Endogeneity in Weakly Separable Models without Monotonicity[*]

Songnian Chen
Zhejiang University

Shakeeb Khan
Boston College

Xun Tang
Rice University

October 20, 2022

Abstract

We identify and estimate treatment effects when potential outcomes are weakly separable with a binary endogenous treatment. Vytlacil and Yildiz (2007) proposed an identification strategy that exploits the mean of observed outcomes, but their approach requires a monotonicity condition. In comparison, we exploit full information in the entire outcome distribution, instead of just its mean. As a result, our method does not require monotonicity and is also applicable to general settings with multiple indices. We provide examples where our approach can identify treatment effect parameters of interest whereas existing methods would fail. These include models where potential outcomes depend on multiple unobserved disturbance terms, such as a Roy model, a multinomial choice model, as well as a model with endogenous random coefficients. We establish consistency and asymptotic normality of our estimators.

JEL Classification: C14, C31, C35

Key Words Weak Separability, Treatment Effects, Monotonicity, Endogeneity

[*] An earlier version of this paper was circulated under the title "Identification and Estimation of weakly separable models without Monotonicity". We are grateful to Jeremy Fox, Sukjin Han, Arthur Lewbel, Whitney Newey, Elie Tamer, Ed Vytlacil, Haiqing Xu, conference participants at the 2018 West Indies Economic Conference, the 2019 SUFE econometrics meetings, the 2020 Texas Camp Econometrics, the 2020 World Congress, and seminar participants at various institutions for helpful comments and suggestions.

1 Introduction

Consider a weakly separable model with a binary endogenous variable:

$$Y = g(v_1(X,D), v_2(X,D), ...v_J(X,D), \varepsilon) \tag{1.1}$$

$$D = 1\{\theta(Z) - U > 0\} \tag{1.2}$$

where $(v_1(X,D), v_2(X,D), ...v_J(X,D)) \equiv v(X,D)$ is a J-vector of unknown linear or nonlinear indices in the outcome equation (1.1) and D is a binary endogenous variable defined by equation (1.2). Here $X \in \mathbb{R}^{d_x}$ and $Z \in \mathbb{R}^{d_z}$ are vectors of observable exogenous variables, which may have overlapping elements. This paper is about the identification and estimation of the average treatment effect (ATE).[1] Similar to conditions in Vytlacil and Yildiz (2007), we require that there is some continuous element in Z excluded from X so that $\theta(Z)$ varies continuously conditional on X, and that we can vary X after conditioning on $\theta(Z)$.[2] It is worth noting that the method we propose also applies directly in a more general setup where the error in the outcome equation is specific to D, i.e., with ε replaced by ε_D, provided the marginal distribution of ε_d is the same for $d \in \{0,1\}$.

The literature on program evaluation abounds in examples in which self-selected, endogenous treatment depends on exogenous instruments that have no direct impact on potential outcomes. For instance, Vytlacil and Yildiz (2007) provided an example where D indicates an individual's enrollment in job training at date 0, and Y is employment status at date 1. In this case, Z may contain variables summarizing local labor market conditions at date 0. They noted these variables affect the opportunity cost of training, and consequently the training decision at date 0, but do not directly affect the individual's employment at date 1.

In the system of equations above, U is the unobservable random variable normalized to follow the standard uniform distribution, and the error term ε in the outcome equation is allowed to be a random vector with a known dimension. We assume (X,Z) are independent of (ε, U). Note that we allow $v(X,D)$ to be a vector of multiple indices, whereas existing

[1] The ATE is one of several parameters of interest in the program evaluation literature. In their seminal work, Imbens and Angrist (1994) focus on a different parameter of interest, the Local Average Treatment Effect (LATE) which is the ATE for the subset of the population referred to as compliers. However, as pointed out in Heckman and Vytlacil (2005), Heckman and Vytlacil (2007a), Carneiro, Vytlacil, and Heckman (2010), and Mogstad, Santos, and Torgovitsky (2018), there are settings where the LATE by itself is not always the focal point of policy studies.

[2] This does not necessarily require a second exclusion restriction that there be an element in X not in Z. As explained in Vytlacil and Yildiz (2007), what is required is that the element in Z not in X be continuously distributed.

methods, such as Vytlacil and Yildiz (2007), can only be applied when it is a single index. Heckman and Vytlacil (2005) and Carneiro and Lee (2009) maintained that (ε, U) is independent of Z conditional on X, and used local instrumental variables to estimate marginal treatment effect (MTE) for realized values of U over the support of $\theta(Z)$ given X. In their case, integrating out MTE to ATE would require the support of $\theta(Z)$ given X to be the complete unit interval, i.e. $[0, 1]$. In comparison, we propose a method that relaxes such a support condition by exploiting more structure in the potential outcome equation and its full distribution. Moreover, while Carneiro and Lee (2009) use $E[1\{Y \leq y\}|X, D, U = p]$ for *each fixed* y to recover the conditional distribution of potential outcomes at y, we use the full distribution of the observed outcomes to find the matching covariates. This helps us to recover ATE without the aforementioned full support condition. As in Vytlacil and Yildiz (2007), our method requires that at least some covariates in the potential outcome equation are exogenous.

Since Vytlacil and Yildiz (2007), important work has considered identification and estimation of similar models, but under alternative conditions, notably on the supports of D, and $\theta(Z)$ as well as the dimension of ε. Imbens and Newey (2009) assume D is continuous and monotone in the error term, $\theta(Z)$ is continuous with large support, and ε is scalar. Kasy (2014) allows ε to be a vector as we do here but imposes that both D and $\theta(Z)$ are continuously distributed and further assumes a monotone relationship between the two. D'Haultfoeuille and Fevrier (2015) and Torgovitsky (2015) assume D is continuous, ε is scalar, and assume monotonicity in the error term in each of the two equations, though they allow for $\theta(Z)$ to be discrete. Vuong and Xu (2017) assume D is discrete and $\theta(Z)$ is continuous, restricts ε to be scalar and requires Y in (1.1) to be monotone in ε. Jun, Pinkse, Xu, and Yildiz (2016) study an extension of Vytlacil and Yildiz (2007), and also require the same monotonicity condition as in the latter. Feng (2020) shows how to identify nonseparable triangular models where the endogenous variable is discrete but restricted to have larger support than the instrument variable.[3]

As in the conventional framework, two potential outcomes Y_1 and Y_0 satisfy

$$Y_D = g(v(X, D), \varepsilon) \text{ for } D = 0, 1.$$

We only observe (Y, D, X, Z), where $Y = DY_1 + (1 - D)Y_0$. In this model, we do not impose

[3] These papers focus on point identification. For partial identification of a model with a binary outcome, see Shaikh and Vytlacil (2011) and Mourifié (2015). Mogstad, Santos, and Torgovitsky (2018) explore partial identification of relevant treatment effect parameters in models without structure imposed on the outcome equation. In such settings, attaining point identification requires support conditions on the propensity score function that are stronger than imposed here.

parametric distribution on the error terms (ε, U) or a linear index structure. Similarly, Vytlacil and Yildiz (2007) also do not require such restrictions, but assume that $v(X, D) \in \mathbb{R}$ is a single index, and

$$E\left[g(v, \varepsilon)|U = u\right] \text{ is strictly increasing in } v \in \mathbb{R} \text{ for all } u. \tag{1.3}$$

We do not impose any such monotonicity structure. That is because our approach exploits the full information in the distribution of the outcome variable, instead of just its mean. Indeed, when the outcome distribution function is more informative than the mean, our method is applicable to more general settings; in particular, not only do we not rely on such a monotonicity assumption, but we also allow for multiple indices.

In Section 2 we present the identification argument and discuss its required conditions. In Sections 3 and 4 we provide some examples in which such a monotonicity condition fails, but the average effect of the binary endogenous variable is still identified. To reiterate, we allow the potential outcomes to be weakly separable in multiple indices, that is, $v(X, D) = (v_1(X, D), v_2(X, D), ..., v_J(X, D)) \in \mathbb{R}^J$. We consider the identification and estimation of the average treatment effect of D on Y, $E(Y_1|X \in A)$, $E(Y_0|X \in A)$, and $E(Y_1 - Y_0|X \in A)$, for some set A, without the aforementioned monotonicity. Indeed, for the case with multiple indices $v(X, D) \in \mathbb{R}^J$, the monotonicity condition is no longer well defined.

2 Identification

Our identification strategy is based on the notion of *matching covariates.* Let $Supp(V)$ denote the support of a generic random vector V. Consider the identification of $E(Y_1|X = x)$ for some $x \in Supp(X)$. Under the assumption that $(\varepsilon, U)\perp(X, Z)$,

$$\begin{aligned} &E(Y_1|X = x) = E(Y_1|X = x, Z = z) \\ &= E(DY_1|X = x, Z = z) + E[(1 - D)Y_1|X = x, Z = z] \\ &= P(z)E(Y|D = 1, X = x, Z = z) + [1 - P(z)]E(Y_1|D = 0, X = x, Z = z) \end{aligned} \tag{2.1}$$

where $P(z) \equiv E(D|Z = z) = \theta(z)$ because U is normalized to follow a standard uniform distribution. The only term not directly identifiable on the right-hand side of (2.1) is:[4]

$$E(Y_1|D = 0, X = x, Z = z) = E[g(v(x, 1), \varepsilon)|U \geq P(z)].$$

[4] If the support of $P(Z)$ given x covers the full closed interval $[0, 1]$, then $E(Y_1|X = x)$ is directly identified as $E(Y|D = 1, X = x, Z = z)$ at z s.t. $P(z) = 1$. However this means point identification hinges on the event "$P(Z) = 1$ given x".

The main idea behind our approach would at first seem to be similar to that of Vytlacil and Yildiz (2007), which is to use exogenous variations of covariates in the outcome equation to find some $(\tilde{x}, \tilde{z}) \in Supp(X, Z)$ such that

$$P(z) = P(\tilde{z}) \text{ and } v(x, 1) = v(\tilde{x}, 0) \tag{2.2}$$

so that

$$\begin{aligned} E(Y|D = 0, X = \tilde{x}, Z = \tilde{z}) &= E(Y_0|D = 0, X = \tilde{x}, Z = \tilde{z}) \\ &= E[g(v(\tilde{x}, 0), \varepsilon)|U \geq P(\tilde{z})] = E[g(v(x, 1), \varepsilon)|U \geq P(z)] \\ &= E[Y_1|D = 0, X = x, Z = z]. \end{aligned} \tag{2.3}$$

However, unlike Vytlacil and Yildiz (2007), we utilize the full distribution of Y (rather than its first moment only) while searching for such pairs of $(x, \tilde{x})$ in (2.2). This allows us to relax the single-index and monotonicity conditions.

To fix ideas, let there be continuous components in Z that are excluded from X. For any p on the support of $P(Z)$ given $X = x$, and for any y, define

$$\begin{aligned} h_1^*(x, y, p) &= E[D1\{Y \leq y\}|X = x, P(Z) = p] \\ &= E[1\{U < P(Z)\}1\{g(v(X, 1), \varepsilon) \leq y\}|X = x, P(Z) = p] \\ &= \int_0^p F_{g|u}(y; v(x, 1))du, \end{aligned} \tag{2.4}$$

where

$$F_{g|u}(y; v(x, d)) \equiv E[1\{g(v(x, d), \varepsilon) \leq y\}|U = u],$$

with $v(x, d)$ being a realized index at $X = x$ and the expectation in the definition of $F_{g|u}$ is with respect to the distribution of ε given $U = u$. The last equality in (2.4) holds because of independence between (ε, U) and (X, Z). Assume that for all y, x and $d \in \{0, 1\}$, the function $F_{g|u}(y; v(x, d))$ is continuous in u over $[0, 1]$. This implies h_1^* is differentiable in p. By construction, $h_1^*(x, y, p)$ is directly identified from the joint distribution of (D, Y, X, Z) in the data-generating process. Furthermore, for any pair $p_1 > p_2$, define:

$$h_1(x, y, p_1, p_2) \equiv h_1^*(x, y, p_1) - h_1^*(x, y, p_2) = \int_{p_2}^{p_1} F_{g|u}(y; v(x, 1))du.$$

Likewise, define

$$\begin{aligned} h_0^*(x, y, p) &= E((1 - D)1\{Y \leq y\}|X = x, P(Z) = p) \\ &= E[1\{U \geq P(Z)\}1\{g(v(X, 0), \varepsilon) \leq y\}|X = x, P(Z) = p] \\ &= \int_p^1 F_{g|u}(y; v(x, 0))du. \end{aligned}$$

and let

$$h_0(x,y,p_1,p_2) \equiv h_0^*(x,y,p_2) - h_0^*(x,y,p_1) = \int_{p_2}^{p_1} F_{g|u}(y;v(x,0))du.$$

Let $\mathcal{P}_x$ denote the support of $P(Z)$ given $X=x$; let $Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})$ denote the interior of the intersection of $\mathcal{P}_x$ and $\mathcal{P}_{\tilde{x}}$. Similar to Heckman and Vytlacil (2005) and Carneiro and Lee (2009), we use continuous, exogenous variation in Z to establish here that for any $x \neq \tilde{x}$ with $Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}}) \neq \emptyset$ and any y,

$$h_1(x,y,p,p') = h_0(\tilde{x},y,p,p') \text{ for all } p > p' \text{ on } Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}}) \tag{2.5}$$

if and only if

$$F_{g|p}(y;v(x,1)) = F_{g|p}(y;v(\tilde{x},0)) \text{ for all } p \in Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}}). \tag{2.6}$$

Sufficiency of (2.6) is immediate from the definition of h_1 and h_0. To see its necessity, note that for all $p > p'$ on $Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})$,

$$\left.\frac{\partial h_1(x,y,\tilde{p},p')}{\partial \tilde{p}}\right|_{\tilde{p}=p} = \left.\frac{\partial h_1^*(x,y,\tilde{p})}{\partial \tilde{p}}\right|_{\tilde{p}=p} = F_{g|p}(y;v(x,1))$$

and

$$\left.\frac{\partial h_0(\tilde{x},y,\tilde{p},p')}{\partial \tilde{p}}\right|_{\tilde{p}=p} = -\left.\frac{\partial}{\partial \tilde{p}} h_0^*(\tilde{x},y,\tilde{p})\right|_{\tilde{p}=p} = F_{g|p}(y;v(\tilde{x},0)).$$

Thus (2.5) and (2.6) are equivalent.

Next, we collect identifying assumptions as follows:

ASSUMPTION A-1: The distribution of U is absolutely continuous with respect to Lebesgue measure, and is normalized to standard uniform, and $\theta(z)$ in (1.2) is continuous in z.

ASSUMPTION A-2: The random vectors (U,ε) and (X,Z) are independent. There exists at least one continuously distributed component in Z that is excluded from X.

ASSUMPTION A-3: Both $g(v(X,1),\varepsilon)$ and $g(v(X,0),\varepsilon)$ have finite first moments conditional on $U=u$ for all $u \in [0,1]$.

ASSUMPTION A-4: For all realized values of y,x and $d \in \{0,1\}$, $E[1\{g(v(x,d),\epsilon) \leq y\}|U=u]$ is continuous in u over $[0,1]$.

ASSUMPTION A-5: For any $x \neq \tilde{x}$ such that $Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})$ is nonempty, $F_{g|p}(y;v(x,1)) = F_{g|p}(y;v(\tilde{x},0))$ for all y and $p \in Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})$ if and only if $v(x,1) = v(\tilde{x},0)$.

Assumptions A-1 and A-3 are common regularity conditions in the literature. Assumption A-2 consists of a standard condition of instrument exogeneity. It allows $P(Z)$ to have continuous, exogenous variation conditional on X. The continuity of instrument is also common in the literature on treatment effects. Examples include Heckman and Vytlacil (2007b) and Vytlacil and Yildiz (2007). As noted above, Assumption A-4 ensures the identifiable functions $h_1^*(x,y,p), h_0^*(x,y,p)$ are both differentiable in the last argument p.

It is important to note that Assumption A-5 allows the support of $P(Z)$ to be a strict subset of the interval $(0,1)$. This is because our method does not use an identification-at-infinity argument, which would require $P(Z)$ to have full support [0,1] given X in order to identify $E(Y_0|X)$ directly. (See Footnote 4 above.) We also note that Assumption A-5 relaxes two limitations of Assumption 4 in Vytlacil and Yildiz (2007). Specifically, to identify pairs $(x,\tilde{x})$ with $v(x,1)=v(\tilde{x},0)$, Vytlacil and Yildiz (2007) rely on two assumptions that $v(X,D)\in\mathbb{R}$ is a single index and that $E\left[g(v(x,1),\varepsilon)|U=u\right]=E\left[g(v(\tilde{x},0),\varepsilon)|U=u\right]$ if and only if $v(x,1)=v(\tilde{x},0)$. The latter holds under a maintained assumption that $E\left[g(v(x,d),\varepsilon)|U=p\right]$ is a strictly monotonic function of $v(x,d)$. In comparison, we construct an identification strategy without the single index and monotonicity restrictions by matching conditional distributions $F_{g|p}(\cdot;v(x,1))$ and $F_{g|p}(\cdot;v(\tilde{x},0))$.

The role of Assumption A-5 in our method can be illustrated by drawing an analogy with a standard identifying condition in nonlinear regression: $Y=f(X,\theta^*)+\varepsilon$. In this case, identification of θ^* requires: "$f(x,\theta^*)=f(x,\theta)$ for all x if and only if $\theta^*=\theta$ ". To see how this is related to Assumption A-5 in our setting, let θ_1,θ_0 be shorthand for $v(x,1),v(\tilde{x},0)$ respectively, and let $m(y,p,\theta)\equiv F_{g|p}(y,\theta)$. Then our method requires: "$m(y,p,\theta_1)=m(y,p,\theta_0)$ for all (y,p) if and only if $\theta_1=\theta_0$ ".

It is worth emphasizing that Assumption A-5 only presents our identification condition in the weakest form, for the sake of generality. Later in Section 3, we exploit the structures embedded in specific examples to show how Assumption A-5 can be satisfied under intuitive, mild primitive conditions.

Next, we specify conditions under which such pairs of covariates exist on the support. Define $\mathcal{S}\equiv\{(x,\tilde{x}):v(x,1)=v(\tilde{x},0)\}$ and $\mathcal{T}\equiv\{(x,\tilde{x}):\exists z,\tilde{z}$ with $(x,z),(\tilde{x},\tilde{z})\in Supp(X,Z)$ and $P(z)=P(\tilde{z})\in Int(\mathcal{P}_x\cap\mathcal{P}_{\tilde{x}})\}$. Let $\mathcal{X}^1\equiv\{x:\exists\tilde{x}$ with $(x,\tilde{x})\in\mathcal{S}\cap\mathcal{T}\}$ and $\mathcal{X}^0\equiv\{x:\exists\tilde{x}$ with $(\tilde{x},x)\in\mathcal{S}\cap\mathcal{T}\}$.

ASSUMPTION A-6: $\Pr(X\in\mathcal{X}^1)>0$ and $\Pr(X\in\mathcal{X}^0)>0$.

This condition is similar to Assumption A-4 in Vytlacil and Yildiz (2007). By definition, for each $x\in\mathcal{X}^1$, we can find $\tilde{x}\in Supp(X)$ such that there exists $(z,\tilde{z})$ with $(x,z),(\tilde{x},\tilde{z})\in$

$Supp(X,Z)$, $v(x,1) = v(\tilde{x},0)$, and $P(z) = P(\tilde{z})$. This requires variation in x while holding $P(Z)$ fixed. As noted earlier (Footnote 2), this does *not* necessarily require a "second exclusion restriction" that there is an element in X that is excluded from Z. In general, there exist multiple such values of $\tilde{x}$ that can be matched with this x. Hence for each $x \in \mathcal{X}^1$, we define the set of such matched values as

$$\lambda_0(x) \equiv \{\tilde{x} : h_1(x,y,p,p') = h_0(\tilde{x},y,p,p') \ \forall p > p' \text{ on } Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})\},$$

and define $\mathcal{P}_0^*(x) \equiv \bigcup_{\tilde{x}\in\lambda_0(x)}(\mathcal{P}_{\tilde{x}} \cap \mathcal{P}_x)$. Note that by definition of $\mathcal{X}^1$, the set $\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}}$ must be non-empty for $x \in \mathcal{X}^1$ and $\tilde{x} \in \lambda_0(x)$ under our maintained assumptions. Likewise, by symmetry, for each $x \in \mathcal{X}^0$, we define

$$\lambda_1(x) \equiv \{\tilde{x} : h_0(x,y,p,p') = h_1(\tilde{x},y,p,p') \ \forall p > p' \text{ on } Int(\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}})\},$$

and let $\mathcal{P}_1^*(x) \equiv \bigcup_{\tilde{x}\in\lambda_1(x)}(\mathcal{P}_{\tilde{x}} \cap \mathcal{P}_x)$.

The theorem below shows how to use such matched values to identify the conditional mean of potential outcomes.

Theorem 2.1. *Suppose Assumptions (A-1)-(A-6) hold. For each $x \in \mathcal{X}^1$,*

$$E(Y_1|X=x) = E(DY|X=x, P(Z)=p) + E[(1-D)Y|X \in \lambda_0(x), P(Z)=p] \tag{2.7}$$

for any $p \in \mathcal{P}_0^*(x)$. For each $x \in \mathcal{X}^0$,

$$E(Y_0|X=x) = E(DY|X \in \lambda_1(x), P(Z)=p) + E[(1-D)Y|X=x, P(Z)=p] \tag{2.8}$$

for any $p \in \mathcal{P}_1^*(x)$.

Proof. As established in the text, (2.5) and (2.6) are equivalent under Assumptions (A-1)-(A-4). Then by Assumption (A-5), we know that for each $x \in \mathcal{X}^1$ and $\tilde{x} \in \lambda_0(x)$, there exists $z, \tilde{z}$ such that $(x,z), (\tilde{x},\tilde{z}) \in Supp(X,Z)$, $v(x,1) = v(\tilde{x},0)$ and $P(z) = P(\tilde{z}) = p \in \mathcal{P}_0^*(x)$. It follows from (2.1) and (2.3) that

$$E(Y_1|X=x) = E(DY|X=x, Z=z) + E[(1-D)Y|X=\tilde{x}, Z=\tilde{z}]. \tag{2.9}$$

By (A-2), the right-hand side of (2.9) equals the right-hand side of (2.7). A symmetric argument proves (2.8). ■

It is worth mentioning that the conditions for this theorem do not directly restrict the support of potential outcomes. As we show in the next section, our method applies in important applications where the potential outcome is either discrete (e.g., determined by multinomial choices), or multi-dimensional with both discrete and continuous components (e.g., determined in a Roy model).

3 Examples

In this section, we present several examples in which the potential outcomes depend on multi-dimensional indices with an endogenous treatment. More importantly, we show how the specific structure embedded in each application naturally leads to transparent, primitive conditions that imply Assumption A-5, thus corroborating the wide applicability of this general approach.

In the first and third examples, the monotonicity condition in Vytlacil and Yildiz (2007) does not hold; in the second example, the identification requires a generalization of the monotonicity condition into an invertibility condition in higher dimensions.

Example 1. (Heteroskaedastic shocks in Uncensored or Censored outcomes) First, consider a triangular system where a continuous *uncensored* outcome is determined by double indices $v(X,D) \equiv (v_1(X,D), v_2(X,D))$:[5]

$$Y = g(v(X,D),\varepsilon) = v_1(X,D) + v_2(X,D)\varepsilon \text{ for } D \in \{0,1\}.$$

The selection equation determining the actual treatment is the same as (1.2). In this case the concept of monotonicity in $v \in \mathbb{R}^2$ is not well-defined, so the procedure proposed in Vytlacil and Yildiz (2007) is not suitable here.[6] Nevertheless, we can apply the method in Section 2 to identify the average treatment effect by using the *distribution* of outcomes to find pairs of x and $\tilde{x}$ such that $v(x,1) = v(\tilde{x},0)$.

To see how Assumption A-5 holds, assume the range of $v_2(\cdot)$ is positive. Note that

$$\begin{aligned} F_{g|u}(y; v(x,d)) &= \Pr\left[v_1(x,d) + v_2(x,d)\varepsilon \leq y | U = u\right] \\ &= F_{\varepsilon|u}\left(\frac{y - v_1(x,d)}{v_2(x,d)}\right) \end{aligned}$$

for $d = 0, 1$. Suppose the distribution of ε conditional on u is increasing over $\mathbb{R}$. Then for

[5] Abrevaya and Xu (2022) adopted a different approach to identify ATE in the same model with uncensored, continuous outcome and multiplicative, heteroskaedastic shocks above, using a binary instrument Z. They showed the conditional mean of the observed outcome Y, when scaled by the conditional covariance of $Y1\{D = d\}$ and Z, is a weighted sum of the conditional means of potential outcomes: $v_1(x,0)$ and $v_1(x,1)$. Thus, using the means of the scaled Y conditional on $Z = 0, 1$ respectively, one can construct a linear system that identifies $v_1(x,0)$ and $v_1(x,1)$, provided the proper rank condition holds. Their method leverages this particular specification of multiplicative endogeneity, but does not generalize to the case of censored outcomes or in the other examples we consider later.

[6] For this particular design, the approach proposed in Vuong and Xu (2017) should be valid. But it will not be for a slightly modified model, such as $Y = v_1(X,D) + (e_2 + v_2(X,D) * e_1)$, whereas ours will be.

all y and $x \in \mathcal{X}^1$ and matched values $\tilde{x} \in \lambda_0(x)$,

$$F_{g|u}(y; v(x,1)) = F_{g|u}(y; v(\tilde{x},0)) \text{ if and only if } \frac{y - v_1(x,1)}{v_2(x,1)} = \frac{y - v_1(\tilde{x},0)}{v_2(\tilde{x},0)}.$$

Differentiating with respect to y yields $v_2(x,1) = v_2(\tilde{x},0)$, which implies $v_1(x,1) = v_1(\tilde{x},0)$.

Next, consider a similar model with continuous *censored* outcomes, where

$$Y = g(v(X,D),\varepsilon) = max\{v_1(X,D) + v_2(X,D)\varepsilon, 0\} \text{ for } D \in \{0,1\}.$$

In this case, our identification argument for the uncensored case above applies almost immediately. The only adjustment needed is to confine the values of y used for constructing matched pairs $(x,\tilde{x})$ over the *uncensored* segment, i.e. $y > 0$. In contrast, other methods for identifying ATE in the uncensored model which rely strictly on the linear structure with multiplicative heterogeneity, such as Abrevaya and Xu (2022), no longer applies.

Example 2. (Multinomial potential outcome) Consider a triangular system where the outcome is multinomial. The multinomial response model has a long and rich history in both applied and theoretical econometrics. Recent examples in the semiparametric literature include Lee (1995), Pakes and Porter (2014), Ahn, Powell, Ichimura, and Ruud (2017), Shi, Shum, and Song (2018), and Khan, Ouyang, and Tamer (2019). None of those papers allow for dummy endogenous variables or potential outcomes.

In this example, let the observed outcome be

$$Y = g(v(X,D),\varepsilon) = \arg\max_{j=0,1,...,J} y^*_{j,D},$$

where

$$y^*_{j,D} = v_j(X,D) + \varepsilon_j \text{ for } j = 1,2,...,J;\ y^*_{0,D} = 0.$$

In this case, the index $v \equiv (v_j)_{j \leq J}$ and the errors $\varepsilon \equiv (\varepsilon_j)_{j \leq J}$ are both J-dimensional. The selection equation that determines D is the same as (1.2). In this case, we can replace $1\{Y \leq y\}$ by $1\{Y = y\}$ in the definition of h_1, h_0, h^*_1, h^*_0 and $F_{g|u}(\cdot; v)$. Then for $d = 0,1$ and $j \leq J$,

$$\begin{aligned} F_{g|u}(j; v(x,d)) &\equiv E[1\{g(v(x,d),\varepsilon) = j\}|U = u] \\ &= \Pr\{v_j(x,d) + \varepsilon_j \geq v_{j'}(x,d) + \varepsilon_{j'}\ \forall j' \leq J \mid U = u\}. \end{aligned}$$

By Ruud (2000) and Ahn, Powell, Ichimura, and Ruud (2017), the mapping from $v \in \mathbb{R}^J$ to $(F_{g|u}(j; v) : j \leq J) \in \mathbb{R}^J$ is smooth and invertible provided that $\varepsilon \in \mathbb{R}^J$ has non-negative density everywhere. This implies Assumption A-5.

Example 3. **(Potential outcome from a Roy model)** Consider a treatment effect model with an endogenous binary treatment D and with the potential outcome determined by a latent Roy model. The Roy model has also been studied extensively from both applied and theoretical perspectives. See, for example, the literature survey in Heckman and Vytlacil (2007a) and the seminal paper in Heckman and Honoré (1990).

Here the observed outcome consists of two pieces: a continuous measure $Y = DY_1 + (1 - D)Y_0$ and a discrete indicator $W = DW_1 + (1 - D)W_0$ for $d = 0, 1$. These potential outcomes are given by

$$Y_d = \max_{j \in \{a,b\}} y^*_{j,d} \text{ and } W_d = \arg \max_{j \in \{a,b\}} y^*_{j,d}$$

where a and b index potential outcomes realized in different sectors, with

$$y^*_{j,d} = v_j(X, d) + \varepsilon_j \text{ for } j \in \{a, b\}.$$

The binary endogenous treatment D is determined as in equation (1.2). For example, $D \in \{1, 0\}$ indicates whether an individual participates in a professional training program, $W_d \in \{a, b\}$ indicates the potential sector in which the individual is employed, $y^*_{j,d}$ is the potential wage from sector j under treatment $D = d$, and $Y_d \in \mathbb{R}$ is the potential wage if the treatment status is $D = d$.[7]

As before, we maintain that $(X, Z) \perp (\varepsilon, U)$. The parameter of interest is

$$\Pr\{Y_1 \leq y, W_1 = a | X\}.$$

By the independence condition that $(X, Z) \perp (\varepsilon, U)$ and an application of the law of total probability, this conditional probability can be expressed in terms of directly identifiable quantities and the following counterfactual quantity

$$\begin{aligned} & \Pr\{Y_1 \leq y, W_1 = a \mid X = x, Z = z, D = 0\} \\ = \; & \Pr\left\{v_b(x, 1) + \varepsilon_b < v_a(x, 1) + \varepsilon_a \leq y \mid U \geq P(z)\right\}. \end{aligned} \tag{3.1}$$

Again, we seek to identify this counterfactual quantity by finding $\tilde{x}$ such that there exists $\tilde{z}$ with $(\tilde{x}, \tilde{z}) \in Supp(X, Z)$, $P(z) = P(\tilde{z})$, and

$$v_a(x, 1) = v_a(\tilde{x}, 0) \text{ and } v_b(x, 1) = v_b(\tilde{x}, 0). \tag{3.2}$$

[7] Note this differs from the classical approach that formulates treatment effects using Roy models (e.g. Heckman, Urzua, and Vytlacil (2006)) in that the potential outcome Y_d itself is determined by a latent Roy model.

This would allow us to recover the counterfactual conditional probability in (3.1) as

$$\Pr\{Y_0 \leq y, W_0 = a \mid X = \tilde{x}, Z = \tilde{z}, D = 0\}.$$

To find such a pair of $(x, \tilde{x})$, define $h_{d,W}(x, p, p'), h^*_{d,W}(x, p)$ by replacing $1\{Y \leq y\}$ with $1\{W = a\}$ in the definition of h_d, h^*_d in Section 2. Similarly, define $h_{d,Y}(x, y, p, p'), h^*_{d,Y}(x, y, p)$ by replacing $1\{Y \leq y\}$ with $1\{Y \leq y, W = a\}$ in the definition of h_d, h^*_d in Section 2. Then

$$\begin{aligned} h_{d,W}(x, p_1, p_2) &= \int_{p_2}^{p_1} \Pr\{v_b(x, d) + \varepsilon_b < v_a(x, d) + \varepsilon_a | U = u\} du; \\ h_{d,Y}(x, y, p_1, p_2) &= \int_{p_2}^{p_1} \Pr\{v_b(x, d) + \varepsilon_b < v_a(x, d) + \varepsilon_a \leq y | U = u\} du; \end{aligned}$$

and $h_{d,W}(x, p_1, p_2)$ and $h_{d,Y}(x, y, p_1, p_2)$ are both identified over their respective domains by construction.

Assume $(\varepsilon_a, \varepsilon_b)$ is continuously distributed with positive density over $\mathbb{R}^2$ conditional on all u, and the distribution of $(\varepsilon_0, \varepsilon_1)$ given $U = u$ is continuous in u over $[0, 1]$. Then the statement

"$h_{1,W}(x, p, p') = h_{0,W}(\tilde{x}, p, p')$, $h_{1,Y}(x, y, p, p') = h_{0,Y}(\tilde{x}, y, p, p')$ for all y and $p > p'$ on $\mathcal{P}_x \cap \mathcal{P}_{\tilde{x}}$"

holds true if and only if (3.2) holds. To see this, first note that matching $h_{1,W}(x, p, p') = h_{0,W}(\tilde{x}, p, p')$ requires

$$v_a(x, 1) - v_b(x, 1) = v_a(\tilde{x}, 0) - v_b(\tilde{x}, 0), \tag{3.3}$$

while matching $h_{1,Y}(x, y, p, p') = h_{0,Y}(\tilde{x}, y, p, p')$ at the same time requires

$$v_a(x, 1) = v_a(\tilde{x}, 0). \tag{3.4}$$

Thus requiring (3.3) and (3.4) to hold jointly is equivalent to (3.2).

It is worth mentioning that the identification strategy above also applies in a more general setup where the error term in the potential outcome is specific to the treatment, i.e., with ε_j replaced by $\varepsilon_{j,d}$ in $y^*_{j,d}$. In such cases, the argument above remains valid under a "rank similarity" condition (that the marginal distribution of $\varepsilon_{j,d}$ is the same for $d \in \{0, 1\}$). The rank similarity condition has been used for identifying treatment effects in instrumental quantile regression models such as Chernozhukov and Hansen (2006).

4 Extension

The identification strategy we use requires finding matched pairs for x in $\mathcal{X}^1$ and $\mathcal{X}^0$. In some cases, with the outcome being continuous, we can construct similar arguments for identifying a counterfactual quantity in a treatment effect model by matching different elements on the support of continuous outcomes. To the best of our knowledge, this approach has not been explored in the literature on the effects of endogenous treatments. The following example illustrates this point.

Example 4. (Potential outcome with random coefficients) Random coefficient models are prominent in both the theoretical and applied econometrics literature. They permit a flexible way to allow for conditional heteroscedasticity and unobserved heterogeneity. For a survey and recent developments, see Hsiao and Pesaran (2008), Hoderlein, Klemelä, and Mammen (2010), Arellano and Bonhomme (2012), and Masten (2018).

We consider a treatment effect model where the potential outcome is determined through random coefficients:

$$Y = DY_1 + (1 - D)Y_0 \text{ where } Y_d = (\alpha_d + X'\beta_d) \text{ for } d = 0, 1$$

and the binary endogenous treatment D is determined as in equation (1.2). The *random* intercepts $\alpha_d \in \mathbb{R}$ and the *random* vectors of coefficients β_d are given by

$$\alpha_d = \bar{\alpha}_d(X) + \eta_d \text{ and } \beta_d = \bar{\beta}_d(X) + \varepsilon_d$$

where for any $x \in Supp(X)$ and $d \in \{0, 1\}$, $(\bar{\alpha}_d(x), \bar{\beta}_d(x)) \in \mathbb{R}^{K+1}$ is a vector of constant parameters while $\eta_d \in \mathbb{R}$ and $\varepsilon_d \in \mathbb{R}^K$ are unobservable noises.

As before, suppose some elements of Z in the treatment equation are excluded from X. We allow the vector of unobservable terms $(\varepsilon_1, \varepsilon_0, \eta_0, \eta_1, U)$ to be arbitrarily correlated, and assume:

$$(X, Z) \perp (\varepsilon_1, \varepsilon_0, \eta_0, \eta_1, U). \tag{4.1}$$

Aslo, normalize the marginal distribution of U to standard uniform, so that $\theta(Z)$ is directly identified as $P(Z) \equiv E(D|Z)$.

Our goal is to identify the distribution of potential outcomes Y_d given $X = x$ for $d = 0, 1$. From this result, we can recover other parameters of interest such as average treatment effects, quantile treatment effects, etc. As a preliminary step, we start by pinpointing a counterfactual item that is crucial for this identification question.

Let $G_{P|x}$ denote the distribution of $P \equiv P(Z)$ given $X = x$ (recall that its support is denoted as $\mathcal{P}_x$). This conditional distribution is directly identifiable from the data-generating process. By construction,

$$\Pr\{Y_1 \leq y|X = x\} = \int \Pr\{Y_1 \leq y|X = x, P = p\} dG_{P|x}(p),$$

where

$$\begin{aligned} & \Pr\{Y_1 \leq y|X = x, P = p\} \\ = \; & E\left[D1\{Y_1 \leq y\}|X = x, P = p\right] + E\left[(1 - D)1\{Y_1 \leq y\}|X = x, P = p\right]. \end{aligned} \tag{4.2}$$

The first term on the right-hand side of (4.2) is identified as

$$E[D1\{Y \leq y\}|X = x, P = p].$$

The second term on the right-hand side of (4.2), denoted by $\phi_0(x, y, p)$, is counterfactual and can be written as

$$\begin{aligned} \phi_0(x, y, p) &\equiv E[1\{U \geq P\}1\{\alpha_1 + X'\beta_1 \leq y\}|X = x, P = p] \\ &= E[1\{U \geq p\}1\{\bar{\alpha}_1(x) + \eta_1 + x'(\bar{\beta}_1(x) + \varepsilon_1) \leq y\}] \\ &= \int_p^1 \Pr\{\eta_1 + x'\varepsilon_1 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U = u\} du. \end{aligned}$$

Hence identification of the conditional distribution of Y_1 amounts to identification of $\phi_0(\cdot)$.

As noted at the beginning of this section, we will identify the counterfactual $\phi_0(\cdot)$ by finding matched elements on the support of the observed outcome Y. This takes two steps. First, we show that if for a given pair (x, y) one can find $t(x, y)$ such that

$$y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) = t(x, y) - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x), \tag{4.3}$$

then one can use $t(x, y)$ to identify $\phi_0(x, y, p)$ for any $p \in \mathcal{P}_x$. Specifically, for any p on the support of P given $X = x$, define

$$\begin{aligned} h_1^*(x, y, p) &\equiv E\left[D1\left\{Y \leq y\right\}|X = x, P = p\right] \\ &= E\left[1\left\{U < P\right\}1\left\{\alpha_1 + X'\beta_1 \leq y\right\}|X = x, P = p\right] \\ &= E\left[1\left\{U < p\right\}1\left\{\alpha_1 + x'\beta_1 \leq y\right\}\right] \\ &= \int_0^p \Pr\{\eta_1 + x'\varepsilon_1 \leq y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U = u\} du, \end{aligned}$$

where the second equality uses (4.1). Likewise, under (4.1) we have:

$$
\begin{aligned}
h_0^*(x,y,p) &\equiv E\left[(1-D)1\left\{Y\le y\right\}|X=x,P=p\right] \\
&= \int_p^1 \Pr\{\eta_0 + x'\varepsilon_0 \le y - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x)|U=u\}du.
\end{aligned}
$$

Assume[8]

$$
F_{(\eta_1,\varepsilon_1)|U=u} = F_{(\eta_0,\varepsilon_0)|U=u} \text{ for all } u\in[0,1]. \tag{4.4}
$$

Under (4.4), we have

$$
\phi_0(x,y,p) = \int_p^1 \Pr\{\eta_0 + x'\varepsilon_0 \le y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du. \tag{4.5}
$$

Then by definition of $t(x,y)$ in (4.3),

$$
\begin{aligned}
h_0^*(x,t(x,y),p) &\equiv \int_p^1 \Pr\{\eta_0 + x'\varepsilon_0 \le t(x,y) - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x)|U=u\}du \\
&= \int_p^1 \Pr\{\eta_0 + x'\varepsilon_0 \le y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du \\
&= \phi_0(x,y,p).
\end{aligned}
$$

Thus the counterfactual $\phi_0(x,y,p)$ would be identified as $h_0^*(x,t(x,y),p)$.

The second step is to show that for each pair (x,y) we can indeed uniquely recover $t(x,y)$ using quantities that are identifiable in the data-generating process. To do so, we define two auxiliary functions as follows: for $p_1 > p_2$ on the support of P given $X=x$, let

$$
\begin{aligned}
h_1(x,y,p_1,p_2) &\equiv h_1^*(x,y,p_1) - h_1^*(x,y,p_2) \\
&= \int_{p_2}^{p_1} \Pr\{\eta_1 + x'\varepsilon_1 < y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x)|U=u\}du;
\end{aligned}
$$

and

$$
\begin{aligned}
h_0(x,y,p_1,p_2) &\equiv h_0^*(x,y,p_2) - h_0^*(x,y,p_1) \\
&= \int_{p_2}^{p_1} \Pr\{\eta_0 + x'\varepsilon_0 < y - \bar{\alpha}_0(x) - x'\bar{\beta}_0(x)|U=u\}du.
\end{aligned}
$$

[8] Such a distributional equality condition has been used to motivate the *rank similarity* condition imposed frequently in the econometrics literature – see, for example, Chernozhukov and Hansen (2005), Vytlacil and Yildiz (2007), Chen and Khan (2014), Frandsen and Lefgren (2018), Dong and Shen (2018).

Suppose $\eta_d + x'\varepsilon_d$ is continuously distributed over $\mathbb{R}$ for all values of x conditional on all $u \in [0,1]$. Then for any fixed pair (x,y) and $p_1 > p_2$,

$$h_1(x,y,p_1,p_2) = h_0(x,t(x,y),p_1,p_2)$$

if and only if

$$t(x,y) = y - \bar{\alpha}_1(x) - x'\bar{\beta}_1(x) + \bar{\alpha}_0(x) + x'\bar{\beta}_0(x).$$

5 Estimation

In this section, we outline estimation procedures from a random sample of the observed variables that are motivated by our identification results. We first describe an estimation procedure for the parameter $E[Y_1]$ in the first three examples. Recall $\mathcal{P}_x$ denotes the support of $P(Z) \equiv P$ given $X = x$. Let $f_P(.|x)$ denote the density of $P(Z)$ given $X = x$, and define

$$\mathcal{P}_x^c \equiv \{p\colon\ f_P(p|x) > c\} \text{ for a known } c > 0.$$

For simplicity, assume

$$1 - c_0 > P(Z) > c_0 \text{ for a known } c_0 > 0 \text{ almost surely.}$$

Define a measure of distance between $h_1(x_1,\cdot)$ and $h_0(x_0,\cdot)$ as follows:

$$\begin{aligned}&\|h_1(x_1,\cdot) - h_0(x_0,\cdot)\|\\ =\ &\left\{\int\int\int\left(\int_{p_2}^{p_1}(F_{g|u}(y;v(x_1,1)) - F_{g|u}(y;v(x_0,0))du\right)^2 I\left(p_1,p_2 \in \mathcal{P}_x^c\right) w(y)dydp_1dp_2\right\}^{1/2}\end{aligned}$$

where $w(y)$ is a chosen weight function.

Consider the case when $h_0(x,y,p_1,p_2)$, $h_1(x,y,p_1,p_2)$ and $P(z)$ are known. For any given x_i, let $\tilde{x}_i$ be such that

$$\|h_0(\tilde{x}_i,\cdot) - h_1(x_i,\cdot)\| = 0,$$

which, under Assumption A-5 in Section 2, is equivalent to

$$v(\tilde{x}_i,0) = v(x_i,1).$$

Let P_i be shorthand for $P(Z_i)$ and define

$$Y_i^* \equiv E[Y|D=0, \|h_0(X,\cdot) - h_1(X_i,\cdot)\| = 0, P = P_i].$$

This conditional expectation equals $E[Y|D = 0, v(X, 0) = v(X_i, 1), P = P_i]$, which in turn equals $E(Y_1|D = 0, X = X_i, P = P_i)$.

The parameter of interest $\Delta \equiv E[Y_1]$ can be written as $\Delta = E[D_iY_i + (1 - D_i)Y_i^*]$. Therefore, we estimate Δ by its sample analog, after replacing Y_i^* with its Nadaraya-Watson estimates. That is, we estimate Δ by

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(D_iY_i + (1 - D_i)\hat{Y}_i\right),$$

where $\hat{Y}_i$ is a kernel regression estimator of Y_i^*, using knowledge of h_0, h_1 and $P(Z)$. Likewise, for estimating the conditional mean $E(Y_1|X \in A)$ where A is a generic subset of the support of X, we use a weighted version

$$\hat{\Delta}_A = \frac{\frac{1}{n}\sum_{i=1}^{n} 1\left\{X_i \in A\right\}\left(D_iY_i + (1 - D_i)\hat{Y}_i\right)}{\frac{1}{n}\sum_{i=1}^{n} 1\left\{X_i \in A\right\}}.$$

Limiting distribution theory for each of these estimators follows from identical arguments in Vytlacil and Yildiz (2007). Here we formally state the theorem for the first estimator:

Theorem 5.1. *Suppose Assumptions (A-1) to (A-6) hold, and Y_1 has positive and finite second moments. Then*

$$\sqrt{n}(\hat{\Delta} - \Delta) \overset{d}{\to} \mathbb{N}(0, \Sigma),$$

where

$$\Sigma = Var(E[Y_1|X, P, D]) + E[PVar(Y_1|X, P, D = 1)].$$

Next, we describe an estimation procedure for the distributional treatment effect in Example 4, where potential outcomes depend on random coefficients. In this case, the parameter of interest is, for a given value $y \in \mathbb{R}$,

$$\Delta_2(y) = \Pr\{Y_1 \leq y\}.$$

For fixed values of y and $p_1 > p_2$, we propose to estimate $t(x, y)$ as

$$\hat{t}(x, y, p_1, p_2) = \arg\min_t (h_1(x, y, p_1, p_2) - h_0(x, t, p_1, p_2))^2,$$

and then average over values of p_1, p_2:

$$\hat{\tau}(x, y) = \frac{1}{n(n-1)}\sum_{i \neq j} I[P_i > P_j]\hat{t}(x, y, P_i, P_j).$$

An infeasible estimator for $\Delta_2(y)$, which assumes $t(x,y)$ is known, would be

$$\hat{\Delta}_2(y) = \frac{1}{n}\sum_{i=1}^{n}\left(D_i 1\{Y_i \leq y\} + (1-D_i)1\{Y_i \leq t(X_i,y)\}\right).$$

In practice, for feasible estimation, one needs to replace $t(x,y)$ with its estimator $\hat{\tau}(x,y)$.

We conclude this section with some discussion about the computational aspects. The computational costs for finding the suitable pairs $(x,\tilde{x})$ are modest in comparison with typical semi- or nonparametric methods in the literature. To see this, consider the case with $J=2$, where $v=(v_1,v_2)\in\mathbb{R}^2$ consists of two index functions $v_j(x,d):\mathbb{R}^K\times\{0,1\}\to R$ for $j=1,2$. The actual dimensions that matter in implementation are: (i) K in the estimation of h_1 and h_0 in the first stage, and (ii) the dimension of the indexes to be matched in the second stage: $J=2$. Therefore, the dimensionality that causes difficulty in estimation is $\max\{J,K\}$, not $J\times K$. In our view, this is no more difficult than implementing estimators that are based on matching or pairwise comparison, such as Blundell and Powell (2003).

6 Simulation Study

This section presents simulation evidence for the performance of the estimator in Section 5, for both the Average Treatment Effect and the Distributional Treatment Effect.

6.1 Designs with continuous outcomes

We report results for both our estimator and that in Vytlacil and Yildiz (2007), for several designs where the potential outcomes are real-valued and continuous. These include designs where the monotonicity condition fails, and designs where the disturbance terms in the outcome equation are multi-dimensional.

Throughout all designs, we model the treatment or dummy endogenous variable as

$$D = I[Z-U>0],$$

where Z,U are independent standard normal. We experiment with the following designs for the outcome.

(Design 1) $Y = X + 0.5\cdot D + \varepsilon$, where X is standard normal while (ε,U) are bivariate normal, with mean 0, variance 1, and correlation $\rho_v\in\{0,0.25,0.5\}$.

(Design 2) $Y = X + 0.5 \cdot D + (X + D) \cdot \varepsilon$, where X is standard normal while (ε, U) are bivariate normal, with mean 0, variance 1, and correlation $\rho_v \in \{0, 0.25, 0.5\}$.

(Design 3) $Y = (X + 0.5 \cdot D + \varepsilon)^2$, where X is standard normal while (ε, U) are bivariate normal, each with mean 0, variance 1, and correlation $\rho_v \in \{0, 0.25, 0.5\}$

We note that the monotonicity condition holds in Design 1 but fails in the other two designs. For each of these designs, we report results for estimating $E[Y_1]$, i.e., the mean potential outcome under treatment $D = 1$. The two estimators reported in the simulation study are our estimator proposed in Section 5 and the one proposed in Vytlacil and Yildiz (2007). The summary statistics, scaled by the true parameter value, Mean Bias, Median Bias, Root Mean Squared Error, (RMSE), and Median Absolute Deviation (MAD) are evaluated for sample sizes of $n = 100, 200, 400$ for 401 replications.

Results for each of these designs are reported in Tables 1 to 3 respectively. In implementing our estimator, we assume the propensity score function is known, and conduct next stage estimation using a nonparametric kernel estimator with normal kernel function, and a bandwidth of $n^{-1/5}$. This rate reflects "undersmoothing" as there are two regressors, the propensity score and the regressor X. For the estimator in Vytlacil and Yildiz (2007), which involves the derivative of conditional expectation functions as well, we also report results for an infeasible version of their estimator, assuming such functions, as well as the propensity scores, are known. To implement the second stage of our estimator, in calculating the distance $\|h_1(x_i, \cdot) - h_0(x_0, \cdot)\|$ we used an evenly spaced grid of values for y, and selected $n/50$ grid points, with n denoting the sample size.

The results indicate the desirable properties of our estimator, generally agreeing with Theorem 5.1. In all designs, our estimator has small values for bias and RMSE, with the value of RMSE decreasing as the sample size grows. In contrast, the procedure based on Vytlacil and Yildiz (2007) only performs well in Design 1, with the sizes of bias and RMSE comparable to those using our method. As in our estimator, these values decrease as the sample size grows, which is expected, as the monotonicity condition they require is satisfied in this design. In this case, their approach has smaller standard errors largely due to the relatively simpler structure of the infeasible version of the estimator, but their biases persist even when the sample size increases.

For Designs 2 and 3, where the monotonicity condition is violated, the estimator proposed in Vytlacil and Yildiz (2007) does not perform well. Table 2 shows that in Design 2 both the bias and RMSE of their estimator are generally decreasing slowly with the sample size. Results for their estimator are better in Design 3 in Table 3, but the bias hardly converges with the sample size, and is much larger compared to our estimator.

Table 1

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	-0.0170	0.0229	-0.0435	-0.1302	-0.1676	-0.2018
MEDIAN BIAS	-0.0137	0.0124	-0.0653	-0.1318	-0.1678	-0.2087
RMSE	0.4936	0.4800	0.4945	0.3308	0.3337	0.3546
MAD	0.3289	0.3328	0.3156	0.2200	0.2271	0.2546
n=200						
MEAN BIAS	0.0032	-0.0024	-0.0069	-0.0864	-0.1299	-0.1766
MEDIAN BIAS	-0.0102	-0.0141	-0.0314	-0.0934	-0.1277	-0.1679
RMSE	0.3355	0.3367	0.3521	0.2293	0.2457	0.2711
MAD	0.2240	0.2228	0.2517	0.1594	0.1676	0.1865
n=400						
MEAN BIAS	-0.0187	0.0101	-0.0055	-0.0584	-0.11134	-0.1593
MEDIAN BIAS	-0.0261	0.0128	-0.0065	-0.0592	-0.1162	-0.1572
RMSE	0.2496	0.2489	0.2578	0.2049	0.1867	0.2167
MAD	0.1523	0.1732	0.1659	0.1197	0.1345	0.1605

Table 2

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	0.0109	0.0397	-0.0671	-0.1509	-0.2875	-0.4207
MEDIAN BIAS	0.0151	0.0227	-0.0939	-0.1590	-0.2918	-0.4262
RMSE	0.5089	0.2737	0.4853	0.3524	0.4199	0.5289
MAD	0.3395	0.2447	0.3105	0.2419	0.30898	0.4310
n=200						
MEAN BIAS	0.0322	0.0143	-0.0311	-0.1273	-0.2559	-0.3875
MEDIAN BIAS	0.0159	0.0054	-0.0543	-0.1310	-0.2553	-0.3884
RMSE	0.3487	0.3444	0.3455	0.2622	0.3407	0.4475
MAD	0.2317	0.2297	0.2552	0.1782	0.2624	0.3884
n=400						
MEAN BIAS	0.0088	0.0269	-0.0294	-0.0962	-0.2247	-0.3708
MEDIAN BIAS	0.0007	0.0244	-0.0309	-0.0982	-0.2255	-0.3769
RMSE	0.2578	0.2557	0.2549	0.1920	0.2764	0.4037
MAD	0.1649	0.1733	0.1606	0.1354	0.2283	0.3769

Table 3

	CKT			VY		
ρ_v	0	1/4	1/2	0	1/4	1/2
n=100						
MEAN BIAS	-0.0097	-0.0070	0.0019	-0.0691	-0.0898	-0.1066
MEDIAN BIAS	-0.0233	-0.0101	-0.0240	-0.0799	-0.0925	-0.1178
RMSE	0.1893	0.2085	0.2126	0.1546	0.1630	0.1701
MAD	0.1398	0.1342	0.1374	0.1125	0.1178	0.1315
n=200						
MEAN BIAS	-0.0108	-0.0069	-0.0068	-0.0609	-0.0765	-0.0968
MEDIAN BIAS	-0.0148	-0.0033	-0.0099	-0.0674	-0.0769	-0.1017
RMSE	0.1372	0.1434	0.1424	0.1163	0.1262	0.1369
MAD	0.0949	0.0989	0.0953	0.0855	0.0887	0.1078
n=400						
MEAN BIAS	-0.0073	-0.0014	-0.0026	-0.0583	-0.0725	-0.0889
MEDIAN BIAS	-0.0149	-0.0023	-0.0029	-0.0610	-0.0751	-0.0887
RMSE	0.1084	0.0994	0.0989	0.0924	0.1007	0.1131
MAD	0.0697	0.0685	0.0654	0.0689	0.0788	0.0901

We also report estimator performance in samples simulated from a model where potential outcomes are determined by random coefficients and dummy endogenous variables. It is important to note that for this design, the estimator in Vytlacil and Yildiz (2007) does not apply. This is because different values of x lead to different distributions of the composite error $\eta_d + x'\epsilon_d$. Our contribution in Section 4 is to propose a new approach based on matching different values of the observed outcome y, rather than the exogenous covariates x. Based on the counterfactual framework discussed in Section 4, here the treatment variable D is modeled as the same way as in the first three designs, with the regressor X being standard normal. For both Y_0, Y_1, the intercepts were modeled as constants (0 and 1, respectively) and the additive error terms were each standard normal. For the random slopes, the means were 1 and 2 respectively, and the additive error terms were also standard normal, independent of all other disturbance terms and each other. Here we use the procedure in Section 4 to estimate the parameter $\Delta_2 = P(Y_1 < y)$, where in the simulation, we set $y = 1$.

Results for this design with random coefficients are reported in Table 4. The same four summary statistics are reported for sample sizes $n \in \{100, 200, 400\}$, based on 401 replications. The estimator proposed in Section 5 performs well. The bias and RMSE are much smaller for a bigger sample with $n = 400$ than for smaller samples with $n = 100$ and $n = 200$, indicating convergence at the parametric rate.

Table 4

	CKT		
ρ_v	0	1/4	1/2
n=100			
MEAN BIAS	0.0109	-0.0086	0.0038
MEDIAN BIAS	0.0000	-0.0064	0.0126
RMSE	0.1011	0.0979	0.0955
MAD	0.0600	0.0648	0.0652
n=200			
MEAN BIAS	-0.0050	-0.0150	0.0095
MEDIAN BIAS	-0.0100	-0.0161	0.0029
RMSE	0.0669	0.0669	0.0665
MAD	0.0400	0.0454	0.0457
n=400			
MEAN BIAS	0.0012	-0.0132	0.0074
MEDIAN BIAS	0.0049	-0.0162	0.0077
RMSE	0.0501	0.0494	0.0495
MAD	0.0349	0.0325	0.0360

6.2 Designs with discrete potential outcomes

We also report the performance of our estimator in a sample drawn from Example 2, where the observed outcomes are determined in a multinomial choice model:

$$Y_i(D_i) = \arg \max_{j \in \{0,1,2\}} Y^*_{i,j}(D_i),$$

where the potential outcomes are $Y^*_{i,j}(d) \equiv v_j(X_{i,j}, d) + \varepsilon_{i,j}$ with

$$v_j(X_{i,j}, d) = \alpha_j(d) + X_{i,j}\beta_j(d)$$

for $j = 1, 2$, and $v_0(X_{i,j}, d) = 0$ by way of normalization. The exogenous covariates $X_{i,j}$ are drawn independently from standard normal, and the intercepts and slope coefficients are:

$$\alpha_1(0) = 0;\ \alpha_2(0) = 1;\ \alpha_1(1) = 1;\ \alpha_2(1) = 2;\ \beta_1(0) = 0.8;\ \beta_2(0) = 1;\ \beta_1(1) = 1;\ \beta_2(1) = 2.$$

For each individual i, the binary treatment D_i is determined as before:

$$D_i = 1\{U_i < Z_i\},$$

where the instrument Z_i is independently drawn from a standard uniform distribution. The marginal distribution of the selection error U_i is standard uniform. Conditional on $U_i = u$, the outcome errors $\varepsilon_{i,j}$ are independent across $j = 0, 1, 2$ and are distributed as type-1 extreme value with unit variance and means $(0, \delta u, 2\delta u)$ respectively, where δ is a parameter to be specified in the data-generating process. We adopt this specification as it allows for substantial dependence between U_i and $\varepsilon_i \equiv (\varepsilon_{i,j})_{j=0,1,2}$.

Our simulation study uses sample sizes $n \in \{250, 500, 1000, 2000\}$. For each sample size n, we generate $S = 400$ independent samples from the data-generating process above. Throughout this section, we focus on estimating a conditional distribution of *potential* outcomes $\Pr\{Y_d = j|X \in \omega\}$ for $d \in \{0, 1\}$ and $j \in \{1, 2\}$, where $\omega \equiv \{x : x_j \in [-1, 1]$ for $j = 1, 2\}$ is a subset of the support of covariates.

As a benchmark, we first implement the *infeasible* version of the estimator in Section 5, where knowledge of $h_1^*(\cdot)$ and $h_0^*(\cdot)$ are used for finding $(x, \tilde{x})$ with $v(x, 1) = v(\tilde{x}, 0)$ for estimating $\Pr\{Y_1 = j|X \in \omega\}$ (or $v(x, 0) = v(\tilde{x}, 1)$ for estimating $\Pr\{Y_0 = j|X \in \omega\}$). In this case, $F_{g|u}$ has a known close form, and we use numerical integration via mid-point approximation to calculate h_1^* and h_0^*. Table 5 shows the mean bias and mean squared error (M.S.E.) for the infeasible estimator calculated from the S simulated samples. We report these measures for $Pr\{Y_d = j|X \in \omega\}$ for $d = 0, 1$ and $j = 1, 2$. In the last two columns of the table, we report the M.S.E. for the *full* vector summarizing the probability mass function of Y_d, i.e., $[\Pr\{Y_d = 1|X \in \omega\}, \Pr\{Y_d = 2|X \in \omega\}]$.

The M.S.E. in Table 5 diminishes at a root-n rate that is proportional to the sample size. While in most cases the mean bias decreases with the sample size, the root-n rate of convergence appears to be substantially driven by the diminishing variance of the estimator. The performance does not vary substantively across different designs with the parameter values that affect the strength of correlation between the structural error $\varepsilon_{i,j}$ and the selection error U_i, i.e., the parameter values $\delta = (1/4, 1/3, 1/2)$.

We then construct a feasible version of the estimator by using h_1^* and h_0^* with their corresponding kernel estimates. Specifically, we use bivariate Gaussian kernels with bandwidths $1.06\hat{\sigma}^{-1/7}$, where $\hat{\sigma}$ denotes the sample standard deviation of components in (X_i, P_i). Table 6 reports the mean bias and M.S.E. of this feasible estimator in the same data-generating process as in Table 5. The estimation errors in Table 6 are overall larger than those reported in Table 5 but demonstrate similar patterns of convergence in most cases, even though the convergence appears to be slower for the MSE of $F_{Y_0|X\in\omega}$ and $F_{Y_1|X\in\omega}$ when δ is large. The difference in the magnitude of estimation errors across Table 5 and Table 6 is attributable to the estimation error in h_1^* and h_2^*. All in all, we conclude our estimator has decent finite-sample performance in these designs.

Table 5. Performance of Infeasible Estimator

		$\Pr\{Y_0 = 1\|\omega\}$		$\Pr\{Y_0 = 2\|\omega\}$		$\Pr\{Y_1 = 1\|\omega\}$		$\Pr\{Y_1 = 2\|\omega\}$		$F[Y_0\|\omega]$	$F[Y_1\|\omega]$
δ	n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	MSE	MSE
1/4	250	0.01305	0.00225	-0.00120	0.00328	-0.00753	0.00254	0.00276	0.00275	0.00554	0.00254
1/4	500	0.00970	0.00113	-0.00215	0.00175	-0.00693	0.00121	0.00227	0.00144	0.00288	0.00121
1/4	1000	0.00910	0.00077	-0.00269	0.00095	-0.00389	0.00078	0.00146	0.00080	0.00172	0.00078
1/4	2000	0.00871	0.00041	-0.00447	0.00051	-0.00376	0.00038	0.00110	0.00043	0.00092	0.00038
1/3	250	0.01825	0.00269	-0.00955	0.00339	-0.00350	0.00248	-0.00262	0.00276	0.00608	0.00275
1/3	500	0.00696	0.00126	-0.00144	0.00166	-0.00656	0.00123	0.00330	0.00130	0.00292	0.00144
1/3	1000	0.01170	0.00079	-0.00842	0.00097	-0.00315	0.00070	0.00007	0.00081	0.00175	0.00080
1/3	2000	0.00918	0.00045	-0.00625	0.00060	-0.00486	0.00043	0.00273	0.00044	0.00105	0.00043
1/2	250	0.01322	0.00235	-0.00770	0.00323	-0.00359	0.00253	-0.00158	0.00295	0.00559	0.00248
1/2	500	0.01426	0.00130	-0.01140	0.00196	-0.00231	0.00120	0.00081	0.00132	0.00326	0.00123
1/2	1000	0.01142	0.00075	-0.00782	0.00094	-0.00512	0.00064	0.00378	0.00071	0.00169	0.00070
1/2	2000	0.01007	0.00048	-0.00772	0.00061	-0.00438	0.00039	0.00353	0.00042	0.00109	0.00043

Table 6. Performance of Feasible Estimator

		$\Pr\{Y_0 = 1\|\omega\}$		$\Pr\{Y_0 = 2\|\omega\}$		$\Pr\{Y_1 = 1\|\omega\}$		$\Pr\{Y_1 = 2\|\omega\}$		$F[Y_0\|\omega]$	$F[Y_1\|\omega]$
δ	n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	MSE	MSE
1/4	250	0.00522	0.00360	0.05045	0.00785	-0.04019	0.00480	-0.00071	0.00401	0.01145	0.00533
1/4	500	0.00931	0.00237	0.04097	0.00489	-0.03585	0.00303	-0.00602	0.00237	0.00726	0.00400
1/4	1000	0.01099	0.00131	0.03946	0.00332	-0.03488	0.00215	-0.00750	0.00130	0.00464	0.00252
1/4	2000	0.01193	0.00093	0.04014	0.00277	-0.03345	0.00167	-0.00784	0.00071	0.00370	0.00171
1/3	250	0.00340	0.00395	0.04905	0.00792	-0.03382	0.00415	-0.00671	0.00415	0.01187	0.00725
1/3	500	0.00409	0.00226	0.04718	0.00525	-0.03287	0.00278	-0.00559	0.00228	0.00751	0.00567
1/3	1000	0.00856	0.00128	0.03914	0.00322	-0.03292	0.00191	-0.00630	0.00125	0.00450	0.00486
1/3	2000	0.01029	0.00086	0.03921	0.00261	-0.02912	0.00137	-0.00819	0.00077	0.00346	0.00468
1/2	250	0.00267	0.00369	0.04291	0.00680	-0.03508	0.00399	0.00350	0.00367	0.01050	0.00643
1/2	500	0.00563	0.00217	0.04076	0.00450	-0.03205	0.00266	0.00067	0.00215	0.00667	0.00417
1/2	1000	0.00426	0.00123	0.04284	0.00357	-0.02865	0.00169	-0.00184	0.00109	0.00481	0.00238
1/2	2000	0.01084	0.00089	0.03538	0.00230	-0.02820	0.00128	-0.00173	0.00058	0.00319	0.00207

7 Concluding Remarks

In this paper, we consider identification and estimation of weakly separable models with endogenous binary treatment. Existing approaches are based on a monotonicity condition, which is violated in models with multiple unobserved idiosyncratic shocks. Such models arise in many important empirical settings, including cases where potential outcomes are determined by Roy models, multinomial choice models, or random coefficients with dummy endogenous variables. We establish new identification results for these models which are constructive and conducive to estimation procedures. A simulation study indicates adequate finite sample performance of our method.

This paper leaves several open questions for future research. For example, it may not be feasible to locate pairs of $(x, \tilde{x})$ that satisfy the matching criterion perfectly due to limited, say, discrete, support of X or Z. In this case, it remains an open question how or whether the partial identification approach proposed in Shaikh and Vytlacil (2011) can be applied in the current setting where multiple indices in potential outcomes defy any notion of monotonicity. Besides, our method requires the selection of the number and location of cutoff points, so a data-driven method for selecting these would be useful. Furthermore, the relative efficiency of our proposed estimation approach needs to be explored, perhaps by deriving efficiency bounds for these new classes of models.

References

Abrevaya, J., and H. Xu (2022): "Estimation of treatment effects under endogenous heteroskedasticity," *Journal of Econometrics*, forthcoming.

Ahn, H., J. Powell, H. Ichimura, and P. Ruud (2017): "Simple Estimators for Invertible Index Models," *Journal of Business Economics and Statistics*, 36, 1–10.

Arellano, M., and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 79(3), 987–1020.

Carneiro, P., and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149(2), 191–208.

Carneiro, P., E. Vytlacil, and J. Heckman (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Chen, S. H., and S. Khan (2014): "Semiparametric Estimation of Program Impacts on Dispersion of Potential Wages," *Journal of Applied Econometrics*, 29, 901–919.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245 – 261.

——— (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132, 491–525.

D'Haultfoeuille, X., and P. Fevrier (2015): "Identification of Nonseparable Triangular Models with Discrete Instruments," *Econometrica*, 3, 1199–1210.

Dong, Y., and S. Shen (2018): "The Empirical Content of the Roy Model," *The Review of Economics and Statistics*, 100, 78–85.

Feng, J. (2020): "Matching Points: Supplementing Instruments with Covariates in Triangular Models," mimeograph, Columbia University.

Frandsen, B., and L. Lefgren (2018): "Testing Rank Similarity," *The Review of Economics and Statistics*, 100, 86–91.

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

Heckman, J., and E. Vytlacil (2007a): “Econometric evaluation of social programs,” in *Handbook of Econometrics, Vol. 6B*, ed. by J. Heckman, and E. Leamer. Amesterdam: North Holland.

Heckman, J., and E. Vytlacil (2007b): “Econometric Evaluation of Social Programs,” *Handbook of Econometrics Volume 6*, pp. 4780–4874.

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

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

Hoderlein, S., J. Klemelä, and E. Mammen (2010): “Analyzing the random coefficient model nonparametrically,” *Econometric Theory*, 26(3), 804–837.

Hsiao, C., and M. H. Pesaran (2008): “Random Coefficient Models,” in *The Econometrics of Panel Data: Advanced Studies in Theoretical and Applied Econonmetrics*, ed. by L. Matyas, and P. Sevestre. Springer.

Imbens, G., and J. Angrist (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–75.

Imbens, G., and W. Newey (2009): “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” *Econometrica*, 77(5), 1481–1512.

Jun, S. J., J. Pinkse, H. Xu, and N. Yildiz (2016): “Multiple Discrete Endogenous Variables in Weakly-Separable Triangular Models,” *Econometrics*, 4 (7).

Kasy, M. (2014): “Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment,” *Review of Economic Studies*, 81, 1614–1636.

Khan, S., F. Ouyang, and E. Tamer (2019): “Inference in Semiparametric Mutinomial Response Models,” Boston College Working Paper.

Lee, L.-F. (1995): “Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models,” *Journal of Econometrics*, 65, 381–428.

Masten, M. A. (2018): “Random coefficients on endogenous variables in simultaneous equations models,” *The Review of Economic Studies*, 85(2), 1193–1250.

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

Mourifié, I. (2015): "Sharp Bounds on Treatment Effects in a Binary Triangular System," *Journal of Econometrics*, 187(1), 74–81.

Pakes, A., and J. Porter (2014): "Moment Inequalties for Multinomial Choice with Fixed Effects," Harvard University Working Paper.

Ruud, P. (2000): "Semiparametric estimation of discrete choice models," mimeograph, University of California at Berkeley.

Shaikh, A. M., and E. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79(3), 949–955.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-Parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

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

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

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779.