{F}_{U|X}^{\mathcal{T}}(x_k)$ is the set of all cdfs which satisfy a constraint on how quickly they increase and which lie between these two cdfs $\underline{h}_k^J$ and $\overline{h}_k^J$.

Theorem 2. Suppose A1, A2, and A3 hold. Let $\mathcal{F}_k$ be the set of all cdfs h_k on $[0,1]$ that satisfy

$$\frac{h_k(u') - h_k(u)}{u' - u} \leq \frac{1}{p_k}$$

for any $0 \leq u < u' \leq 1$. Then for any $k = 1, \ldots, K$,

$$\mathcal{F}_{U|X}^{\mathcal{T}}(x_k) = \left\{ h_k \in \mathcal{F}_k : \underline{h}_k^J(u) \leq h_k(u) \leq \overline{h}_k^J(u) \text{ for all } u \in [0,1] \right\},$$

and this set is sharp. Moreover, $(\underline{h}_k^J, \overline{h}_l^J), (\overline{h}_k^J, \underline{h}_l^J) \in \mathcal{F}_{U|X}^{\mathcal{T}}(x_k, x_l)$, for any $x_k, x_l \in \text{supp}(X)$.

The last part of this theorem shows that the upper and lower bounds can be jointly attained.

This will be useful for obtaining bounds on treatment effect parameters like the average treatment effect. The constraint that $h_k \in \mathcal{F}_k$ prevents the cdf from increasing too quickly. This is necessary because a cdf h_k that violates this constraint implies, by the law of total probability, that the corresponding cdf $\mathbb{P}(U \leq u \mid X \neq x_k)$ would be decreasing in u. A consequence of this restriction is that cdfs h_k with discontinuities—distributions of $U \mid X$ with mass points—are ruled out when X is discrete.

Examples

To illustrate this result, we consider two special cases: $\mathcal{T} = \emptyset$ and $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$. Consider the first case. Even with no quantile independence constraints, the normalization A1 and discrete support A2 deliver a nontrivial set of cdfs $F_{U|X}(\cdot \mid x)$, because of the law of total probability:

$$u = F_U(u) = \sum_{k=1}^{K} h_k(u) p_k.$$

Here we also see the relationship to the decomposition of mixtures problem: We know the marginal cdf $F_U(u)$ and the weights p_k, and are interested in identifying the mixing distributions $h_k(u)$, subject to various constraints.

Equations (7) and (8) simplify to the following functions.

$$\overline{h}_k^{\emptyset}(u) = \begin{cases} \dfrac{u}{p_k} & \text{if } 0 \leq u \leq p_k \\ 1 & \text{if } p_k \leq u \leq 1 \end{cases} \qquad \text{and} \qquad \underline{h}_k^{\emptyset}(u) = \begin{cases} 0 & \text{if } 0 \leq u \leq 1 - p_k \\ \dfrac{u-1}{p_k} + 1 & \text{if } 1 - p_k \leq u \leq 1. \end{cases} \tag{9}$$

$\overline{h}_k^{\emptyset}$ and $\underline{h}_k^{\emptyset}$ are both valid cdfs. They are plotted in figure 5, along with their corresponding propensity scores. Both cdfs are piecewise linear functions with slopes 0 and $1/p_k$. Here and below we use the following helpful result.

Lemma 2. Suppose A1 and A2 hold. Then for each $k = 1, \ldots, K$,

$$\mathbb{P}(X = x_k \mid U = u) = \mathbb{P}(X = x_k) f_{U|X}(u \mid x_k).$$

This lemma shows that the propensity score is just the scaled density of $U \mid X$. Hence by computing the densities corresponding to $\overline{h}_k^{\emptyset}$ and $\underline{h}_k^{\emptyset}$ and then rescaling them we obtain the corresponding propensity scores.

Let $\mathcal{F}_{U|X}^{\emptyset}(x_k)$ denote the sharp set of cdfs consistent with A1 and A2, but with no further independence constraints, as obtained via theorem 2. This set does exclude some cdfs. As $p_k \to 0$, however, this set approaches the set of all possible cdfs. In theorem 8 on page 82 in the appendix we formally derive bounds on h_k for continuous X and show that they are the limit of the discrete X bounds as $p_k \to 0$.

Next consider the case $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$ with $\tau_0 = 0$, $\tau_{J+1} = 1$. Equations (7) and (8) simplify

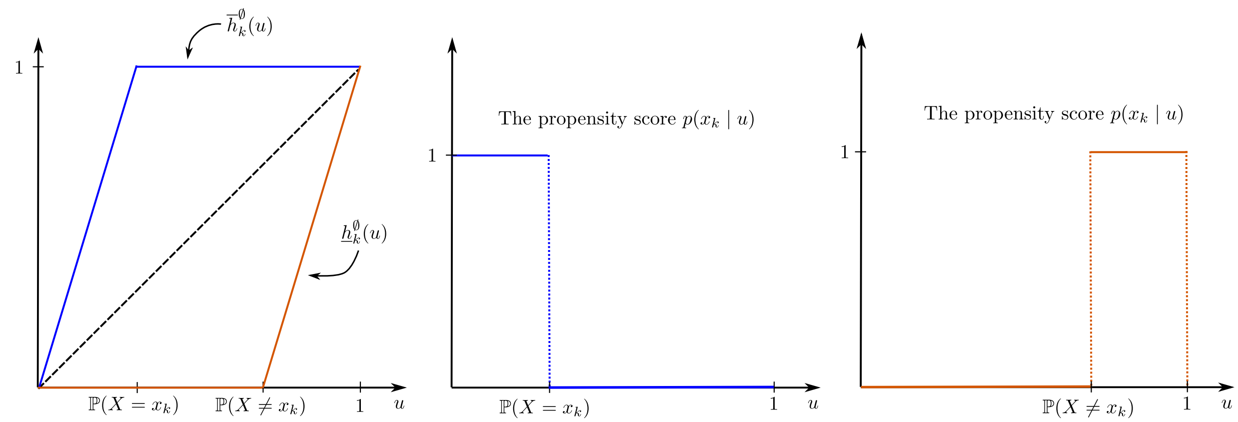


Figure 5: Left: Upper and lower bounds on $h_k(u) = \mathbb{P}(U \leq u \mid X = x_k)$ under the normalization A1 and discrete support A2, but with no further independence constraints. Middle and Right: The corresponding propensity scores.

to the following functions:

$$\overline{h}^J_k(u) = \begin{cases} \dfrac{u - \tau_0}{p_k} + \tau_0 & \text{if } \tau_0 \leq u \leq (1-p_k)\tau_0 + p_k\tau_1 \\ \tau_1 & \text{if } (1-p_k)\tau_0 + p_k\tau_1 \leq u \leq \tau_1 \\ \dfrac{u - \tau_1}{p_k} + \tau_1 & \text{if } \tau_1 \leq u \leq (1-p_k)\tau_1 + p_k\tau_2 \\ \tau_2 & \text{if } (1-p_k)\tau_1 + p_k\tau_2 \leq u \leq \tau_2 \\ \quad\vdots & \\ \tau_{J+1} & \text{if } (1-p_k)\tau_J + p\tau_{J+1} \leq u \leq \tau_{J+1} \end{cases} \tag{10}$$

and

$$\underline{h}^J_k(u) = \begin{cases} \tau_0 & \text{if } \tau_0 \leq u \leq p_k\tau_0 + (1-p_k)\tau_1 \\ \dfrac{u - \tau_1}{p_k} + \tau_1 & \text{if } p_k\tau_0 + (1-p_k)\tau_1 \leq u \leq \tau_1 \\ \tau_1 & \text{if } \tau_1 \leq u \leq p_k\tau_1 + (1-p_k)\tau_2 \\ \dfrac{u - \tau_2}{p_k} + \tau_2 & \text{if } p_k\tau_1 + (1-p_k)\tau_2 \leq u \leq \tau_2 \\ \quad\vdots & \\ \dfrac{u - \tau_{J+1}}{p_k} + \tau_{J+1} & \text{if } p_k\tau_J + (1-p_k)\tau_{J+1} \leq u \leq \tau_{J+1}. \end{cases} \tag{11}$$

$\overline{h}^J_k$ and $\underline{h}^J_k$ are both valid cdfs. An example is plotted in figure 6, along with their corresponding propensity scores, for $\mathcal{T} = \{\tau_1, \tau_2\}$ with just two points. As before, these cdfs are piecewise linear functions with slopes equal to either 0 or $1/p_k$. Finally, notice that the propensity scores oscillate between the values zero and one.

These bounds $\underline{h}^J_k$ and $\overline{h}^J_k$ collapse to a single point for values $u \in \mathcal{T}$, and are wider when u is

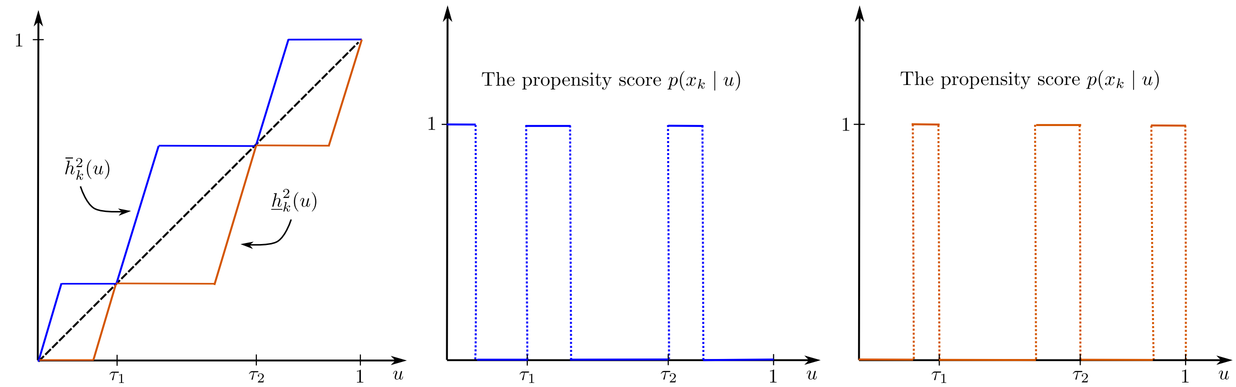


Figure 6: Left: Upper and lower bounds on $h_k(u) = \mathbb{P}(U \leq u \mid X = x_k)$ under the normalization A1, discrete support A2, and $\mathcal{T}$-independence with $\mathcal{T} = \{\tau_1, \tau_2\}$. Middle and Right: The corresponding propensity scores.

farther from the points in $\mathcal{T}$. As the set $\{\tau_1, \ldots, \tau_J\}$ becomes dense in $[0,1]$, both $\overline{h}_k^J(u)$ and $\underline{h}_k^J(u)$ converge to $F_U(u) = u$ uniformly. This occurs since full statistical independence of U and X is fully characterized by quantile independence of U from X at *all* quantiles. The exact analytical expressions we've obtained for these bounds will be helpful later in computing analytical identified sets.

4.2 Bounds under c-independence

We now characterize bounds under c-independence. In this section we assume X is binary. Define

$$\overline{h}_k^c(u) = \begin{cases} \left(1 + \dfrac{c}{p_k}\right) u & \text{if } 0 \leq u \leq 1/2 \\ \left(1 - \dfrac{c}{p_k}\right) u + \dfrac{c}{p_k} & \text{if } 1/2 < u \leq 1 \end{cases} \quad \text{and} \quad \underline{h}_k^c(u) = \begin{cases} \left(1 - \dfrac{c}{p_k}\right) u & \text{if } 0 \leq u \leq 1/2 \\ \left(1 + \dfrac{c}{p_k}\right) u - \dfrac{c}{p_k} & \text{if } 1/2 < u \leq 1. \end{cases} \tag{12}$$

for $k = 0, 1$. These bounds are valid cdfs if and only if $c \leq \min\{p_0, p_1\}$.

Theorem 3. Suppose A1 holds, X is binary, and U is c-independent of X for some $c \in [0, \min\{p_0, p_1\}]$. Let $\mathcal{F}_k^c$ be the set of all cdfs h_k on $[0,1]$ that satisfy

$$\frac{h_k(u') - h_k(u)}{u' - u} \in \left[1 - \frac{c}{p_k}, 1 + \frac{c}{p_k}\right]$$

for any $0 \leq u < u' \leq 1$. Then for $k = 0, 1$,

$$\mathcal{F}_{U|X}^c(k) = \left\{ h_k \in \mathcal{F}_k^c : \underline{h}_k^c(u) \leq h_k(u) \leq \overline{h}_k^c(u) \text{ for all } u \in [0,1] \right\}.$$

For each $k = 0, 1$, this set is sharp. Moreover, $(\underline{h}_0^c, \overline{h}_1^c), (\overline{h}_0^c, \underline{h}_1^c) \in \mathcal{F}_{U|X}^c(0,1)$.

As c gets closer to zero, these bounds collapse to the unconditional cdf $F_U(u) = u$. When c exceeds $\max\{p_0, p_1\}$, the c-independence constraint is not binding and consequently, the cdf bounds are the no-assumption bounds of equation (9). When c is between p_0 and p_1, the above cdf bounds are not proper cdfs on some segments. We derive the appropriate cdf bounds for this case in proposition 7 on page 84 in the appendix. Under c-independence the slope of the cdf—its density—is restricted by the constraint imposed by the set $\mathcal{F}_k^c$. The law of total probability constraints are imposed by the set $\mathcal{F}_k$. Theorem 3 shows that, as with $\mathcal{T}$-independence, the set of cdfs allowed under c-independence is bounded by two piecewise linear functions with simple, analytical expressions. Figure 7 plots an example.

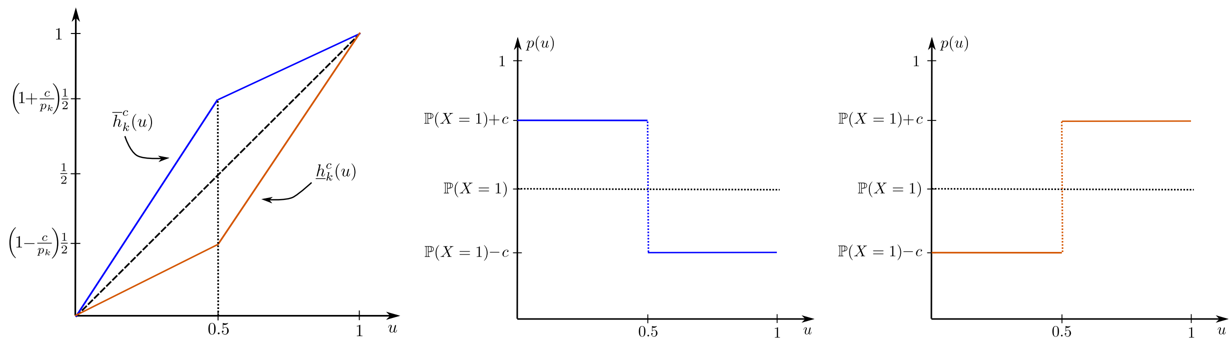


Figure 7: Left: Upper and lower bounds on $h_k(u) = \mathbb{P}(U \leq u \mid X = x_k)$ under the normalization A1, binary X, and c-independence. Middle and Right: The corresponding propensity scores for $x_k = 1$.

4.3 Bounds under d-independence

We now characterize bounds under d-independence. We again consider only the binary X case. We furthermore consider only the case $d(0) = d/p_0$ and $d(1) = d/p_1$ for $d \in [0, 1]$. This is the form implied by c-independence in proposition 2, and we give additional motivation for this choice on page 81 in the appendix. A final reason to prefer this choice of $d(x)$ is that it yields lower and upper bound cdfs that are jointly attainable. Other choices of $d(x)$ generally do not have such properties.

For $d \leq p_0 p_1$, define

$$\overline{h}_k^d(u) = \begin{cases} \dfrac{u}{p_k} & \text{if } 0 \leq u \leq \dfrac{d}{1-p_k} \\ u + \dfrac{d}{p_k} & \text{if } \dfrac{d}{1-p_k} \leq u \leq 1 - \dfrac{d}{p_k} \\ 1 & \text{if } 1 - \dfrac{d}{p_k} \leq u \leq 1 \end{cases} \quad \text{and} \quad \underline{h}_k^d(u) = \begin{cases} 0 & \text{if } 0 \leq u \leq \dfrac{d}{p_k} \\ u - \dfrac{d}{p_k} & \text{if } \dfrac{d}{p_k} \leq u \leq 1 - \dfrac{d}{1-p_k} \\ \dfrac{u-1}{p_k} + 1 & \text{if } 1 - \dfrac{d}{1-p_k} \leq u \leq 1. \end{cases} \tag{13}$$

When $d = p_0 p_1$ these bounds are precisely the no-assumption bounds of equation (9). When $d > p_0 p_1$, the d-independence constraint is not binding and hence only the no-assumptions bounds hold.

Theorem 4. Suppose A1 holds, X is binary, and U is d-independent of X with $d(0) = d/p_0$ and

$d(1) = d/p_1$ with $d \in [0, p_0 p_1]$. Let $\mathcal{F}_k$ be defined as in theorem 2. Then for $k = 0, 1$,

$$\mathcal{F}^d_{U|X}(k) = \left\{ h_k \in \mathcal{F}_k : \underline{h}^d_k(u) \leq h_k(u) \leq \overline{h}^d_k(u) \text{ for all } u \in [0,1] \right\}.$$

For each $k = 0, 1$, this set is sharp. Moreover, $(\underline{h}^d_0, \overline{h}^d_1), (\overline{h}^d_1, \underline{h}^d_0) \in \mathcal{F}^d_{U|X}(0,1)$.

Again these bounds are simple piecewise linear functions. Figure 8 plots an example. The corresponding propensity scores for $x_k = 1$ are shown as well. These propensity scores vary from 0 to 1 regardless of the value of d. As $d \to 0$, the propensity scores converge pointwise to $\mathbb{P}(X = 1)$, but not uniformly.

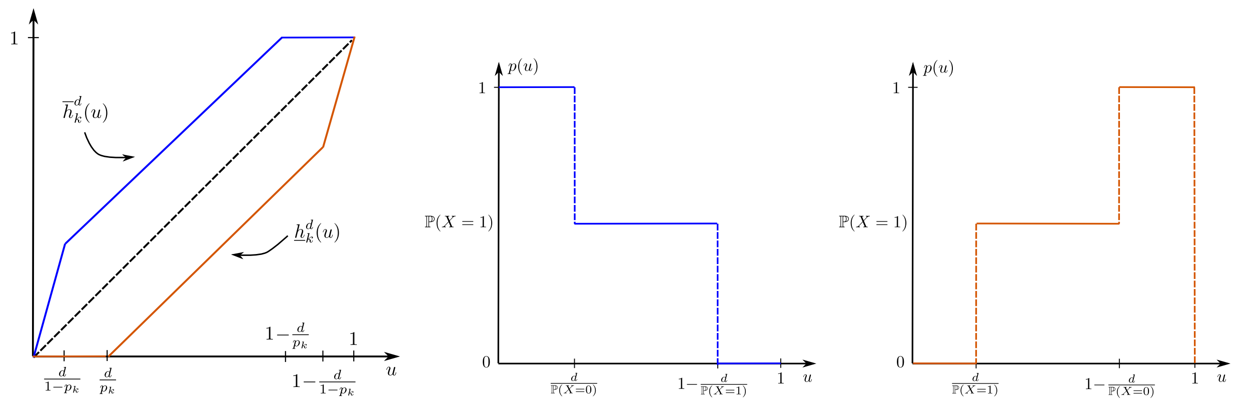


Figure 8: Left: Upper and lower bounds on $h_k(u) = \mathbb{P}(U \leq u \mid X = x_k)$ under the normalization A1, binary X, and d-independence. Middle and Right: The corresponding propensity scores, for $x_k = 1$.

5 A nonseparable regression model

In the previous section we characterized sets of cdfs $F_{U|X}$ consistent with the normalization A1 and the partial independence assumptions. As our third main contribution, in this and the following two sections we use these results to study identification of nonseparable models. In this section we study the nonseparable regression model

$$Y = m(X, U) \tag{14}$$

where Y and X are scalar observable random variables, U is a scalar unobservable random variable, and m is an unknown function. This outcome equation generalizes both classical linear regression models

$$Y = X\beta + U$$

and nonparametric additively separable models

$$Y = g(X) + U.$$

The key benefit of allowing for nonseparability of X and U is that the marginal effect of X on Y can vary across observationally equivalent individuals. Both the classical linear model and the nonparametric additively separable model have homogeneous effects of X on Y.

Matzkin (2003) studied identification of this model in detail. In particular, she imposed a monotonicity assumption similar to the following.

Assumption A4 (Monotonicity). $m(x,\cdot)$ is strictly increasing and continuous for all $x \in \text{supp}(X)$.

She then showed that a normalization is necessary for point identification of m; see Torgovitsky (2015a) pages 1186–1187 for additional discussion. Assumption A1 that $U \sim \text{Unif}[0,1]$ is one such convenient normalization. Notice that there are really two normalizations in A1: that U has a known distribution, and that this known distribution is uniform. The specific choice of a uniform is convenient, but not necessary. See Matzkin (2003) for other possible normalizations. Since U is a scalar, this model imposes rank invariance; see Heckman, Smith, and Clements (1997) as well as Torgovitsky (2015a) page 1187 for additional discussion. Finally, Matzkin assumed full statistical independence between X and U.

Proposition 3. Consider the outcome equation (14) where the joint distribution of (Y,X) is observed. Suppose A1, A4, and $X \perp\!\!\!\perp U$ hold. Then the structural function $m(x,u)$ is point identified at all points $(x,u) \in \text{supp}(X) \times [0,1]$.

We include the proof here because it will be informative for our later results.

Proof of proposition 3. Let $(x,\tau) \in \text{supp}(X) \times [0,1]$. Consider the τth conditional quantile of Y given X,

$$\begin{aligned} Q_{Y|X}(\tau \mid x) &= Q_{m(X,U)|X}(\tau \mid x) && \text{by the outcome equation (14)} \\ &= Q_{m(x,U)|X}(\tau \mid x) && \\ &= m[x, Q_{U|X}(\tau \mid x)] && \text{by A4 and quantile equivariance} \\ &= m[x, Q_U(\tau)] && \text{by } X \perp\!\!\!\perp U \\ &= m(x,\tau) && \text{by A1.} \end{aligned}$$

The left hand side is known from the data and hence the right hand side is point identified. □

Given the normalization A1 and point identification of the structural function m, the entire structure of the model is point identified. Hence all functionals of interest are also point identified. For example, the quantile structural function

$$\text{QSF}(x,\tau) = Q_{m(x,U)}(\tau) = m(x,\tau),$$

the average structural function

$$\text{ASF}(x) = \mathbb{E}[m(x,U)] = \int_0^1 m(x,u)\,du,$$

the average treatment effect for a shift from x to x',

$$\text{ATE}(x \to x') = \text{ASF}(x') - \text{ASF}(x),$$

and the distribution of unit level treatment effects for a shift from x to x',

$$\text{DTE}(t; x \to x') = \mathbb{P}[m(x', U) - m(x, U) \leq t].$$

In this section, we study the same model, except we relax full independence $X \perp\!\!\!\perp U$. Instead, we impose the weaker partial independence conditions defined in section 2.

Identification under partial independence

In this subsection we analyze identification of the structural function itself under partial independence. To do this, we first generalize Matzkin's result to the case where the distribution of $U \mid X$ falls in a known set of distributions $\mathcal{F}_{U|X}$. Her result under full independence is the special case where $\mathcal{F}_{U|X}$ is the singleton $\{F_U(u)\}$. Our later results obtain by setting $\mathcal{F}_{U|X}$ to be the various sets characterized in section 4.

The key intuition for our generalization comes from partway through the proof of proposition 3. There we used A4 to show that

$$Q_{Y|X}(\tau \mid x) = m[x, Q_{U|X}(\tau \mid x)].$$

Assuming that $U \mid X = x$ is continuously distributed on $[0, 1]$, we have

$$\tau = Q_{U|X}(\tau \mid x) \qquad \Leftrightarrow \qquad F_{U|X}(\tau \mid x) = \tau$$

and therefore

$$m(x, \tau) = Q_{Y|X}[F_{U|X}(\tau \mid x) \mid x]. \tag{15}$$

Under full statistical independence, $F_{U|X}(\tau \mid x) = F_U(\tau) = \tau$ and hence we again obtain the result of proposition 3. Without statistical independence, however, $F_{U|X}$ is an unknown function. As we showed in section 4, partial independence restrictions impose bounds on this unknown cdf. These bounds in turn yield bounds on the structural function m.

In the above argument we assumed $U \mid X = x$ was continuously distributed on $[0, 1]$. This does not necessarily hold for the cdfs bounds we obtained in section 4. In particular, we will allow the support of $U \mid X = x$ to be different for different values of x. This requires adjusting the argument above to allow for non-invertibility of $F_{U|X}$. For this we use left and right quantiles. Let $[y_L, y_U]$ denote known bounds on Y, possibly $y_L = -\infty$ and $y_U = \infty$. For any $\tau \in [0, 1]$ and $x \in \text{supp}(X)$, define the left quantile as the left-continuous inverse of the cdf:

$$Q^-_{Y|X}(\tau \mid x) = \inf\{y \in [y_L, y_U] : F_{Y|X}(y \mid x) \geq \tau\}.$$

The left quantile is the typical definition of 'the' quantile used in statistics and econometrics. Notice $Q^-_{Y|X}(1 \mid x) = \sup \operatorname{supp}(Y \mid X = x)$ and $Q^-_{Y|X}(0 \mid x) = y_L$. Similarly, define the right quantile as the right-continuous inverse of the cdf:

$$\begin{aligned} Q^+_{Y|X}(\tau \mid x) &= \sup\{y \in [y_L, y_U] : F_{Y|X}(y \mid x) \leq \tau\} \\ &= \inf\{y \in [y_L, y_U] : F_{Y|X}(y \mid x) > \tau\}. \end{aligned}$$

Notice $Q^+_{Y|X}(0 \mid x) = \inf \operatorname{supp}(Y \mid X = x)$ and $Q^+_{Y|X}(1 \mid x) = y_U$. Hosseini (2010) and Embrechts and Hofert (2013) study properties of left and right quantiles. We collect some helpful results in lemma 5 on page 75 in the appendix.

Finally, we define the *set*-quantile to be the closed interval with bounds equal to the left and right quantiles:

$$Q^*_{Y|X}(\tau \mid x) = \left[Q^-_{Y|X}(\tau \mid x), Q^+_{Y|X}(\tau \mid x)\right].$$

When the distribution of $Y \mid X$ is strictly increasing, the left and right quantiles are equal to each other and the set-quantile is a single point. When the support of $U \mid X$ is non-connected or generally not equal to $[0, 1]$, the conditional cdf is non-increasing on subsets of $U \mid X$ and therefore the left and right quantiles will differ at some values τ.

We now characterize the identified set for the structural function $m(x, u)$. This result generalizes Matzkin's (2003) result under full independence.

Theorem 5. Consider the outcome equation (14). Suppose A1 and A4 hold. Let $x \in \operatorname{supp}(X)$. Suppose the true cdf $F_{U|X}(\cdot \mid x)$ lies in a known set $\mathcal{F}_{U|X}(x)$. Let $\mathcal{H}$ be the set of continuous and strictly increasing functions on $\operatorname{supp}(U)$. Then the identified set for the function $m(x, \cdot)$ is

$$\mathcal{M}(x) = \left\{m(x, \cdot) \in \mathcal{H} : m(x, \cdot) \in Q^*_{Y|X}[F_{U|X}(\cdot \mid x) \mid x] \text{ for some } F_{U|X}(\cdot \mid x) \in \mathcal{F}_{U|X}(x)\right\}. \quad (16)$$

This result shows that our characterization of the set of cdfs allows us to characterize the identified set for the structural function. Notice that even if the conditional distributions $F_{U|X}(\cdot \mid x)$ are known, the structural function might not be point identified for all values $(x, \tau) \in \operatorname{supp}(X) \times [0, 1]$. This follows since the left and right quantiles might be different from one another.

For example, suppose the conditional distributions $U \mid X = 0 \sim \text{Unif}[0, 1/2]$ and $U \mid X = 1 \sim \text{Unif}[1/2, 1]$ are known, and also that $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$. Suppose $-\infty < y_L < y_U < \infty$. In this example, the structural function is not point identified at $m(0, \tau)$ for $\tau \in (1/2, 1)$ because the conditional distribution of $U \mid X = 0$ does not take values above 1/2. By theorem 5,

$$Q^-_{Y|X}(F_{U|X}(\tau \mid 0) \mid 0) \leq m(0, \tau) \leq Q^+_{Y|X}(F_{U|X}(\tau \mid 0) \mid 0)$$

for all $\tau \in [0,1]$. Let $\tau \in (1/2, 1)$. Then $F_{U|X}(\tau \mid 0) = 1$ by $U \mid X = 0 \sim \text{Unif}[0, 1/2]$. Hence

$$\begin{aligned} Q^{+}_{Y|X}(F_{U|X}(\tau \mid 0) \mid 0) &= Q^{+}_{Y|X}(1 \mid 0) \\ &= y_U. \end{aligned}$$

For the lower bound, by quantile equivariance,

$$\begin{aligned} Q^{-}_{Y|X}(1 \mid 0) &= m(0, Q^{-}_{U|X}(1 \mid 0)) \\ &= m(0, 1/2). \end{aligned}$$

The third line follows since $U \mid X = 0 \sim \text{Unif}[0, 1/2]$. Hence for any $\tau \in (1/2, 1]$,

$$\begin{aligned} m(0, \tau) &\in [Q^{-}_{Y|X}(1 \mid 0), y_U] \\ &= [m(0, 1/2), y_U]. \end{aligned}$$

This example is precisely what occurs in our first Roy model example; see the left plot of figure 4 on page 18. Here we have perfect sorting of people with $U \leq 1/2$ to $X = 0$ and people with $U \geq 1/2$ to $X = 1$. Since the data contain no observations for whom $u > 1/2$ and $X = 0$, it is not surprising that we cannot point identify the value of the function $m(x, u)$ at these points. Instead, the identified set for the function is bounded above by the a priori logical bound on outcomes and bounded below by $m(0, 1/2)$. The lower bound holds by monotonicity A4 and the fact that $m(0, 1/2)$ is point identified as the largest outcome among people with $X = 0$.

Theorem 5 is an abstract result that applies for any choice of $\mathcal{F}_{U|X}(x)$. By combining monotonicity of $Q^{-}_{Y|X}(\cdot \mid x)$ and $Q^{+}_{Y|X}(\cdot \mid x)$ with our results in section 4 we obtain a more concrete characterization of the set $\mathcal{M}(x)$, along with identified sets for other functionals of interest. For example, suppose $\mathcal{T}$-independence A3 holds. Then substituting our largest and smallest cdfs $\underline{h}^{J}_{k}(\cdot)$ and $\overline{h}^{J}_{k}(\cdot)$ from theorem 2 into the set quantile bounds of equation (16) and using monotonicity of the left and right quantiles we obtain the bounds

$$m(x_k, \tau) \in \left[Q^{-}_{Y|X}(\underline{h}^{J}_{k}(\tau) \mid x_k),\ Q^{+}_{Y|X}(\overline{h}^{J}_{k}(\tau) \mid x_k)\right] \tag{17}$$

for any $x_k \in \text{supp}(X)$ and $\tau \in [0,1]$. Importantly for our results below, these bounds are uniform in τ, since the cdf bounds depend on τ only via their argument. For brevity, throughout the rest of this section we primarily present formal results for $\mathcal{T}$-independence. But equation (17) and the other results in this section can be straightforwardly modified to c- and d-independence by replacing the relevant cdf bounds by those derived in section 4.

Chesher (2003) showed that $m(x_k, \tau)$ is point identified for all τ such that quantile independence holds at τ. Our results in this section extend his result in three ways. First, his result can be used to construct bounds as follows. Suppose $\mathcal{T}$-independence holds and $\tau_j, \tau_l \in \mathcal{T}$, $\tau_j < \tau_l$, but $\tau \notin \mathcal{T}$

for all $\tau \in (\tau_j, \tau_l)$. Then $m(x_k, \tau_j)$ and $m(x_k, \tau_l)$ are point identified and by monotonicity A4,

$$m(x_k, \tau_j) \leq m(x_k, \tau) \leq m(x_k, \tau_l)$$

for all $\tau \in (\tau_j, \tau_l)$. Bounds like this can be pieced together to obtain a bound on the entire function $m(x_k, \cdot)$. When X is discrete, these bounds are not sharp. This follows because the upper and lower bounds correspond to the cdfs of $U \mid X$ given in figure 13 on page 83 in the appendix, which are not admissible when X is discrete, since they violate the law of total probability. Our theorem 5 provides the sharp identified set for the entire structural function. Second, while Chesher (2003) focused on the structural function at a point, we also derive bounds for various functionals of the entire structural function. Finally, unlike Chesher, we also consider c- and d-independence. Under c- and d-independence, the simple bounds approach above yields trivial bounds because *no* quantile independence conditions hold exactly and hence there are no points at which the structural function is point identified.

Bounds on the average structural function

Because the bounds in equation (17) are uniform over τ, integrating them from 0 to 1 yields bounds on the average structural function. By performing this integration and substituting in the specific functional forms for the cdf bounds obtained in section 4, we obtain simple, analytical bounds for the average structural function. We begin with the case where $\mathcal{T}$-independence holds with $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$.

Corollary 4. Consider the outcome equation (14). Suppose A1, A2, and A4 hold. Suppose U is $\mathcal{T}$-independent of X with $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$ and let $\tau_0 = 0$ and $\tau_{J+1} = 1$. Let $x_k \in \text{supp}(X)$. Then

$$\text{ASF}(x_k) \in \Bigg[(1 - p_k) \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x_k) + p_k \mathbb{E}[Y \mid X = x_k],$$
$$(1 - p_k) \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^+_{Y|X}(\tau_{j+1} \mid x_k) + p_k \mathbb{E}[Y \mid X = x_k]\Bigg].$$

Moreover, the interior of this set is sharp.

Since the cdf bounds $\overline{h}_k^J$ and $\underline{h}_k^J$ are flat in some regions, the endpoints of the bounds in equation (17) are not strictly increasing in τ everywhere, which means they violate A4. Hence the endpoints of the ASF bounds are not be attainable. A similar remark holds for the analogous d-independence result. For the analogous c-independence result, however, the endpoints might be attainable, depending on the value of c.

In corollary 11 on page 85 in the appendix we provide a corresponding result for continuous X. There we see that as $p_k \to 0$ in the discrete X bounds we obtain the bounds for the continuous X

case. Next, under full statistical independence $X \perp\!\!\!\perp U$, it can be shown that

$$\text{ASF}(x_k) = \mathbb{E}[Y \mid X = x_k]$$

and hence the ASF is point identified under $X \perp\!\!\!\perp U$. Corollary 4 generalizes this result in the following sense. Consider the lower bound. The term

$$\sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x_k)$$

is a lower Riemann sum over the partition of $[0,1]$ defined by $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$. The corresponding term in the upper bound is the upper Riemann sum over this same partition. Assuming $\mathbb{E}[Y|X = x_k]$ exists, the functions $Q^-_{Y|X}(\cdot \mid x_k)$ and $Q^+_{Y|X}(\cdot \mid x_k)$ will be integrable and both these sums converge to

$$\int_0^1 Q^-_{Y|X}(u \mid x) \, du = \mathbb{E}[Y \mid X = x]$$

as the partition becomes dense in $[0,1]$ by lemma 5 part 3.[2] But as this partition becomes dense in $[0,1]$, our assumptions get 'closer' to the full statistical independence assumption $X \perp\!\!\!\perp U$. Hence our result is consistent with, and generalizes this well-known result that the conditional mean function point identifies the average structural function in a nonseparable model under full statistical independence. Also note that, since full independence is always allowed as a possibility by our three partial independence concepts, any sample analog estimator based on full independence will converge to an element of the sets we derive.

The bounds in corollary 4 depend on $Q^-_{Y|X}(0 \mid x) = y_L$ and $Q^+_{Y|X}(1 \mid x) = y_U$. If the support of $Y \mid X$ is the entire real line then these quantiles will be $-\infty$ and $+\infty$, respectively. Hence the bounds will be non-informative. This point applies to the bounds we derive for the ATE in a later subsection as well, but not for the bounds we later derive on the DTE. Hence we generally need bounded support outcomes to obtain non-trivial identified sets for the ASF and ATE. This is a common finding in the literature on partial identification of mean parameters. For example, see Manski (2007) pages 43–44.

Next we consider identification of the ASF under the general $\mathcal{T}$-partial independence assumption A3.

Corollary 5. Consider the outcome equation (14). Suppose A1, A2, A3, and A4 hold. Let $x_k \in \text{supp}(X)$. Then

$$\text{ASF}(x_k) \in \big[\text{LB}[\text{ASF}(x_k)], \text{UB}[\text{ASF}(x_k)]\big]$$

[2]This integral is the same for left and right quantiles since left and right quantiles only differ on a countable set; see lemma 5 part 2 on page 75 in the appendix.

where

$$\text{LB}[\text{ASF}(x_k)] = (1-p_k)\sum_{j=0}^{J}(a_{j+1}-b_j)Q^{-}_{Y|X}(b_j \mid x_k) + (1-p_k)\sum_{j=1}^{J}\int_{a_j}^{b_j} Q^{-}_{Y|X}(u \mid x_k)\,du + p_k\mathbb{E}[Y \mid X = x_k]$$

and

$$\text{UB}[\text{ASF}(x_k)] = (1-p_k)\sum_{j=0}^{J}(a_{j+1}-b_j)Q^{+}_{Y|X}(a_{j+1} \mid x_k) + (1-p_k)\sum_{j=1}^{J}\int_{a_j}^{b_j} Q^{+}_{Y|X}(u \mid x_k)\,du + p_k\mathbb{E}[Y \mid X = x_k].$$

Moreover, the interior of this set is sharp.

These bounds simplify to those given in corollary 4 if the intervals $[a_j, b_j]$ are singletons since the second term will drop out. In both the upper and lower bounds, the first term is again a Riemann sum which converges to zero as $\mathcal{T}$ gets 'close' to $[0,1]$, while the second term will converge to $(1-p_k)\mathbb{E}[Y|X = x_k]$. This shows that these bounds again collapse to the point $\mathbb{E}[Y \mid X = x_k]$ in the limiting case of full statistical independence.

ASF bounds under c- and d-independence

We now briefly describe the bounds on the average structural function under c- and d-independence. For c-independence, $\text{ASF}(x_k) \in [\text{LB}, \text{UB}]$ where

$$\text{LB} = \frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \leq Q^{-}_{Y|X}\left(\frac{1}{2}\left(1-\frac{c}{p_k}\right)\middle| x_k\right)\right] + \frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \geq Q^{-}_{Y|X}\left(\frac{1}{2}\left(1-\frac{c}{p_k}\right)\middle| x_k\right)\right]$$

and

$$\text{UB} = \frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \leq Q^{+}_{Y|X}\left(\frac{1}{2}\left(1+\frac{c}{p_k}\right)\middle| x_k\right)\right] + \frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \geq Q^{+}_{Y|X}\left(\frac{1}{2}\left(1+\frac{c}{p_k}\right)\middle| x_k\right)\right].$$

As $c \to 0$, the bounds collapse to the single point $\mathbb{E}[Y \mid X = x_k]$, as under $\mathcal{T}$-independence.

For d-independence with the choice $d(x_k) = d/p_k$, the ASF bounds are

$$\text{ASF}(x_k) \in \Bigg[(p_k - 1)\int_{1-\frac{d}{p_k(1-p_k)}}^{1} Q^-_{Y|X}(u \mid x_k)\, du + \frac{d}{p_k} Q^-_{Y|X}(0 \mid x_k) + \mathbb{E}[Y \mid X = x_k],$$
$$(p_k - 1)\int_0^{\frac{d}{p_k(1-p_k)}} Q^+_{Y|X}(u \mid x_k)\, du + \frac{d}{p_k} Q^+_{Y|X}(1 \mid x_k) + \mathbb{E}[Y \mid X = x_k]\Bigg].$$

Again these bounds converge to $\mathbb{E}[Y \mid X = x_k]$ as $d \to 0$. The interiors of these bounds under c- and d-independence are sharp.

Bounds on ATE, the distribution of treatment effects, and spread parameters

Next we derive bounds on the average treatment effect, for a shift from x_l to x_k, defined by

$$\begin{aligned} \text{ATE}(x_l \to x_k) &= \text{ASF}(x_k) - \text{ASF}(x_l) \\ &= \mathbb{E}[m(x_k, U) - m(x_l, U)]. \end{aligned}$$

Corollary 6. Consider the outcome equation (14). Suppose A1, A2, A3, and A4 hold. Then

$$\text{ATE}(x_l \to x_k) \in \Big[\text{LB}[\text{ASF}(x_k)] - \text{UB}[\text{ASF}(x_l)],\ \text{UB}[\text{ASF}(x_k)] - \text{LB}[\text{ASF}(x_l)]\Big].$$

Moreover, the interior of this set is sharp.

Sharpness of the interior holds here because our results in section 4 showed that we can simultaneously achieve the upper bound cdf $\overline{h}_k^J$ and the smallest cdf for $\underline{h}_l^J$. When the set $\mathcal{T}$ approaches $(0, 1)$, the bounds for $\text{ATE}(x_l \to x_k)$ collapses to the difference in conditional means,

$$\mathbb{E}(Y \mid X = x_k) - \mathbb{E}(Y \mid X = x_l).$$

An analogous result to corollary 6 holds for c- and d-independence.

Besides the ASF and the ATE, there are many other parameters one might be interested in. In particular, as discussed earlier, a key reason to consider nonseparable models like (14) is that they allow for heterogeneous treatment effects. Hence a natural object of interest is the distribution (or cdf) of the unit level treatment effects

$$\Delta(x_l \to x_k) = m(x_k, U) - m(x_l, U).$$

We show that we can bound the unobserved random variable $\Delta(x_l \to x_k)$ using two point identified random variables as in the following lemma.

Lemma 3. Consider the outcome equation (14). Suppose A1, A2, A3, and A4 hold. Let $x_k \in$

supp(X). Then the following bounds hold U-almost surely:

$$Q^-_{Y|X}(\underline{h}^J_k(U) \mid x_k) \leq m(x_k, U) \leq Q^+_{Y|X}(\overline{h}^J_k(U) \mid x_k)$$

and

$$Q^-_{Y|X}(\underline{h}^J_k(U) \mid x_k) - Q^+_{Y|X}(\overline{h}^J_l(U) \mid x_l) \leq \Delta(x_l \to x_k) \leq Q^+_{Y|X}(\overline{h}^J_k(U) \mid x_k) - Q^-_{Y|X}(\underline{h}^J_l(U) \mid x_l).$$

In corollaries 4, 5, and 6 we implicitly used these bounds to derive bounds on the means of $m(x_k, U)$ and $\Delta(x_l \to x_k)$. The mean is sometimes called a D_1-parameter (Stoye 2010 page 327; Manski 2003 calls them D-parameters), because it respects the stochastic dominance order relation. As discussed in Manski (2003), the cdf evaluated at a point and quantiles are also examples of D_1-parameters. Hence the bounds in lemma 3 can immediately be applied to derive bounds on quantiles of $m(x_k, U)$ and $\Delta(x_l \to x_k)$. For example, for the median of the treatment effect $\Delta(x_l \to x_k)$ we obtain

$$\begin{aligned}
&\text{Med}[Q^-_{Y|X}(\underline{h}^J_k(U) \mid x_k) - Q^+_{Y|X}(\overline{h}^J_l(U) \mid x_l)] \\
&\quad \leq \text{Med}[\Delta(x_l \to x_k)] \\
&\qquad \leq \text{Med}[Q^+_{Y|X}(\overline{h}^J_k(U) \mid x_k) - Q^-_{Y|X}(\underline{h}^J_l(U) \mid x_l)].
\end{aligned}$$

Furthermore, we can also immediately derive bounds on the cdfs $\mathbb{P}[m(x_k, U) \leq t]$ and $\mathbb{P}[m(x_k, U) - m(x_l, U) \leq t]$. This latter cdf is called the distribution of unit level treatment effects.

Corollary 7. Consider the outcome equation (14). Suppose A1, A2, A3, and A4 hold. Let $x_k \in \text{supp}(X)$ and $t \in \mathbb{R}$. Then

$$\mathbb{P}\big(Q^-_{Y|X}(\underline{h}^J_k(U) \mid x_k) \leq t\big) \leq \mathbb{P}\big(m(x_k, U) \leq t\big) \leq \mathbb{P}\big(Q^+_{Y|X}(\overline{h}^J_k(U) \mid x_k) \leq t\big)$$

and

$$\begin{aligned}
&\mathbb{P}\big(Q^-_{Y|X}(\underline{h}^J_k(U) \mid x_k) - Q^+_{Y|X}(\overline{h}^J_l(U) \mid x_l) \leq t\big) \\
&\quad \leq \mathbb{P}\big(\Delta(x_l \to x_k) \leq t\big) \\
&\qquad \leq \mathbb{P}\big(Q^+_{Y|X}(\overline{h}^J_k(U) \mid x_k) - Q^-_{Y|X}(\underline{h}^J_l(U) \mid x_l) \leq t\big).
\end{aligned}$$

Moreover, the interiors of these sets are sharp.

Spread parameters, like the variance and other measures of dispersion, are not D_1 parameters. Nonetheless, Stoye (2010) shows how to use bounds like those in corollary 7 to derive bounds on such spread parameters, which he calls D_2-parameters. This class of parameters includes not only the variance but also inequality measures like the Gini coefficient. Using the results of corollary 7 combined with corollary 5 we can directly apply theorem 2 of Stoye (2010) and derive explicit bounds on the variance of treatment effects and other D_2 parameters.

Numerical illustration

We conclude this section with a brief numerical illustration. Suppose

$$Y = \pi X + (\gamma X + 1)\Phi^{-1}_{[-4,4]}(U)$$

where $U \sim \text{Unif}[0,1]$, $U \perp\!\!\!\perp X$, $\pi = 0.4$, $\gamma = 0.4$, X is binary with $\mathbb{P}(X = 1) = 1/2$, and $\Phi_{[-4,4]}$ is the cdf for the truncated standard normal on $[-4, 4]$. This dgp has heterogeneous treatment effects: units with U larger than about 0.16 have positive treatment effects, with the magnitude of the effect increasing in U, while units with U smaller than about 0.16 have negative treatment effects. The true ASF is

$$\text{ASF}(x) = \mathbb{E}(Y \mid X = x) = \pi x$$

and the true conditional quantile function is

$$Q_{Y|X}(\tau \mid x) = \pi x + (\gamma x + 1)\Phi^{-1}_{[-4,4]}(\tau).$$

Figure 9 shows identified sets for ATE under various partial independence assumptions. In all plots we see that the ATE under full independence is positive, and that this conclusion is robust to minor deviations from full independence, but not to larger deviations. For $\mathcal{T} = (0, \tau]$ the identified set is fully in the positive region for approximately all $\tau \geq 0.80$. With a roughly uniform grid of quantile independence conditions we only need about 15 grid points for the identified set to be fully in the positive region. For c-independence, deviations from the unconditional propensity score of up to $c = 0.1$ are allowed while still point identifying the sign of ATE. Also note that $\overline{c} = \max\{p_0, p_1\} = 0.5$ here. For d-independence, only very minor deviations of up to about $d = 0.02$ are allowed while still point identifying the sign of ATE. Keep in mind also that c-independence can hold for c arbitrarily close to zero even while no τ-cdf independence constraints hold, and vice versa.

Finally, note that the general shape of all these identified sets depends on the specific dgp. Some dgps may yield results where conclusions under full independence are robust to partial independence deviations, while other dgps may not. For example, in the dgp we consider the identified sets shift upward as π increases. Hence, holding all else fixed, a larger ATE implies that the sign of ATE will be point identified under weaker independence assumptions. Here we merely illustrate the kinds of objects empirical researchers can compute using the results we develop in this paper.

6 A nonseparable instrumental variables model

In section 5 we studied a regression model where X is 'exogenous' in the sense that X is partially independent of U. In this section, we consider a nonseparable instrumental variables model where X is endogenous and U is partially independent of an instrument Z. We also impose a relevance condition on the joint distribution of X and Z. Throughout this section we focus on the case where

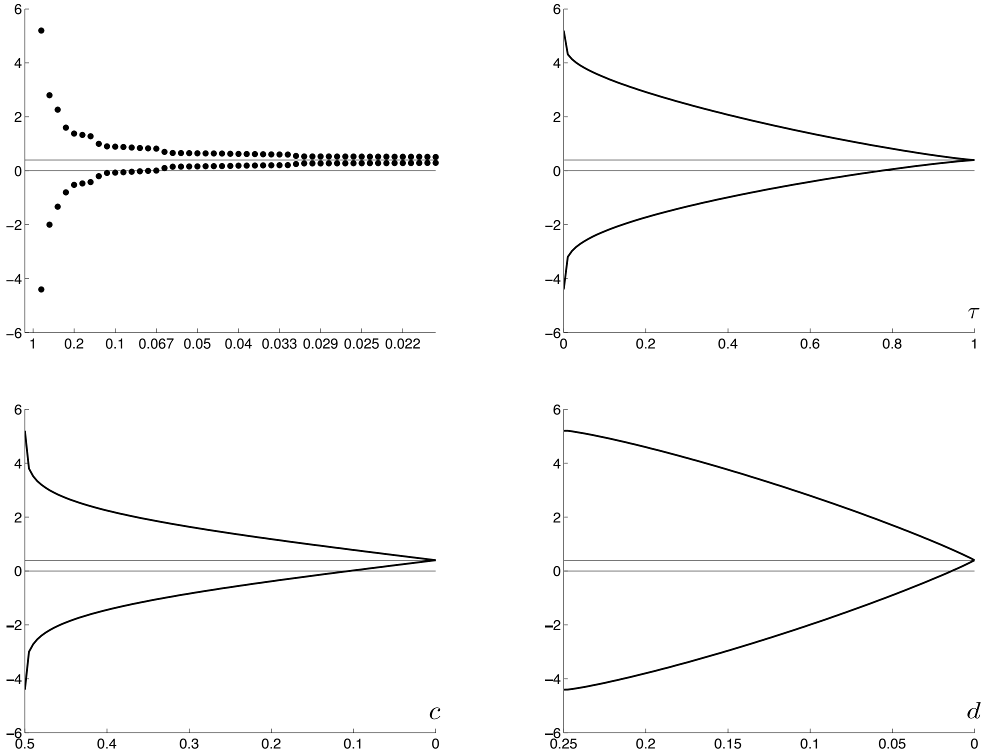


Figure 9: Identified sets for ATE. Top left: $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$ for a sequence of Halton draws. Horizontal axis measures space between grid points. Top right: $\mathcal{T} = (0, \tau]$. Bottom left: c-independence. Bottom right: d-independence. In all plots the true ATE is plotted as a solid horizontal line.

X and Z are binary. Under any of the three partial independence assumptions, we derive identified sets for the structural function. We also derive sharp random variable bounds on the unit level treatment effects which can be used to compute sharp bounds on functionals like ATE and the DTE.

We continue to consider the nonseparable outcome equation

$$Y = m(X, U) \tag{14}$$

where U is a scalar unobservable random variable and m is an unknown function. We also observe an instrument Z. We build on the analysis of Chernozhukov and Hansen (2005), who suppose the instrument is fully independent: $Z \perp\!\!\!\perp U$. Their analysis yields the following result.[3]

Proposition 4 (Chernozhukov and Hansen (2005) Theorem 2)**.** Consider the outcome equation (14). Suppose X and Z are binary and that A1, A4, and $Z \perp\!\!\!\perp U$ hold. Let $\mathcal{L}$ be the parameter

[3] As in section 5, we only consider the model under rank invariance. Chernozhukov and Hansen (2005) also consider the weaker rank similarity assumption.

space as in Chernozhukov and Hansen (2005) equation (2.12). Suppose the rank of the matrix

$$\Pi'(m(0,\tau), m(1,\tau)) = \begin{pmatrix} f_{Y|X,Z}(m(0,\tau) \mid 0,0)\mathbb{P}(X=0 \mid Z=0) & f_{Y|X,Z}(m(1,\tau) \mid 1,0)\mathbb{P}(X=1 \mid Z=0) \\ f_{Y|X,Z}(m(0,\tau) \mid 0,1)\mathbb{P}(X=0 \mid Z=1) & f_{Y|X,Z}(m(1,\tau) \mid 1,1)\mathbb{P}(X=1 \mid Z=1) \end{pmatrix} \tag{18}$$

is full for any $\tau \in [0,1]$ and that $(m(0,\tau), m(1,\tau)) \in \mathcal{L}$. Further assume that $\Pi'(m(0,\tau), m(1,\tau))$ is continuous in $(m(0,\tau), m(1,\tau))$. Then $m(0,\tau)$ and $m(1,\tau)$ are point identified for any $\tau \in [0,1]$.

As with the nonseparable regression model, it is helpful to examine a simplified version of the proof here. We have

$$\begin{aligned} \mathbb{P}[Y \leq m(X,\tau) \mid Z=z] &= \mathbb{P}[m(X,U) \leq m(X,\tau) \mid Z=z] \\ &= \mathbb{P}[U \leq \tau \mid Z=z] \\ &= \mathbb{P}[U \leq \tau] \\ &= \tau. \end{aligned}$$

The first line follows by the outcome equation (14). The second by A4. The third by full independence. The fourth by A1. Since X is binary,

$$m(X,\tau) = m(1,\tau)X + m(0,\tau)(1-X).$$

So we get the moment condition

$$\mathbb{P}[Y \leq m(1,\tau)X + m(0,\tau)(1-X) \mid Z=z] = \tau$$

for all $z \in \text{supp}(Z)$. Fix τ and let $m_x = m(x,\tau)$. Since Z is binary, we have two moment conditions:

$$\begin{aligned} \mathbb{P}[Y \leq m_1 X + m_0(1-X) \mid Z=0] - \tau = 0 \\ \mathbb{P}[Y \leq m_1 X + m_0(1-X) \mid Z=1] - \tau = 0 \end{aligned}$$

and two unknown parameters, m_1 and m_0. This is a system of two nonlinear (in m_0, m_1) equations. The rank condition and the proof of theorem 2 in Chernozhukov and Hansen (2005) shows that these equations yield a unique solution.

We next replace full statistical independence of U and Z with the assumption that U is either $\mathcal{T}$-, c-, or d-independent of Z. Most of the proof above continues to hold. The key change is that our two moment conditions become

$$\begin{aligned} \mathbb{P}[Y \leq m_1 X + m_0(1-X) \mid Z=0] - F_{U|Z}(\tau \mid 0) = 0 \\ \mathbb{P}[Y \leq m_1 X + m_0(1-X) \mid Z=1] - F_{U|Z}(\tau \mid 1) = 0 \end{aligned}$$

for some $F_{U|Z} \in \mathcal{F}_{U|Z}$.

If the cdf $F_{U|Z}$ were known, this would again be a system of two equations with two unknowns. By our results in section 4, partial independence yields an identified set $\mathcal{F}_{U|Z}(z)$ for the cdf $F_{U|Z}(\cdot \mid z)$. We can thus combine this result with the above system to derive the identified set for (m_0, m_1), as follows. First, remember that τ is fixed in this analysis. Next, define

$$\begin{aligned}
\Pi_1(m_0, m_1) &= \mathbb{P}[Y \leq m_1 X + m_0(1 - X) \mid Z = 0] \\
\Pi_2(m_0, m_1) &= \mathbb{P}[Y \leq m_1 X + m_0(1 - X) \mid Z = 1].
\end{aligned}$$

So our moment conditions are

$$\begin{aligned}
\Pi_1(m_0, m_1) &= F_{U|Z}(\tau \mid 0) \\
\Pi_2(m_0, m_1) &= F_{U|Z}(\tau \mid 1)
\end{aligned}$$

or

$$\begin{aligned}
\Pi_1(m_0, m_1) &= \frac{\tau - \mathbb{P}(X = 1) h_1(\tau)}{1 - \mathbb{P}(X = 1)} \\
\Pi_2(m_0, m_1) &= h_1(\tau)
\end{aligned}$$

where

$$h_1(\tau) = F_{U|Z}(\tau \mid 1).$$

For a fixed h_1, let $\theta_0(h_1(\tau)), \theta_1(h_1(\tau))$ denote the unique solution to this system. The rank assumption guarantees uniqueness. The union of solutions over the set $h_1 \in \mathcal{F}_{U|Z}(1)$ is the identified set, which we characterize in the following result.

Theorem 6. Consider the outcome equation (14). Suppose X and Z are binary. Suppose A1 and A4 hold. Assume $(m(0, \tau), m(1, \tau)) \in \mathcal{L}$ and that $\Pi'(m(0, \tau), m(1, \tau))$ has full rank over $\mathcal{L}$. Suppose U is either $\mathcal{T}$-, c-, or d-independent of Z, with corresponding cdf bounds $\overline{h}_1$ and $\underline{h}_1$ from section 4. Then the following bounds are sharp:

$$\begin{aligned}
m(0, \tau) &\in [\theta_0(\underline{h}_1(\tau)), \theta_0(\overline{h}_1(\tau))] \\
m(1, \tau) &\in [\theta_1(\overline{h}_1(\tau)), \theta_1(\underline{h}_1(\tau))]
\end{aligned}$$

when $\theta_0(\underline{h}_1(\tau)) \leq \theta_0(\overline{h}_1(\tau))$. Both bounds are reversed if $\theta_0(\underline{h}_1(\tau)) > \theta_0(\overline{h}_1(\tau))$.

In the proof we show that $\theta_0(h_1(\tau))$ is monotonic in $h_1(\tau)$. Similarly, $\theta_1(h_1(\tau))$ is monotonic in h_1 in the opposite direction of θ_0. When full independence $Z \perp\!\!\!\perp U$ holds, the point identification result proposition 4 is not fully constructive. This follows because the parameters $m(0, \tau)$ and $m(1, \tau)$ are obtained implicitly as the solution to a system of equations. The same property holds in our analysis under partial independence, theorem 6. Our result, however, only requires computing the solution to this system twice, compared with once in the point identified case. Hence, in practice computing the identified set under partial independence is not any easier or harder than computing the point estimate under full independence.

Bounds on unit level treatment effects

As in section 5 there are many parameters we might be interested in. In particular, we are often interested in treatment effects. Given the bounds on the structural function, we can find bounds on all its functionals, such as the ASF, ATE, or DTE, by integrating over the bounds of the structural function. For example, recall the definition of the unit level treatment effect: $\Delta(0 \to 1) = m(1,U) - m(0,U)$. Then since the bounds in theorem 6 hold uniformly over τ, the following bounds hold U-almost surely:

$$\theta_1(\overline{h}_1(U)) - \theta_0(\overline{h}_1(U)) \leq \Delta(0 \to 1) \leq \theta_1(\underline{h}_1(U)) - \theta_0(\underline{h}_1(U))$$

when the lower bound here is smaller than the upper bound. Otherwise, the bounds are reversed. Here $\overline{h}_1$ and $\underline{h}_1$ are the relevant partial independence cdf bounds. These random variable bounds can be combined with general purpose results on bounds on D_1 and D_2 parameters, exactly as in section 5.

7 A binary outcome model

A third and final example we discuss is the binary outcome model

$$Y = \mathbb{1}[m(X,U) \leq 0] \tag{19}$$

where $U \sim \text{Unif}[0,1]$ and $m(x,\cdot)$ is strictly increasing. Then

$$\begin{aligned} Y &= \mathbb{1}[U \leq m^{-1}(X,0)] \\ &= \mathbb{1}[U \leq g(X)] \end{aligned}$$

where $g(x) \equiv m^{-1}(x,0)$. The average structural function is

$$\text{ASF}(x) = \mathbb{P}[U \leq g(x)] = F_U[g(x)] = g(x).$$

This model was studied by Matzkin (1992). Under the full statistical independence assumption $X \perp\!\!\!\perp U$, g is point identified.

Proposition 5. Consider the binary outcome model (19). Suppose A1 and $X \perp\!\!\!\perp U$ hold. Then g is point identified.

As in the two previous models we've studied, it is helpful to examine the proof.

Proof of proposition 5.

$$\begin{aligned}
\mathbb{E}(Y \mid X = x) &= \mathbb{P}(Y = 1 \mid X = x) \\
&= \mathbb{P}(U \leq g(x) \mid X = x) \\
&= F_{U|X}[g(x) \mid x] \\
&= F_U[g(x)] && \text{by } X \perp\!\!\!\perp U \\
&= g(x) && \text{by A1.}
\end{aligned}$$

□

Next, we relax the full independence assumption and replace it by a partial independence assumption. It is still the case that

$$\mathbb{E}(Y \mid X = x) = F_{U|X}[g(x) \mid x].$$

Hence

$$g(x) \in Q^*_{U|X}\big(\mathbb{E}(Y \mid X = x) \mid x\big).$$

Since $F_{U|X}(\cdot \mid x) \in \mathcal{F}_{U|X}(x)$, we can characterize the set of conditional quantiles of U given X, and plug in the identified conditional expectation $\mathbb{E}[Y \mid X = x]$ to obtain the bounds on the ASF $g(x)$. This leads to the following result.

Theorem 7. Consider the binary outcome model (19). Suppose A1 and A2 hold. Suppose U is either $\mathcal{T}$-, c-, or d-independent of X, with corresponding cdf bounds $\overline{h}_k$ and $\underline{h}_k$ from section 4. Then the identified set for $g(x_k) = \text{ASF}(x_k)$ is

$$\Big[\overline{Q}^-_{U|X}(\mathbb{E}(Y \mid X = x_k) \mid x_k), \quad \underline{Q}^+_{U|X}(\mathbb{E}(Y \mid X = x_k) \mid x_k)\Big]$$

where $\overline{Q}^-_{U|X}(v \mid x_k) = \inf\{y \in [0,1] : \overline{h}_k(y) \geq v\}$ and $\underline{Q}^+_{U|X}(v \mid x_k) = \sup\{y \in [0,1] : \underline{h}_k(y) \leq v\}$, the left and right quantiles of the cdf bounds.

This result allows us to derive simple, explicit bounds for the average structural function. For example, consider the case where $\mathcal{T}$-independence holds on a grid of points $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$. Then the quantile bounds equal

$$\overline{Q}^-_{U|X}(v \mid x_k) = \begin{cases} p_k v & \text{if } 0 \leq v \leq \tau_1 \\ p_k(v - \tau_1) + \tau_1 & \text{if } \tau_1 < v \leq \tau_2 \\ \vdots & \\ p_k(v - \tau_J) + \tau_J & \text{if } \tau_J < v \leq 1 \end{cases}$$

and

$$\underline{Q}^{+}_{U|X}(v \mid x_k) = \begin{cases} (v-\tau_1)p_k + \tau_1 & \text{if } 0 \leq v < \tau_1 \\ (v-\tau_2)p_k + \tau_2 & \text{if } \tau_1 \leq v < \tau_2 \\ \vdots & \\ (v-1)p_k + 1 & \text{if } \tau_J \leq v \leq 1. \end{cases}$$

Substituting $v = \mathbb{E}[Y \mid X = x_k]$ yields the ASF bounds. As we can see, these functions are piecewise linear functions and therefore the ASF bounds are easy to compute under partial independence. The bounds under c- and d-independence have similar simple expressions. Finally, notice that in the 'no assumptions' case $\mathcal{T} = \emptyset$ these bounds simplify to the classic Manski (1990) bounds:

$$[p_k \mathbb{E}(Y \mid X = x_k), \quad p_k \mathbb{E}(Y \mid X = x_k) + (1 - p_k)].$$

8 Conclusion

In this paper we studied three concepts of *partial independence*, as ways to weaken the full statistical independence assumption. We formalized them as either a set of quantile independence conditions ($\mathcal{T}$-independence), a distance between conditional and unconditional propensity scores (c-independence), or a distance between conditional and unconditional cdfs (d-independence). Our first contribution was to characterize the set of propensity scores consistent with $\mathcal{T}$-independence, thus providing a treatment assignment interpretation of quantile independence conditions. We showed that nonconstant propensity scores consistent with quantile independence must be nonmonotonic and oscillatory. These properties motivated our consideration of c- and d-independence as alternative concepts. Our second contribution was to characterize sets of cdfs $F_{U|X}$ consistent with each of the partial independence concepts. These results do not depend on the specific choice of econometric model. Our third contribution was to apply these results to derive simple, analytical characterizations of identified sets in three popular nonseparable models.

Although we have focused on identification results in this paper, our analytical identified sets lend themselves to sample analog estimation and inference via the existing literature on inference under partial identification (see Canay and Shaikh 2016 for a survey). We have also omitted covariates from our analysis. In principle we can condition on them nonparametrically, but in practice one will typically also impose dimension reduction assumptions. We plan to examine these issues of estimation, inference, and incorporating covariates in future work.

Finally, all of the models we studied have a single, scalar unobservable U. Many models, however, have multiple unobservable random variables. Two important examples are the classical selection model with an explicit first stage equation (e.g., Heckman and Vytlacil 2007) and simultaneous equations models (e.g., Matzkin 2008). As in the models with a scalar unobservable, these models typically make full statistical independence assumptions. While it is not clear how to best generalize $\mathcal{T}$- and d-independence to vector U, c-independence generalizes immediately. Extending our results to such models is an important next step.

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151–184.

——— (2008): "Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization," *American Economic Review P&P*, 98, 345–350.

Altonji, J. G. and R. L. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053–1102.

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74, 431–497.

Berger, J. O. (1985): *Statistical decision theory and Bayesian analysis*, Springer-Verlag, second ed.

Brown, B. (1983): "The identification problem in systems nonlinear in the variables," *Econometrica*, 51, 175–196.

Canay, I. A. and A. M. Shaikh (2016): "Practical and theoretical advances for inference in partially identified models," *Working paper*.

Chamberlain, G. and E. E. Leamer (1976): "Matrix weighted averages and posterior bounds," *Journal of the Royal Statistical Society, Series B*, 73–84.

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

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

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

——— (2007a): "Identification of nonadditive structural functions," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Cambridge University Press, vol. 3, 1–16.

——— (2007b): "Instrumental values," *Journal of Econometrics*, 139, 15–34.

——— (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78, 575–601.

Chesher, A. and A. Rosen (2015): "Generalized instrumental variable models," *Working paper*.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94, 260–272.

Cross, P. J. and C. F. Manski (2002): "Regressions, short and long," *Econometrica*, 70, 357–368.

de Paula, A. (2008): "Conditional moments and independence," *The American Statistician*, 62, 219–221.

Doukhan, P. (1994): *Mixing: Properties and examples*, Springer.

Embrechts, P. and M. Hofert (2013): "A note on generalized inverses," *Mathematical Methods of Operations Research*, 77, 423–432.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191–1206.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, 64, 487–535.

Heckman, J. J. and E. J. Vytlacil (2007): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," *Handbook of Econometrics*, 6, 4779–4874.

Hosseini, R. (2010): "Quantiles equivariance," *arXiv preprint arXiv:1004.0533*.

Hotz, V. J., C. H. Mullin, and S. G. Sanders (1997): "Bounding causal effects using data from a contaminated natural experiment: Analysing the effects of teenage childbearing," *The Review of Economic Studies*, 64, 575–603.

Ichino, A., F. Mealli, and T. Nannicini (2008): "From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?" *Journal of Applied Econometrics*, 23, 305–327.

Imbens, G. W. (2003): "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review P&P*, 126–132.

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

Imbens, G. W. and D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Judd, K. L. (1998): *Numerical methods in economics*, MIT press.

Kline, P. and A. Santos (2013): "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 4, 231–267.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 33–50.

Leamer, E. E. (1978): *Specification searches: Ad hoc inference with nonexperimental data*, vol. 53, John Wiley & Sons Incorporated.

——— (1982): "Sets of posterior means with bounded variance priors," *Econometrica*, 725–736.

——— (1983): "Model choice and specification analysis," *Handbook of Econometrics*, 1, 285–330.

——— (1985): "Sensitivity analyses would help," *American Economic Review*, 308–313.

Lin, D. Y., B. M. Psaty, and R. A. Kronmal (1998): "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, 948–963.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205–228.

——— (1983): "Closest empirical distribution estimation," *Econometrica*, 305–319.

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83, 729–738.

——— (1990): "Nonparametric bounds on treatment effects," *American Economic Review P&P*, 80, 319–323.

——— (1999): "Choice as an alternative to control in observational studies: Comment," *Statistical Science*, 14, 279–281.

——— (2003): *Partial identification of probability distributions*, Springer.

——— (2007): *Identification for prediction and decision*, Harvard University Press.

——— (2013): "Response to the review of 'Public policy in an uncertain world'," *The Economic Journal*, 123, F412–F415.

——— (2016): "Credible interval estimates for official statistics with survey nonresponse," *Journal of Econometrics*, 191, 293–301.

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010.

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12, S200–S216.

Matzkin, R. L. (1992): "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models," *Econometrica*, 239–270.

——— (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1339–1375.

——— (2008): "Identification in nonparametric simultaneous equations models," *Econometrica*, 76, 945–978.

——— (2012): "Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity," *Journal of Econometrics*, 166, 106–115.

——— (2015): "Estimation of nonparametric models with simultaneity," *Econometrica*, 83, 1–66.

Molinari, F. and M. Peski (2006): "Generalization of a result on "Regressions, short and long"," *Econometric Theory*, 22, 159–163.

Nevo, A. and A. M. Rosen (2012): "Identification with imperfect instruments," *Review of Economics and Statistics*, 94, 659–671.

Powell, J. L. (1994): "Estimation of semiparametric models," *Handbook of Econometrics*, 4, 2443–2521.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000): "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical models in epidemiology, the environment, and clinical trials*, Springer, 1–94.

Roehrig, C. S. (1988): "Conditions for identification in nonparametric and parametric models," *Econometrica*, 433–447.

Rosenbaum, P. R. (1995): *Observational Studies*, Springer.

——— (2002): *Observational Studies*, Springer, second ed.

Rosenbaum, P. R. and D. B. Rubin (1983): "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society, Series B*, 212–218.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998): "Semiparametric regression for repeated outcomes with nonignorable nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339.

Rotnitzky, A., D. O. Scharfstein, T.-L. Su, and J. M. Robins (2001): "Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring," *Biometrics*, 57, 103–113.

Rudin, W. (1976): *Principles of Mathematical Analysis*, McGraw-Hill, third ed.

Scharfstein, D. O., C. F. Manski, and J. C. Anthony (2004): "On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the good behavior game trial," *Biometrics*, 60, 154–164.

Small, D. S. and P. R. Rosenbaum (2008): "War and wages: The strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association*, 103, 924–933.

Stoye, J. (2010): "Partial identification of spread parameters," *Quantitative Economics*, 1, 323–357.

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

——— (2015b): "Partial identification by extending subdistributions," *Working paper.*

van der Vaart, A. W. (2000): *Asymptotic statistics*, Cambridge University Press.

A Additional literature discussion

In this section we provide additional details explaining how our results compare to several approaches in the literature. We begin with a general comparison before moving on to the specifics of several papers.

Sensitivity analysis: Parametric paths versus nonparametric neighborhoods

In this section we discuss how using our identification results to do sensitivity analysis compares to the previous literature. The most common approach, as in Rosenbaum and Rubin (1983), has two key features: (1) a specific parametric deviation r from a baseline assumption of $r = 0$ and (2) a parameter $\theta(r)$ that is point identified given that deviation. We call this the *parametric path* approach. The object of interest is the function $\theta(r)$, and how it changes with r.

Our approach, following Manski (2007) (also see section 3 of Manski 2013), weakens the parametric path approach in two ways. First, we consider nonparametric neighborhoods of the baseline assumption. The neighborhood is still indexed by a parameter r, but the set of allowed models is no longer a singleton. Consequently, in the second step, our parameter of interest is typically partially identified. Let $\Theta(r)$ denote the identified set. The goal is then to see how $\Theta(r)$ changes with r. We call this the *nonparametric neighborhood* approach.

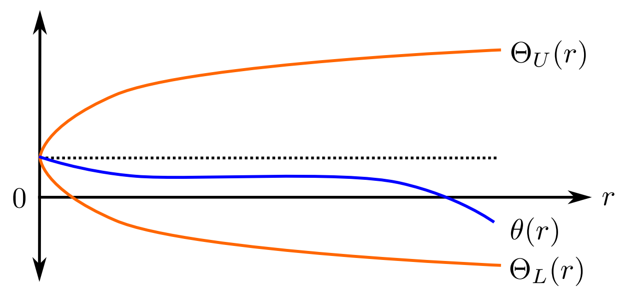


Figure 10: Comparing two approaches to sensitivity analysis: Parametric paths and nonparametric neighborhoods.

In general, the nonparametric neighborhood approach will show that one's findings are more sensitive to deviations than the parametric path approach. This follows simply because the parametric path approach picks a specific sequence of submodels from the neighborhood surrounding the baseline models. For example, suppose we are interested in a scalar parameter. Suppose $\Theta(r)$ equals an interval $[\Theta_L(r), \Theta_U(r)]$. Figure 10 compares the sequence of identified sets under nonparametric neighborhoods to a sequence of parametric submodels. Suppose we are interested in the sensitivity of our conclusion that the parameter is positive. Then a sensitivity analysis based on the parametric path $\theta(r)$ shown in the figure suggests that the results are quite robust to deviations, while a sensitivity analysis based on $\Theta(r)$ suggests the opposite. Now, there are many different parametric submodels and at least one of them may correspond to, say, the lower bound $\Theta_L(r)$. Our point is merely that the choice of parametric submodel may substantively matter. The current literature using parametric path approaches, including Rosenbaum and Rubin (1983) and others we discuss below, do not justify their choice of parametric paths, a concern pointed out by Manski (1999). Without such justification, we find a nonparametric neighborhood approach more plausible.

It is often the case that $\{\theta(r) : \text{ all } r\} = \Theta(R)$ where $\Theta(R)$ denotes the identified set when no assumptions on the dependence are made; for example, those of Manski (1990) in the analysis of treatment response. See Imbens and Rubin (2015) section 22.4. As we emphasize above, however, in the parametric path approach it is the function $\theta(r)$ which is reported, and how it varies for r close to 0. Hence the shape of the function, rather than just its range, is substantively important. Our nonparametric neighborhood approach helps guard against the shape of this function $\theta(r)$ being driven solely by parametric assumptions.

In addition to considering nonparametric neighborhoods rather than parametric paths, we consider three different kinds of neighborhoods, corresponding to our $\mathcal{T}$-, c-, and d-independence. We show that the particular choice of neighborhood has substantive implications on what parameters are and are not tightly identified. For example, in various nonseparable models Chesher (2003) showed that a specific quantile structural function can be point identified under a single quantile independence condition. Hence he showed that even with large deviations from independence, in the $\mathcal{T}$-independence sense, some parameters may still be point identified. But if we change the notion of deviation to c-independence those same parameters become partially identified even for arbitrarily small deviations from independence. Hence there are two dimensions along which one can do sensitivity analysis: the choice of metric for measuring deviations and our choice of the size of the neighborhood. Approaches that look at a single parametric path are restrictive in both steps, which may lead empirical researchers to have unjustified confidence in their results.

Rosenbaum and Rubin (1983)

Here we briefly review the results of Rosenbaum and Rubin (1983). We follow the exposition given in section 22.4 of Imbens and Rubin (2015). Consider a binary outcome and a binary treatment. Let $Y(0)$ and $Y(1)$ denote the potential outcomes. Let X denote received treatment. Under $(Y(0), Y(1)) \perp\!\!\!\perp X$ we can point identify ATE. But we are concerned that this independence assumption may not hold because of an omitted variable U. Suppose we believe that $(Y(0), Y(1)) \perp\!\!\!\perp X \mid U$. If we observed U then we would once again be able to point identify ATE. Unfortunately, we do not observe U. Instead, Rosenbaum and Rubin specify a parametric logit model for the propensity score:

$$\begin{aligned} p(u) &= \mathbb{P}(X = 1 \mid U = u) \\ &= \Lambda(\gamma_0 + \gamma_1 u) \end{aligned}$$

where $\Lambda(s) = \exp(s)/[1+\exp(s)]$ is the logit link function. $\gamma_1 = 0$ corresponds to full independence while any $\gamma_1 \neq 0$ corresponds to deviations from full independence. Our c-independence is a nonparametric generalization of this parametric model—c-independence for a small c implies that γ_1 is also small. Rosenbaum and Rubin do not study $\mathcal{T}$- or d-independence. Moreover, in their analysis they assume U is binary while we allow for continuously distributed U's.

Letting $q = \mathbb{P}(U = 1)$, they next note that, for fixed (q, γ_1), γ_0 is point identified by the law of total probability constraint

$$\mathbb{P}(X = 1) = \Lambda(\gamma_0 + \gamma_1)q + \Lambda(\gamma_0)(1 - q).$$

Hence, to perform their sensitivity analysis, they will vary q over $[0, 1]$ and γ_1 over $\mathbb{R}$.

In the second step of their analysis, they specify a parametric logit model for the distributions of $Y(0) \mid U$ and $Y(1) \mid U$:

$$\begin{aligned} \mathbb{P}[Y(1) = 1 \mid U = u] &= \Lambda(\alpha_0 + \alpha_1 u) \\ \mathbb{P}[Y(0) = 1 \mid U = u] &= \Lambda(\beta_0 + \beta_1 u). \end{aligned}$$

Similarly as above, they show that if we fixed (α_1, β_1) along with (q, γ_1) then (α_0, β_0) will be point identified. From this we can compute ATE, which is point identified. Thus their full approach is to compute the function $\text{ATE}(q, \gamma_1, \alpha_1, \beta_1)$ and see how this varies over these sensitivity parameters. If we specified that $(q, \gamma_1, \alpha_1, \beta_1)$ lay inside some known set, then the above analysis yields bounds on ATE. As Imbens and Rubin (2015) note (also see page 158 of Rosenbaum 2002), these bounds equal Manski's (1990) no assumptions bounds if we allow $(q, \gamma_1, \alpha_1, \beta_1) \in [0,1] \times \overline{\mathbb{R}}^3$.

Although we use nonparametric neighborhoods rather than a parametric path for the first step, our second step analysis (sections 5, 6, and 7) is similar in spirit to Rosenbaum and Rubin's second step. There are several differences, however. We do not require a parametric model for the outcome distribution. We study both binary and continuous outcomes. We do not restrict attention to the average treatment effect. We do, however, make the rank invariance assumption which Rosenbaum and Rubin do not need. Finally, our analysis is not restricted to a modified selection on observables model—we have studied models with and without instruments.

Robins et al. (2000)

Robins et al. (2000) assume X is discrete with

$$\mathbb{P}(X = x \mid Y(x) = y) = 1 - T[h(x) + q(x,y)]$$

for all $x \in \text{supp}(X)$, where T is a known cdf, such as $T(s) = \Lambda(s)$, the logit link function. Equivalently,

$$\mathbb{P}(X \neq x \mid Y(x) = y) = T[h(x) + q(x,y)].$$

In their theorem 1 they show that for a fixed a priori q, h is point identified and the joint distribution of $(X, Y(x))$ is point identified for all $x \in \text{supp}(X)$. $q \equiv 0$ implies that treatment is marginally independent of each potential outcome, and hence this corresponds to a full independence assumption. $q \neq 0$ represents a deviation from full independence. As with Rosenbaum and Rubin (1983), they propose a parametric path approach by examining how the point identified parameter of interest varies as q varies. Although q can be nonparametric in principle, they focus on parametric models for q. Hence in practice they study specific parametric deviations from independence. Note, however, that unlike Rosenbaum and Rubin (1983), they do not require a parametric model for outcomes.

Also unlike Rosenbaum and Rubin (1983), they do not explicitly focus on the omitted variable explanation for failure of independence. Instead they directly model the relationship between potential outcomes and treatment assignment. In this sense, our analysis is closer to Robins et al. (2000) than to Rosenbaum and Rubin (1983). The reason is that, under rank invariance and strict monotonicity of the outcome equation m in the unobservable U, the propensity score we study is

$$\begin{aligned} p(u) &= \mathbb{P}(X = x \mid U = u) \\ &= \mathbb{P}[X = x \mid m(x,U) = m(x,u)] \\ &= \mathbb{P}[X = x \mid Y(x) = m(x,u)]. \end{aligned}$$

c-independence places constraints directly on this propensity score. This is conceptually similar to the Robins et al. (2000) approach of modeling this propensity score. Nonetheless, as emphasized in our introduction, our analysis differs in several important ways. We consider nonparametric neighborhoods of full independence, rather than deviations along a parametric path. Consequently, our deviations do not point identify the parameters of interest, and therefore we compute the corresponding sharp identified sets. This latter analysis is aided by the structure of the nonseparable models in the literature which we build on.

Also, as Manski (2016) emphasizes at the beginning of his section 4.2, simply specifying a parametric selection model does not guarantee that the parameter of interest is point identified. As our lemma 5 and the surrounding discussion shows, knowledge of the distribution of $U \mid X$ is not necessarily sufficient for point identification of the distribution of potential outcomes. Although the selection model considered by Robins et al. (2000) delivers point identification, they provide no motivation or discussion to justify their choice of parametric model.

Rosenbaum (1995, 2002)

Rosenbaum (1995, 2002) essentially considered the assumption

$$\sup_{u \in \text{supp}(U)} \left| \log \frac{\mathbb{P}(X=1)}{\mathbb{P}(X=0)} - \log \frac{p(u)}{1-p(u)} \right| \leq G \tag{20}$$

for some constant G. This approach no longer requires parametric assumptions on the propensity score. However, Rosenbaum (1995, 2002) did not derive identified sets under this assumption. Nor did he perform any alternative sensitivity similar to that of Rosenbaum and Rubin (1983). Instead, he only studied the effect of the above form of deviations from independence on the finite sample properties of inference procedures for testing the sharp null of zero treatment effect $H_0 : Y_i(0) = Y_i(1)$ for $i = 1, \ldots, N$ in a randomization inference framework. Imbens and Rubin (2015) chapter 5 provide further discussion of this framework; section 22.5 also discusses the Rosenbaum (1995, 2002) approach. Rosenbaum rewrote the assumption (20) as

$$\frac{1}{\Gamma} \leq \frac{p(U_i)[1-p(U_j)]}{p(U_j)[1-p(U_i)]} \leq \Gamma.$$

for the constant $\Gamma = \exp(2G) \geq 1$ and all $i, j = 1, \ldots, N$. Since the function p is unrestricted we can let $p_i = p(U_i)$ without loss of generality. For any specified values of $p_1, \ldots, p_N$, the randomization distribution of any statistic can be computed, under the maintained sharp null hypothesis. Allowing the values of $p(U_i)$ to vary within the constraints above produces a set of randomization distributions. From this set we can compute a set of p-values, for example. We can then see how this set of p-values changes with Γ. This is a very different approach to both Rosenbaum and Rubin (1983) and to what we do in the present paper. An interesting question for future work would be to extend our identification analysis to cover the log odds ratio distance (20), in addition to the three concepts of $\mathcal{T}$-, c-, and d-independence we study in the paper.

B Proofs

Proofs for section 3

Lemma 4. Suppose A1 and A2 hold. Then $h_k(\cdot)$ is a continuous function for all $k = 1, \ldots, K$.

Proof of Lemma 4. Suppose by way of contradiction that $h_k(\cdot)$ is not continuous at some point u^*. Since cdfs are right-continuous, we must have

$$\lim_{u \nearrow u^*} h(u) < h(u^*).$$

This implies

$$\mathbb{P}(U = u^* \mid X = x_k) > 0.$$

Therefore, by the law of total probability,

$$\begin{aligned}
0 &= \mathbb{P}(U = u^*) \\
&= \sum_{j=1}^{K} \mathbb{P}(U = u^* \mid X = x_j) p_j \\
&\geq \mathbb{P}(U = u^* \mid X = x_k) p_k \\
&> 0.
\end{aligned}$$

This is a contradiction. □

Proof of proposition 1. Suppose $\mathcal{T}$-independence holds. Fix $u \in [0,1]$ and $x \in \text{supp}(X)$. Then

$$\begin{aligned}
\mathbb{P}(U \leq u \mid X = x, U \in \mathcal{T}) &= \frac{\mathbb{P}(0 \leq U \leq u, U \in \mathcal{T} \mid X = x)}{\mathbb{P}(U \in \mathcal{T} \mid X = x)} \\
&= \frac{\int_{s \in [0,u] \cap \mathcal{T}} \; dF_{U|X}(s \mid x)}{\int_{s \in \mathcal{T}} \; dF_{U|X}(s \mid x)} \\
&= \frac{\int_{s \in [0,u] \cap \mathcal{T}} \; dF_{U}(s)}{\int_{s \in \mathcal{T}} \; dF_{U}(s)} \\
&= \frac{\mathbb{P}(0 \leq U \leq u, U \in \mathcal{T})}{\mathbb{P}(U \in \mathcal{T})} \\
&= \mathbb{P}(U \leq u \mid U \in \mathcal{T}).
\end{aligned}$$

The first line follows by the definition of conditional probability, and since $U \sim \text{Unif}[0,1]$ so that $U \geq 0$. The third line follows by $\mathcal{T}$-independence. The last line follows again by the definition of conditional probability. Since the last line does not depend on x, we have shown that $X \perp\!\!\!\perp U \mid \{U \in \mathcal{T}\}$. □

Proof of theorem 1. This result actually holds for discretely distributed X, so we work with that case in this proof.

(⇒) Suppose U is $\mathcal{T}$-independent of X. Let $t_1, t_2 \in \mathcal{T} \cup \{0,1\}$ with $t_1 < t_2$. Then, for any $x \in \text{supp}(X)$,

$$\begin{aligned}
\mathbb{P}(X = x \mid U \in [t_1, t_2]) &= \frac{\mathbb{P}(X = x, U \in [t_1, t_2])\mathbb{P}(X = x)}{\mathbb{P}(U \in [t_1, t_2])\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}(U \in [t_1, t_2] \mid X = x)\mathbb{P}(X = x)}{t_2 - t_1} \\
&= \frac{(\mathbb{P}(U \leq t_2 \mid X = x) - \mathbb{P}(U < t_1 \mid X = x))\mathbb{P}(X = x)}{t_2 - t_1} \\
&= \frac{(\mathbb{P}(U \leq t_2 \mid X = x) - \mathbb{P}(U \leq t_1 \mid X = x))\mathbb{P}(X = x)}{t_2 - t_1} \\
&= \frac{(t_2 - t_1)\mathbb{P}(X = x)}{t_2 - t_1} \\
&= \mathbb{P}(X = x).
\end{aligned}$$

The second line follows since $U \sim \text{Unif}[0,1]$. The fourth line follows since $U \mid X$ is continuously distributed, since X is discretely distributed (by lemma 4). The fifth line follows from $\mathcal{T}$-independence.

($\Leftarrow$) Suppose that for any $x \in \text{supp}(X)$,

$$\mathbb{P}(X = x \mid U \in [t_1, t_2]) = \mathbb{P}(X = x)$$

for all $t_1, t_2 \in \mathcal{T} \cup \{0, 1\}$ with $t_1 < t_2$. Then,

$$\begin{aligned}
\mathbb{P}(U \in [t_1, t_2] \mid X = x) &= \frac{\mathbb{P}(X = x \mid U \in [t_1, t_2])\mathbb{P}(U \in [t_1, t_2])}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}(X = x)\mathbb{P}(U \in [t_1, t_2])}{\mathbb{P}(X = x)} \\
&= \mathbb{P}(U \in [t_1, t_2]).
\end{aligned}$$

The second line follows by assumption. Setting $t_1 = 0$ and using $U \sim \text{Unif}[0, 1]$ gives the result.

The result now follows by lemma 1. □

Proof of lemma 1. We have

$$\begin{aligned}
\mathbb{P}(X = 1 \mid U \in [t_1, t_2]) &= \frac{\mathbb{P}(X = 1, t_1 \leq U \leq t_2)}{\mathbb{P}(t_1 \leq U \leq t_2)} \\
&= \frac{1}{t_2 - t_1}\mathbb{P}(X = 1, t_1 \leq U \leq t_2) \\
&= \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} f_{X,U}(1, u)\, du \\
&= \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{f_{X,U}(1, u)}{f_U(u)}\, du \\
&= \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \mathbb{P}(X = 1 \mid U = u)\, du.
\end{aligned}$$

The fourth line follows since $f_U(u) = 1$ by $U \sim \text{Unif}[0, 1]$. □

Proof of corollary 1. Without loss of generality, suppose p is weakly increasing. Then for any τ, the average value of the propensity score to the left of τ is weakly smaller than the average value to the right:

$$\frac{1}{\tau}\int_0^\tau p(u)\, du \leq \frac{1}{1 - \tau}\int_\tau^1 p(u)\, du.$$

Since p is also not constant, this inequality must actually be strict for all $\tau \in (0, 1)$. The result follows by theorem 1. □

Proof of corollary 2. For each interval $\mathcal{U}_k$, we just repeat the argument of corollary 1, conditional on $U \in \mathcal{U}_k$, noting that a nontrivial τ-cdf independence conditional will still hold conditional on $U \in \mathcal{U}_k$. □

Proof of corollary 3. Let $[a,b] \subseteq (0,1) \setminus \mathcal{T}$. Also, let the propensity score be equal to

$$p(u) = \begin{cases} 1 & \text{if } u \in [a, a + \mathbb{P}(X=1)(b-a)) \\ 0 & \text{if } u \in [a + \mathbb{P}(X=1)(b-a), b] \\ \mathbb{P}(X=1) & \text{if } u \notin [a,b]. \end{cases}$$

Let t_1 and t_2 be any two values in $\mathcal{T}$ such that $t_1 < t_2$. Then

$$\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} p(u)\, du = \mathbb{P}(X=1)$$

if $t_1 < t_2 < a$ or $b < t_1 < t_2$. This condition also holds if $t_1 < a$ and $b < t_2$ since

$$\begin{aligned} \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} p(u)\, du &= \frac{1}{t_2 - t_1}\Big(\mathbb{P}(X=1)(a - t_1) + 1\big(\mathbb{P}(X=1)(b-a) + a - a\big) + \mathbb{P}(X=1)(t_2 - b)\Big) \\ &= \mathbb{P}(X=1). \end{aligned}$$

Therefore, by theorem 1, $\mathcal{T}$-independence holds and the propensity score attains the values 0 and 1 over non-degenerate intervals. □

Proof of proposition 2. We have

$$\begin{aligned} \sup_{u \in [0,1]} |\mathbb{P}(U \leq u \mid X=1) - \mathbb{P}(U \leq u)| &= \sup_{u \in [0,1]} \left| \int_0^u f_{U|X}(s \mid 1)\, ds - \int_0^u 1\, ds \right| \\ &= \sup_{u \in [0,1]} \left| \int_0^u \frac{\mathbb{P}(X=1 \mid U=s)}{\mathbb{P}(X=1)}\, ds - \int_0^u \frac{\mathbb{P}(X=1)}{\mathbb{P}(X=1)}\, ds \right| \\ &= \frac{1}{\mathbb{P}(X=1)} \sup_{u \in [0,1]} \left| \int_0^u (\mathbb{P}(X=1 \mid U=s) - \mathbb{P}(X=1))\, ds \right| \\ &\leq \frac{1}{\mathbb{P}(X=1)} \sup_{u \in [0,1]} \int_0^u |\mathbb{P}(X=1 \mid U=s) - \mathbb{P}(X=1)|\, ds \\ &\leq \frac{1}{\mathbb{P}(X=1)} \sup_{u \in [0,1]} \int_0^u c\, ds \\ &= \frac{c}{\mathbb{P}(X=1)}. \end{aligned}$$

The fifth line follows by c-independence. Next observe that

$$\mathbb{P}(U \leq u \mid X=0) = \frac{\mathbb{P}(U \leq u) - \mathbb{P}(U \leq u \mid X=1)\mathbb{P}(X=1)}{\mathbb{P}(X=0)}.$$

Hence

$$\begin{aligned}\sup_{u\in[0,1]} |\mathbb{P}(U \leq u \mid X=0) - \mathbb{P}(U \leq u)| &= \sup_{u\in[0,1]} \left| \frac{\mathbb{P}(U \leq u) - \mathbb{P}(U \leq u \mid X=1)\mathbb{P}(X=1)}{\mathbb{P}(X=0)} - \mathbb{P}(U \leq u) \right| \\ &= \frac{\mathbb{P}(X=1)}{\mathbb{P}(X=0)} \sup_{u\in[0,1]} |\mathbb{P}(U \leq u) - \mathbb{P}(U \leq u \mid X=1)| \\ &\leq \frac{\mathbb{P}(X=1)}{\mathbb{P}(X=0)} \frac{c}{\mathbb{P}(X=1)} \\ &= \frac{c}{\mathbb{P}(X=0)}.\end{aligned}$$

The first line follows by our derivations above, the third by our result for $X = 1$ above. □

Proofs for section 4

Proof of theorem 2. To show $\overline{h}_k^J(\cdot)$ is the uniform upper bound, let

$$h_k(u) = \mathbb{P}(U \leq u \mid X = x_k)$$

and suppose there exists a value $u^* \in [0,1]$ such that $h_k(u^*) > \overline{h}_k^J(u^*)$. Since $\overline{h}_k^J$ is a step function, we divide the proof into three cases based on the value of u^*:

$$\begin{array}{lll}\text{Case 1:} & u^* \in [b_j, (1-p_k)b_j + p_k a_{j+1}] & \text{for } j = 0,\ldots,J \\ \text{Case 2:} & u^* \in [(1-p_k)b_j + p_k a_{j+1}, a_{j+1}] & \text{for } j = 0,\ldots,J \\ \text{Case 3:} & u^* \in [a_j, b_j] & \text{for } j = 1,\ldots,J.\end{array}$$

Case 1. Fix $j = 0,\ldots,J$. Then

$$\begin{aligned}h_k(u^*) &> \overline{h}_k^J(u^*) \\ &= \frac{u^* - b_j}{p_k} + b_j.\end{aligned}$$

Therefore, by the law of total probability,

$$\begin{aligned}\mathbb{P}(U \leq u^* \mid X \neq x_k) &= \frac{\mathbb{P}(U \leq u^*) - h_k(u^*)p_k}{1-p_k} \\ &< \frac{u^* - (u^* - b_j) - b_j p_k}{1-p_k} \\ &= b_j.\end{aligned}$$

By A3,

$$\mathbb{P}(U \leq b_j \mid X \neq x_k) = b_j.$$

Since $u^* \geq b_j$,

$$\begin{aligned}b_j &= \mathbb{P}(U \leq b_j \mid X \neq x_k) \\ &\leq \mathbb{P}(U \leq u^* \mid X \neq x_k) \\ &< b_j,\end{aligned}$$

a contradiction.

Case 2. Fix $j = 0, \ldots, J$. We have

$$\begin{aligned} h_k(u^*) &> \overline{h}_k^J(u^*) \\ &= a_{j+1}. \end{aligned}$$

But $u^* \leq a_{j+1}$, implying that $h_k(u^*) \leq h_k(a_{j+1})$. Thus

$$\begin{aligned} a_{j+1} &< h_k(u^*) \\ &\leq h_k(a_{j+1}) \\ &= a_{j+1}, \end{aligned}$$

a contradiction.

Case 3. If $u^* \in [a_j, b_j]$ for $j = 1, \ldots, J$, then $h_k(u^*) = u^*$ by A3 and therefore $h_k(u^*) > u^*$ presents a contradiction.

To show $\underline{h}_k^J(\cdot)$ is the uniform lower bound, we suppose there exists a $u^* \in [0, 1]$ such that $h_k(u^*) < \underline{h}_k^J(u^*)$. We again consider three cases for the value u^*:

Case 1:	$u^* \in [b_j, p_k b_j + (1 - p_k)a_{j+1}]$	for $j = 0, \ldots, J$
Case 2:	$u^* \in [p_k b_j + (1 - p_k)a_{j+1}, a_{j+1}]$	for $j = 0, \ldots, J$
Case 3:	$u^* \in [a_j, b_j]$	for $j = 1, \ldots, J$.

Case 1. Fix $j = 0, \ldots, J$. We have

$$h_k(u^*) < \underline{h}_k^J(u^*) = b_j.$$

Since $u^* \geq b_j$,

$$\begin{aligned} b_j &> h_k(u^*) \\ &\geq h_k(b_j) \\ &= b_j, \end{aligned}$$

a contradiction.

Case 2. Fix $j = 0, \ldots, J$. we have

$$\begin{aligned} h_k(u^*) &< \underline{h}_k^J(u^*) \\ &= \frac{u^* - a_{j+1}}{p_k} + a_{j+1}. \end{aligned}$$

Therefore, by the law of total probability,

$$\begin{aligned}\mathbb{P}(U \le u^* \mid X \ne x_k) &= \frac{\mathbb{P}(U \le u^*) - h_k(u^*)p_k}{1 - p_k} \\ &> \frac{u^* - (u^* - a_{j+1}) - a_{j+1}p_k}{1 - p_k} \\ &= a_{j+1}.\end{aligned}$$

By assumption A3,

$$\mathbb{P}(U \le a_{j+1} \mid X \ne x_k) = a_{j+1}.$$

Since $u^* \le a_{j+1}$,

$$\begin{aligned}a_{j+1} &= \mathbb{P}(U \le a_{j+1} \mid X \ne x_k) \\ &\ge \mathbb{P}(U \le u^* \mid X \ne x_k) \\ &> a_{j+1},\end{aligned}$$

a contradiction.

Case 3. Similar to the previous case 3.

To show that the cdf h_k must be in the set $\mathcal{F}_k$, suppose there exists $0 \le u < u' \le 1$ for which

$$\frac{h_k(u') - h_k(u)}{u' - u} > \frac{1}{p_k}.$$

Then

$$\begin{aligned}\mathbb{P}(U \in (u, u'] \mid X \ne x_k) &= \mathbb{P}(U \le u' \mid X \ne x_k) - \mathbb{P}(U \le u \mid X \ne x_k) \\ &= \frac{u' - u - p_k(h_k(u') - h_k(u))}{1 - p_k} \\ &< 0,\end{aligned}$$

a contradiction. The second line follows by applying the law of total probability twice.

To show that the set $\mathcal{F}_{U|X}(x_k)$ is sharp, fix k and let h_k be an arbitrary element of $\mathcal{F}_{U|X}(x_k)$. Define

$$h_l(u) = \frac{u - p_k h_k(u)}{1 - p_k}$$

for all $l \ne k$. The function $h_l(u)$ is a proper cdf on $[0, 1]$ since (1) $h_l(0) = 0$ and $h_l(1) = 1$, (2) it is right continuous since h_k is right continuous, and (3) it is non-decreasing since for $u' > u$

$$\begin{aligned}h_l(u') - h_l(u) &= \frac{u' - u - p_k(h_k(u') - h_k(u))}{1 - p_k} \\ &\ge 0\end{aligned}$$

by h_k being an element of $\mathcal{F}_k$. It also satisfies A3 since h_k does and $h_l \in \mathcal{F}_l$ since

$$\begin{aligned}
\frac{h_l(u') - h_l(u)}{u' - u} &= \frac{u' - u - p_k(h_k(u') - h_k(u))}{1 - p_k} \\
&= \frac{1}{1 - p_k} - \frac{p_k}{1 - p_k} \frac{h_k(u') - h_k(u)}{u' - u} \\
&\leq \frac{1}{1 - p_k} \\
&\leq \frac{1}{p_l}.
\end{aligned}$$

Therefore, letting

$$\mathbb{P}(U \leq u \mid X = x_k) = h_k(u) \qquad \text{and} \qquad \mathbb{P}(U \leq u \mid X = x_l) = h_l(u)$$

for $l \neq k$ is a proper distribution for the unobservables and we can verify that

$$h_k(u)p_k + \sum_{l \neq k} h_l(u)p_l = u = \mathbb{P}(U \leq u).$$

To show $(\overline{h}_k^J, \underline{h}_l^J) \in \mathcal{F}_{U|X}(x_k, x_l)$, we construct valid cdfs $h_j(\cdot)$ for $j \notin \{k, l\}$ such that

$$\sum_{j=1}^{K} h_j(u)p_j = u.$$

If X is binary this step is not required. Let

$$\begin{aligned}
h_{lk}^*(u) &= \frac{u - p_k \overline{h}_k^J(u) - p_l \underline{h}_l^J(u)}{1 - p_k - p_l} \\
&= \begin{cases}
b_0 & \text{if } b_0 \leq u \leq (1 - p_k)b_0 + p_k a_1 \\
\dfrac{u - p_k a_1 - p_l b_0}{1 - p_k - p_l} & \text{if } (1 - p_k)b_0 + p_k a_1 \leq u \leq p_l b_0 + (1 - p_l)a_1 \\
a_1 & \text{if } p_l b_0 + (1 - p_l)a_1 \leq u \leq a_1 \\
u & \text{if } a_1 \leq u \leq b_1 \\
b_1 & \text{if } b_1 \leq u \leq (1 - p_k)b_1 + p_k a_2 \\
\dfrac{u - p_k a_2 - p_l b_1}{1 - p_k - p_l} & \text{if } (1 - p_k)b_1 + p_k a_2 \leq u \leq p_l b_1 + (1 - p_l)a_2 \\
a_2 & \text{if } p_l b_1 + (1 - p_l)a_2 \leq u \leq a_2 \\
u & \text{if } a_2 \leq u \leq b_2 \\
\quad \vdots & \\
a_{J+1} & \text{if } p_l b_J + (1 - p_l)a_{J+1} \leq u \leq a_{J+1}
\end{cases}
\end{aligned}$$

We can verify that $h_{lk}^*(0) = 0$, $h_{lk}^*(1) = 1$, h_{lk}^* is right-continuous, non-decreasing and also that it satisfies assumption A3. Therefore, setting $h_j = h_{lk}^*$ for $j \notin \{k, l\}$, we have

$$\sum_{j=1}^{K} h_j(u) p_j = \mathbb{P}(U \le u)$$

and all conditional cdfs satisfy A3. We can show that $(\underline{h}_k^J, \overline{h}_l^J) \in \mathcal{F}_{U|X}(x_k, x_l)$ by symmetry. □

Proof of lemma 2. Note that

$$\begin{aligned} f_{U|X}(u \mid x_k) &= \frac{\mathbb{P}(X = x_k \mid U = u) f_U(u)}{\mathbb{P}(X = x_k)} \\ &= \frac{\mathbb{P}(X = x_k \mid U = u) \cdot 1}{\mathbb{P}(X = x_k)}. \end{aligned}$$

In the last line we used that $U \sim \text{Unif}[0, 1]$ and hence its density is equal to 1. □

Proof of theorem 3. To show $\overline{h}_k^c$ is the uniform upper bound, let

$$h_k(u) = \mathbb{P}(U \le u \mid X = k)$$

and suppose there exists a value u^* such that $h_k(u^*) > \overline{h}_k^c(u^*)$.

Case 1. Suppose $u^* \in [0, 1/2]$. Then,

$$\begin{aligned} h_k(u^*) &= \mathbb{P}(U \le u^* \mid X = k) \\ &= \int_0^{u^*} f_{U|X}(v \mid k)\, dv \\ &= \int_0^{u^*} \frac{\mathbb{P}(X = k \mid U = v)}{p_k}\, dv \\ &\le \int_0^{u^*} \frac{p_k + c}{p_k}\, dv \\ &= u^* \left(1 + \frac{c}{p_k}\right) \\ &= \overline{h}_k^c(u^*), \end{aligned}$$

a contradiction of $h_k(u^*) > \overline{h}_k^c(u^*)$. The third line follows by lemma 2. The fourth line follows by c-independence.

Case 2. Suppose $u^* > 1/2$. Then as in the previous case,

$$\begin{aligned}
h_k(u^*) &= \int_0^{u^*} \frac{\mathbb{P}(X = k \mid U = v)}{p_k}\, dv \\
&= 1 - \int_{u^*}^{1} \frac{\mathbb{P}(X = k \mid U = v)}{p_k}\, dv \\
&\leq 1 - \int_{u^*}^{1} \frac{p_k - c}{p_k}\, dv \\
&= u^* \left(1 - \frac{c}{p_k}\right) + \frac{c}{p_k} \\
&= \overline{h}_k^c(u^*),
\end{aligned}$$

a contradiction of $h_k(u^*) > \overline{h}_k^c(u^*)$.

We can use the same arguments to show that $\underline{h}_k^c$ is the uniform (over u) lower bound for cdfs.

To show that $h_k \in \mathcal{F}_k^c$. Then

$$\begin{aligned}
\frac{h_k(u') - h_k(u)}{u' - u} &= \frac{1}{u' - u} \int_u^{u'} \frac{\mathbb{P}(X = k \mid U = v)}{p_k}\, dv \\
&\leq \frac{p_k + c}{p_k} \\
&= 1 + \frac{c}{p_k}.
\end{aligned}$$

The second line follows by c-independence. The lower bound holds similarly.

To show this set is sharp, let $h_k \in \mathcal{F}_{U|X}^c(k)$ be arbitrary. Without loss of generality, let $k = 1$. Define

$$h_0(u) = \frac{u - p_1 h_1(u)}{1 - p_1}.$$

Clearly, $h_0(0) = 0$, $h_1(1) = 1$ and $h_0(u)$ is right-continuous. h_0 is non-decreasing since for $u' \geq u$,

$$\begin{aligned}
h_0(u') - h_0(u) &= \frac{u' - u}{p_0} - \frac{p_1}{p_0}(h_1(u') - h_1(u)) \\
&\geq \frac{u' - u}{p_0} - (u' - u)\frac{\left(1 + \frac{c}{p_1}\right) p_1}{p_0} \\
&= \frac{u' - u}{p_0}(1 - p_1 - c) \\
&= \frac{u' - u}{p_0}(p_0 - c) \\
&\geq 0.
\end{aligned}$$

The second line follows by $h_1 \in \mathcal{F}_k^c$. The last line follows since $c \leq p_0$. Also,

$$\begin{aligned}
\frac{h_0(u') - h_0(u)}{u' - u} &= \frac{1}{p_0} - \frac{p_1}{p_0} \frac{h_1(u') - h_1(u)}{u' - u} \\
&\in \left[\frac{1}{p_0} - \frac{p_1}{p_0}\left(1 + \frac{c}{p_1}\right), \frac{1}{p_0} - \frac{p_1}{p_0}\left(1 - \frac{c}{p_1}\right) \right] \\
&= \left[1 - \frac{c}{p_0}, 1 + \frac{c}{p_0} \right],
\end{aligned}$$

The second line follows by $h_1 \in \mathcal{F}_k^c$. Therefore the set is sharp.

Finally, to see that $(\underline{h}_0^c, \overline{h}_1^c) \in \mathcal{F}_{U|X}^c(0,1)$, note that if $h_1 = \overline{h}_1^c$, then

$$\begin{aligned}
h_0(u) &= \frac{u - p_1 \overline{h}_1^c(u)}{p_0} \\
&= \begin{cases} \dfrac{u - (p_1 + c)u}{p_0} & \text{if } 0 \leq u \leq 1/2 \\ \dfrac{u - (p_1 - c)u - c}{p_0} & \text{if } 1/2 < u \leq 1. \end{cases} \\
&= \begin{cases} \left(1 - \dfrac{c}{p_0}\right) u & \text{if } 0 \leq u \leq 1/2 \\ \left(1 + \dfrac{c}{p_0}\right) u - \dfrac{c}{p_0} & \text{if } 1/2 < u \leq 1. \end{cases} \\
&= \underline{h}_0^c(u).
\end{aligned}$$

We can show that $(\overline{h}_0^c, \underline{h}_1^c) \in \mathcal{F}_{U|X}^c(0,1)$ similarly. $\square$

Proof of theorem 4. Suppose there exists a u^* such that $h_k(u^*) > \overline{h}_k^d(u^*)$.

Case 1. If $u^* \in [0, d/(1-p_k)]$ then

$$h_k(u^*) > \frac{u^*}{p_k} = \overline{h}_k^{\emptyset}(u^*),$$

the no-assumptions upper bound, a contradiction.

Case 2. If

$$\frac{d}{1 - p_k} \leq u \leq 1 - \frac{d}{p_k}$$

then

$$h_k(u^*) > u^* + \frac{d}{p_k},$$

which violates the d-independence assumption.

Case 3. If $u^* > 1 - d/p_k$ then $h_k(u^*) > 1$, a contradiction.

A similar argument can be used to show that $\underline{h}_k^d(u)$ is the uniform lower bound.

To show the set is sharp, fix $h_1 \in \mathcal{F}^d_{U|X}(1)$ and let

$$h_0(u) = \frac{u - p_1 h_1(u)}{1 - p_1}.$$

Hence

$$\begin{aligned} \sup_{u \in [0,1]} |h_0(u) - u| &= \sup_{u \in [0,1]} \left| \frac{p_1}{1-p_1}(h_1(u) - u) \right| \\ &\leq \frac{p_1}{1-p_1} \frac{d}{p_1} \\ &= \frac{d}{p_0}. \end{aligned}$$

Therefore it satisfies d-independence. Also, $h_0 \in \mathcal{F}^d_{U|X}(0)$ because it is a proper cdf since it is non-decreasing, right-continuous and $h_0(0) = 0$ and $h_0(1) = 1$.

Finally we show that $(\underline{h}^d_0, \overline{h}^d_1) \in \mathcal{F}^d(0,1)$. To see this, note that if $h_0 = \underline{h}^d_0$ then

$$\begin{aligned} h_1(u) &= \frac{u - (1-p_1)\underline{h}^d_0(u)}{p_1} \\ &= \begin{cases} \dfrac{u}{p_1} & \text{if } 0 \leq u \leq \dfrac{d}{p_0} \\ u + \dfrac{d}{p_1} & \text{if } \dfrac{d}{p_0} \leq u \leq 1 - \dfrac{d}{1-p_0} \\ 1 & \text{if } 1 - \dfrac{d}{1-p_0} \leq u \leq 1. \end{cases} \\ &= \overline{h}^d_1(u). \end{aligned}$$

Likewise, we can show that $(\overline{h}^d_0, \underline{h}^d_1) \in \mathcal{F}^d_{U|X}(0,1)$. □

Proofs for section 5

Proof of theorem 5. By A4 and quantile equivariance (lemma 5 part 7), for $u \in [0,1]$,

$$\begin{aligned} Q^-_{Y|X}(u \mid x) &= m(x, Q^-_{U|X}(u \mid x)) \\ Q^+_{Y|X}(u \mid x) &= m(x, Q^+_{U|X}(u \mid x)). \end{aligned}$$

Let $\tau \in [0,1]$. Evaluating at $u = F_{U|X}(\tau \mid x)$ yields

$$\begin{aligned} Q^-_{Y|X}(F_{U|X}(\tau \mid x) \mid x) &= m(x, Q^-_{U|X}(F_{U|X}(\tau \mid x) \mid x)) \\ &\leq m(x, \tau) \end{aligned}$$

and

$$\begin{aligned} Q^+_{Y|X}(F_{U|X}(\tau \mid x) \mid x) &= m(x, Q^+_{U|X}(F_{U|X}(\tau \mid x) \mid x)) \\ &\geq m(x, \tau). \end{aligned}$$

The inequalities follow by lemma 5 part 1 and A4. Therefore, for a fixed $F_{U|X}(\tau \mid x)$,

$$m(x,\tau) \in Q^*_{Y|X}(F_{U|X}(\tau \mid x) \mid x).$$

Considering the union over all $F_{U|X}(\cdot \mid x) \in \mathcal{F}_{U|X}(x)$ gives the desired bounds.

To show this set is sharp, fix $m(x,\cdot) \in \mathcal{M}(x)$. Then there exists $F_{U|X} \in \mathcal{F}_{U|X}(x)$ such that

$$m(x,\tau) \in Q^*_{Y|X}(F_{U|X}(\tau \mid x) \mid x).$$

Let $m^{-1}(x,\cdot)$ be the inverse of $m(x,\cdot)$, which exists by A4. Define $\tilde{U} = m^{-1}(X,Y)$. Then

$$\begin{aligned}
\mathbb{P}(\tilde{U} \leq \tau \mid X = x) &= \mathbb{P}(Y \leq m(x,\tau) \mid X = x) \\
&\geq \mathbb{P}(Y \leq Q^-_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid X = x) \\
&= \mathbb{P}(Q^-_{Y|X}(V \mid x) \leq Q^-_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid X = x) \\
&= \mathbb{P}(V \leq F_{Y|X}(Q^-_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid x) \mid X = x) \\
&= F_{Y|X}(Q^-_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid x) \\
&\geq F_{U|X}(\tau \mid x).
\end{aligned}$$

The first line follows by invertibility of $m(x,\cdot)$ and the definition of $\tilde{U}$. The second since $m(x,\tau) \in Q^*_{Y|X}(F_{U|X}(\tau \mid x) \mid x)$. In the third we have $V \mid X = x$ distributed Unif$[0,1]$ by lemma 5 part 3. The fourth by lemma 5 part 6. The fifth by $V \mid X \sim$ Unif$[0,1]$. The sixth by lemma 5 part 4.

Similarly,

$$\begin{aligned}
\mathbb{P}(\tilde{U} < \tau \mid X = x) &= \mathbb{P}(Y < m(x,\tau) \mid X = x) \\
&\leq \mathbb{P}(Y < Q^+_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid X = x) \\
&= \mathbb{P}(Q^+_{Y|X}(V \mid x) < Q^+_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid X = x) \\
&= \mathbb{P}(V < F^-_{Y|X}(Q^+_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid x) \mid X = x) \\
&= F^-_{Y|X}(Q^+_{Y|X}(F_{U|X}(\tau \mid x) \mid x) \mid x) \\
&\leq F_{U|X}(\tau \mid x).
\end{aligned}$$

The first, second, and fifth lines follow as before. In the third line $V \mid X = x \sim$ Unif$[0,1]$ by lemma 5 part 3. The fourth line follows by lemma 5 part 6. The sixth by lemma 5 part 5.

Therefore

$$\begin{aligned}
F_{U|X}(\tau \mid x) &\leq \mathbb{P}(\tilde{U} \leq \tau \mid X = x) \\
&= \mathbb{P}(\tilde{U} < \tau \mid X = x) + \mathbb{P}(\tilde{U} = \tau \mid X = x) \\
&\leq F_{U|X}(\tau \mid x) + \mathbb{P}(\tilde{U} = \tau \mid X = x).
\end{aligned}$$

Since $\mathbb{P}(\tilde{U} = \tau \mid X = x) = 0$ except on a measure zero set,

$$\mathbb{P}(\tilde{U} \leq \tau \mid X = x) = F_{U|X}(\tau \mid x)$$

except on a measure zero set. Let $\tau^* \in [0,1]$ be a point such that

$$\mathbb{P}(\tilde{U} \leq \tau^* \mid X = x) \neq F_{U|X}(\tau^* \mid x).$$

Since these functions are equal on a set of measure one, we can find a sequence $\{\epsilon_n\}_{n\geq 1}$ where $\epsilon_n > 0$ and

$$\mathbb{P}(\tilde{U} \leq \tau^* + \epsilon_n \mid X = x) = F_{U|X}(\tau^* + \epsilon_n \mid x)$$

for any $n \geq 1$. By the right-continuity of both functions, taking the limit as $\epsilon_n \searrow 0$ yields

$$\mathbb{P}(\tilde{U} \leq \tau^* \mid X = x) = F_{U|X}(\tau^* \mid x),$$

a contradiction. Therefore,

$$\mathbb{P}(\tilde{U} \leq \tau \mid X = x) = F_{U|X}(\tau \mid x)$$

for all $\tau \in [0, 1]$.

Therefore all $m(x, \cdot) \in \mathcal{M}(x)$ are attainable since we can find a random variable $\tilde{U} \mid X = x$ which generates such a structural function and has cdf $F_{U|X}$. □

Proof of corollary 4. We have

$$\begin{aligned}
\text{ASF}(x_k) &= \mathbb{E}[m(x_k, U)] \\
&= \int_0^1 m(x_k, u)\, du \\
&\in \left[\int_0^1 Q^-_{Y|X}(\underline{h}^J_k(u) \mid x_k)\, du, \int_0^1 Q^+_{Y|X}(\overline{h}^J_k(u) \mid x_k)\, du \right].
\end{aligned}$$

For this corollary we are assuming X is discrete. First consider the lower bound:

$$\begin{aligned}
&\int_0^1 Q^-_{Y|X}(\underline{h}^J_k(u) \mid x_k)\, du \\
&= \sum_{j=0}^{J} \int_{\tau_j}^{p_k\tau_j + \tau_{j+1}(1-p_k)} Q^-_{Y|X}(\tau_j \mid x_k)\, du + \sum_{j=0}^{J} \int_{p_k\tau_j + \tau_{j+1}(1-p_k)}^{\tau_{j+1}} Q^-_{Y|X}\left(\frac{u - \tau_{j+1}}{p_k} + \tau_{j+1} \middle| x_k \right) du \\
&= (1 - p_k) \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x_k) + p_k \sum_{j=0}^{J} \int_{\tau_j}^{\tau_{j+1}} Q^-_{Y|X}(u \mid x_k)\, du \\
&= (1 - p_k) \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x_k) + p_k \int_0^1 Q^-_{Y|X}(u \mid x_k)\, du \\
&= (1 - p_k) \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x_k) + p_k \mathbb{E}[Y \mid X = x_k].
\end{aligned}$$

Next consider the upper bound:

$$\begin{aligned}
&\int_0^1 Q^+_{Y|X}(\overline{h}^J_k(u \mid x_k) \mid x_k)\, du \\
&= \sum_{j=0}^J \int_{\tau_j}^{(1-p_k)\tau_j + p_k\tau_{j+1}} Q^+_{Y|X}\left(\frac{u-\tau_j}{p_k} + \tau_j \middle| x_k\right) du + \sum_{j=0}^J \int_{(1-p_k)\tau_j+p_k\tau_{j+1}}^{\tau_{j+1}} Q^+_{Y|X}(\tau_{j+1} \mid x_k)\, du \\
&= p_k \sum_{j=0}^J \int_{\tau_j}^{\tau_{j+1}} Q^+_{Y|X}(u \mid x_k)\, du + (1-p_k)\sum_{j=0}^J (\tau_{j+1}-\tau_j) Q^+_{Y|X}(\tau_{j+1} \mid x_k) \\
&= (1-p_k)\sum_{j=0}^J (\tau_{j+1}-\tau_j) Q^+_{Y|X}(\tau_{j+1} \mid x_k) + p_k \mathbb{E}[Y \mid X = x_k].
\end{aligned}$$

Next we prove sharpness of the interior of this interval. For each value A in this interior, we must find a continuous, strictly increasing structural function inside the identified set of theorem 5 which integrates to A.

First note that the functions $Q^-_{Y|X}(\cdot \mid x_k)$ and $Q^+_{Y|X}(\cdot \mid x_k)$ are strictly increasing. This follows since the conditional cdf of $Y \mid X = x_k$ contains no flat regions. To see this, note that

$$\begin{aligned}
\mathbb{P}(Y \leq y \mid X = x_k) &= \mathbb{P}(m(x_k, U) \leq y \mid X = x_k) \\
&= \mathbb{P}(U \leq m^{-1}(x_k, y) \mid X = x_k) \\
&= F_{U|X}(m^{-1}(x_k, y) \mid x_k).
\end{aligned}$$

$F_{U|X}(\cdot \mid x_k)$ is continuous by lemma 4 and $m^{-1}(x_k, \cdot)$ is continuous and strictly increasing by A4. Hence the composition is continuous.

We split our sharpness proof into two cases.

Case 1. Suppose the functions $Q^-_{Y|X}(\cdot \mid x_k)$ and $Q^+_{Y|X}(\cdot \mid x_k)$ are continuous, and therefore are equal to each other.

The functions $\underline{h}^J_k$ and $\overline{h}^J_k$ are both valid cdfs that satisfy A3. Likewise,

$$h_k(u, c) = c\underline{h}^J_k(u) + (1-c)\overline{h}^J_k(u)$$

for $c \in [0, 1]$ is a valid cdf that satisfies assumption A3. Define

$$\text{ASF}(x_k, c) = \int_0^1 Q^-_{Y|X}(h_k(u, c) \mid x_k)\, du$$

for $c \in [0, 1]$. This function spans the entire closed interval bounds as c varies from 0 to 1, but the functions $h_k(u, c)$ are not strictly increasing in u for all values of c. Therefore $Q^-_{Y|X}(h_k(u, c) \mid x_k)$ is not strictly increasing in u. Hence they violate A4 and so are not valid structural functions.[4] Nonetheless, we can find a strictly increasing structural function that is arbitrarily close to the invalid one. To see this, we can slightly perturb the function $h_k(u, c)$ by

$$h_k(u, c, \epsilon) = (1-\epsilon)h_k(u, c) + \epsilon F_U(u)$$

[4]They are never strictly increasing for $c = 0, 1$, but may or may not be strictly increasing for other values of c depending on the value of p_k.

such that $h_k(u,c,\epsilon) \in \mathcal{F}^{\mathcal{T}}_{U|X}(x_k)$ for small ϵ. This follows from $F_U \in \mathcal{F}^{\mathcal{T}}_{U|X}(x_k)$. These perturbations can have positive or negative ϵ when $c \in (0,1)$, but only positive (negative) ϵ when $c = 1$ ($c = 0$). These perturbations ensure that $h_k(u,c,\epsilon)$ is strictly increasing in u and therefore $Q^-_{Y|X}(h_k(u,c,\epsilon) \mid x_k)$ is strictly increasing and continuous, and therefore a valid structural function.

The integral

$$\text{ASF}(x_k, c, \epsilon) = \int_0^1 Q^-_{Y|X}(h_k(u,c,\epsilon) \mid x_k)\, du$$

can be made arbitrarily close to $\text{ASF}(x_k, c)$ for $c \in [0,1]$ by continuity in ϵ. Therefore, by varying $c \in [0,1]$ and $\epsilon \in [-\delta, \delta]$ for some small $\delta > 0$, we can attain any element in $(\text{ASF}(x_k, 0), \text{ASF}(x_k, 1))$, as desired.

Notice that the sign constraint on ϵ when $c = 0$ or 1 is precisely why the endpoints cannot be attained.

Case 2. An additional complication arises if $Q^-_{Y|X}(\cdot \mid x_k)$ is discontinuous at a point, since the implied structural functions $Q^-_{Y|X}(h_k(u,c) \mid x_k)$ will not be proper structural functions due to both their failure to be strictly increasing and their failure to be continuous.

The proof of sharpness in this case is best understood by considering an example, although the idea can be extended to cover the general case. In particular, this example has a single discontinuity, but the same idea applies for multiple discontinuities. On to the example: Suppose

$$U \mid X = 1 \sim \text{Unif}([0,1/4] \cup [3/4,1]) \qquad \text{and} \qquad U \mid X = 0 \sim \text{Unif}[1/4,3/4]$$

and $p_0 = p_1 = 1/2$. Without loss of generality we focus on the ASF at $x_k = 1$. The set quantile function is

$$Q^*_{Y|X}(u \mid 1) = \begin{cases} m(1, u/2) & \text{for } u \in [0, 1/2) \\ [m(1,1/4), m(1,3/4)] & \text{for } u = 1/2 \\ m(1,(u+1)/2) & \text{for } u \in (1/2, 1]. \end{cases}$$

For example, the left plot of figure 11 shows $Q^*_{Y|X}(u \mid 1)$ when $m(x,u) = u$. Hence the conditional quantile function $Q^-_{Y|X}(\tau \mid 1)$ is discontinuous at the value $\tau = 1/2$ and therefore $Q^-_{Y|X}(h_1(u) \mid 1)$ will be a discontinuous function for any $h_1 \in \mathcal{F}^{\mathcal{T}}_{U|X}(1)$. This leaves us in a similar situation as case 1: Although $\text{ASF}(1,c)$ sweeps out the entire interval as c varies from 0 to 1, the function $Q^-_{Y|X}(h_1(u,c) \mid 1)$ is discontinuous, and hence is not a valid structural function. An example of this discontinuity is given in the bottom left plot of figure 12. Here we consider $c = 1$ and $J = 0$, so that we are plugging the no-assumption bound (as shown in the top left plot of this figure) into $Q^-_{Y|X}(\cdot \mid 1)$.

Again as in case 1, the solution is to perturb the cdf $h_1(u,c)$ to $h_1(u,c,\epsilon)$ in such a way that there exists a continuous selection from $Q^*_{Y|X}(h_1(u,c,\epsilon) \mid 1)$. For small $\epsilon > 0$, consider the perturbation

$$h_1(u,c,\epsilon) = \begin{cases} h_1(u+\epsilon, c) & \text{for } u \text{ with } h_1(u+\epsilon,c) \leq 1/2 \\ 1/2 & \text{for } u \text{ with } h_1(u,c) \leq 1/2 \leq h_1(u+\epsilon,c) \\ h_1(u,c) & \text{for } u \text{ with } h_1(u,c) \geq 1/2. \end{cases}$$

The top right plot of figure 12 shows an example, with $c = 1$ and $J = 0$. This function $h_1(u,c,\epsilon)$ is a valid cdf, is in $\mathcal{F}^{\mathcal{T}}_{U|X}(1)$ for small enough ϵ and, crucially, can approximate $h_1(u,c)$ arbitrarily well uniformly from above for $c \in (0,1]$. When we compute the set of identified functions

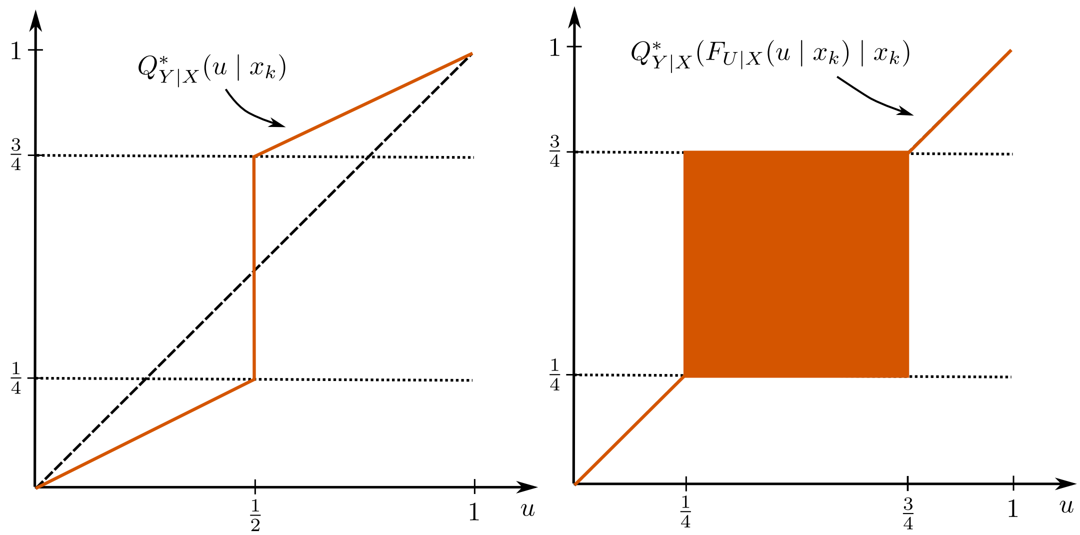


Figure 11: Both plots consider the example with $U \mid X = 1 \sim \text{Unif}([0,1/4] \cup [3/4,1])$ and $m(x,u) = u$. Left: The set quantile function $Q^*_{Y|X}(u \mid x_k)$ for $x_k = 1$. Right: $Q^*_{Y|X}(F_{U|X}(u \mid x_k) \mid x_k)$ where $F_{U|X}$ is the true cdf of $U \mid X$.

$Q^*_{Y|X}(h_1(u,c,\epsilon) \mid 1)$ for $\epsilon > 0$, we obtain

$$Q^*_{Y|X}(h_1(u,c,\epsilon) \mid 1) = \begin{cases} m(1, h_1(u+\epsilon,c)/2) & \text{for } u \text{ with } h_1(u+\epsilon,c) < 1/2 \\ [m(1,1/4), m(1,3/4)] & \text{for } u \text{ with } h_1(u,c) \leq 1/2 \leq h_1(u+\epsilon,c) \\ m(1,(h_1(u,c)+1)/2) & \text{for } u \text{ with } h_k(u,c) > 1/2. \end{cases}$$

The bottom right plot of figure 12 shows an example, with $c = 1$ and $J = 0$. The key observation here is that a continuous function can be traced out in the graph of the set function $Q^*_{Y|X}(h_1(\cdot,c,\epsilon) \mid 1)$ that is arbitrarily close to selections from $Q^*_{Y|X}(h_1(u,c) \mid 1)$, which may not be continuous.

Although we have just dealt with the discontinuity problem, it is possible that the continuous selection we have constructed has flat regions, and thus violates the monotonicity constraint. For example, this happens in the bottom right plot of figure 12. If this happens we can apply ideas from case 1 to find another perturbed cdf which leads to selections from the set quantile which are both continuous and strictly increasing and arbitrarily close to $Q^*_{Y|X}(h_1(u,c) \mid 1)$.

Finally, as in case 1 we consider integrals over the continuous selections from $Q^*_{Y|X}(h_1(u,c,\epsilon) \mid 1)$ and by varying c from 0 to 1 and ϵ appropriately we sweep out the entire interior of the ASF bounds, noting that $h_k(u,c,\epsilon)$ is continuous in c and ϵ. □

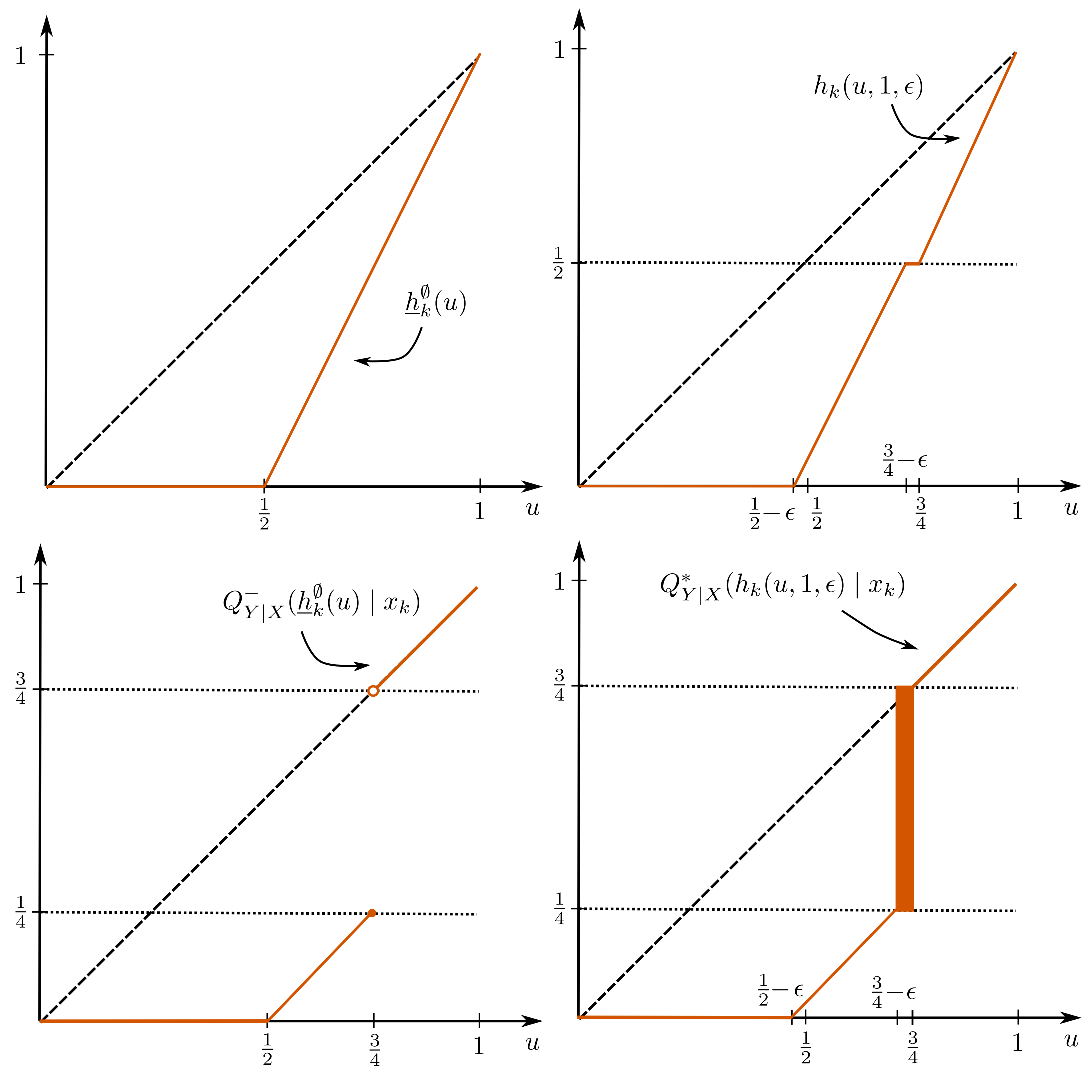


Figure 12: Top left: The no-assumptions lower bound cdf. Top right: an approximation to the no-assumptions lower bound cdf. Bottom left: The left quantile of $Y \mid X$ evaluated at the no-assumptions lower bound cdf. Bottom right: The set quantile of $Y \mid X$ evaluated at the approximation to the no-assumptions lower bound cdf.

Proof of corollary 5. First consider the lower bound:

$$\begin{aligned}
&\int_0^1 Q^-_{Y|X}(\underline{h}^J_k(u) \mid x_k)\, du \\
&= \sum_{j=0}^J \int_{b_j}^{p_k b_j + a_{j+1}(1-p_k)} Q^-_{Y|X}(b_j \mid x_k)\, du + \sum_{j=1}^J \int_{a_j}^{b_j} Q^-_{Y|X}(u \mid x_k)\, du \\
&\quad + \sum_{j=0}^J \int_{p_k b_j + a_{j+1}(1-p_k)}^{a_{j+1}} Q^-_{Y|X}\left(\frac{u - a_{j+1}}{p_k} + a_{j+1} \middle| x_k\right) du \\
&= (1-p_k)\sum_{j=0}^J (a_{j+1} - b_j) Q^-_{Y|X}(b_j \mid x_k) + \sum_{j=1}^J \int_{a_j}^{b_j} Q^-_{Y|X}(u \mid x_k)\, du + p_k \sum_{j=0}^J \int_{b_j}^{a_{j+1}} Q^-_{Y|X}(u \mid x_k)\, du \\
&= (1-p_k)\sum_{j=0}^J (a_{j+1} - b_j) Q^-_{Y|X}(b_j \mid x_k) + (1-p_k)\sum_{j=1}^J \int_{a_j}^{b_j} Q^-_{Y|X}(u \mid x_k)\, du + p_k \mathbb{E}[Y \mid X = x_k].
\end{aligned}$$

Next consider the upper bound:

$$\begin{aligned}
&\int_0^1 Q^+_{Y|X}(\overline{h}^J_k(u) \mid x_k)\, du \\
&= \sum_{j=0}^J \int_{b_j}^{(1-p_k)b_j + a_{j+1}p_k} Q^+_{Y|X}\left(\frac{u - b_j}{p_k} + b_j \middle| x_k\right) du + \sum_{j=1}^J \int_{a_j}^{b_j} Q^+_{Y|X}(u \mid x_k)\, du \\
&\quad + \sum_{j=0}^J \int_{(1-p_k)b_j + a_{j+1}p_k}^{a_{j+1}} Q^+_{Y|X}\left(a_{j+1} \mid x_k\right)\, du \\
&= (1-p_k)\sum_{j=0}^J (a_{j+1} - b_j) Q^+_{Y|X}(a_{j+1} \mid x_k) + (1-p_k)\sum_{j=1}^J \int_{a_j}^{b_j} Q^+_{Y|X}(u \mid x_k)\, du + p_k \mathbb{E}[Y \mid X = x_k].
\end{aligned}$$

Sharpness of the interior follows as in the proof of corollary 4. □

Sketch of c-independence ASF bounds. We have

$$\begin{aligned}
\text{LB} &= \int_0^{\frac{1}{2}} Q^-_{Y|X}\left(\left(1 - \frac{c}{p_k}\right) u \mid x_k\right) du + \int_{\frac{1}{2}}^1 Q^-_{Y|X}\left(\left(1 + \frac{c}{p_k}\right) u - \frac{c}{p_k} \mid x_k\right) du \\
&= \frac{1}{2}\left(\frac{1}{2}\left(1 - \frac{c}{p_k}\right)\right)^{-1} \int_0^{\frac{1}{2}\left(1 - \frac{c}{p_k}\right)} Q^-_{Y|X}\left(v \mid x_k\right)\, dv \\
&\quad + \frac{1}{2}\left(\frac{1}{2}\left(1 + \frac{c}{p_k}\right)\right)^{-1} \int_{\frac{1}{2}\left(1 - \frac{c}{p_k}\right)}^1 Q^-_{Y|X}\left(v \mid x_k\right)\, dv \\
&= \frac{1}{2}\mathbb{E}\left[Q^-_{Y|X}(V \mid x_k) \mid V \leq \frac{1}{2}\left(1 - \frac{c}{p_k}\right)\right] + \frac{1}{2}\mathbb{E}\left[Q^-_{Y|X}(V \mid x_k) \mid V \geq \frac{1}{2}\left(1 - \frac{c}{p_k}\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[Q^-_{Y|X}(V \mid X) \mid X = x_k, Q^-_{Y|X}(V \mid X) \leq Q^-_{Y|X}\left(\frac{1}{2}\left(1 - \frac{c}{p_k}\right) \mid x_k\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[Q^-_{Y|X}(V \mid X) \mid X = x_k, Q^-_{Y|X}(V \mid X) \geq Q^-_{Y|X}\left(\frac{1}{2}\left(1 - \frac{c}{p_k}\right) \mid x_k\right)\right],
\end{aligned}$$

where we defined $V \mid X \sim \text{Unif}[0,1]$ (in particular, $V \perp\!\!\!\perp X$, as used in the fourth line). By lemma 5 part 3, $Q^-_{Y|X}(V \mid X) \sim Y$ and therefore the previous expression equals

$$\frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \leq Q^-_{Y|X}\left(\frac{1}{2}\left(1 - \frac{c}{p_k}\right) \mid x_k\right)\right] \\ + \frac{1}{2}\mathbb{E}\left[Y \mid X = x_k, Y \geq Q^-_{Y|X}\left(\frac{1}{2}\left(1 - \frac{c}{p_k}\right) \mid x_k\right)\right].$$

Similar techniques can be used to derive the upper bound.

We next briefly consider sharpness of these ASF bounds. When $c < \min\{p_0, p_1\}$, the distribution

$$\begin{aligned} \mathbb{P}(Y \leq y \mid X = k) &= F_{U|X}(m^{-1}(k,y) \mid k) \\ &= \frac{1}{\mathbb{P}(X = k)} \int_0^{m^{-1}(k,y)} \mathbb{P}(X = k \mid U = u)\, du \end{aligned}$$

is a continuous and strictly increasing function because $\mathbb{P}(X = k \mid U = u)$ is bounded away from zero and $m^{-1}(k, \cdot)$ is continuous and strictly increasing. Therefore its inverses $Q^-_{Y|X}(\cdot \mid k)$ and $Q^+_{Y|X}(\cdot \mid k)$ are strictly increasing and continuous everywhere. The cdf bounds $\underline{h}^c_k(\cdot)$ and $\overline{h}^c_k(\cdot)$ are also strictly increasing and continuous. Hence the quantile bounds $Q^-_{Y|X}(\underline{h}^c_k(u) \mid k)$ and $Q^+_{Y|X}(\overline{h}^c_k(u) \mid k)$ are strictly increasing, continuous functions. Thus the endpoints are attainable. We can show the interior is sharp by taking linear combinations of the upper and lower cdf bounds. As with the endpoints, this linear combination function $h_k(u,c)$ leads to $Q^-_{Y|X}(h_k(u,c) \mid x_k)$ being a valid structural function, so that the complications that arose for $\mathcal{T}$-independence in the proof of corollary 5 do not arise for c-independence with $c < \min\{p_0, p_1\}$. Thus the entire closed interval is attainable and hence is the identified set.

If $c \geq \min\{p_0, p_1\}$, only the interior is sharp. This can be shown as in the proof of corollary 5. □

Sketch of d-independence ASF bounds. We have

$$\begin{aligned} \text{LB} &= \int_0^{d/p_k} Q^-_{Y|X}(0 \mid x_k)\, du \\ &\quad + \int_{d/p_k}^{1-d/(1-p_k)} Q^-_{Y|X}\left(u - \frac{d}{p_k} \mid x_k\right) du + \int_{1-d/(1-p_k)}^{1} Q^-_{Y|X}\left(\frac{u-1}{p_k} + 1 \mid x_k\right) du \\ &\equiv T_1 + T_2 + T_3. \end{aligned}$$

The first piece is

$$\begin{aligned} T_1 &= \int_0^{d/p_k} Q^-_{Y|X}(0 \mid x_k)\, du \\ &= \frac{d}{p_k} Q^-_{Y|X}(0 \mid x_k)\, du. \end{aligned}$$

The second and third pieces are

$$
\begin{aligned}
T_2 + T_3 &= \int_{d/p_k}^{1-d/(1-p_k)} Q^-_{Y|X}\left(u - \frac{d}{p_k} \mid x_k\right) du + \int_{1-d/(1-p_k)}^{1} Q^-_{Y|X}\left(\frac{u-1}{p_k} + 1 \mid x_k\right) du \\
&= \int_0^{1-d/p_k(1-p_k)} Q^-_{Y|X}(u \mid x_k)\, du + p_k \int_{1-d/p_k(1-p_k)}^{1} Q^-_{Y|X}(u \mid x_k)\ du \\
&= \int_0^1 Q^-_{Y|X}(u \mid x_k)\ du + (p_k - 1) \int_{1-d/p_k(1-p_k)}^{1} Q^-_{Y|X}(u \mid x_k)\ du \\
&= \mathbb{E}[Y \mid X = x_k] + (p_k - 1) \int_{1-d/p_k(1-p_k)}^{1} Q^-_{Y|X}(u \mid x_k)\ du.
\end{aligned}
$$

In the last line we used

$$
\int_0^1 Q^-_{Y|X}(u \mid x_k)\, du = \mathbb{E}[Y \mid X = x_k].
$$

Similar techniques can be used to derive the upper bound. We can use results in the proof of corollary 5 to show the interior of the bounds is sharp. □

Proof of corollary 6. The key step is that by theorem 2 it is possible to jointly achieve the cdf bounds $(\underline{h}_k^J, \overline{h}_l^J)$ and $(\underline{h}_l^J, \overline{h}_k^J)$. Hence the bounds on the ATE follow immediately from the bounds on the ASF's obtained in corollary 5.

To show sharpness of the interior, let

$$
(F_{U|X}(u \mid x_k), F_{U|X}(u \mid x_l)) = (h_k(u,c), h_l(u, 1-c))
$$

where

$$
\begin{aligned}
h_k(u,c) &= c\underline{h}_k^J(u) + (1-c)\overline{h}_k^J(u) \\
h_l(u, 1-c) &= (1-c)\underline{h}_l^J(u) + c\overline{h}_l^J(u).
\end{aligned}
$$

These two cdfs satisfy assumption A3. Moreover, they are jointly attainable. When X is binary, this holds since

$$
p_k h_k(u,c) + p_l h_l(u,c) = c\left(p_k \underline{h}_k^J(u) + p_l \overline{h}_l^J(u)\right) + (1-c)\left(p_k \overline{h}_k^J(u) + p_l \underline{h}_l^J(u)\right) = u
$$

by the last part of theorem 2. To see that the two cdfs are jointly attainable when $\text{supp}(X)$ contains more than two points, consider

$$
\begin{aligned}
\mathbb{P}(U \leq u \mid X \notin \{x_k, x_l\}) &= \frac{u - p_k h_k(u,c) - p_l h_l(u, 1-c)}{1 - p_k - p_l} \\
&= \frac{u - c(p_k \underline{h}_k^J(u) + p_l \overline{h}_l^J(u)) - (1-c)(p_k \overline{h}_k^J(u) + p_l \underline{h}_l^J(u))}{1 - p_k - p_l} \\
&= c\frac{u - p_k \underline{h}_k^J(u) - p_l \overline{h}_l^J(u)}{1 - p_k - p_l} + (1-c)\frac{u - p_k \overline{h}_k^J(u) - p_l \underline{h}_l^J(u)}{1 - p_k - p_l} \\
&\equiv c h_{kl}^*(u) + (1-c) h_{lk}^*(u).
\end{aligned}
$$

The first line follows by the law of total probability. Here $h_{kl}^*(u), h_{lk}^*(u)$ are defined as in the proof of theorem 2. These are proper cdfs that satisfy assumption A3, and therefore $(h_k(\cdot, c), h_l(\cdot, 1-c)) \in$

$\mathcal{F}^{\mathcal{T}}_{U|X}(x_k, x_l)$. The proof of sharpness now continues similarly to the proof of corollary 4. $\square$

Proof of lemma 3. By theorem 5, for any $\tau \in [0,1]$,

$$Q^{-}_{Y|X}(\underline{h}^J_k(\tau) \mid x_k) \leq m(x_k, \tau) \leq Q^{+}_{Y|X}(\overline{h}^J_k(\tau) \mid x_k).$$

Since $U \in [0,1]$, the bounds hold for U as well. By corollary 6, the bound for $\Delta(x_l \to x_k)$ similarly holds. $\square$

Proof of corollary 7. These bounds follow directly from lemma 3 and sharpness follows from sharpness of theorem 5 and arguments similar to the proof of corollary 4. $\square$

Proofs for section 6

Proof of proposition 4. This is the rank invariance case of theorem 2 in Chernozhukov and Hansen (2005). $\square$

Proof of theorem 6. For a given cdf $h_1 \in \mathcal{F}_{U|Z}(1)$, the corresponding cdf $h_0 \in \mathcal{F}_{U|Z}(0)$ is given by

$$h_0(u) = \frac{u - p_1 h_1(u)}{1 - p_1}$$

for any u. For a fixed cdf $h_1 \in \mathcal{F}_{U|Z}(1)$ and given τ, Chernozhukov and Hansen's (2005) proof of proposition 4 shows that $(m(0,\tau), m(1,\tau)) = (\theta_0(h_1(\tau)), \theta_1(h_1(\tau)))$ are uniquely identified in $\mathcal{L}$. For a fixed τ, the identified set is the collection of solutions

$$\left\{(\theta_0(h_1(\tau)), \theta_1(h_1(\tau))) : h_1 \in \mathcal{F}_{U|Z}(1)\right\}.$$

We characterize this set next.

We first show that $\theta_0(\cdot)$ and $\theta_1(\cdot)$ are both monotonic in their arguments. The identifying moment conditions are:

$$\begin{aligned} h_0(\tau) &= \frac{\tau - p_1 h_1(\tau)}{1 - p_1} \\ &= \mathbb{P}(Y \leq \theta_0(h_1(\tau)) \mid X = 0, Z = 0) P_{X|Z}(0 \mid 0) \\ &\quad + \mathbb{P}(Y \leq \theta_1(h_1(\tau)) \mid X = 1, Z = 0) P_{X|Z}(1 \mid 0) \end{aligned} \tag{21}$$

and

$$\begin{aligned} h_1(\tau) &= \mathbb{P}(Y \leq \theta_0(h_1(\tau)) \mid X = 0, Z = 1) P_{X|Z}(0 \mid 1) \\ &\quad + \mathbb{P}(Y \leq \theta_1(h_1(\tau)) \mid X = 1, Z = 1) P_{X|Z}(1 \mid 1) \end{aligned} \tag{22}$$

by the law of total probability. Using the implicit function theorem, we differentiate both sides of (21) and (22) with respect to $h_1(\tau)$ and obtain:

$$\begin{aligned} \begin{pmatrix} -\dfrac{p_1}{1-p_1} \\ \\ 1 \end{pmatrix} &= \begin{pmatrix} f_{Y|X,Z}(\theta_0(h_1(\tau)) \mid 0,0) P_{X|Z}(0 \mid 0) & f_{Y|X,Z}(\theta_1(h_1(\tau)) \mid 1,0) P_{X|Z}(1 \mid 0) \\ f_{Y|X,Z}(\theta_0(h_1(\tau)) \mid 0,1) P_{X|Z}(0 \mid 1) & f_{Y|X,Z}(\theta_1(h_1(\tau)) \mid 1,1) P_{X|Z}(1 \mid 1) \end{pmatrix} \begin{pmatrix} \theta_0'(h_1(\tau)) \\ \theta_1'(h_1(\tau)) \end{pmatrix}. \\ &= \Pi'(m(0,\tau), m(1,\tau)) \begin{pmatrix} \theta_0'(h_1(\tau)) \\ \theta_1'(h_1(\tau)) \end{pmatrix}. \end{aligned}$$

By the full rank assumption of the matrix $\Pi'(m(0,\tau), m(1,\tau))$, we find

$$\begin{aligned}
\begin{pmatrix} \theta_0'(h_1(\tau)) \\ \theta_1'(h_1(\tau)) \end{pmatrix} &= \begin{pmatrix} f_{Y|X,Z}(m(0,\tau) \mid 0,0) P_{X|Z}(0 \mid 0) & f_{Y|X,Z}(m(1,\tau) \mid 1,0) P_{X|Z}(1 \mid 0) \\ f_{Y|X,Z}(m(0,\tau) \mid 0,1) P_{X|Z}(0 \mid 1) & f_{Y|X,Z}(m(1,\tau) \mid 1,1) P_{X|Z}(1 \mid 1) \end{pmatrix}^{-1} \begin{pmatrix} -\dfrac{p_1}{1-p_1} \\ \\ 1 \end{pmatrix} \\
&= \frac{1}{\det(\Pi'(m(0,\tau), m(1,\tau)))} \\
&\quad \times \begin{pmatrix} -\dfrac{p_1 f_{Y|X,Z}(m(1,\tau) \mid 1,1) P_{X|Z}(1 \mid 1)}{1-p_1} - f_{Y|X,Z}(m(1,\tau) \mid 1,0) P_{X|Z}(1 \mid 0) \\ \\ \dfrac{p_1 f_{Y|X,Z}(m(0,\tau) \mid 0,1) P_{X|Z}(0 \mid 1)}{1-p_1} + f_{Y|X,Z}(m(0,\tau) \mid 0,0) P_{X|Z}(0 \mid 0) \end{pmatrix}.
\end{aligned}$$

Hence the signs of the derivatives of the two components are opposites.

By the rank condition,

$$\det(\Pi'(m(0,\tau), m(1,\tau))) \neq 0$$

for any τ. By the continuity of the determinant operator, the matrix Π', and the functions $(m(0,\tau), m(1,\tau))$, the determinant is a continuous function of τ. Therefore, by the intermediate value theorem, this determinant is either positive for all τ, or negative for all τ.

Thus, we have shown that θ_0 is increasing (or decreasing) in $h_1(\tau)$, and that θ_1 is decreasing (increasing) in $h_1(\tau)$. Solving the system for

$$\mathbb{P}(U \leq \tau \mid Z = 1) = \underline{h}_1(\tau)$$

will yield the lower (upper) bound for $m(0,\tau)$ and the upper (lower) bound for $m(1,\tau)$, while solving it for

$$\mathbb{P}(U \leq \tau \mid Z = 1) = \overline{h}_1(\tau)$$

will yield the upper (lower) bound for $m(0,\tau)$ and the lower (upper) bound for $m(1,\tau)$. For sharpness, consider

$$h_1(u,c) = c\underline{h}_1(u) + (1-c)\overline{h}_1(u) \in \mathcal{F}_{U|Z}(1)$$

for $c \in [0,1]$. Since this cdf varies from the lower to the upper bound, $(\theta_0(h_1(\tau,c)), \theta_1(h_1,\tau,c))$ varies from the lower to the upper bound as c varies from zero to one. By continuity of $h_1(\tau,\cdot)$ and of $(\theta_0(\cdot), \theta_1(\cdot))$, and by the intermediate value theorem, for any element of the bounds we derived above there exists a c which attains it. □

Proofs for section 7

Proof of theorem 7. For a fixed $F_{U|X}(\cdot \mid x_k) \in \mathcal{F}_{U|X}(x_k)$, the identified set for $g(x_k)$ is determined by the set of solutions to the equation

$$\mathbb{E}[Y \mid X = x_k] = F_{U|X}(g(x_k) \mid x_k).$$

Taking $Q^-_{U|X}$ and $Q^+_{U|X}$ of both sides yields

$$\begin{aligned}
Q^-_{U|X}(\mathbb{E}[Y \mid X = x_k] \mid x_k) &= Q^-_{U|X}(F_{U|X}(g(x_k) \mid x_k) \mid x_k) \\
&\leq g(x_k)
\end{aligned}$$

and

$$\begin{aligned} Q^{+}_{U|X}(\mathbb{E}[Y \mid X = x_k] \mid x_k) &= Q^{+}_{U|X}(F_{U|X}(g(x_k) \mid x_k) \mid x_k) \\ &\geq g(x_k). \end{aligned}$$

The inequalities here hold by lemma 5 part 1. Thus

$$g(x) \in Q^{*}_{U|X}\big(\mathbb{E}[Y \mid X = x_k] \mid x_k\big).$$

These bounds are sharp as in the proof of theorem 5. Next, letting $F_{U|X}$ range over $\mathcal{F}_{U|X}(x_k)$ allows us to construct the identified set. Since $\underline{h}_k$ and $\overline{h}_k$ are pointwise and uniform cdf bounds for $\mathcal{F}_{U|X}(x_k)$, inverting these cdf bounds provides the pointwise and uniform bounds for the left and right quantile functions $Q^{-}_{U|X}(\cdot \mid x_k)$ and $Q^{+}_{U|X}(\cdot \mid x_k)$. Hence

$$g(x_k) \in \Big[\overline{Q}^{-}_{U|X}(\mathbb{E}(Y \mid X = x_k) \mid x_k), \quad \underline{Q}^{+}_{U|X}(\mathbb{E}(Y \mid X = x_k) \mid x_k)\Big].$$

Sharpness follows by defining

$$h_k(u, c) = \min\{\underline{h}_k(u + c), \overline{h}_k(u)\}$$

for $c \in [0, 1]$, where $\underline{h}_k(u) = 1$ for $u > 1$. Note that

$$h_k(u, 0) = \underline{h}_k(u) \qquad \text{and} \qquad h_k(u, 1) = \overline{h}_k(u).$$

For each $c \in [0, 1]$, $h_k(u, c)$ is a proper cdf since it is non-decreasing, right-continuous and $h_k(0, c) = 0$ and $h_k(1, c) = 1$. We can verify that

$$\underline{h}_k(u) \leq h_k(u, c) \leq \overline{h}_k(u)$$

for all $u, c \in [0, 1]$. Furthermore, it can be shown that $h_k(\cdot, c) \in \mathcal{F}_{U|X}(x_k)$ for any $c \in [0, 1]$ for each of the sets $\mathcal{F}_{U|X}(x_k)$ defined in section 4.

We can also see that the left-inverse of $h_k(u, c)$ can be written as

$$\begin{aligned} h^{-}_k(\tau, c) &= \inf\{u \in [0, 1] : \min\{\underline{h}_k(u + c), \overline{h}_k(u)\} \geq \tau\} \\ &= \inf\{u \in [0, 1] : \underline{h}_k(u + c) \geq \tau \text{ and } \overline{h}_k(u) \geq \tau\} \\ &= \max\{\overline{h}^{-}_k(\tau), \underline{h}^{-}_k(\tau) - c\}. \end{aligned}$$

Using a similar argument, we also have

$$h^{+}_k(\tau, c) = \max\{\overline{h}^{+}_k(\tau), \underline{h}^{+}_k(\tau) - c\}.$$

Note that $h^{-}(\tau, 0) = \max\{\overline{h}^{-}_k(\tau), \underline{h}^{-}_k(\tau)\} = \underline{h}^{-}_k(\tau)$ and $h^{-}(\tau, 1) = \max\{\overline{h}^{-}_k(\tau), \underline{h}^{-}_k(\tau) - 1\} = \overline{h}^{-}_k(\tau)$ (and similarly for right-quantiles) and therefore we can sweep out the identified set by using the left- and right-inverses of $h_k(u, c)$ and letting vary c between zero and one. □

Left and right quantile lemmas

Lemma 5. Let X be a scalar random variable with $x_L \leq X \leq x_U$ almost surely, where x_L may be $-\infty$ and x_U may be $+\infty$. Let $F(x) = \mathbb{P}(X \leq x)$ and $F^-(x) = \mathbb{P}(X < x)$. For $\tau \in [0,1]$, define

$$Q^-(\tau) = \inf\{x \in [x_L, x_U] : F(x) \geq \tau\} \qquad \text{and} \qquad Q^+(\tau) = \sup\{x \in [x_L, x_U] : F(x) \leq \tau\}.$$

Let $\tau \in [0,1]$ and $x \in [x_L, x_U]$. Then the following properties hold.

1. $Q^-(F(x)) \leq x$ and $Q^+(F(x)) \geq x$.
2. $Q^-(t) = Q^+(t)$ for all $t \in [0,1]$ except for a set of measure zero.
3. Let Y be a random variable with cdf F. Let $U \sim \text{Unif}[0,1]$. Then the cdf of $Q^-(U)$ and $Q^+(U)$ is also F.
4. $F(Q^-(\tau)) \geq \tau$ and $F(Q^+(\tau)) \geq \tau$.
5. $F^-(Q^+(\tau)) \leq \tau$.
6. $Q^-(\tau) \leq x \Leftrightarrow \tau \leq F(x)$ and $Q^+(\tau) < x \Leftrightarrow \tau < F^-(x)$.
7. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a strictly increasing continuous function. Then for all $\tau \in [0,1]$,

$$\begin{aligned} Q^-_{\phi(X)}(\tau) &\equiv \inf\{y \in [\phi(x_L), \phi(x_U)] : \mathbb{P}(\phi(X) \leq y) \geq \tau\} \\ &= \phi(Q^-(\tau)) \end{aligned}$$

   and

$$\begin{aligned} Q^+_{\phi(X)}(\tau) &\equiv \sup\{y \in [\phi(x_L), \phi(x_U)] : \mathbb{P}(\phi(X) \leq y) \leq \tau\} \\ &= \phi(Q^+(\tau)). \end{aligned}$$

Proof.

1. The first part is lemma 21.1 part (iv) of van der Vaart (2000). The second part follows from $x \in \{y \in [x_L, x_U] : F(y) \leq F(x)\}$, therefore $Q^+(F(x)) = \sup\{y \in [x_L, x_U] : F(y) \leq F(x)\} \geq x$ by the definition of the supremum.
2. By definition, $Q^-(t)$ and $Q^+(t)$ differ if and only if the set $\{y \in [x_L, x_U] : F(y) = t\}$ has nonzero measure. Also, the functions $Q^-(t)$ and $Q^+(t)$ will be discontinuous at t if and only if $\{y \in [x_L, x_U] : F(y) = t\}$ is a flat region of the cdf, which happens if and only if $\{y \in [x_L, x_U] : F(y) = t\}$ has measure non-zero. Therefore, $Q^-(t) \neq Q^+(t)$ only on their respective set of discontinuity points. Since Q^- and Q^+ are monotone functions, the Darboux-Froda theorem (e.g., theorem 4.30 of Rudin 1976) shows that the set of discontinuities of each function is at most countable, and hence has measure zero.
3. We have

$$\begin{aligned} \mathbb{P}(Q^-(U) \leq y) &= \mathbb{P}(U \leq F(y)) \\ &= F(y). \end{aligned}$$

   The first line follows by lemma 21.1 part (i) of van der Vaart (2000). The second by $U \sim \text{Unif}[0,1]$. By part 2 of this lemma, $Q^-(U) = Q^+(U)$ a.s. and therefore $\mathbb{P}(Q^+(U) \leq y) = F(y)$.

4. The first part is lemma 21.1 part (ii) of van der Vaart (2000). The second part follows by $Q^+(\tau) \geq Q^-(\tau)$ and monotonicity of $F(y)$.

5. This is lemma 2.1f of Hosseini (2010).

6. The first part is lemma 21.1 part (i) of van der Vaart (2000). The second part follows from proposition 1 point (5) in Embrechts and Hofert (2013) and from the fact that F^- is the generalized inverse of Q^+, a right-continuous function.

7. Both statements hold for $\tau \in (0,1)$ by theorem 3.1 of Hosseini (2010). For the first statement, equality at $\tau = 0$ holds because $Q^-_{\phi(X)}(0) = \phi(x_L)$ and $Q^-(0) = x_L$ by the support restrictions. Equality at $\tau = 1$ holds because of

$$\begin{aligned} Q^-_{\phi(X)}(1) &= \inf\{y \in [\phi(x_L), \phi(x_U)] : \mathbb{P}(\phi(X) \leq y) \geq 1\} \\ &\leq \phi(Q^-(1)) \end{aligned}$$

since $\mathbb{P}(\phi(X) \leq \phi(Q^-(1)) = \mathbb{P}(X \leq Q^-(1)) = 1 \geq 1$, and also because

$$\begin{aligned} Q^-_{\phi(X)}(1) &= \phi(\phi^{-1}(Q^-_{\phi(X)}(1))) \\ &\leq \phi(\phi^{-1}(\phi(Q^-(1)))) \\ &= \phi(Q^-(1)) \end{aligned}$$

using the first inequality with ϕ^{-1}. This shows that $Q^-_{\phi(X)}(1) = \phi(Q^-(1))$. A similar proof can be used to show that the statement holds for right quantiles for $\tau \in \{0,1\}$.

□

C Additional results

In this appendix we give additional results relating to various parts of the paper.

On interpretation of $\mathcal{T}$-independence

In this subsection we give additional results relating to section 3. The first corollary specializes theorem 1 to the case when $\mathcal{T}$ is a finite set.

Corollary 8. Suppose A1 holds. Let

$$\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$$

and $\tau_0 = 0$ and $\tau_{J+1} = 1$. Define

$$\mathcal{U}_j = [\tau_{j-1}, \tau_j)$$

for $j = 1, \ldots, J$ and $\mathcal{U}_{J+1} = [\tau_J, \tau_{J+1}]$. Suppose X is binary with $\mathbb{P}(X = 1) \in (0,1)$. Then U is $\mathcal{T}$-independent of X if and only if

$$\frac{1}{|\mathcal{U}_j|} \int_{\mathcal{U}_j} p(u)\, du = \mathbb{P}(X = 1) \tag{23}$$

for all $j = 1, \ldots, J+1$, where $|\mathcal{U}_j| = \tau_j - \tau_{j-1}$.

Corollary 8 says that $\mathcal{T}$-independence for finite $\mathcal{T}$ holds if and only if there exists a partition of the domain $[0,1]$ such that the average value of the propensity score within each partition equals the overall average of the propensity score. An equivalent way to phrase this result is that the average value of the propensity score has to be equal across all partition sets:

$$\frac{1}{|\mathcal{U}_j|} \int_{\mathcal{U}_j} p(u) \; du = \frac{1}{|\mathcal{U}_k|} \int_{\mathcal{U}_k} p(u) \; du$$

for all j and k from $1, \ldots, J+1$.

Applying lemma 1, when $\mathcal{T}$ is finite, corollary 8 shows that $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$-independence holds if and only if

$$\mathbb{P}(X = 1 \mid U \in \mathcal{U}_j) = \mathbb{P}(X = 1)$$

for all $j = 1, \ldots, J+1$. Or, equivalently, if

$$\mathbb{P}(X = 1 \mid U \in \mathcal{U}_j) = \mathbb{P}(X = 1 \mid U \in \mathcal{U}_k)$$

for $j \neq k$.

Proof of corollary 8. To give additional perspective on these results, we provide a different proof of this corollary than the one used in theorem 1. This proof begins with lemma 6 below and then extends it to the case where multiple cdf independence conditions hold.

($\Rightarrow$) Suppose U is $\mathcal{T}$-independent of X. By lemma 6,

$$\int_0^{\tau_{j+1}} p(u) \; du = \tau_{j+1} \cdot \mathbb{P}(X = 1) \qquad \text{and} \qquad \int_0^{\tau_j} p(u) \; du = \tau_j \cdot \mathbb{P}(X = 1)$$

for all $j = 0, \ldots, J$. Subtracting the second equation from the first yields

$$\int_{\tau_j}^{\tau_{j+1}} p(u) \; du = (\tau_{j+1} - \tau_j)\mathbb{P}(X = 1)$$

for all $j = 0, \ldots, J$, as desired.

($\Leftarrow$) Suppose

$$\int_{\tau_j}^{\tau_{j+1}} p(u) \; du = (\tau_{j+1} - \tau_j)\mathbb{P}(X = 1)$$

for all $j = 0, \ldots, J$. Letting $j = 0$, we have

$$\int_0^{\tau_1} p(u) \; du = \tau_1 \cdot \mathbb{P}(X = 1).$$

By lemma 6,

$$\mathbb{P}(U \leq \tau_1 \mid X = 1) = \tau_1 \qquad \text{and} \qquad \mathbb{P}(U \leq \tau_1 \mid X = 0) = \tau_1.$$

Letting $j = 1$, we have

$$\int_{\tau_1}^{\tau_2} p(u) \; du = (\tau_2 - \tau_1)\mathbb{P}(X = 1)$$

by assumption. Adding our average integral equality for $j = 0$ yields

$$\int_0^{\tau_2} p(u)\, du = \tau_2 \cdot \mathbb{P}(X = 1).$$

By lemma 6, this implies that

$$\mathbb{P}(U \leq \tau_2 \mid X = 1) = \tau_2 \qquad \text{and} \qquad \mathbb{P}(U \leq \tau_2 \mid X = 0) = \tau_2.$$

Repeating this argument shows that

$$\mathbb{P}(U \leq \tau_j \mid X = 1) = \tau_j \qquad \text{and} \qquad \mathbb{P}(U \leq \tau_j \mid X = 0) = \tau_j$$

for all $j = 0, \ldots, J$.

□

Lemma 6. Suppose X is binary with $\mathbb{P}(X = 1) \in (0, 1)$. Then U is τ-cdf independent of X if and only if

$$\frac{1}{\tau} \int_0^{\tau} p(u)\, du = \mathbb{P}(X = 1).$$

Proof of lemma 6. Recall that by lemma 2 we have

$$p(u) = f_{U|X}(u \mid 1)\mathbb{P}(X = 1).$$

Hence

$$\begin{aligned}
\int_0^{\tau} p(u)\, du &= \int_0^{\tau} f_{U|X}(u \mid 1)\mathbb{P}(X = 1)\, du \\
&= \left(\int_0^{\tau} f_{U|X}(u \mid 1)\, du \right) \mathbb{P}(X = 1) \\
&= \mathbb{P}(U \leq \tau \mid X = 1)\mathbb{P}(X = 1).
\end{aligned}$$

($\Rightarrow$) Suppose τ-cdf independence holds: $\mathbb{P}(U \leq \tau \mid X = 1) = \tau$ and $\mathbb{P}(U \leq \tau \mid X = 0) = \tau$. Then

$$\int_0^{\tau} p(u)\, du = \tau \cdot \mathbb{P}(X = 1)$$

as desired.

($\Leftarrow$) Next suppose

$$\int_0^{\tau} p(u)\, du = \tau \cdot \mathbb{P}(X = 1)$$

holds. Then

$$\tau \cdot \mathbb{P}(X = 1) = \mathbb{P}(U \leq \tau \mid X = 1)\mathbb{P}(X = 1)$$

and hence

$$\tau = \mathbb{P}(U \leq \tau \mid X = 1)$$

since $\mathbb{P}(X = 1) > 0$. Similarly to above,

$$\int_0^{\tau} [1 - p(u)]\, du = \mathbb{P}(U \leq \tau \mid X = 0)\mathbb{P}(X = 0).$$

Consider the left hand side:

$$\int_0^\tau [1 - p(u)]\, du = \tau - \int_0^\tau p(u)\, du$$
$$= \tau - \tau \mathbb{P}(X = 1)$$
$$= \tau[1 - \mathbb{P}(X = 1)]$$
$$= \tau \cdot \mathbb{P}(X = 0).$$

Hence

$$\tau \cdot \mathbb{P}(X = 0) = \mathbb{P}(U \leq \tau \mid X = 0)\mathbb{P}(X = 0)$$

and thus

$$\tau = \mathbb{P}(U \leq \tau \mid X = 0)$$

since $\mathbb{P}(X = 0) > 0$. Thus τ-cdf independence holds.

□

When examining whether a plotted propensity score satisfies the average value condition (23), it may be easier to visualize areas above and below the propensity score. The following corollary shows that with $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$, $\mathcal{T}$-independence is equivalent to there being a partition of the domain such that the ratio of the area below the propensity score to the area above the propensity score is the same across all partitions.

Corollary 9. Suppose X is binary with $\mathbb{P}(X = 1) \in (0, 1)$. Let $\mathcal{T}$ and $\mathcal{U}_j$ be defined as in corollary 8. Then $\mathcal{T}$-independence holds if and only if

$$\frac{\int_{\mathcal{U}_j} p(u)\, du}{\int_{\mathcal{U}_j} [1 - p(u)]\, du} = \frac{\int_{\mathcal{U}_k} p(u)\, du}{\int_{\mathcal{U}_k} [1 - p(u)]\, du} \tag{24}$$

for all $j, k = 1, \ldots, J$.

We furthermore have that, under $\mathcal{T}$-independence of U and X, these ratios equal the odds ratio:

$$\frac{\int_{\mathcal{U}_j} p(u)\, du}{\int_{\mathcal{U}_j} [1 - p(u)]\, du} = \frac{\mathbb{P}(X = 1)}{\mathbb{P}(X = 0)}$$

for all $j = 1, \ldots, J$.

Proof of corollary 9. The proportional balance condition (24) is equivalent to

$$\left(\int_{\mathcal{U}_j} p(u)\, du \right) \left(\int_{\mathcal{U}_k} [1 - p(u)]\, du \right) = \left(\int_{\mathcal{U}_j} [1 - p(u)]\, du \right) \left(\int_{\mathcal{U}_k} p(u)\, du \right).$$

Distributing terms yields

$$- \left(\int_{\mathcal{U}_j} p(u)\, du \right) \left(\int_{\mathcal{U}_k} p(u)\, du \right) + |\mathcal{U}_k| \int_{\mathcal{U}_j} p(u)\, du$$
$$= - \left(\int_{\mathcal{U}_j} p(u)\, du \right) \left(\int_{\mathcal{U}_k} p(u)\, du \right) + |\mathcal{U}_j| \int_{\mathcal{U}_k} p(u)\, du.$$

Hence equation (24) is equivalent to

$$\frac{1}{|\mathcal{U}_j|}\int_{\mathcal{U}_j} p(u)\ du = \frac{1}{|\mathcal{U}_k|}\int_{\mathcal{U}_k} p(u)\ du$$

for all $j, k = 1, \ldots, J$. The result then follows immediately from corollary 8. □

The following result shows that if p is consistent with $\mathcal{T}$-independence then we can shift and scale p to obtain new propensity scores that are also consistent with $\mathcal{T}$-independence. The first two propensity scores in figure 1 illustrate this result.

Corollary 10. Suppose p satisfies the average value condition (23). Define

$$\tilde{p}(u) = C_1 p(u) + C_2$$

where C_1 and C_2 are constants chosen such that $\tilde{p}(\cdot) \in [0, 1]$ and

$$\int_0^1 \tilde{p}(u)\ du = \mathbb{P}(X = 1).$$

Then $\tilde{p}$ also satisfies the average value condition (23).

Proof of corollary 10. First note that

$$\int_0^1 \tilde{p}(u)\ du = C_1 \mathbb{P}(X = 1) + C_2.$$

So enforcing the integration constraint

$$\int_0^1 \tilde{p}(u)\ du = \mathbb{P}(X = 1)$$

yields

$$C_1 \mathbb{P}(X = 1) + C_2 = \mathbb{P}(X = 1)$$

and hence

$$\frac{C_2}{1 - C_1} = \mathbb{P}(X = 1).$$

So suppose C_1 and C_2 satisfy this equation, and also that $\tilde{p}(\cdot) \in [0, 1]$. By assumption,

$$\frac{1}{|\mathcal{U}_j|}\int_{\mathcal{U}_j} p(u)\ du = \mathbb{P}(X = 1)$$

for all $j = 1, \ldots, J$. Consider

$$\begin{aligned}\frac{1}{|\mathcal{U}_j|}\int_{\mathcal{U}_j} \tilde{p}(u)\ du &= C_1 \frac{1}{|\mathcal{U}_j|}\int_{\mathcal{U}_j} p(u)\ du + C_2 \frac{1}{|\mathcal{U}_j|} \cdot |\mathcal{U}_j| \\ &= C_1 \mathbb{P}(X = 1) + C_2 \\ &= \mathbb{P}(X = 1).\end{aligned}$$

The last line follows by the integration constraint on the choice of C_1 and C_2. □

On the definitions of c- and d-independence

The following lemma shows that, for binary X, our definition of c-independence using $\mathbb{P}(X = 1 \mid U = u)$ is equivalent to a definition using $\mathbb{P}(X = 0 \mid U = u)$ instead.

Lemma 7. U is c-independent of X if and only if

$$\sup_{u \in [0,1]} |\mathbb{P}(X = 0 \mid U = u) - \mathbb{P}(X = 0)| \leq c.$$

Proof of lemma 7. We have

$$\begin{aligned} |\mathbb{P}(X = 1 \mid U = u) - \mathbb{P}(X = 1)| &= |(1 - \mathbb{P}(X = 0 \mid U = u)) - (1 - \mathbb{P}(X = 0))| \\ &= |\mathbb{P}(X = 0) - \mathbb{P}(X = 0 \mid U = u)|. \end{aligned}$$

□

Next we consider an alternative definition of d-independence. Let $\tilde{d} \in [0, 1]$. Say U is $\tilde{d}$-independent of X if

$$\sup_{u \in [0,1]} |F_{U|X}(u \mid 1) - F_{U|X}(u \mid 0)| \leq \tilde{d}.$$

This approach to measuring distance from independence is also used in Kline and Santos (2013). While d-independence compares the distributions of $U \mid X$ to the unconditional distribution of U, $\tilde{d}$-independence compares the distributions of $U \mid X$ with each other. The following lemma shows that $\tilde{d}$-independence implies our d-independence.

Lemma 8. Suppose U is $\tilde{d}$-independent of X for $\tilde{d} = e/\mathbb{P}(X = 0)\mathbb{P}(X = 1)$ where $e \in [0, \mathbb{P}(X = 0)\mathbb{P}(X = 1)]$. Then U is d-independent of X with

$$d(0) = \frac{e}{\mathbb{P}(X = 0)} \qquad \text{and} \qquad d(1) = \frac{e}{\mathbb{P}(X = 1)}.$$

Proof of lemma 8. By the law of total probability,

$$F_U(u) = F_{U|X}(u \mid 0)\mathbb{P}(X = 0) + F_{U|X}(u \mid 1)\mathbb{P}(X = 1).$$

Hence

$$\begin{aligned} \left|F_{U|X}(u \mid 1) - F_U(u)\right| &= \left|F_{U|X}(u \mid 1) - \left(F_{U|X}(u \mid 0)\mathbb{P}(X = 0) + F_{U|X}(u \mid 1)\mathbb{P}(X = 1)\right)\right| \\ &= \left|F_{U|X}(u \mid 1)[1 - \mathbb{P}(X = 1)] - F_{U|X}(u \mid 0)\mathbb{P}(X = 0)\right| \\ &= \left|F_{U|X}(u \mid 1)\mathbb{P}(X = 0) - F_{U|X}(u \mid 0)\mathbb{P}(X = 0)\right| \\ &= |F_{U|X}(u \mid 1) - F_{U|X}(u \mid 0)|\mathbb{P}(X = 0). \end{aligned}$$

Thus, by $\tilde{d}$-independence,

$$\begin{aligned} \sup_{u \in [0,1]} \left|F_{U|X}(u \mid 1) - F_U(u)\right| &\leq \tilde{d}\mathbb{P}(X = 0) \\ &= \frac{e}{\mathbb{P}(X = 1)} \end{aligned}$$

where the last line follows by our choice of $\tilde{d}$. Similar arguments show

$$\sup_{u\in[0,1]} \left|F_{U|X}(u \mid 0) - F_U(u)\right| \leq \frac{e}{\mathbb{P}(X=0)}.$$

□

$\mathcal{T}$-independence and continuous X

In this subsection we briefly consider the continuous X case for $\mathcal{T}$-independence.

Assumption A2′ (Continuous support). X has an everywhere continuous density $f_X(x)$ with respect to the Lebesgue measure.

In this case, the conditional distributions $\mathbb{P}(U \leq u \mid X = x)$ are not necessarily continuous in u. For example, if $U = X$ with probability 1,

$$\mathbb{P}(U \leq u \mid X = x) = \mathbb{1}[x \leq u],$$

a discontinuous function. Unlike when X is discrete, with continuous X the trivial bounds on the cdf are sharp when no assumptions are imposed on the dependence between U and X.

Proposition 6. Suppose A1 and A2′ hold. Then for any $x \in \text{supp}(X)$ the trivial bounds

$$\mathbb{1}[u \geq 1] \leq F_{U|X}(u \mid x) \leq \mathbb{1}[u \geq 0]$$

hold and are sharp for all $u \in [0,1]$.

Proof of proposition 6. The result follows by applying theorem 8 with $\mathcal{T} = \emptyset$. □

That is, without any assumptions, the cdfs $\mathbb{P}(U \leq u \mid X = x) = \mathbb{1}[u \geq 0]$ and $\mathbb{P}(U \leq u \mid X = x) = \mathbb{1}[u \geq 1]$ are both attainable. These cdfs correspond to point masses at 0 and 1, respectively, and are trivial bounds on the space of cdfs since no cdf can be larger than $\mathbb{1}[u \geq 0]$, or smaller than $\mathbb{1}[u \geq 1]$.

Next consider A3 and define the functions

$$\overline{h}^J(u) = \begin{cases} a_1 & \text{if } b_0 \leq u < a_1 \\ u & \text{if } a_1 \leq u \leq b_1 \\ a_2 & \text{if } b_1 < u < a_2 \\ u & \text{if } a_2 \leq u \leq b_2 \\ \vdots & \\ 1 & \text{if } b_J < u \leq a_{J+1} \end{cases} \quad \text{and} \quad \underline{h}^J(u) = \begin{cases} 0 & \text{if } b_0 \leq u < a_1 \\ u & \text{if } a_1 \leq u \leq b_1 \\ b_1 & \text{if } b_1 < u < a_2 \\ u & \text{if } a_2 \leq u \leq b_2 \\ \vdots & \\ 1 & \text{if } u = a_{J+1}. \end{cases} \tag{25}$$

While $\underline{h}^J$ is a valid cdf, $\overline{h}^J$ is not since it is not right-continuous at the points b_j, $j = 1, \ldots, J$. Nevertheless, there exists a sequence of valid cdfs which converges pointwise to the upper bound $\overline{h}^J$ all points. The next theorem shows that these are sharp bounds for cdfs under $\mathcal{T}$-independence.

Theorem 8. Suppose A1, A2′, and A3 hold. Let $\mathcal{F}$ be the set of all cdfs on $[0,1]$. Then for any $x \in \text{supp}(X)$,

$$\mathcal{F}^{\mathcal{T}}_{U|X}(x) = \left\{h \in \mathcal{F} : \underline{h}^J(u) \leq h(u) \leq \overline{h}^J(u) \text{ for all } u \in [0,1]\right\}.$$

For any $x \in \text{supp}(X)$, this set is sharp. Moreover, $\mathcal{F}^{\mathcal{T}}_{U|X}(x, x') = \mathcal{F}^{\mathcal{T}}_{U|X}(x) \times \mathcal{F}^{\mathcal{T}}_{U|X}(x')$, for any $x, x' \in \text{supp}(X)$.

These cdf bounds are step functions which correspond to distributions with point masses at the points in $\{\tau_0, \tau_1, \ldots, \tau_J\}$ for the upper bounds, and at $\{\tau_1, \ldots, \tau_J, \tau_{J+1}\}$ for the lower bound. Figure 13 plots an example of these bounds for $\mathcal{T} = \{\tau_1, \tau_2\}$ with just two points. In this example, $\underline{h}^J(u)$ places mass τ_1 at $u = \tau_1$, mass $\tau_2 - \tau_1$ at $u = \tau_2$ and mass $1 - \tau_2$ at $u = 1$.

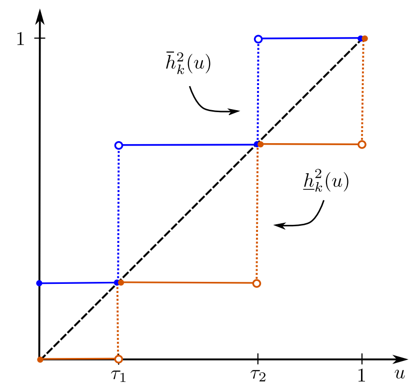


Figure 13: Upper and lower bounds on $h(u \mid x) = \mathbb{P}(U \le u \mid X = x)$ under the normalization A1, discrete support A2, and $\mathcal{T}$-independence with $\mathcal{T} = \{\tau_1, \tau_2\}$.

Proof of theorem 8. First we show that $\overline{h}^J(u)$ is an upper bound for cdfs that satisfy A3.

Case 1. Let $u^* \in (b_j, a_{j+1})$ for $j = 0, \ldots, J$, and suppose $h(\cdot)$ is a cdf such that

$$h(u^*) > \overline{h}^J(u^*).$$

Then

$$\begin{aligned} h(u^*) &> a_{j+1} \\ &= h(a_{j+1}) \end{aligned}$$

The second line follows by A3. This implies that $h(\cdot)$ is a decreasing function, a contradiction.

Case 2. Let $u^* \in [a_j, b_j]$ for $j = 1, \ldots, J$. Then

$$\begin{aligned} h(u^*) &= u^* \\ &= \overline{h}^J(u^*). \end{aligned}$$

The first line follows by A3. The second by definition of $\overline{h}^J$.

Therefore any cdf h that satisfies A3 must also satisfy $h(u) \le \overline{h}^J(u)$ for all u.

Now consider the lower bound $\underline{h}^J(u)$. Suppose there exists an h such that

$$h(u^*) < u^*$$

for some $u^* \in (b_j, a_{j+1})$, $j = 0, \dots, J$. Then

$$\begin{aligned} h(u^*) &< b_j \\ &= h(b_j). \end{aligned}$$

The second line follows by A3. This implies that h is decreasing, a contradiction. This implies that all h satisfying A3 satisfy $h(u) \geq \underline{h}^J(u)$ for all $u \in [0,1]$.

To show that $\mathcal{F}_{U|X}(x)$ is sharp, let $h(u \mid x)$ be an arbitrary element of $\mathcal{F}_{U|X}(x)$. Let

$$h(u \mid x') = \mathbb{P}(U \leq u \mid X = x') = u$$

for $x' \neq x$. Then $h(u \mid x')$ is a proper cdf that satisfies A3 and we can show that the marginal distribution of U follows that of assumption A1 since

$$\int_{\text{supp}(X)} \mathbb{P}(U \leq u \mid X = x) \, dF_X(x) = u.$$

Similarly, we can show that any two cdfs $h(\cdot \mid x) \in \mathcal{F}_{U|X}(x)$ and $h(\cdot \mid x') \in \mathcal{F}_{U|X}(x')$ can be attained by letting

$$\mathbb{P}(U \leq u \mid X = x^*) = u$$

for any $x^* \notin \{x, x'\}$. □

c-independence bounds for $\min\{p_0, p_1\} < c \leq \max\{p_0, p_1\}$

The following proposition extends theorem 3 to the case where $\min\{p_0, p_1\} < c \leq \max\{p_0, p_1\}$.

Proposition 7. Suppose A1 holds, X is binary, and U is c-independent of X for some $\min\{p_0, p_1\} < c \leq \max\{p_0, p_1\}$. Let

$$\overline{h}_1^c(u) = \begin{cases} \dfrac{u}{p_1} & \text{if } 0 \leq u \leq \dfrac{c}{p_0 + c} \\ \left(1 - \dfrac{c}{p_1}\right) u + \dfrac{c}{p_1} & \text{if } \dfrac{c}{p_0 + c} < u \leq 1, \end{cases} \qquad \underline{h}_1^c(u) = \begin{cases} \left(1 - \dfrac{c}{p_1}\right) u & \text{if } 0 \leq u \leq \dfrac{p_0}{p_0 + c} \\ \dfrac{u - p_0}{p_1} & \text{if } \dfrac{p_0}{p_0 + c} < u \leq 1 \end{cases}$$

and

$$\underline{h}_0^c(u) = \frac{u - p_1 \overline{h}_1^c(u)}{p_0} = \begin{cases} 0 & \text{if } 0 \leq u \leq \dfrac{c}{p_0 + c} \\ \left(1 + \dfrac{c}{p_0}\right) u - \dfrac{c}{p_0} & \text{if } \dfrac{c}{p_0 + c} < u \leq 1, \end{cases}$$

$$\overline{h}_0^c(u) = \frac{u - p_1 \underline{h}_1^c(u)}{p_0} = \begin{cases} \left(1 + \dfrac{c}{p_0}\right) u & \text{if } 0 \leq u \leq \dfrac{p_0}{p_0 + c} \\ 1 & \text{if } \dfrac{p_0}{p_0 + c} < u \leq 1. \end{cases}$$

Then

$$\mathcal{F}_{U|X}^c(k) = \left\{ h_k \in \mathcal{F}_k^c : \underline{h}_k^c(u) \leq h_k(u) \leq \overline{h}_k^c(u) \text{ for all } u \in [0,1] \right\}$$

for $k = 0, 1$, where $\mathcal{F}_k^c$ is defined as in theorem 3. For each $k = 0, 1$, this set is sharp.

Unlike the cdf bounds in theorem 3, the knot at which these cdf bounds change is no longer 1/2. Consequently, the propensity scores corresponding to the bounds no longer change their value at 1/2—they are asymmetric.

Proof of proposition 7. Assume $p_0 < c \leq p_1$ without loss of generality. To show these are the upper and lower bounds for $h_1(u)$, we first assume there exists u^* such that $h_1(u^*) > \overline{h}_1^c(u^*)$. If $u^* \in [0, c/(p_0+c)]$, then $h_1(u^*) > u^*/p_1$, the no-assumption bounds, hence we have a contradiction. If $u^* \in (c/(p_0+c), 1]$, then

$$\begin{aligned} h_1(u^*) &= \int_0^{u^*} \frac{p(u)}{p_1}\, du \\ &= 1 - \int_{u^*}^1 \frac{p(u)}{p_1}\, du \\ &\leq 1 - \int_{u^*}^1 \frac{p_1 - c}{p_1}\, du \\ &= \left(1 - \frac{c}{p_1}\right) u^* + \frac{c}{p_1} \\ &= \overline{h}_1^c(u^*), \end{aligned}$$

a contradiction. The same arguments can be used to show that $\underline{h}_1^c$ is the cdf lower bound for h_1, and similarly for the bounds for when $k = 0$. Sharpness can be established as in the proof of theorem 3. □

Bounds for the nonseparable regression model with continuous X

Corollary 11. Consider the outcome equation (14). Suppose A1 and A4 hold. Suppose U is $\mathcal{T}$-independent of X with $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$ and let $\tau_0 = 0$ and $\tau_{J+1} = 1$. Let $x \in \text{supp}(X)$ and suppose the continuous support assumption A2$'$ holds. Then

$$\text{ASF}(x) \in \left[\sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x), \quad \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^+_{Y|X}(\tau_{j+1} \mid x) \right].$$

Moreover, the interior of this set is sharp.

Proof of corollary 11. The bounds on the average structural functions are

$$\text{ASF}(x) \in \left[\int_0^1 Q^-_{Y|X}(\underline{h}^J(u) \mid x)\, du, \int_0^1 Q^+_{Y|X}(\overline{h}^J(u) \mid x)\, du \right],$$

where

$$\begin{aligned} \int_0^1 Q^-_{Y|X}(\underline{h}^J(u) \mid x)\, du &= \sum_{j=0}^{J} \int_{\tau_j}^{\tau_{j+1}} Q^-_{Y|X}(\tau_j \mid x)\, du \\ &= \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^-_{Y|X}(\tau_j \mid x) \end{aligned}$$

and

$$\begin{aligned}\int_0^1 Q^+_{Y|X}(\overline{h}^J(u) \mid x)\, du &= \sum_{j=0}^{J} \int_{\tau_j}^{\tau_{j+1}} Q^+_{Y|X}(\tau_{j+1} \mid x)\, du \\ &= \sum_{j=0}^{J} (\tau_{j+1} - \tau_j) Q^+_{Y|X}(\tau_{j+1} \mid x).\end{aligned}$$

The proof of sharpness follows as in the proof of corollary 4. □