Identifying Effects of Multivalued Treatments

Sokbae Lee[*] **Bernard Salanié**[†]

November 25, 2015

Abstract

Multivalued treatment models have only been studied so far under restrictive assumptions: ordered choice, or more recently unordered monotonicity. We show how marginal treatment effects can be identified in a more general class of models. Our results rely on two main assumptions: treatment assignment must be a measurable function of threshold-crossing rules; and enough continuous instruments must be available. On the other hand, we do not require any kind of monotonicity condition. We illustrate our approach on several commonly used models; and we also discuss the identification power of discrete instruments.

[*]Seoul National University.

[†]Columbia University, bsalanie@columbia.edu. We are grateful to Stéphane Bonhomme, Rodrigo Pinto, Christoph Rothe, Azeem Shaikh, Alex Torgovitsky and especially Jim Heckman for their useful comments.

1 Introduction

Since the seminal work of Heckman (1979), selection problems have been one of the main themes in both empirical economics and econometrics. One popular approach in the literature is to rely on instruments to uncover the patterns of the self-selection into different levels of treatments, and thereby to identify treatment effects. The main branches of this literature are the local average treatment effect (LATE) framework of Imbens and Angrist (1994) and the local instrumental variables (LIV) framework of Heckman and Vytlacil (2005).

The LATE and LIV frameworks emphasize different parameters of interest and suggest different estimation methods. However, they both focus on binary treatments, and restrict selection mechanisms to be "monotonic". Vytlacil (2002) establishes that the LATE and LIV approaches rely on the same monotonicity assumption. For binary treatment models, it requires that selection into treatment be governed by a single index crossing a threshold.

Many real-world selection problems are not adequately described by single-crossing models. The literature has developed ways of dealing with less restrictive models of asssignment to treatment. Angrist and Imbens (1995) analyze ordered choice models. Heckman, Urzua, and Vytlacil (2006, 2008) show how (depending on restrictions and instruments) a variety of treatment effects can be identified in discrete choice models; and more recently, Heckman and Pinto (2015) define an "unordered monotonicity" condition that is weaker than monotonicity for multivalued treatment. They show that given unordered monotonicity, several treatment effects can be identified.

The most generally applicable of these approaches still can only deal with models of treatment that are formally analogous to an additively separable discrete choice model,

as proved in section 6 of Heckman and Pinto (2015). But many transfer programs, for instance (or many tests in education) rely on several criteria and combine them in complex ways to assign agents to treatment; and agents add their own objectives and criteria to the list. A discrete choice model may then be too restrictive. To see this, consider a very simple and useful application: the double hurdle model with partial observability[1], which treats agents only if each of *two* indices passes a threshold. As we will see, it cannot be represented by a discrete choice model; it is formally equivalent to a discrete choice model with three alternatives in which the analyst only observes partitioned choices (e.g. the analyst only observes whether alternative 1 is chosen or not.) Our identification results apply to the double hurdle model. In fact, one way to describe our contribution is that it encompasses all additively separable discrete choice models in which the analyst only observes a partition of the set of alternatives.

Our analysis allows selection to be determined by a vector of unobservables, instead of a scalar random variable; and these unobservables can be correlated with potential outcomes. We rely on the control function approach, but we use a vector of control variables to deal with multidimensional unobserved heterogeneity. We establish conditions under which one can identify the probability distribution of unobservables governing the selection mechanism, as well as a generalized version of the marginal treatment effects (MTE) of Heckman and Vytlacil (2005). Furthermore, we clarify the extent to which a general form of selection equation is identified and we discuss a few applications to illustrate the usefulness of our approach.

We will give a detailed comparison between our paper and the existing literature in Section 6. Let us only mention at this stage a few points in which our paper differs from

[1]See, e.g. Poirier (1980) for a parametric version of this model.

the literature. Unlike Imbens (2000), Hirano and Imbens (2004), and Cattaneo (2010), we allow for selection on unobservables. We focus on the point identification of marginal treatment effects, unlike the research on partial identification (see e.g. Manski (1990), Manski (1997) and Manski and Pepper (2000)). Hoderlein and Mammen (2007), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), D'Haultfœuille and Février (2015), and Torgovitsky (2015) study models with continuous endogenous regressors. Each of these papers develops identification results for various parameters of interest. Our paper complements this literature by considering multivalued (but not continuous) treatments with more general types of selection mechanisms.

Heckman and Vytlacil (2007, Appendix B) and Heckman, Urzua, and Vytlacil (2008) and more recently Heckman and Pinto (2015) and Pinto (2015) are perhaps closest to our paper. But they focus on the selection induced by multinomial discrete choice models, whereas our paper allows for more general selection problems.

The paper is organized as follows. Section 2 sets up our framework; it motivates our central assumptions by way of examples. We present and prove our identification results in section 3. Section 4 discusses several important classes of applications, including the two-hurdle model mentioned in this introduction. Section 5 discusses the case in which instruments are discrete-valued, using the two-hurdle model as an illustration. Finally, we relate our contributions to the literature in section 6. Some details of the proofs are collected in an appendix.

2 The Model and our Assumptions

We assume throughout that treatments take values in a finite set of treatments $\mathcal{K}$. This set may be naturally ordered, as with different tax rates. But it may not be, as when welfare recipients enroll in different training schemes for instance; this makes no difference to our results. We assume that treatments are exclusive; this involves no loss of generality as treatment values could easily be redefined otherwise. We denote $K = |\mathcal{K}|$ the number of treatments, and we map the set $\mathcal{K}$ into $\{0, \ldots, K-1\}$ for notational convenience.

Potential outcomes $\{Y_k : k \in \mathcal{K}\}$ are generated by

$$Y_k = \mu_k(\boldsymbol{X}, U_k),$$

where $\boldsymbol{X}$ is a vector of covariates, U_k is an unobserved random variable, and $\mu_k(\cdot,\cdot)$ is an unknown function of $\boldsymbol{X}$ and U_k for each $k \in \mathcal{K}$. We denote $D_k = 1$ if the k treatment is realized and $D_k = 0$ otherwise. The observed outcome and treatment are $Y := \sum_{k\in\mathcal{K}} Y_k D_k$ and $D := \sum_{k\in\mathcal{K}} k D_k$, respectively.

In addition to the covariates $\boldsymbol{X}$, observed treatment D and outcomes Y, the data contain a random vector $\boldsymbol{Z}$ that will serve as instruments. We will always condition on the value of $\boldsymbol{X}$ in our analysis of identification; and we suppress it from the notation. Observed data consist of a sample $\{(Y_i, D_i, \boldsymbol{Z}_i) : i = 1, \ldots, N\}$ of $(Y, D, \boldsymbol{Z})$, where N is the sample size. We will denote the generalized propensity scores by $P_k(\boldsymbol{Z}) := \Pr(D = k|\boldsymbol{Z})$; they are directly identified from the data.

Let G denote a function defined on the support $\mathcal{Y}$ of Y. We focus on identification of $EG(Y_k)$. For example, if we take $G(Y_k) = Y_k$, then the object of interest is the mean

of the counterfactual outcome Y_k (conditional on the omitted covariates $\boldsymbol{X}$). Once we identify $EG(Y_k)$ for each k, we also identify the average treatment effect $E(G(Y_k) - G(Y_j))$ between any two treatments k and j. Alternatively, if we let $G(Y_k) = \mathbb{1}(Y_k \leq y)$ for some y, where $\mathbb{1}(\cdot)$ is the usual indicator function, then the object of interest is the marginal distribution of Y_k. This leads to the identification of quantile treatment effects.

One of our aims is to relax the usual monotonicity assumption that underlies LATE and LIV. Consider the following, simple example where $K = 3$, and treatment assignment is driven by a pair of random variables V_1 and V_2 whose marginal distributions are normalized to be $U[0, 1]$.

Example 1. Assume that there are two thresholds $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$ such that

- $D = 0$ iff $V_1 < Q_1(\boldsymbol{Z})$ and $V_2 < Q_2(\boldsymbol{Z})$,
- $D = 1$ iff $V_1 > Q_1(\boldsymbol{Z})$ and $V_2 > Q_2(\boldsymbol{Z})$,
- $D = 2$ iff $(V_1 - Q_1(\boldsymbol{Z}))$ and $(V_2 - Q_2(\boldsymbol{Z}))$ have opposite signs.

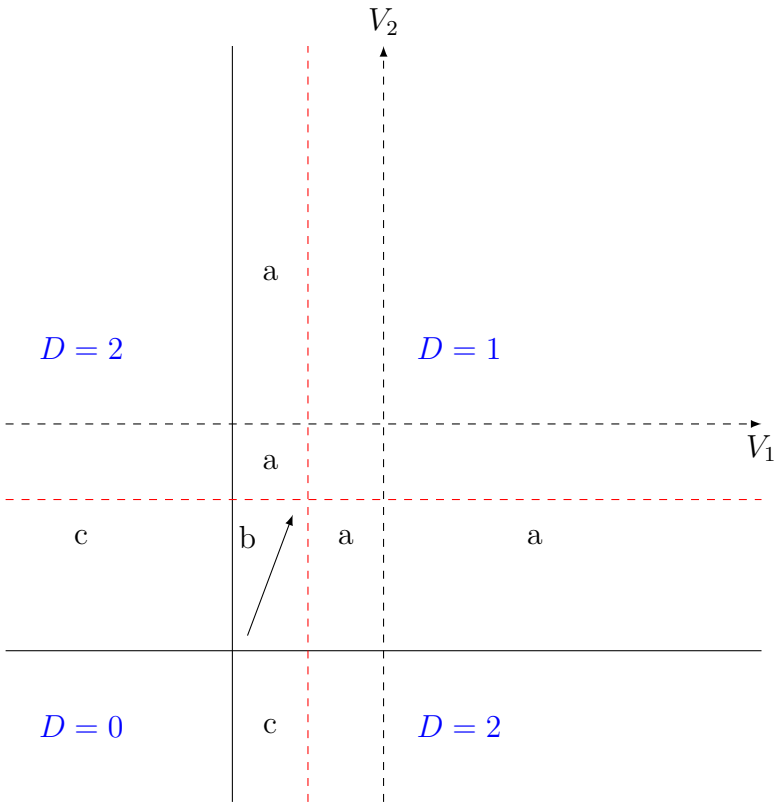
We could interpret Q_1 and Q_2 as minimum grades or scores in a two-part exam or an eligibility test based on two criteria: failing both parts/criteria assigns you to $D = 0$, passing both to $D = 1$, and failing only one to $D = 2$.

If F is the joint cdf of (V_1, V_2), it follows that the generalized propensity scores are

$$
\begin{aligned}
P_0(\boldsymbol{Z}) &= F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right), \\
P_1(\boldsymbol{Z}) &= 1 - Q_1(\boldsymbol{Z}) - Q_2(\boldsymbol{Z}) + F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right), \\
P_2(\boldsymbol{Z}) &= Q_1(\boldsymbol{Z}) + Q_2(\boldsymbol{Z}) - 2F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right).
\end{aligned}
$$

Take a change in the values of the instruments that increases both $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$, as represented in Figure 1: both criteria, or both parts of the exam, become more demanding. Then some observations (a) will move from $D = 1$ to $D = 2$, some (b) from $D = 1$ to $D = 0$, and some (c) will move from $D = 2$ to $D = 0$. This violates monotonicity, and even the weaker assumption that generalized propensity scores are monotonic in the instruments. Note also that some observations leave $D = 2$ and some move into $D = 2$: these *two-way flows* violate the even weaker requirement of unordered monotonicity of Heckman and Pinto (2015), which we describe in section 4.3. □

Figure 1: Example 1



To take a slightly more complicated example, consider the following entry game.

Example 2. Two firms $j = 1, 2$ are considering entry into a new market. Firm j has profit π_j^m if it becomes a monopoly, and $\pi_j^d < \pi_j^m$ if both firms enter. The static Nash equilibria are simple:

- if both $\pi_j^m < 0$, then no firm enters;
- if $\pi_j^m > 0$ and $\pi_k^m < 0$, then only firm j enters;
- if both $\pi_j^d > 0$, then both firms enter;
- if $\pi_j^d > 0$ and $\pi_k^d < 0$, then only firm j enters;
- if $\pi_j^m > 0 > \pi_j^d$ for both firms, then there are two symmetric equilibria, with only one firm operating.

Now let $\pi_j^m = V_j - Q_j(\boldsymbol{Z})$ and $\pi_j^d = \bar{V}_j - \bar{Q}_j(\boldsymbol{Z})$, and suppose we only observe the number $D = 0, 1, 2$ of entrants. Then

- $D = 0$ iff $V_1 < Q_1$ and $V_2 < Q_2$
- $D = 2$ iff $\bar{V}_1 > \bar{Q}_1$ and $\bar{V}_2 > \bar{Q}_2$
- $D = 1$ otherwise.

This is very similar to the structure of Example 1; in fact it coincides with it in the degenerate case when for each firm, π_m^j and π_d^j have the same sign with probability one. □

2.1 The Selection Mechanism

These two examples motivate the weak assumption we impose on the underlying selection mechanism. In the following we use $\boldsymbol{J}$ to denote the set $\{1, \ldots, J\}$.

Assumption 2.1 (Selection Mechanism)**.** *There exist a finite number J, a vector of unobserved random variables $\boldsymbol{V} := \{V_j : j \in \boldsymbol{J}\}$, and a vector of known functions $\{\boldsymbol{Q}_j(\boldsymbol{Z}) : j \in \boldsymbol{J}\}$ such that, equivalently:*

(i) the treatment variable D is measurable with respect to the σ-field generated by the events

$$E_j(\boldsymbol{Z}) := \{V_j < Q_j(\boldsymbol{Z})\} \text{ for } j \in \boldsymbol{J};$$

(ii) each event $\{D = k\} = \{D_k = 1\}$ is a member of this σ-field;

(iii) for each k, there exists a function g_k that is measurable with respect to this σ-field such that $D_k = 1$ iff $g_k(\boldsymbol{V}, \boldsymbol{Q}(Z)) = 0$.

Moreover, every treatment value k has positive probability for all $\boldsymbol{Z}$.

Note that the fact that every observation belongs to one and only one treatment group imposes further constraints; we will not need to spell them out at this stage, but we will show later how they can be used for overidentification tests.

In this notation, the validity of the instruments translates into:

Assumption 2.2 (Conditional Independence of Instruments)**.** *Y_k and $\boldsymbol{V}$ are independent of $\boldsymbol{Z}$ for each $k = 0, \ldots, K-1$.*

2.2 Atoms and Indices

To describe the class of selection mechanisms defined in Assumption 2.1 more concretely, we focus on a treatment value k. We define $S_j(\boldsymbol{Z}) := \mathbf{1}(V_j < Q_j(\boldsymbol{Z}))$ for $j = 1, \ldots, J$. Any element of the σ-field generated by the $\{E_j(\boldsymbol{Z}) : j = 1, \ldots, J\}$ can be written

uniquely as a finite union of the 2^J disjoint sets

$$F_1 \cap \ldots \cap F_J,$$

where F_j is either E_j or its complement $\bar{E}_j$. We will call them the *atoms* of the σ-field. Note that any such atom has an indicator function of the form

$$\prod_{j=1}^{J} T_j,$$

where T_j is either S_j (when $F_j = E_j$) or $(1 - S_j)$ (when $F_j = \bar{E}_j$). The event $\{D = k\}$ is a finite union of such atoms $n = 1, \ldots, N$; denote M_n the subset of indices $j = 1, \ldots, J$ for which $F_j = \bar{E}_j$ in atom n. Then the indicator function Π_n of atom n is

$$\Pi_n = \prod_{j \notin M_n} S_j \left(1 - \sum_{l \in M_n} S_l + \sum_{(l,m) \in M_n^2,\, l \neq m} S_l S_m + \ldots + (-1)^{p_n} \prod_{l \in M_n} S_l \right),$$

where $p_n = |M_n|$ and the dots represent Newton's symmetric polynomials over M_n of degrees between 2 and $(p_n - 1)$, with alternating coefficients ± 1. The highest degree term of the indicator function of this atom is

$$(-1)^{p_n} \prod_{j=1}^{J} S_j.$$

To illustrate, suppose that $J = 4$ and take atom n to be $E_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap E_4$: its M_n subset is $\{2, 3\}$, its $p_n = 2$, and its indicator function is

$$\Pi_n = S_1 S_4 (1 - S_2)(1 - S_3) = S_1 S_4 \left(1 - (S_2 + S_3) + S_2 S_3\right).$$

Now consider $\{D = k\}$ as the union of these N atoms. Since they are disjoint, its indicator function is simply the sum of their indicator functions. By construction, it is a multivariate polynomial in $(S_1, \ldots, S_J)$. Consider any subset $(j_1, \ldots, j_m)$ of indices in $\boldsymbol{J}$. Then it is easy to see that the coefficient of the product

$$\prod_{l=1,\ldots,m} S_{j_l}$$

in the indicator function of treatment value k is

$$\sum_{n=1}^{N} (-1)^{|(j_1,\ldots,j_m)\cap M_n|}.$$

The highest degree term of this polynomial will play a central role in our analysis. Note that if we choose $(j_1, \ldots, j_m) = \boldsymbol{J}$, then $(j_1, \ldots, j_m) \cap M_n = M_n$ for any atom n. It follows that the coefficient of the full product $\prod_{j=1}^{J} S_j$ is

$$a_k = \sum_{n=1}^{N} (-1)^{p_n}.$$

We call this number the *index* of treatment k. It can be any integer between $-N$ and N, including zero. To illustrate this, let us return to Example 1, with $J = 2$ and $K = 3$. For $D = 0$, the selection mechanism is described by the intersection $E_1(\boldsymbol{Z}) \cap E_2(\boldsymbol{Z})$. Hence, this case corresponds to $\Pi_{0,1}(\boldsymbol{Z}) = S_1(\boldsymbol{Z})S_2(\boldsymbol{Z})$. Similarly, $\Pi_{1,1}(\boldsymbol{Z}) = (1 - S_1(\boldsymbol{Z}))(1 - S_2(\boldsymbol{Z}))$. Finally, for $k = 2$ we have

$$\Pi_{2,1}(\boldsymbol{Z}) = S_1(\boldsymbol{Z})(1 - S_2(\boldsymbol{Z}))$$
$$\Pi_{2,2}(\boldsymbol{Z}) = (1 - S_1(\boldsymbol{Z}))S_2(\boldsymbol{Z}).$$

In this example the indices are $a_0 = a_1 = 1$ and $a_2 = (-1)^1 + (-1)^1 = -2$.

Appendix A.1 gives some results on indices. With $J = 2$ as in Example 1, the only treatments with a zero index are those which only depend on one threshold: e.g. $D_1 = \mathbb{1}(V_1 < Q_1)$. But for $J > 2$ it is not hard to generate cases in which a treatment value k depends on all J thresholds and still has $a_k = 0$, as shown in Example 3.

Example 3. Assume that $J = K = 3$ and take treatment 0 such that

$$\begin{aligned} D_0 = \mathbb{1}(V_1 < Q_1(\boldsymbol{Z}), V_2 < Q_2(\boldsymbol{Z}), V_3 < Q_3(\boldsymbol{Z})) \\ + \mathbb{1}(V_1 > Q_1(\boldsymbol{Z}), V_2 > Q_2(\boldsymbol{Z}), V_3 > Q_3(\boldsymbol{Z})). \end{aligned}$$

This has two atoms; the atom on the first line has $p_{0,1} = 0$, and the second one has $p_{0,2} = 3$. The index is $a_0 = 1 - 1 = 0$. Another way to see this is that the indicator function for $\{D_0 = 1\}$ is

$$S_1S_2S_3 + (1 - S_1)(1 - S_2)(1 - S_3) = 1 - S_1 - S_2 - S_3 + S_1S_2 + S_1S_3 + S_2S_3,$$

which has no degree three term. □

When the index is zero as in Example 3, the indicator function of the corresponding treatment k has degree strictly smaller than J. Since Assumption 2.1 rules out the uninteresting cases when treatment k has probability zero or one, its indicator function cannot be constant; and its leading terms have degree $m \geq 1$. We call m the *degree* of treatment k, and we summarize this discussion in a lemma:

Lemma 2.1. *Under Assumption 2.1, for each $k \in \mathcal{K}$ there exist an integer $N > 0$ such*

that for each $n = 1, \ldots, N$,

$$D_k = \sum_{n=1}^{N} \Pi_n(\boldsymbol{Z}),$$

with

$$\Pi_n(\boldsymbol{Z}) := \prod_{j \in \boldsymbol{J} - M_n} S_j(\boldsymbol{Z}) \prod_{l \in M_n} (1 - S_l(\boldsymbol{Z})).$$

The leading terms of the multivariate polynomial $D_k(\boldsymbol{S}) = \prod_{j \in \boldsymbol{J} - M_n} S_j \prod_{l \in M_n} (1 - S_l)$ *have degree* $m \geq 1$, *which we also call the* degree *of treatment* k.

Define $p_n \equiv |M_n|$, *and* $a_k \equiv \sum_{n=1}^{N} (-1)^{p_n}$ *the* index *of treatment* k. *Treatment* k *has degree* $m = J$ *if and only if* $a_k \neq 0$; *and then the leading term of* $D_k(\boldsymbol{S})$ *is*

$$a_k \prod_{j=1}^{J} S_j.$$

It is useful to think of atoms as alternatives in a discrete choice model. Any of the 2^J atoms can be interpreted as the choice of alternative n, where the binary representation of n has a one for digit j if $F_j = E_j$ and a zero if $F_j = \bar{E}_j$. The assignment of an observation to treatment k, which is a union of atoms, then is formally equivalent to the choice of an alternative whose number matches that of one of these atoms. In essence, we are dealing with discrete choice models with only partially observed choices.

3 Identification Results

In this section we fix $\boldsymbol{x}$ in the support of $\boldsymbol{X}$ and we suppress it from the notation. All the results obtained below are local to this choice of $\boldsymbol{x}$. Global (unconditional) identification results follow immediately if our assumptions hold for almost every $\boldsymbol{x}$ in the support of

$\boldsymbol{X}$.

We will treat separately the non-zero index and the zero index cases. We make this explicit in the following assumption.

Assumption 3.1 (Nonzero index)**.** *The index a_k defined in Lemma 2.1 is nonzero.*

We will return to zero-index treatments in section 3.2.

We require that $\boldsymbol{V}$ have full support:

Assumption 3.2 (Continuously Distributed Unobserved Heterogeneity in the Selection Mechanism)**.** *The joint distribution of $\boldsymbol{V}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^J$ and its support is $[0,1]^J$.*

Normalization: We normalize the marginal distribution of each $V_j \in \boldsymbol{V}$ to be $U[0,1]$.

Note that when $J = 1$, Assumptions 2.1 and 3.2 define the usual threshold-crossing model that underlies the LATE and LIV approaches. However, our assumptions allow for a much richer class of selection mechanisms when $J > 1$. Our Example 1 illustrates that our "multiple thresholds model" does not impose any multidimensional extension of the monotonicity condition that is implicit with a single threshold model. Even when $K = 2$ so that treatment is binary, J could be larger than one, allowing for flexible treatment assignment: just modify Example 1 to

$$D = \mathbb{1}\left(V_1 < Q_1(\boldsymbol{Z}) \text{ and } V_2 < Q_2(\boldsymbol{Z})\right).$$

Let $f_{\boldsymbol{V}}(\boldsymbol{v})$ denote the joint density function of $\boldsymbol{V}$ at $\boldsymbol{v} \in [0,1]^J$. Our identification argument relies on continuous instruments that generate enough variation in the

thresholds. This motivates the following three assumptions.

For any function ψ of $\boldsymbol{q}$, define "local equicontinuity at $\overline{\boldsymbol{q}}$" by the following property: for any subset $I \subset \boldsymbol{J}$, the family of functions $\boldsymbol{q}_I \mapsto \psi(\boldsymbol{q}_I, \boldsymbol{q}_{-I})$ indexed by $\boldsymbol{q}_{-I} \in [0,1]^{|\boldsymbol{J}-I|}$ is equicontinuous in a neighborhood of $\overline{\boldsymbol{q}}_I$.

Assumption 3.3 (Local equicontinuity at $\boldsymbol{q}$)**.** *The functions* $\boldsymbol{v} \mapsto f_{\boldsymbol{V}}(\boldsymbol{v})$ *and* $\boldsymbol{v} \mapsto E\left(G(Y_k)|\boldsymbol{V} = \boldsymbol{v}\right)$ *are locally equicontinuous at* $\boldsymbol{v} = \boldsymbol{q}$.

Assumption 3.3 will allow us to differentiate the relevant expectation terms. It is fairly weak: Lipschitz-continuity for instance implies local equicontinuity.

The next two assumptions apply to the functions $\boldsymbol{Q}(\boldsymbol{Z})$. These are unknown in most cases, and need to be identified; in this part of the paper we assume that they are known. We will return to identification of the $\boldsymbol{Q}$ functions in section 3.3.

Assumption 3.4 (Open Mapping at $\boldsymbol{q}$)**.** *The function* $\boldsymbol{Q}$ *is an open map at every point* $\boldsymbol{z}$ *such that* $\boldsymbol{Q}(\boldsymbol{z}) = \boldsymbol{q}$.

Assumption 3.4 requires that the image by $\boldsymbol{Q}$ of every small neighborhood of $\boldsymbol{z}$ contain a neighborhood of $\boldsymbol{Q}(\boldsymbol{z})$. It ensures that we can generate any small variation in $\boldsymbol{Q}(\boldsymbol{Z})$ by varying the instruments around $\boldsymbol{z}$. This makes the instruments strong enough to deal with multidimensional unobserved heterogeneity $\boldsymbol{V}$. It is crucial to our approach. In Example 1 for instance, it would fail if Q_1 and Q_2 were functionally dependent around $\boldsymbol{z}$, with say $Q_1 \equiv Q_2^3$. More generally, Assumption 3.4 ensures both that there are "enough instruments" and that they have enough variation locally[2]. In its absence, we would only get partial identification of the marginal treatment effects.

[2]Note that it does not require a rank condition at $\boldsymbol{z}$. If $\boldsymbol{Q}$ has a Jacobian, this could have reduced rank at $\boldsymbol{z}$ as long as it has full row rank in small neighborhood of $\boldsymbol{z}$—as it must if $\boldsymbol{Q}$ is an open map at $\boldsymbol{z}$. Since critical points of non-constant maps are typically isolated, this is a much weaker requirement.

We also consider a global version of Assumption 3.4.

Assumption 3.5 (Global Condition)**.** *Assumption 3.4 holds at all* $\boldsymbol{q} \in (0,1)^J$.

3.1 Identification with a non-zero index

We are now ready to prove identification of $EG(Y_k)$ when treatment k has a non-zero index. In the following theorem, for any real-valued function $\boldsymbol{q} \mapsto h(\boldsymbol{q})$, the notation

$$Th(\boldsymbol{q}) \equiv \frac{\partial^J h}{\prod_{j=1}^J \partial q_j}(\boldsymbol{q})$$

refers to the J-order derivative that obtains by taking derivatives of the function h at $\boldsymbol{q}$ in each direction of $\boldsymbol{J}$ in turn.

Theorem 3.1 (Identification with a non-zero index)**.** *Let Assumptions 2.1, 2.2, 3.1, and 3.2 hold. Fix a value* $\boldsymbol{q}$ *in the support of* $\boldsymbol{Q}(\boldsymbol{Z})$ *and assume that 3.3 and 3.4 hold at* $\boldsymbol{q}$*. Then the density of* $\boldsymbol{V}$ *and conditional expectation of* $G(Y_k)$ *are given by*

$$f_{\boldsymbol{V}}(\boldsymbol{q}) = \frac{1}{a_k} T \Pr(D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$$

$$E[G(Y_k)|\boldsymbol{V} = \boldsymbol{q}] = \frac{TE\left(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right)}{T \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})}.$$

If in addition Assumption 3.5 holds, then

$$EG(Y_k) = \frac{1}{a_k} \int_{[0,1]^J} TE\left(G(Y)D_k|\boldsymbol{Q} = \boldsymbol{q}\right) d\boldsymbol{q}. \tag{3.1}$$

Proof of Theorem 3.1. Our proof has three steps. We first write conditional moments as integrals with respect to indicator functions. Then we show that these integrals are

differentiable and we compute their multidimensional derivatives. Finally, we impose Assumption 3.1 and we derive the equalities in the theorem.

Step 1:

Under the assumptions imposed in the theorem, for any $\boldsymbol{q}$ in the range of $\boldsymbol{Q}$,

$$
\begin{aligned}
&E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \\
&= E[G(Y_k)|D = k, \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}) \\
&= E[G(Y_k)|g_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = 0, \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \Pr(g_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = 0|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}) \\
&= E[G(Y_k)|g_k(\boldsymbol{V}, \boldsymbol{q}) = 0] \Pr\left(g_k(\boldsymbol{V}, \boldsymbol{q}) = 0\right) \\
&= E[G(Y_k)\mathbb{1}\left(g_k(\boldsymbol{V}, \boldsymbol{q}) = 0\right)] \\
&= E\left(E[G(Y_k)\mathbb{1}\left(g_k(\boldsymbol{V}, \boldsymbol{q}) = 0\right)|\boldsymbol{V}\right) \\
&= E\left(E[G(Y_k)|\boldsymbol{V}]\mathbb{1}\left(g_k(\boldsymbol{V}, \boldsymbol{q}) = 0\right)\right),
\end{aligned}
$$

where the third equality follows from Assumption 2.2 and the others are obvious. As a consequence,

$$
\begin{aligned}
&E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \\
&= \int \mathbb{1}\left(g_k(\boldsymbol{v}, \boldsymbol{q}) = 0\right) E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}. \qquad (3.2)
\end{aligned}
$$

Let $b_k(\boldsymbol{v}) \equiv E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$. Then (3.2) takes the form

$$
B_k(\boldsymbol{q}) = \int \mathbb{1}(g_k(\boldsymbol{v}, \boldsymbol{q}) = 0) b_k(\boldsymbol{v}) d\boldsymbol{v}.
$$

Now remember from Lemma 2.1 that the indicator function of $D = k$ is a multivariate

polynomial of the indicator functions S_j for $j \in \boldsymbol{J}$. Moreover, $S_j(\boldsymbol{Z}) = \mathbb{1}(V_j < Q_j(\boldsymbol{Z})) = H(Q_j(\boldsymbol{Z}) - V_j)$, where $H(t) = \mathbb{1}(t > 0)$ is the one-dimensional Heaviside function. Therefore we can rewrite the selection of treatment k as

$$\mathbb{1}(g_k(\boldsymbol{v}, \boldsymbol{q}) = 0) = \sum_{n=1}^{N} \prod_{j \in \boldsymbol{J} - M_n} H(q_j - v_j) \prod_{j \in M_n} (1 - H(q_j - v_j)), \tag{3.3}$$

and it follows that

$$B_k(\boldsymbol{q}) = \sum_{n=1}^{N} \int \left(\prod_{j \in \boldsymbol{J} - M_n} H(q_j - v_j) \prod_{j \in M_n} (1 - H(q_j - v_j)) \right) b_k(\boldsymbol{v}) d\boldsymbol{v}.$$

Expanding the products, the right-hand side can be written as a sum

$$B_k(\boldsymbol{q}) = \sum_{l} c_l \int \prod_{j \in I_l} H(q_j - v_j) b_k(\boldsymbol{v}) d\boldsymbol{v}, \tag{3.4}$$

where for each l, the set $I_l \subset \boldsymbol{J}$ and c_l is an algebraic integer.

Step 2:

By Assumption 3.3, the function $\boldsymbol{b}$ is locally equicontinuous. This implies that all terms in (3.4) are differentiable along all dimensions of $\boldsymbol{q}$. To see this, start with dimension $j = 1$. Any term l in (3.4) such that I_l does not contain 1 is constant in q_1 and obviously differentiable. Take any other term and rewrite it as

$$A_l(q_1) \equiv c_l \int_0^{q_1} \int \prod_{j \in I_l, j \neq 1} H(q_j - v_j) b_k(v_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1} dv_1,$$

where $\boldsymbol{v}_{-1}$ collects all directions of $\boldsymbol{v}$ in $I_l - \{1\}$.

Then for any $\varepsilon \neq 0$,

$$\begin{aligned}\frac{A_l(q_1+\varepsilon)-A_l(q_1)}{\varepsilon} - c_l \int \prod_{j\in I_l, j\neq 1} H(q_j - v_j) b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1} \\ = \frac{c_l}{\varepsilon}\int_{q_1}^{q_1+\varepsilon} \int \prod_{j\in I_l, j\neq 1} H(q_j - v_j)\left(b_k(v_1, \boldsymbol{v}_{-1}) - b_k(q_1, \boldsymbol{v}_{-1})\right) d\boldsymbol{v}_{-1} dv_1.\end{aligned}$$

Since the functions $(b_k(\cdot, \boldsymbol{v}_{-1}))$ are locally equicontinuous at q_1, for any $\eta > 0$ we can choose ε such that if $|q_1 - v_1| < \varepsilon$,

$$|b_k(q_1, \boldsymbol{v}_{-1}) - b_k(v_1, \boldsymbol{v}_{-1})| < \eta;$$

and since the Heaviside functions are bounded above by one, we will have

$$\left|\frac{A_l(q_1+\varepsilon)-A_l(q_1)}{\varepsilon} - c_l \int \prod_{j\in I_l, j\neq 1} H(q_j - v_j) b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1}\right| < |c_l|\,\eta.$$

This proves that A_l is differentiable in q_1 and that its derivative with respect to q_1, which we denote A_l^1, is

$$A_l^1 = c_l \int \prod_{j\in I_l, j\neq 1} H(q_j - v_j)\; b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1}.$$

But this derivative itself has the same form as A. Letting $\boldsymbol{v}_{-1,2}$ collect all components of $\boldsymbol{v}$ except (q_1, q_2), the same argument would prove that since the functions $(b_k(\cdot, \boldsymbol{v}_{-1,2}))$ are locally equicontinuous at (q_1, q_2), the function A_l^1 is differentiable with respect to q_2

and its derivative is

$$c_l \int \prod_{j \in I_l, j \neq 1,2} H(q_j - v_j)\; b_k(q_1, q_2, \boldsymbol{v}_{-1,2}) d\boldsymbol{v}_{-1,2}.$$

Continuing this argument finally gives us the cross-derivative with respect to $(\boldsymbol{q}^{I_l})$ as

$$c_l \int b_k(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l},$$

where $\boldsymbol{v}_{-I_l}$ collects all components of $\boldsymbol{v}$ whose indices are not in I_l.

Step 3:

Lemma 2.1 and Assumption 3.1 also imply that the leading term in the H_j's is

$$a_k \prod_{j=1}^{J} H(q_j - p_j).$$

Now take the J-order derivative of $B(\boldsymbol{q})$ with respect to all q_j in turn. By Lemma 2.1, the highest-degree term of B in $\boldsymbol{q}$ is

$$a_k \int \left(\prod_{j=1}^{J} H(q_j - v_j) \right) b_k(\boldsymbol{v}) d\boldsymbol{v}$$

as $a_k \neq 0$ under Assumption 3.1; all other terms involve a smaller number of indices j.

This term contributes a cross-derivative

$$a_k b_k(\boldsymbol{q}),$$

and all other terms generate null contributions since each of them is constant in at least

one of the directions j.

More formally,

$$TB_k(\boldsymbol{q}) = \frac{\partial^J B_k(\boldsymbol{q})}{\prod_{j \in \boldsymbol{J}} \partial q_j} = a_k b_k(\boldsymbol{q}). \tag{3.5}$$

Given Assumption 3.3, equation (3.5) also applies to the pair of functions

$$\bar{B}_k(\boldsymbol{q}) = \Pr[D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \text{ with } \bar{b}_k(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v}).$$

This gives the first equality in the theorem. To obtain the second equality, we use

$$\tilde{B}_k(\boldsymbol{q}) = E[G(Y) D_k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \text{ and } \tilde{b}_k(\boldsymbol{v}) = E[G(Y_k) | \boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v}),$$

which again is locally equicontinuous by Assumption 3.3.

Under Assumption 3.5, the final conclusion of the theorem follows by using

$$EG(Y_k) = \int E\left(G(Y_k) | \boldsymbol{V} = \boldsymbol{v}\right) f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}.$$

□

It follows from Theorem 3.1 that if k and k' are two treatments to which all of our assumptions apply, then we can identify the average treatment effect, as well as the marginal treatment effect and the quantile treatment effect of moving between these two treatments.

To identify the average treatment effect, we need the full support condition in Assumption 3.5. This is a stringent assumption that may not hold in many applications.

In such cases we can extend Carneiro, Heckman, and Vytlacil (2010) to identify the marginal policy relevant treatment effect (MPRTE) and the average marginal treatment effect (AMTE). The MPRTE is a marginal version of the policy relevant treatment effect (PRTE) of Heckman and Vytlacil (2001), which measures the average effect of moving from a baseline policy to an alternative policy. The AMTE is the average benefit of treatment for people at the margin of indifference between participation in treatment and nonparticipation. We could obtain identification results for a generalized version of the MPRTE by specifying marginal changes for our selection mechanism. In applications, the analyst may not have continuous instruments to identify even the MRPTE or AMTE. We consider the case of discrete instruments in section 5.

3.2 Identification with a zero index

Theorem 3.1 required that the index of treatment k be non-zero (Assumption 3.1). Therefore it does not apply to Example 3 for instance. Recall that in that example,

$$D_0 = 1 - S_1 - S_2 - S_3 + S_1S_2 + S_1S_3 + S_2S_3$$

and treatment 0 has degree $m^0 = 2 < J^0 = 3$.

Note, however, that steps 1 and 2 of the proof of Theorem 3.1 apply to zero-index treatments as well; the relevant polynomial of Heaviside functions has leading term

$$H(q_1 - v_1)H(q_2 - v_2) + H(q_1 - v_1)H(q_3 - v_3) + H(q_2 - v_2)H(q_3 - v_3),$$

and we can take the derivative in (q_1, q_2) for instance to obtain an equation that replaces

(3.5):

$$\frac{\partial^2}{\partial q_1 \partial q_2} B_0(\boldsymbol{q}) = \int b_0(q_1, q_2, v_3) dv_3.$$

Applying this to $B_0(\boldsymbol{q}) = \Pr[D = 0 | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ and $b_0(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v})$, and then to $B_0(\boldsymbol{q}) = E[Y D_0 | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ and $b_0(\boldsymbol{v}) = E[G(Y_0) | \boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})$, identifies

$$\int f_{V_1, V_2, V_3}(q_1, q_2, v_3) dv_3 = f_{V_1, V_2}(v_1, v_2)$$

and

$$\int E[G(Y_0) | V_1 = q_1, V_2 = q_2, V_3 = q_3] f_{V_1, V_2, V_3}(q_1, q_2, v_3) dv_3 = E[G(Y_0) \mathbb{1}(V_1 = q_1, V_2 = q_2)];$$

and dividing through identifies a sort of local counterfactual outcome:

$$E[G(Y_0) | V_1 = q_1, V_2 = q_2].$$

Under assumption 3.5, this also identifies $EG(Y_0)$. Moreover, we can apply the same logic to the pairs (q_1, q_3) and (q_2, q_3) to get further information on the treatment effects.

This argument is quite general. It allows us to state the following theorem:

Theorem 3.2 (Identification with a zero index)**.** *Let Assumptions 2.1, 2.2, and 3.2 hold. Fix a value $\boldsymbol{q}$ in the support of $\boldsymbol{Q}(\boldsymbol{Z})$ and assume that 3.3 and 3.4 hold at $\boldsymbol{q}$. Let m be the degree of treatment k, and $c_l \prod_{i=1,\ldots,m} S_{j_i}$ be any of the leading terms of the*

indicator function of $\{D = k\}$. *Denote* $I = \{j_1, \ldots, j_m\}$, *and* $\widetilde{T}$ *the differential operator*

$$\widetilde{T} = \frac{\partial^m}{\prod_{i=1,\ldots,m} \partial_{j_i}}.$$

Then for $\boldsymbol{q} = (\boldsymbol{q}^I, \boldsymbol{q}^{\boldsymbol{J}-I})$,

$$f_{\boldsymbol{V}^I}(\boldsymbol{q}^I) = \frac{1}{c_l} \widetilde{T} \Pr[D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$$

$$E[G(Y_k) | \boldsymbol{V}^I = \boldsymbol{q}^I] = \frac{\widetilde{T}\, E[G(Y) D_k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]}{\widetilde{T} \Pr[D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]}.$$

If in addition Assumption 3.5 holds, then

$$EG(Y_k) = \frac{1}{c_l} \int_{[0,1]^J} \widetilde{T}\, E[G(Y) D_k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}) d\boldsymbol{q}.$$

Proof of Theorem 3.2. The proof of Theorem 3.2 is basically the same as that of Theorem 3.1; it is included in Appendix A.2. □

Theorem 3.2 is a generalization of Theorem 3.1 (just take $m = J$). It calls for three remarks. First, we could weaken its hypotheses somewhat. We could for instance replace $(0,1)^J$ with $(0,1)^{|I|}$ in the statement of Assumption 3.5.

Second, when $m < J$ the treatment effects are overidentified. This is obvious from the equalities in Theorem 3.2, in which the right-hand side depends on $\boldsymbol{q}$ but the left-hand side only depends on $\boldsymbol{q}^I$.

Finally, considering several treatment values can identify even more, since $\boldsymbol{V}$ is assumed to be the same across k. Theorem 3.1 implies for instance that if there is any treatment value k with a nonzero index, then the joint density $f_{\boldsymbol{V}}$ is identified from that

treatment value.

3.3 Identification of Q

So far we assumed that the functions $\{\boldsymbol{Q}_j(\boldsymbol{Z}) : j = 1, \ldots, J\}$ were known (see Assumption 2.1). In practice we will most often need to identify them from the data before applying Theorems 3.1 or 3.2. The most natural way to do so starts from the generalized propensity scores $\{P_j(\boldsymbol{Z}) : j = 1, \ldots, J\}$, which are identified as the conditional probabilities of treatment[3].

First note that by definition (and by Assumption 2.2),

$$\begin{aligned} P_k(\boldsymbol{z}) &= \Pr(D = k | \boldsymbol{Z} = \boldsymbol{z}) \\ &= \int \mathbb{1}\left(g_k\left(\boldsymbol{v}, \boldsymbol{Q}(\boldsymbol{z})\right) = 0\right) f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}. \end{aligned}$$

Note that this is a J-index model. Ichimura and Lee (1991) consider identification of multiple index models with the indices are specified parametrically. Matzkin (1993, 2007) obtains nonparametric identification results for discrete choice models[4]; we build on her results in section 4.5 to obtain the identification of $\boldsymbol{Q}$ for multiple hurdle models. Matzkin's results only apply to a subset of the types of selection mechanisms we consider (discrete choice models when all choices are observed). Section 4 discusses identification of the $\boldsymbol{Q}$'s in some specific models in more detail.

[3]It would also be possible to seek identification jointly from the generalized propensity scores and from the cross-derivatives that appear in Theorems 3.1 or 3.2, especially when they are over-identified. We do not pursue this here.

[4]See Heckman and Vytlacil (2007, Appendix B) for an application to treatment models.

4 Applications

4.1 Monotone Treatment

Angrist and Imbens (1995) consider two-stage least-squares estimation of a model in which the ordered treatment takes a finite number of values, and a discrete-valued instrument is available. Let $z = 0, \ldots, M-1$ be the possible values of the instrument, ordered so that $E(D|Z=z)$ increases with z; and $D = 0, \ldots, K-1$. Angrist and Imbens show that the TSLS estimator obtained by regressing outcome Y on a preestimated $E(D|Z)$ converges to $\beta_{TSLS} \equiv \sum_{m=1}^{M-1} \mu_m \beta_m$, where β_m's are called the *average causal responses*, defined by

$$\beta_m \equiv \frac{E(Y|Z=m) - E(Y|Z=m-1)}{E(D|Z=m) - E(D|Z=m-1)}$$

for $m = 1, \ldots, M-1$, and the family of weights $\{\mu_m\}_{m=1}^{M-1}$ is given by the joint distribution of D and Z.

The average causal response β_m itself can only be interpreted as causal under a stronger monotonicity assumption. Denote D_z the counterfactual treatment for $Z = z$, and assume that $D_m \geq D_{m-1}$ with probability one. Angrist and Imbens (1995) prove that under these assumptions, β_m is a weighted average of the effects of treatment on the various groups of compliers:

$$\beta_m = \sum_{k=1}^{K} \omega_k E\left(Y_k - Y_{k-1} | D_m \geq k > D_{m-1}\right).$$

Here, the weights (ω_k) are given by the joint distribution of D_{m-1} and D_m, and they can be estimated under the monotonicity assumption. On the other hand, the individual

terms

$$E\left(Y_k - Y_{k-1} | D_m \geq k > D_{m-1}\right)$$

cannot be identified; only their weighted average β_{TSLS} is.

Given the results in Vytlacil (2002), the monotonicity assumption is essentially equivalent to the existence of a family of threshold crossing rules. Take a family of thresholds $(Q_1(\boldsymbol{Z}), \ldots, Q_{K-1}(\boldsymbol{Z}))$ and unobserved random variables $(V_1, \ldots, V_{K-1})$ whose marginal distributions are $U[0,1]$. First, assign $D = 0$ when $V_1 > Q_1(\boldsymbol{Z})$; and for every $k = 1, \ldots, K-1$ and given that $D \geq k-1$, let the model assign $D \geq k$ if and only if

$$V_k < Q_k(\boldsymbol{Z}).$$

This generates a model of treatment that satisfies our Assumption 2.1. It has $J = K-1$ and a very specific structure:

$$D = \arg\min\{k = 0, \ldots, K-2 \mid V_{k+1} > Q_{k+1}(\boldsymbol{Z})\},$$

with $D = K-1$ if $V_k < Q_k(\boldsymbol{Z})$ for all $k = 1, \ldots, K-1$.

Remark 4.1. Note that the conventional ordered choice model only uses a common scalar random variable v, which we can normalize to be $U[0,1]$: for $k = 0, \ldots, K-1$,

$$D_k = 1 \text{ iff } F_k(\boldsymbol{Z}) < v < F_{k+1}(\boldsymbol{Z}),$$

with $F_0 \equiv 0$ and $F_K \equiv 1$. This model of assignment to treatment is observationally equivalent to ours, provided that the probabilities of treatment $P_k = F_k - F_{k-1}$ coincide. If the functions Q_k are unconstrained, there are many ways to achieve this. We could

for instance impose independence of the V_k's and define the Q_k functions recursively by $1 - Q_1(\boldsymbol{Z}) = F_1(\boldsymbol{Z})$ and

$$(1 - Q_k(\boldsymbol{Z}))(1 - F_{k-1}(\boldsymbol{Z})) = F_k(\boldsymbol{Z}) - F_{k-1}(\boldsymbol{Z}).$$

for $k = 2, \ldots, K-1$. $\square$

Going back to the original nonparametric model, the thresholds are easily identified from

$$Q_k(\boldsymbol{Z}) = \Pr(D \geq k | D \geq k-1, \boldsymbol{Z}).$$

Each treatment value k is defined by k atoms E_j (for $j = 1, \ldots, k-1$) and one event $\bar{E}_{k+1}$, with the exceptions of $k = 0$ which only has atom $\bar{E}_1$ and $k = K-1$ which has atoms E_j for $j = 1, \ldots, K-1$. Therefore only treatment values $(K-2)$ and $(K-1)$ have a nonzero index, with $a_{K-2} = -1$ and $a_{K-1} = 1$. Treatment values $k = 0, \ldots, K-2$ (if $K > 2$) have $a_k = 0$ and their leading coefficient is $c_l = -1$.

To apply Theorems 3.1 and Theorems 3.2, we need to deviate from Angrist and Imbens (1995) and assume the existence of enough continuous instruments $\boldsymbol{Z}$. Using the generic notation $\boldsymbol{x}^n = (x_1, \ldots, x_n)$, we then obtain a series of formulæ for $k = 0, \ldots, K-2$ and all $\boldsymbol{v} \in (0,1)^{K-1}$:

$$f_{\boldsymbol{V}^{k+1}}(\boldsymbol{v}^{k+1}) = -\frac{\partial^{k+1}}{\partial q_1 \ldots \partial q_{k+1}} \Pr(D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{v})$$

$$E(Y_k | \boldsymbol{V}^{k+1} = \boldsymbol{v}^{k+1}) f_{\boldsymbol{V}^{k+1}}(\boldsymbol{v}^{k+1}) = -\frac{\partial^{k+1}}{\partial q_1 \ldots \partial q_{k+1}} E(Y D_k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{v})$$

along with two slightly different formulæ for $k = K - 1$:

$$f_{\boldsymbol{V}}(\boldsymbol{v}) = \frac{\partial^{K-1}}{\partial q_1 \dots \partial q_{K-1}} \Pr(D = K-1 | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{v})$$
$$E(Y_{K-1} | \boldsymbol{V} = \boldsymbol{v}) f_{\boldsymbol{V}}(\boldsymbol{v}) = \frac{\partial^{K-1}}{\partial q_1 \dots \partial q_{K-1}} E(Y D_{K-1} | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{v}).$$

These formulæ can be used to estimate marginal treatment effects, and to run overidentifying tests.

Now take for instance the unconditional average treatment effect of moving to treatment value $(K-1)$ from treatment value $(K-2)$. Assume that $\boldsymbol{Z}$ contains at least $(K-1)$ continuous instruments that generate full support variation in $\boldsymbol{Q}(\boldsymbol{Z})$. Then by integrating we obtain

$$E(Y_{K-1} - Y_{K-2}) = \int_{(0,1)^{K-1}} \frac{\partial^{K-1}}{\partial q_1 \dots \partial q_{K-1}} E\left(Y(D_{K-1} + D_{K-2}) | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right) d\boldsymbol{q}.$$

For the binary treatment model $K = 2$, this is simply

$$E(Y_1 - Y_0) = \int_0^1 \frac{\partial E(Y | Q(\boldsymbol{Z}) = q)}{\partial q} dq$$

since $D_0 + D_1 = 1$. This is the standard formula that derives the ATE from the MTE .

This illustrates the trade-offs in our approach. We get much more identifying power than the standard method; but this comes at the price of requiring multivariate continuous variation in instruments. On the other hand, we do not need to assume monotonicity: each treatment value can be any measurable function of the events $E_l = \{V_l < Q_l(\boldsymbol{Z})\}$, as long as the thresholds (Q_l) are identified from the data—as they clearly are here.

Heckman, Urzua, and Vytlacil (2006, 2008) go beyond Angrist and Imbens (1995)

by showing how the 2SLS estimate can be reinterpreted in more transparent ways in the MTE framework. They also analyze a family of discrete choice models, to which we now turn.

4.2 Discrete Choice Models

Heckman, Urzua, and Vytlacil (2008, see also Heckman and Vytlacil (2007)) consider a multinomial discrete choice model of treatment. They posit

$$D = k \iff R_k(\boldsymbol{Z}) - U_k > R_l(\boldsymbol{Z}) - U_l \text{ for } l = 0, \ldots, K-1 \text{ such that } l \neq k,$$

where the U's are continuously distributed and independent of $\boldsymbol{Z}$.

Define

$$\boldsymbol{R}(\boldsymbol{Z}) = (R_k(\boldsymbol{Z}) - R_l(\boldsymbol{Z}))_{l \neq k} \quad \text{and} \quad \boldsymbol{U} = (U_k - U_l)_{l \neq k}\,.$$

Then $D_k = \mathbb{1}(\boldsymbol{R}(\boldsymbol{Z}) > \boldsymbol{U})$; and defining $\boldsymbol{Q}_l(\boldsymbol{Z}) = \Pr(\boldsymbol{U}_l < \boldsymbol{R}_l(\boldsymbol{Z}))$ allows us to write the treatment model as

$$D = k \text{ iff } \boldsymbol{V} < \boldsymbol{Q}(\boldsymbol{Z}),$$

where each V_l is distributed as $U[0,1]$.

Heckman, Urzua, and Vytlacil (2008) then study the identification of marginal and local average treatment effects under assumptions that are similar to ours: continuous instruments that generate enough dimensions of variation in the thresholds.

As they note, the discrete choice model with an additive structure implicitly imposes monotonicity, in the following form: if the instruments $\boldsymbol{Z}$ change in a way that increases $R_k(\boldsymbol{Z})$ relative to all other $R_l(\boldsymbol{Z})$, then no observation with treatment value k will be

assigned to a different treatment. In our notation, D_k is an increasing function of $\boldsymbol{Q}(\boldsymbol{Z})$. We make no such assumption, as Example 1 and Figure 1 illustrate. Our results extend those of Heckman, Urzua, and Vytlacil (2008) to any model with identified thresholds.

4.3 Unordered Monotonicity

In an important recent paper, Heckman and Pinto (2015) introduce a new concept of monotonicity. Their "unordered monotonicity" assumption can be rephrased in our notation in the following way. Take two values $\boldsymbol{z}$ and $\boldsymbol{z}'$ of the instruments $\boldsymbol{Z}$. We want to study the treatment effect of moving from k to k' by exploiting the change of instruments from $\boldsymbol{z}$ to $\boldsymbol{z}'$.

Assumption 4.1 (Unordered Monotonicity)**.** *Denote $D_{\boldsymbol{z}}$ and $D_{\boldsymbol{z}'}$ the counterfactual treatments. Then for $l = k, k'$, there cannot be two-way flows in and out of treatment value l as the instruments change. More succinctly,*

$$\Pr(D_{\boldsymbol{z}} = l \ \ \textit{and} \ \ D_{\boldsymbol{z}'} \neq l) \times \Pr(D_{\boldsymbol{z}} \neq l \ \ \textit{and} \ \ D_{\boldsymbol{z}'} = l) = 0.$$

Unordered monotonicity for treatment value l requires that if some observations move out of (resp. into) treatment value l when instruments change value from $\boldsymbol{z}$ to $\boldsymbol{z}'$, then no observation can move into (resp. out of) treatment value l. For binary treatments, unordered monotonicity is equivalent to the usual monotonicity assumption: there cannot be both compliers and defiers. When $K > 2$, it is much weaker, and also weaker than ordered choice.

Heckman and Pinto (2015) show that unordered monotonicity (for well-chosen changes in instruments) is equivalent to a treatment model based on rules that are additively

nonseparable in the unobserved variables. That is,

$$D_k = \mathbb{1}\left(\phi_k(\boldsymbol{V}) \leq \psi_k(\boldsymbol{Z})\right)$$

for some functions ϕ_k and ψ_k that assign all observations to a unique treatment value. This is almost, but not quite, equivalent to a discrete choice model with additively separable utilities; the useful changes in instruments include increasing the mean utility of an alternative relative to all others. We refer the reader to section 6 of Heckman and Pinto (2015) for a more rigorous discussion, and to Pinto (2015) for an application to the Moving to Opportunity program.

Unlike us, Heckman and Pinto (2015) do not require continuous instruments; all of their analysis is framed in terms of discrete-valued instruments and treatments. Beyond this (important) difference, unordered monotonicity clearly obeys our assumptions—just redefine $\phi_k(\boldsymbol{V})$ and $\psi_k(\boldsymbol{Z})$ above so that the unobserved variable is distributed as $U[0,1]$. On the other hand, we allow for much more general models of treatment. It would be impossible, for instance, to rewrite our Examples 1, 2 and 3 so that they obey unordered monotonicity: to use the terminology of Heckman and Pinto (2015), they are both unordered and non-monotonic. We illustrate this point using Example 1 below.

Example 1 (continued)**.** In Example 1, we have that $D = 2$ iff $(V_1 - Q_1(\boldsymbol{Z}))$ and $(V_2 - Q_2(\boldsymbol{Z}))$ have opposite signs. To make a comparison between our model and that of Heckman and Pinto (2015), first note that within $D = 2$ there are two categories:

$$D = 2a \text{ iff } V_1 < Q_1 \text{ and } V_2 > Q_2,$$
$$D = 2b \text{ iff } V_1 > Q_1 \text{ and } V_2 < Q_2.$$

Each one is unordered monotonic; but because we only observe their union, $D = 2$ is not unordered monotonic—increasing Q_1 brings more people into $2a$ but moves some out of $2b$, so that in the end we have two-way flows, contradicting unordered monotonicity. To put it differently, the selection mechanism in Example 1 becomes a discrete choice model when each of four alternatives $d = 0, 1, 2a, 2b$ is observed; however, we only observe whether alternative $d = 0$, $d = 1$ or $d = 2$ is chosen in Example 1. Hence, we allow for a "coarse partition" of unordered monotonicity. □

4.4 Selection into Schooling and Employment

Let S denote a binary schooling decision (say, going to college or not) and E a binary employment decision. We observe the outcome Y (wages) only when an individual is employed (say, $E = 1$). We are interested in the returns to a college education in the form of higher wages. Table 1 summarizes the selection problem in this example. Crossing E and S gives four treatment values $D = 0, 1, 2, 3$. We observe the value of D for each individual, and their wages iff $E = 1$; we denote Y_0 (resp. Y_1) the wages of an employee without (resp. with) a college education, and our parameters of interest are the moments of the college premium $(Y_1 - Y_0)$.

Table 1: Schooling, employment, and wages

	$E = 0$ (non-employed)	$E = 1$ (employed)
$S = 0$ (no college education)	D_0	(D_2, Y_0)
$S = 1$ (college education)	D_1	(D_3, Y_1)

In line with our general model, we assume that both assignments S and E are characterized by a single crossing model based on a one-dimensional unobserved heterogeneity

term:

$$S = 1 \text{ iff } V_1 \leq Q_1(\boldsymbol{Z})$$
$$E = 1 \text{ iff } V_2 \leq Q_2(\boldsymbol{Z}),$$

where the unobservables V_1 and V_2 are independent of $\boldsymbol{Z}$, marginally distributed as $U[0,1]$ and their codependence structure is unknown. In this example, Q_1 and Q_2 are identified from the population directly by $Q_1(\boldsymbol{Z}) = \Pr(S = 1|\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z}) = \Pr(E = 1|\boldsymbol{Z})$.

To use the notation of Section 3, we have

$$D_2 = \mathcal{S}_2(\boldsymbol{Z})(1 - \mathcal{S}_1(\boldsymbol{Z}))$$
$$D_3 = \mathcal{S}_1(\boldsymbol{Z})\mathcal{S}_2(\boldsymbol{Z}).$$

Note that the indices for both treatment values 2 and 3 are non-zero: $a_2 = -1$ and $a_3 = 1$. Therefore Theorem 3.1 applies to $k = 2, 3$, provided in particular that $\boldsymbol{Q}_1(\boldsymbol{Z})$ and $\boldsymbol{Q}_2(\boldsymbol{Z})$ are functionally independent—which is generically true if $\boldsymbol{Z}$ contains two continuous instruments. Under these assumptions,

$$E(Y_0|V_1 = p_1, V_2 = p_2) = \frac{\partial^2 E[Y D_2|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}) = p_2]/\partial p_1 \partial p_2}{\partial^2 \Pr[D_2 = 1|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}) = p_2]/\partial p_1 \partial p_2}$$
$$E(Y_1|V_1 = p_1, V_2 = p_2) = \frac{\partial^2 E[Y D_3|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}) = p_2]/\partial p_1 \partial p_2}{\partial^2 \Pr[D_3 = 1|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}) = p_2]/\partial p_1 \partial p_2}$$

and the marginal treatment effect obtains by simple difference.

To identify the average treatment effect $E(Y_1 - Y_0)$, we use Theorem 3.1 again under

the "full support" Assumption 3.5.[5] Since $a_2 = -1$ and $a_3 = 1$, we obtain

$$EY_0 = -\int_0^1 \int_0^1 \frac{\partial^2 E[YD_2|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2} dq_1 dq_2$$
$$EY_1 = \int_0^1 \int_0^1 \frac{\partial^2 E[YD_3|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2} dq_1 dq_2$$

so that, since $D_2 + D_3 = E$,

$$E(Y_1 - Y_0) = \int_0^1 \int_0^1 \frac{\partial^2 E[YE|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2} dq_1 dq_2.$$

This formula is very intuitive: integrating the right hand side of the equation above gives

$$\begin{aligned} E(Y_1 - Y_0) = & E[YE|Q_1(\boldsymbol{Z}) = 1, Q_2(\boldsymbol{Z}) = 1] \\ & - E[YE|Q_1(\boldsymbol{Z}) = 0, Q_2(\boldsymbol{Z}) = 1] \\ & - E[YE|Q_1(\boldsymbol{Z}) = 1, Q_2(\boldsymbol{Z}) = 0] \\ & + E[YE|Q_1(\boldsymbol{Z}) = 0, Q_2(\boldsymbol{Z}) = 0]. \end{aligned}$$

The last two terms are zero since the probability of employment is zero when $Q_2(\boldsymbol{Z}) = 0$; and conversely, the probability of employment is one when $Q_2(\boldsymbol{Z}) = 1$. That leaves us

[5] Remember that all of our analysis is conditional on covariates $\boldsymbol{X}$. In practice, it is often impossible to do so nonparametrically. In their study of returns to schooling, Carneiro, Heckman, and Vytlacil (2011) circumvent this difficulty by assuming that both the covariates $\boldsymbol{X}$ and instruments $\boldsymbol{Z}$ are independent of the error terms U_k and $\boldsymbol{V}_k$. Then $\boldsymbol{Q}$ can be constructed as a function of both $\boldsymbol{X}$ and $\boldsymbol{Z}$. Such an assumption would allow us to obtain full support even if $\boldsymbol{Z}$ is discrete, by interacting $\boldsymbol{Z}$ with continuous components of $\boldsymbol{X}$.

with

$$E(Y_1 - Y_0) = E[Y|Q_1(\boldsymbol{Z}) = 1, Q_2(\boldsymbol{Z}) = 1] - E[Y|Q_1(\boldsymbol{Z}) = 0, Q_2(\boldsymbol{Z}) = 1],$$

the difference in average wages between the surely-employed populations who are surely college-educated or surely not.

Our approach goes much beyond this fairly trivial result, since it identifies the whole function $(q_1, q_2) \mapsto E(Y_1 - Y_0|V_1 = q_1, V_2 = q_2)$, as well as the joint density. The joint density $f_{V_1,V_2}(q_1, q_2)$ is of interest in itself, as (conditioning on the instruments) it reveals the dependence structure between the likelihood of graduation and the likelihood of employment. Note that $f_{V_1,V_2}(q_1, q_2)$ is over-identified, since it can be obtained from taking cross partial derivatives of $\Pr[D_2 = 1|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]$ or of $\Pr[D_3 = 1|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]$:

$$\begin{aligned} f_{V_1,V_2}(q_1, q_2) &= \frac{\partial^2 \Pr[E = 1, S = 1|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2} \\ &= -\frac{\partial^2 \Pr[E = 1, S = 0|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2}. \end{aligned}$$

Comparing the two resulting estimators provides a specification check.

To conclude this example, note that we could allow for a direct effect of schooling on employment, by adding an argument in Q_2:

$$E = 1 \text{ iff } V_2 \leq Q_2(\boldsymbol{Z}, S).$$

We could try to rewrite this selection rule as

$$E = 1 \text{ iff } V_2' \leq Q_2'(\boldsymbol{Z})$$

for a different unobserved heterogeneity term V_2'; but since S is a discontinuous function of V_1, this would violate the continuity requirements that drive Theorem 3.1.

On the other hand, we may still be able to apply our results since we deal with $D = 2$ and $D = 3$ separately. The threshold Q_1 is still directly identified from the probability of graduation. The probability of employment now depends on both $Q_2(\cdot, 0)$ and $Q_2(\cdot, 1)$; we will assume here that their variations are restricted so that they are still identified. With obvious changes in notation, we now have

$$\begin{aligned} D_2 &= (1 - \mathcal{S}_1(\boldsymbol{Z}))\mathcal{S}_2(\boldsymbol{Z}, 0) \\ D_3 &= \mathcal{S}_1(\boldsymbol{Z})\mathcal{S}_2(\boldsymbol{Z}, 1); \end{aligned}$$

and the conditional expectations are identified by

$$\begin{aligned} E(Y_0|V_1 = p_1, V_2 = p_2) &= \frac{\partial^2 E[YD_2|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}, 1) = p_2]/\partial p_1 \partial p_2}{\partial^2 \Pr[D_2 = 1|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}, 1) = p_2]/\partial p_1 \partial p_2} \\ E(Y_1|V_1 = p_1, V_2 = p_2) &= \frac{\partial^2 E[YD_3|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}, 0) = p_2]/\partial p_1 \partial p_2}{\partial^2 \Pr[D_3 = 1|Q_1(\boldsymbol{Z}) = p_1, Q_2(\boldsymbol{Z}, 0) = p_2]/\partial p_1 \partial p_2}, \end{aligned}$$

from which we can compute marginal and average treatment effects. This shows that interesting models that do not seem to fit our assumptions at first sight can still yield to our approach.

We conclude this section by mentioning a recent paper by Fricke, Frölich, Huber, and Lechner (2015), who consider a model with both treatment endogeneity and non-

response bias that has a structure similar to this schooling-employment example. They use a discrete instrument for the binary treatment and a continuous instrument for attrition to identify the average treatment effect for both the compliers and the total population. In contrast, we identify the marginal treatment effects with two continuous instruments.

4.5 Double Hurdles with Partial Observability

In Example 1, passing a test, or being eligible for a treatment, required crossing two thresholds. They can be eligibility criteria, or simply the combination of "being willing" and "being able." The simplest version of such a "multiple hurdles with partial observability" model has a binary treatment. Suppose that the selection mechanism is governed by

$$D = 1 \text{ iff } V_1 < Q_1(\boldsymbol{Z}) \text{ and } V_2 < Q_2(\boldsymbol{Z}), \tag{4.1}$$

and $D = 0$ otherwise. This is a nonparametric version of the binary probit model with partial observability (Poirier (1980)). Both treatment values have two relevant thresholds ($J^0 = J^1 = 2$); their indices are again nonzero: $a_1 = 1$ and $a_0 = -1$.

Identification of Q_1 and Q_2, which is a requisite to applying Theorem 3.1, is not as straightforward as in the schooling/employment model of the previous section. We observe

$$\Pr(D = 1|\boldsymbol{Z}) = F_{V_1,V_2}\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right),$$

which is a nonparametric double index model in which both the link function F_{V_1,V_2} and the indices Q_1 and Q_2 are unknown. This is clearly underidentified without stronger

restrictions. Matzkin (1993, 2007) considers nonparametric identification and estimation of polychotomous choice models. Our multiple hurdle model has a similar but not identical structure. We build on Lewbel (2000) and on Matzkin's results to identify $\boldsymbol{Q}$. To do so, we assume that the thresholds have the following structure:

$$Q_1(\boldsymbol{Z}) = G_1\left(Z_1 + q_2(\boldsymbol{Z}_2)\right)$$
$$Q_2(\boldsymbol{Z}) = G_3\left(Z_3 + q_4(\boldsymbol{Z}_4)\right)$$

where G_1, G_3, q_2 and q_4 are unknown functions; we also allow for $q_2 = q_4 = 0$. We impose that

Assumption 4.2 (Identifying the Thresholds)**.** *The density of* (V_1, V_2) *is continuous on* $[0,1]^2$*, with marginal uniform distributions. Furthermore,*

1. G_1 *and* G_3 *are strictly increasing* C^1 *functions from possibly unbounded intervals* $[a_1, b_1]$ *and* $[a_3, b_3]$ *onto* $[0,1]$*;*

2. *there exists a point* $(\bar{\boldsymbol{z}}_2, \bar{\boldsymbol{z}}_4)$ *in the support of* $(\boldsymbol{Z}_2, \boldsymbol{Z}_4)$ *such that*

   (a) *the support of* (Z_1, Z_3) *conditional on* $\boldsymbol{Z}_2 = \bar{\boldsymbol{z}}_2, \boldsymbol{Z}_4 = \bar{\boldsymbol{z}}_4$ *is the rectangle* $R_{13} = [a_1, b_1] \times [a_3, b_3]$*;*

   (b) *the support of* $\boldsymbol{Z}_2$ *conditional on* $\boldsymbol{Z}_4 = \bar{\boldsymbol{z}}_4$ *equals its unconditional support;*

   (c) *the support of* $\boldsymbol{Z}_4$ *conditional on* $\boldsymbol{Z}_2 = \bar{\boldsymbol{z}}_2$ *equals its unconditional support.*

3. *if* q_2 *and/or* q_4 *are known to be zero, drop the corresponding conditioning statements in 2.*

Proposition 4.1 (Identification in the double-hurdle model)**.** *Under Assumption 4.2, the functions* $F_{\boldsymbol{V}}, G_1, G_3$ *and (if nonzero)* q_2 *and* q_4 *are identified from the propensity score* $\Pr(D = 1|\boldsymbol{Z})$.

Proof. Without loss of generality[6], we normalize $q_2(\bar{\boldsymbol{z}}_2) = q_4(\bar{\boldsymbol{z}}_4) = 0$. Define H by

$$H(z_1, z_3) = \Pr\left(D = 1|Z_1 = z_1, Z_3 = z_3, \boldsymbol{Z}_2 = \bar{\boldsymbol{z}}_2, \boldsymbol{Z}_4 = \bar{\boldsymbol{z}}_4\right)$$

for any $(z_1, z_3) \in \mathbb{R}^2$.

Let $f_{\boldsymbol{V}}(v_1, v_2)$ denote the density of $\boldsymbol{V}$. By construction,

$$H(z_1, z_3) = F_{\boldsymbol{V}}(G_1(z_1), G_3(z_3)) = \int_0^{G_1(z_1)} \int_0^{G_3(z_3)} f_{\boldsymbol{V}}(v_1, v_2) dv_1 dv_2. \tag{4.2}$$

Differentiating both sides of (4.2) with respect to z_1 gives

$$\frac{\partial H}{\partial z_1}(z_1, z_3) = G_1'(z_1) \int_0^{G_3(z_3)} f_{\boldsymbol{V}}(G_1(z_1), v_2) dv_2. \tag{4.3}$$

Now letting $z_3 \to b_3$ on the both sides of (4.3) yields

$$\lim_{z_3 \to b_3} \frac{\partial H}{\partial z_1}(z_1, z_3) = G_1'(z_1) \left[\lim_{z_3 \to b_3} \int_0^{G_3(z_3)} f_{\boldsymbol{V}}(G_1(z_1), v_2) dv_2 \right]. \tag{4.4}$$

Note that the expression inside the brackets on the right side side of (4.4) is 1 since $\lim_{z_3 \to b_3} G_3(z_3) = 1$ and the marginal distribution of V_2 is $U[0, 1]$. Therefore we identify

[6] We can always adjust G_1 and G_3 to compensate.

G_1 by

$$G_1(z_1) = \int_{a_1}^{z_1} \lim_{t_3 \to b_3} \frac{\partial H}{\partial z_1}(t_1, t_3) dt_1. \tag{4.5}$$

Analogously, we identify G_3 by

$$G_3(z_3) = \int_{a_3}^{z_3} \lim_{t_1 \to b_1} \frac{\partial H}{\partial z_3}(t_1, t_3) dt_3. \tag{4.6}$$

Returning to (4.2), since G_1 and G_3 are strictly increasing we also identify $F_{\boldsymbol{V}}$ by

$$F_{\boldsymbol{V}}(v_1, v_2) = H(G_1^{-1}(v_1), G_3^{-1}(v_2)).$$

Once $F_{\boldsymbol{V}}, G_1$ and G_3 are identified, we fix any point $(\bar{z}_1, \bar{z}_3)$ and we identify $q_2(\boldsymbol{z}_2)$ by choosing $Z_1 = \bar{z}_1; Z_3 = \bar{z}_3; Z_2 = \boldsymbol{z}_2$; and $Z_4 = \bar{\boldsymbol{z}}_4$. This gives

$$\Pr\left(D = 1 | Z_1 = \bar{z}_1; Z_3 = \bar{z}_3; Z_2 = \boldsymbol{z}_2, Z_4 = \bar{\boldsymbol{z}}_4\right) = F_{\boldsymbol{V}}\left(G_1(\bar{z}_1 + q_2(\boldsymbol{z}_2)), G_3(\bar{z}_3)\right)$$

which inverts to give the value of $q_2(\boldsymbol{z}_2)$. We proceed in the same way for $q_4(\boldsymbol{z}_4)$. □

While the proof above requires at least four continuous instruments when q_2 and q_4 are nonzero, various additional restrictions would relax this requirement. If for instance the functional forms of q_2 and q_4 were known, then $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_4$ could be identical. And if q_2 and q_4 were linear, we would be back to the linear multiple index models of Ichimura and Lee (1991).

Once $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$ are identified, then under our assumptions we identify the

joint density by

$$f_{V_1,V_2}(q_1,q_2) = \frac{\partial^2 \Pr[D=1|Q_1(\boldsymbol{Z})=q_1, Q_2(\boldsymbol{Z})=q_2]}{\partial q_1 \partial q_2}. \tag{4.7}$$

Note that under Assumption 4.2, F_{V_1,V_2} is already identified; so that we have overidentification of f_{V_1,V_2}. The marginal treatment effect is given by

$$E(Y_1 - Y_0|V_1=q_1, V_2=q_2) f_{V_1,V_2}(q_1,q_2) = \frac{\partial^2 E[Y|Q_1(\boldsymbol{Z})=q_1, Q_2(\boldsymbol{Z})=q_2]}{\partial q_1 \partial q_2}. \tag{4.8}$$

Under Assumption 4.2, both $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$ have full support, so that the average treatment effect is identified by

$$E[Y_1 - Y_0] = \int_0^1 \int_0^1 \frac{\partial^2 E[Y|Q_1(\boldsymbol{Z})=p_1, Q_2(\boldsymbol{Z})=p_2]}{\partial p_1 \partial p_2}\Big|_{(p_1,p_2)=(q_1,q_2)} dq_1 dq_2. \tag{4.9}$$

Example 4. As another illustration, consider the following model of employment, adapted from Laroque and Salanié (2002). An employee ($D = 1$) must be employable, in the sense that her unobserved productivity ρ must be above the minimum wage $\underline{Y}$. Specify productivity as

$$\rho = R_1(\boldsymbol{Z}) - v_1,$$

where v_1 is independent of $\boldsymbol{Z}$. This gives a first hurdle $v_1 < R_1(\boldsymbol{Z}) - \underline{Y}$; and transforming both sides by the cdf of v_1 gives $V_1 < Q_1(\boldsymbol{Z})$.

In addition, employees must be willing to work at the offered wage. Assume that each employee receives her full productivity. Then with a disutility of work specified as

$$d = R_2(\boldsymbol{Z}) - v_2,$$

with again v_2 independent of $\boldsymbol{Z}$, the second hurdle $\rho > d$ translates to $v_1 + v_2 < R_1(\boldsymbol{Z}) - R_2(\boldsymbol{Z})$. Again, this can be transformed into $V_2 < Q_2(\boldsymbol{Z})$ using the cdf $F_{v_1+v_2}$.

The impact of employment on outcomes Y can then be assessed using (4.7), (4.8) and (4.9). Note that this particular structure naturally suggests ways of identifying Q_1 and Q_2, as Q_1 only depends on R_1 and Q_2 depends on both R_1 and R_2. □

Example 5. Finally, consider a common parental choice problem: the choice of a school for a child. The child will go to a private school ($D = 1$) if both parents agree that she should: $V_1 < Q_1(\boldsymbol{Z})$ and $V_2 < Q_2(\boldsymbol{Z})$. Otherwise the child will attend a public school ($D = 0$). If Y is any child outcome, then the effect of attending a private school can be identified from (4.7), (4.8) and (4.9). □

4.6 Fully Partitioned Treatment Assignment

Sometimes the combination of J criteria determines 2^J different treatments[7], according to the value of the binary vector $(V_1 < Q_1(\boldsymbol{Z}), \ldots, V_J < Q_J(\boldsymbol{Z}))$. Each of these treatment values is what we called earlier an atom, with an index ± 1. Identification of $(Q_1, \ldots, Q_J)$ is straightforward; if for instance the first 2^{J-1} treatment values have $V_1 < Q_1$ and the last 2^{J-1} have $V_1 > Q_1$, then $Q_1(\boldsymbol{Z}) = \Pr(D \leq 2^{m-1}|\boldsymbol{Z})$.

To identify the treatment effects and joint density, we need J continuous instruments in $\boldsymbol{Z}$. To illustrate, order treatment values in the standard binary order, coding $V_j < Q_j$ as a 1. The joint density is multiply overidentified: for each treatment value $d = 0, \ldots, 2^J - 1$ of index a_d,

$$f_{V_1,\ldots,V_J}(q_1, \ldots, q_J) = \frac{1}{a_d} \frac{\partial^J}{\partial q_1 \ldots \partial q_J} \Pr(D = d | Q_1(\boldsymbol{Z}) = q_1, \ldots, Q_J(\boldsymbol{Z}) = q_J).$$

[7]We thank Rodrigo Pinto for suggesting this example to us.

Say that $J \geq 4$ and we want to identify the treatment effect of moving from treatment value $d = 1011$ (with index $a_d = -1$) to treatment value $d' = 0101$ (which has index $a_d = 1$). The marginal treatment effect is given by

$$\begin{aligned} &E(Y_d - Y_{d'}|V_1 = q_1, \ldots, V_J = q_J) f_{V_1,\ldots,V_J}(q_1, \ldots, q_J) \\ &= \frac{\partial^J}{\partial q_1 \ldots \partial q_J} E(Y(D_d + D_{d'})|Q_1(\boldsymbol{Z}) = q_1, \ldots, Q_J(\boldsymbol{Z}) = q_J). \end{aligned}$$

5 Discrete Instruments

Continuous instruments are a luxury that may not be available to the analyst. While our method sems to be extremely dependent on them, it is sometimes possible to use it with discrete-valued instruments, in the same way that LATE is an integrated version of the MTE. To see this, take the nonzero index case, for which Theorem 3.1 gave us an equality

$$f_{\boldsymbol{V}}(\boldsymbol{q}) = \frac{1}{a_k} T \Pr\left(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right),$$

where T is the linear differential operator

$$TH = \frac{\partial^J H}{\prod_{j=1}^J \partial q_j}.$$

5.1 Partial Identification

With discrete-valued instruments, we cannot make sense of the right-hand side; on the other hand, a close examination of the steps of our proof shows that we can apply the

inverse operator T^{-1} to both sides, to obtain

$$F_{\boldsymbol{V}}(\boldsymbol{q}) = \Pr\left(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right) + F_0(\boldsymbol{q}),$$

where F_0 is any function with $TF_0 = 0$; that is,

$$F_0(\boldsymbol{q}) = \sum_{j=1}^{J} F_{0j}(\boldsymbol{q}_{-j}),$$

where each term in the sum excludes one of the components of $\boldsymbol{q}$.

Given discrete-valued instruments that generates values of $\boldsymbol{Q}(\boldsymbol{Z})$ in $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^s\}$, we can apply the finite-difference $\bar{T}$ version of T. The terms F_{0j} generate null finite differences, and we are left with point identification of the finite differences of the cdf $F_{\boldsymbol{V}}$:

$$\bar{T}F_{\boldsymbol{V}}(\boldsymbol{q}) = \bar{T}_k \Pr\left(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right).$$

In many models this will allow us to identify the average effect of a treatment on a family of observations that comprises several groups of "compliers." We illustrate this in the following subsection, using the double hurdle model of section 4.5.

5.2 The case of the double hurdle model

Recall the double hurdle model of section 4.5, which has two thresholds so that

$$\tilde{T}H(q_1, q_2) = \frac{\partial^2 H}{\partial q_1 \partial q_2}(q_1, q_2) \text{ and } \bar{T}H(\boldsymbol{q}^i, \boldsymbol{q}^l) = H(\boldsymbol{q}^i) + H(\boldsymbol{q}^l) - H(q_1^i, q_2^l) - H(q_1^l, q_2^i).$$

Now assume that $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^s\}$ contains two vectors $\boldsymbol{q}^i$ and $\boldsymbol{q}^l$ such that $q_1^i < q_1^l$ and

$q_2^i < q_2^l$: both thresholds are higher under $\boldsymbol{q}^l$ than under $\boldsymbol{q}^i$. Since both thresholds increase, no observation moves from $D = 1$ to $D = 0$; and three groups move from $D = 0$ to $D = 1$:

1. those with $q_1^i < V_1 < q_1^l$ and $V_2 < q_2^i$
2. those with $V_1 < q_1^i$ and $q_2^i < V_2 < q_2^l$
3. those with $q_1^i < V_1 < q_1^l$ and $q_2^i < V_2 < q_2^l$.

To borrow from the language of the LATE literature, there are no defiers, and three different groups of compliers. The relative weights of these groups cannot be estimated from the data without further assumptions. If we form the Wald estimator

$$\frac{E(Y|\boldsymbol{Q} = \boldsymbol{q}^l) - E(Y|\boldsymbol{Q} = \boldsymbol{q}^i)}{\Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^l) - \Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^i)},$$

we only identify a weighted treatment effect for all three groups combined together. This illustrates the limitation of discrete instruments and the difficulty of interpreting Wald estimands or their extensions when the selection mechanism is more complex than in the usual single-threshold model.

Let us now move to a more favorable case: we assume that the sample contains four values $(\boldsymbol{q}^i, \boldsymbol{q}^l, \boldsymbol{q}^m, \boldsymbol{q}^n)$ that form a rectangle in (q_1, q_2) space. This could arise if the thresholds are varied independently and then the variations are combined.[8] Of course, it requires $s \geq 4$.

[8]The thresholds $\boldsymbol{q}$ may correspond to historical policies. Then the average treatment of moving from one value of $\boldsymbol{q}$ to another can be interpreted as the policy relevant treatment effect (PRTE) of Heckman and Vytlacil (2001).

We reorder i, l, m, n if necessary so that as before $q_1^l > q_1^i$ and $q_2^l > q_2^i$. We form

(5.1)

$$\begin{aligned}\Delta_1 &= \Pr(D=1|\boldsymbol{Q}=\boldsymbol{q}^i) + \Pr(D=1|\boldsymbol{Q}=\boldsymbol{q}^l) - \Pr(D=1|\boldsymbol{Q}=\boldsymbol{q}^m) - \Pr(D=1|\boldsymbol{Q}=\boldsymbol{q}^n) \\ &= F_{V_1,V_2}(\boldsymbol{q}^i) + F_{V_1,V_2}(\boldsymbol{q}^l) - F_{V_1,V_2}(\boldsymbol{q}^m) - F_{V_1,V_2}(\boldsymbol{q}^n),\end{aligned}$$

which identifies the last term. Note in passing that the identified quantity

$$\frac{\Delta_1}{(q_1^i - q_1^l)(q_2^i - q_2^l)}$$

is the value of the density $f_{\boldsymbol{V}}$ at some point between[9] $\boldsymbol{q}^i$ and $\boldsymbol{q}^l$.

Similar arguments give us

$$\begin{aligned}\Delta_2 &= E(Y|\boldsymbol{Q}=\boldsymbol{q}^i) + E(Y|\boldsymbol{Q}=\boldsymbol{q}^l) - E(Y|\boldsymbol{Q}=\boldsymbol{q}^m) - E(Y|\boldsymbol{Q}=\boldsymbol{q}^n) \\ &= \int_{q_1^i}^{q_1^l}\int_{q_2^i}^{q_2^l} E(Y_1 - Y_0|V_1=t_1, V_2=t_2) f_{V_1,V_2}(t_1,t_2)\,dt_1\,dt_2, \qquad (5.2)\end{aligned}$$

which identifies the last term. Again,

$$\frac{\Delta_2}{(q_1^i - q_1^l)(q_2^i - q_2^l)} = E(Y_1 - Y_0|V_1=t_1, V_2=t_2) f_{V_1,V_2}(t_1,t_2)$$

for some point between $\boldsymbol{q}^i$ and $\boldsymbol{q}^l$.

If the rectangle is small enough, it will be a good first approximation to say that

[9] More precisely, at some point on each arc that links these two points.

Δ_2/Δ_1 identifies the MTE locally. If it is not, then we identify

$$\frac{\Delta_2}{\Delta_1} = \int_{q_1^i}^{q_1^l} \int_{q_2^i}^{q_2^l} E(Y_1 - Y_0 | V_1 = t_1, V_2 = t_2) w(t_1, t_2) dt_1 dt_2$$

where the function

$$w(t_1, t_2) = \frac{f_{V_1,V_2}(t_1, t_2)}{\Delta_1}$$

gives unknown positive weights[10] that integrate to one. Note that the weights are simply the density of $\boldsymbol{V}$ truncated to the rectangle.

This is an integrated MTE, just like LATE. Note that Δ_1 corresponds to the size of group 3 (see page 46); in fact the ratio Δ_2/Δ_1 is a density-weighted average of the effect of the treatment for group 3. If we are lucky enough to observe such a "rectangular" variation in the thresholds, then we can estimate the effect of treatment on the group of compliers who failed both criteria and now pass both.

Note that we could also construct other Wald estimators in the rectangular case. For instance, assume that $q_1^n > q_1^i$ (so that $q_2^n = q_2^i$), and consider the identified ratio

$$\frac{E(Y|\boldsymbol{Q} = \boldsymbol{q}^n) - E(Y|\boldsymbol{Q} = \boldsymbol{q}^i)}{\Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^n) - \Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^i)}.$$

The denominator equals

$$\int_{(0,1)^2} \left(\mathbb{1}(v_1 < q_1^n, v_2 < q_2^n) - \mathbb{1}(v_1 < q_1^i, v_2 < q_2^i) \right) f_{V_1,V_2}(v_1, v_2) dv_1 dv_2;$$

[10] Δ_1 is positive given our ordering of i, l, m, n.

and since $q_2^n = q_2^i$, this can be rewritten as

$$\int_{(0,1)^2} \mathbb{1}(v_2 < q_2^n)\mathbb{1}(q_1^i < v_1 < q_1^n) f_{V_1,V_2}(v_1, v_2)$$

which is the size of group 1 of compliers. It is easy to see that this new Wald estimator estimates the treatment effect on this group. We could similarly define

$$\frac{E(Y|\boldsymbol{Q} = \boldsymbol{q}^m) - E(Y|\boldsymbol{Q} = \boldsymbol{q}^i)}{\Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^m) - \Pr(D = 1|\boldsymbol{Q} = \boldsymbol{q}^i)}$$

and identify the average effect of treatment on group 2.

The rectangular case therefore identifies the sizes of the three groups of compliers, as well as the average effect of treatment on each of them.

6 Relation to the Existing Literature

Several papers have analyzed multivalued treatments under the unconfoundedness assumption. Imbens (2000) and Hirano and Imbens (2004) develop generalizations of the propensity score to discrete treatments and to continuous treatments, respectively. Cattaneo (2010) show that the semiparametric efficiency bound can be achieved in discrete treatment models by first estimating the generalized propensity score, then applying an inverse probability weighted estimator.

Since we do not assume conditional independence between potential outcomes and unobservables governing the selection mechanism, the rest of this section discusses selection on unobservables in models with multivalued treatment. The most popular approaches rely on instruments, like ours. We already discussed Angrist and Imbens (1995)

in section 4.1 and Heckman and Pinto (2015) in section 4.3. In addition, Florens, Heckman, Meghir, and Vytlacil (2008) consider a potential outcome model with a continuous treatment. They assume a stochastic polynomial restriction such that the counterfactual outcome Y_d corresponding to the continuous treatment value d has the form:

$$Y_d = EY_d + \sum_{k=0} d\varepsilon_k,$$

where $K > \infty$ is known. They show that the average treatment effect can be identified if a control function $\tilde{V}$ can be found such that

$$E(\varepsilon_k|D, Z) = E(\varepsilon_k|\tilde{V}) \equiv r_k(\tilde{V}),\ k = 0, \ldots, K.$$

Imbens and Newey (2009) also consider selection on unobservables with a continuous treatment. They assume that the treatment (more generally in their paper, an endogenous variable) is given $D = g(Z, V)$, with g increasing in a scalar unobserved V. Then they normalize by $V = F_{D|Z}(D|Z)$ which is $U[0, 1]$. If counterfactual outcomes are assumed to follow $Y_d = \phi(d) + U_d$ with $EU_d \equiv 0$, then

$$EY_d = EE(Y_d|V) = EE(Y|D = d, V)$$

is identified by Imbens and Newey (2009)'s control function approach. Other more recent identification results along this line can be found in Torgovitsky (2015) and D'Haultfœuille and Février (2015) among others. One key restriction in this group of papers is the monotonicity in the scalar V in the selection equation. We do not rely on this type of restriction, but we only focus on the case of multivalued treatments.

Hence, our approach and those of Imbens and Newey (2009), Torgovitsky (2015) and D'Haultfœuille and Février (2015) are complementary.

Finally, our approach shares some similarities with Hoderlein and Mammen (2007). They consider the identification of marginal effects in nonseparable models without monotonicity:

$$Y = \phi(X, Z, U)$$

where Z is continuous multivariate and $U \perp\!\!\!\perp X|Z$. They show that

$$E\left(\frac{\partial \phi}{\partial x}(x, z, U)|X = x, Z = z, Y = q_\alpha(x, z)\right) = \frac{\partial q_\alpha}{\partial x}(Y|X = x, Z = z).$$

In this equation, $q_\alpha(Y|X, Z)$ represents the α-quantile of the distribution of Y conditional on X and Z; and the left-hand side is a local average structural derivative. Since the quantiles are clearly identified from the data, so is the left-hand side. Their approach based on differentials is reminiscent of our method of taking derivatives. The parameters of interest they study are quite different, however; and their selection mechanism is not as explicit as ours.

References

Angrist, J. D., and G. W. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating marginal

policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 78(1), 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 275481.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

D'Haultfœuille, X., and P. Février (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 83(3), 1199–1210.

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(5), 1191–1206.

Fricke, H., M. Frölich, M. Huber, and M. Lechner (2015): "Endogeneity and non-response bias in treatment evaluation: nonparametric identification of causal effects by instruments," Discussion paper, IZA Discussion Paper No. 9428.

Heckman, J., and R. Pinto (2015): "Unordered monotonicity," University of Chicago, mimeo.

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

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding instrumental variables in models with essential heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

——— (2008): "Instrumental variables in models with multiple outcomes: The general unordered case," *Annales d'économie et de statistique*, pp. 151–174.

Heckman, J. J., and E. Vytlacil (2001): "Policy-relevant treatment effects," *American Economic Review*, pp. 107–111.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Heckman, J. J., and E. Vytlacil (2007): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," in *Handbook of econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 6B, chap. 70, pp. 4779–4874. Elsevier, Amsterdam.

Hirano, K., and G. W. Imbens (2004): "The propensity score with continuous treatments," *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164, 73–84.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, 75(5), 1513–1518.

Ichimura, H., and L.-f. Lee (1991): "Semiparametric least squares estimation of multiple index models: single equation estimation," in *International Symposia in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge University Press.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

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

Imbens, G. W., and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Laroque, G., and B. Salanié (2002): "Labour market institutions and employment in France," *Journal of Applied Econometrics*, 17(1), 25–48.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97(1), 145–177.

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

——— (1997): "Monotone treatment response," *Econometrica*, 65, 1311–1334.

Manski, C. F., and J. V. Pepper (2000): "Monotone instrumental variables: with an application to the returns to schooling," *Econometrica*, 68(4), 997–1010.

Matzkin, R. L. (1993): "Nonparametric identification and estimation of polychotomous choice models," *Journal of Econometrics*, 58(1), 137–168.

——— (2007): "Heterogeneous choice," in *Advances in economics and econometrics: theory and applications*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 2, chap. 4, pp. 75–110. Cambridge University Press.

Pinto, R. (2015): "Selection bias in a controlled experiment: the case of Moving to Opportunity," University of Chicago, mimeo.

Poirier, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12(2), 209–217.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83(3), 1185–1197.

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

A Appendix

A.1 Some Results on Indices

Assume that the model of treatment has J thresholds. This generates 2^J atoms. A treatment value is defined by the union of any number of atoms, that is by a subset of $\{1, \ldots, 2^J\}$. There are no fewer than $(2^{2^J} - 2)$ possible selection rules (excluding the two trivial cases). The number of treatment models with t treatment values is the number of partitions of the set $\{1, \ldots, 2^J\}$ into t non-empty sets, which is an exponentially increasing number.

For $m = 0, \ldots, J$, the number of atoms with m terms $\bar{E}_j$ is $\binom{J}{m}$; and such atoms have an index $(-1)^m$. Since $\sum_{m=0}^{J} \binom{J}{m}(-1)^m = (1-1)^m = 0$, the sum of the indices of all atoms is zero; and so is that of the indices of all treatment values since each atom belongs to one treatment and to one only. Moreover, for every atom with index 1 there is one with index -1, and vice versa (just take complements of the E_j and $\bar{E}_j$ sets). It follows that there are 2^{J-1} atoms with index 1 and 2^{J-1} with index -1.

To create a treatment value with all J thresholds relevant and a zero index, we need to combine (at least) an atom with index 1 and one with index -1. Take any such pair of atoms. They must differ on an odd number of threshold-crossing rules. They can differ on *only one* threshold j: but then their union would combine E_j or $\bar{E}_j$, and threshold j would not be relevant any more. It follows that the two-threshold case is very special: for $J = 2$ no treatment value that responds to both thresholds can have zero index.

On the other hand, with $J = 3$ thresholds one can simply take the complement of the three E_j or $\bar{E}_j$ in any atom; combining the resulting two atoms creates a zero-index treatment value, as in Example 3. And for $J > 3$, we can leave all other threshold

crossings unchanged.

A.2 Proof of Theorem 3.2

As explained in the text, steps 1 and 2 of the proof of Theorem 3.1 do not rely on any assumption about indices. They show that if we define

$$W_l(\boldsymbol{q}) = \int \prod_{j \in I_l} H(q_j - v_j) b_k(\boldsymbol{v}) d\boldsymbol{v}$$

where the set $I_l \subset \boldsymbol{J}$, then its cross-derivative with respect to $(\boldsymbol{p}^{I_l})$ is

$$\int b_k(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l},$$

where $\boldsymbol{v}_{-I_l}$ collects all components of $\boldsymbol{v}$ whose indices are not in I_l.

Now let m be the degree of treatment k. In the sum (3.4), take any term l such that $|I_l| = m$. Recall that $\widetilde{T}$ denotes the differential operator

$$\widetilde{T} = \frac{\partial^m}{\prod_{i=1,\ldots,m} \partial_{j_i}}.$$

By the formula above, applying $\widetilde{T}$ to term l gives

$$c_l \int b_k(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l}.$$

Moreover, applying $\widetilde{T}$ to any other term l' obviously gives zero if term l' has degree lower than m. Turning to terms l' of degree m, any such term must have a $I_{l'} \neq I_l$, or it would be collected in term l. But then $\widetilde{T}$ takes at least one derivative along a direction that is

not in l', and that term contributes zero too.

This proves that

$$\widetilde{T}B_k(\boldsymbol{q}) = c_l \int b_k(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l};$$

note that it also implies that $\widetilde{T}B_k(\boldsymbol{q})$ only depends on $\boldsymbol{q}^{I_l}$.

Applying this first to $b_k(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$, then to $b_k(\boldsymbol{v}) = E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}]f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ exactly as in the proof of Theorem 3.1, we get

$$\int f_{\boldsymbol{V}}(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l} = \frac{1}{c_l}\widetilde{T}\Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$$
$$\int E[G(Y_k)|\boldsymbol{V} = (\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l})] f_{\boldsymbol{V}}(\boldsymbol{q}^{I_l}, \boldsymbol{v}_{-I_l}) d\boldsymbol{v}_{-I_l} = \frac{1}{c_l}\widetilde{T}E(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}).$$

Since the left-hand sides are simply $f_{\boldsymbol{V}^{I_l}}(\boldsymbol{v}^{I_l})$ and $E[G(Y_k)|\boldsymbol{V}^{I_l} = \boldsymbol{q}^{I_l}]f_{\boldsymbol{V}^{I_l}}(\boldsymbol{v}^{I_l})$, the conclusion of the theorem follows immediately.