Discrete Endogenous Variables in Weakly Separable Models*

Sung Jae Jun,† Joris Pinkse‡ and Haiqing Xu§
Center for Auctions, Procurements and Competition Policy
Department of Economics
The Pennsylvania State University

July 2011

This paper contains an extension of the identification method proposed in Jun, Pinkse, and Xu (2009, JPX), which is based on a generated collection of sets, i.e. a *Dynkin system*. We demonstrate the usefulness of this extension in the context of the model proposed by Vytlacil and Yıldız (2007, VY). VY formulate a fully nonparametric model featuring a nested weakly separable structure in which an endogenous regressor is binary–valued. The extension of the JPX approach considered here allows for nonbinary–valued discrete endogenous regressors and requires weaker support conditions than VY in the binary case, which substantially broadens the range of potential applications of the VY model. In this paper we focus on the binary case for which we provide several alternative simpler sufficient conditions and outline an estimation strategy. Our results allow for the possibility of a nonbinary endogenous regressor, but we provide examples for the nonbinary case in a separate paper.

Key Words: Nonparametric Identification, Weak Separability, Monotonicity

JEL Classification Codes: C14; C31.

*This paper is based on research supported by NSF grant SES–0922127. We thank the Human Capital Foundation (www.hcfoundation.ru), and especially Andrey P. Vavilov, for their support of CAPCP (http://capcp.psu.edu) at Penn State University. We thank Yuanyuan Wan, Neșe Yıldız, the managing editor, and two anonymous referees for constructive comments and suggestions.

†(corresponding author) 608 Kern Graduate Building, University Park 16802, sjun@psu.edu

‡joris@psu.edu; Joris Pinkse is an extramural fellow at Center, Tilburg University.

§hux100@psu.edu

1. Motivation

This paper contains an extension of the identification method proposed in Jun, Pinkse, and Xu (2009, JPX), which is based on a generated collection of sets, i.e. a *Dynkin system*. We demonstrate the usefulness of this extension in the context of the model proposed by Vytlacil and Yıldız (2007, VY). VY formulate a fully nonparametric model featuring a nested weakly separable structure in which an endogenous regressor is binary–valued. The extension of the JPX approach considered here allows for nonbinary–valued discrete endogenous regressors and requires weaker support conditions than VY in the binary case, which substantially broadens the range of potential applications of the VY model. In this paper we focus on the binary case for which we provide several alternative simpler sufficient conditions and outline an estimation strategy. Our results allow for the possibility of a nonbinary endogenous regressor, but we provide examples for the nonbinary case in a separate paper. The nonbinary case is also discussed in the working paper version.

The Dynkin system method provides a general framework for extracting information available in a model. The original Dynkin system method was used in JPX in the context of obtaining bounds on values of a structural function in a nonparametric triangular model with discrete endogenous regressors. The method proposed here extends the original procedure. So the method proposed here can be applied to the model in JPX, where it yields bounds equivalent to those obtained there. However, the results one would obtain using the basic method of JPX (without the extension proposed here) in the VY model are generally weaker than the ones obtained in the current paper.

The VY model is attractive in that it can be interpreted as providing nonparametric alternatives to parametric procedures proposed in Heckman (1978) for systems of equations featuring one or more binary endogenous variables. VY contains a full and eloquent description of the merits of the VY model relative to alternatives available in the literature to which we have little to add.

Attractive as the VY model is, the VY approach makes use of strong support restrictions and excludes the possibility of discrete–valued endogenous regressors which can take more than two values, e.g. when the endogenous regressor of interest is the level of education.

The Dynkin system method resolves these issues. Both the VY method of identification and ours depend on a conditioning argument, but ours employs a larger class of conditioning possibilities. As mentioned, this removes the requirement that the endogenous regressor be binary. With binary endogenous regressors the larger class allows for weaker support conditions to obtain identification and the potential for more efficient, albeit more complicated, estimators under the same conditions as in VY. The difference between the VY method and ours in the binary case is highlighted in examples in sections 4.1 and 4.2. In section 4.3 we generalize the examples to illustrate the simpler–to–verify sufficient conditions. We then compare our sufficient conditions to those in VY.

The Dynkin system method has other applications. In a second follow–up paper (Jun, Pinkse, Xu, and Yıldız, 2010) we use it to obtain tight bounds in a multi–equation triangular system with discrete endogenous regressors and weak separability. It can also be used to obtain tighter identification bounds when the outcome variable is also binary as in Shaikh and Vytlacil (2011). A similar objective is achieved in Chiburis (2010) but unlike Chiburis our method relies on matching arguments, only.

We finish our paper by sketching an estimation method which exploits a subset of the possibilities provided by our Dynkin system. The estimation discussion can be found in section 5.

2. Model

Consider the model

$$\begin{cases} \boldsymbol{y} &= g(\boldsymbol{\ell}, \boldsymbol{u}), \\ \boldsymbol{\ell} &= m(\boldsymbol{x}, \boldsymbol{w}), \\ \boldsymbol{x} &= h(\boldsymbol{z}, \boldsymbol{v}), \end{cases} \tag{1}$$

where $\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are observed, $\boldsymbol{u}, \boldsymbol{v}$ are unobserved, and g, m, h are unknown functions. For the intuitive interpretation of the model we will use the labels 'earnings,' '(structural labor market) potential,' 'education,' 'market uncertainty,' and 'talent' for $\boldsymbol{y}, \boldsymbol{\ell}, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}$, respectively, and 'demographics' for both $\boldsymbol{w}$ and $\boldsymbol{z}$. The difference between $\boldsymbol{w}$ and $\boldsymbol{z}$ is that $\boldsymbol{w}$ affects potential directly whereas $\boldsymbol{z}$ only affects potential via its effect on education. We will allow $\boldsymbol{w}$ and $\boldsymbol{z}$ to have elements in common, but at least one element of $\boldsymbol{z}$ must be absent from $\boldsymbol{w}$ for reasons to become apparent.

The parameter of interest is

$$\delta(x^*, w^*) = \mathbb{E}\{\boldsymbol{y}(x^*) | \boldsymbol{w} = w^*\} \tag{2}$$

for given values of x^* and w^*, where $\boldsymbol{y}(x^*) = g\{m(x^*, \boldsymbol{w}), \boldsymbol{u}\}$, i.e. the counterfactual value of $\boldsymbol{y}$ if $\boldsymbol{x}$ is exogenously fixed at x^*; hence $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{y}$. Then $\delta(x^*, w^*)$ represents average earnings for individuals with w–demographics w^* who were exogenously assigned education x^*. The value of a level one education relative to a level zero education for someone with demographics w^* is then $\delta(1, w^*) - \delta(0, w^*)$. We make the following assumptions.

Assumption A. *$\boldsymbol{u}, \boldsymbol{v}$ are independent of $\boldsymbol{w}, \boldsymbol{z}$.*

Assumption B. *$\boldsymbol{v}$ has a uniform distribution on $\mathscr{U}$, where $\mathscr{U} = (0, 1]$.*

Assumption C. *h is nondecreasing in its second argument and takes values in $\mathscr{S}_{\boldsymbol{x}} = \{0, 1, \ldots, J\}$.*

Assumption D. *$\theta(\ell, V) = \mathbb{E}\{g(\ell, \boldsymbol{u})|\boldsymbol{v} \in V\}$ is strictly increasing in ℓ for all (Borel–measurable) $V \subseteq \mathscr{U}$.*

The model in (1) with the above assumptions is the same as the VY model except that here $\boldsymbol{x}$ need not be binary. The key restriction is the index structure of the function g, which depends on the scalar–valued index function m and the scalar–valued random variable $\boldsymbol{u}$. Assumption D relies on the index structure of g. This type of strong monotonicity can be avoided by employing an *identification–at–infinity* argument (Chamberlain, 1986; Heckman, 1990), which we discuss later in this section. The remaining assumptions are fairly standard in this literature; assumption B is effectively a normalization, assumption A is strong but difficult to avoid in a fully nonparametric environment, and assumption C is partly a normalization and partly expresses the fact that $\boldsymbol{x}$ is discrete–valued.

Let $p_x(z) = \mathbb{P}(\boldsymbol{x} = x|\boldsymbol{z} = z)$ and $\eta_x(z) = \sum_{j=0}^{x-1} p_j(z)$ for $x = 1, \cdots, J+1$; we define $\eta_0(z) = 0$. Then without loss of generality, h in (1) can be expressed as

$$h(z, v) = \inf\{x \in \mathscr{S}_{\boldsymbol{x}} : \ \mathbb{P}(\boldsymbol{x} \leq x|\boldsymbol{z} = z) \geq v\} = \sum_{x=1}^{J} \mathbb{1}\{v > \eta_x(z)\}. \tag{3}$$

Let further $V(x, z) = \{v \in \mathscr{U} : h(z, v) = x\}$ and

$$\mathscr{E}(x, w, V) = \mathbb{E}\{\boldsymbol{y}(x)|\boldsymbol{w} = w, \boldsymbol{v} \in V\} = \theta\{m(x, w), V\}, \tag{4}$$

which can be interpreted as the average earnings for someone with demographics w, education level x, and talent level in the set V. Since $\boldsymbol{x} = x$ and $\boldsymbol{z} = z$ is equivalent to $\boldsymbol{z} = z$ and $\boldsymbol{v} \in V(x, z)$, we have $\mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x, \boldsymbol{w} = w, \boldsymbol{z} = z) = \mathbb{E}\big[g\{m(x, w), \boldsymbol{u}\}|\boldsymbol{w} = w, \boldsymbol{z} = z, \boldsymbol{v} \in V(x, z)\big]$, which in turn is equal to $\mathbb{E}\big[g\{m(x, \boldsymbol{w}), \boldsymbol{u}\}|\boldsymbol{w} = w, \boldsymbol{v} \in V(x, z)\big] = \mathscr{E}\{x, w, V(x, z)\}$ by assumption A. Therefore, $\mathscr{E}\{x, w, V(x, z)\}$ is identified by a conditional expectation of observables. Moreover, for any disjoint $V_0, \cdots, V_K$ for which $\cup_{k=0}^{K} V_k = \mathscr{U}$,

$$\delta(x^*, w^*) = \theta\{m(x^*, w^*), \mathscr{U}\} = \mathscr{E}(x^*, w^*, \mathscr{U}) = \sum_{k=0}^{K} \mathscr{E}(x^*, w^*, V_k)\mu(V_k), \tag{5}$$

where μ is the Lebesgue measure.

Equation (5) shows how identification at infinity obtains. Let $\mathscr{S}_{\boldsymbol{z}}(x,w)$ be the support of the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{x}=x$ and $\boldsymbol{w}=w$. Suppose that there exists a sequence $\{z_t\}$ belonging to $\mathscr{S}_{\boldsymbol{z}}(x^*,w^*)$ such that $V(x^*,z_t)\subset V(x^*,z_{t+1})$ and $\cup_{t=1}^{\infty}V(x^*,z_t)=\mathscr{U}$. It then follows that for $\psi(x,w,z)=\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=x,\boldsymbol{w}=w,\boldsymbol{z}=z)$,

$$\begin{aligned}\delta(x^*,w^*)&=\mathscr{E}\{x^*,w^*,V(x^*,z_t)\}\mu\{V(x^*,z_t)\}+\mathscr{E}\{x^*,w^*,\mathscr{U}-V(x^*,z_t)\}\mu\{\mathscr{U}-V(x^*,z_t)\}\\&=\lim_{t\to\infty}\mathscr{E}\{x^*,w^*,V(x^*,z_t)\}=\lim_{t\to\infty}\psi(x^*,w^*,z_t).\end{aligned}$$

However, as argued in VY, using identification at infinity does not make efficient use of the available data if assumption D is satisfied. In fact, the objective of this paper and of VY is to show that the identification at infinity argument can be avoided by using the monotonicity of θ. Details are explained in section 3.

3. Identification

We now describe our new identification strategy. Let, as before, $\mathscr{S}_{\boldsymbol{z}}(x,w)$ be the support of the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{x}=x$ and $\boldsymbol{w}=w$. Let further $\mathscr{V}(x,w)=\{\emptyset\neq V\subset\mathscr{U}:\exists z\in\mathscr{S}_{\boldsymbol{z}}(x,w):V=V(x,z)\}$. In the previous section we showed that $\mathscr{E}(x,w,V)$ is identified whenever $V\in\mathscr{V}(x,w)$. We now show that $\mathscr{E}(x,w,V)$ is in fact identified for a much larger class of V–sets This fact can be exploited to identify $\delta(x^*,w^*)$ without resorting to an identification at infinity argument. We adapt a definition from our earlier work.

Definition 1 (Jun, Pinkse, and Xu, 2009)**.** *$\mathscr{D}(x,w)$ is the collection $\mathscr{D}_\infty(x,w)$ in the following iterative scheme. Let $\mathscr{D}_0(x,w)=\mathscr{V}(x,w)$. Then for all $t\geq 0$, $\mathscr{D}_{t+1}(x,w)$ consists of all sets A^* such that at least one of the following four conditions is satisfied.*

(i) $A^*\in\mathscr{D}_t(x,w)$,

(ii) $\exists A_1,A_2\in\mathscr{D}_t(x,w):A_1\subset A_2,\ \mu(A_2-A_1)>0,\ A^*=A_2-A_1$,

(iii) $\exists A_1,A_2\in\mathscr{D}_t(x,w):A_1\cap A_2=\emptyset,\ \mu(A_1\cup A_2)>0,\ A^*=A_1\cup A_2$.

(iv) $\exists(\tilde{x},\tilde{w})\in\mathscr{S}_{\boldsymbol{xw}}:m(\tilde{x},\tilde{w})=m(x,w),\mathscr{D}_t(x,w)\cap\mathscr{D}_t(\tilde{x},\tilde{w})\neq\emptyset,A^*\in\mathscr{D}_t(\tilde{x},\tilde{w})$. □

The first three conditions are very similar to those in Jun, Pinkse, and Xu (2009); (iv) is new. The reason the first three conditions are helpful is both illustrated in Jun, Pinkse, and Xu (2009) and is apparent from the proof of lemma 1. The usefulness of (iv) can be seen in the proof of lemma 1. Examples that illustrate the use of these conditions follow in the next section.

Lemma 1. *Suppose that assumptions A to C are satisfied. For all $(x, w) \in \mathscr{S}_{xw}$, every $V \in \mathscr{D}(x, w)$ is identified. Further, $\mathscr{E}(x, w, V)$ is identified whenever $(x, w) \in \mathscr{S}_{xw}$ and $V \in \mathscr{D}(x, w)$.*

Proof. We use mathematical induction. For all $(x, w) \in \mathscr{S}_{xw}$, if $A_0 \in \mathscr{D}_0(x, w)$ then there exists a $z \in \mathscr{S}_z(x, w)$ such that $A_0 = V(x, z) \neq \emptyset$, which is identified. The identification of $\mathscr{E}(x, w, A_0)$ follows from $\mathscr{E}\{x, w, V(x, z)\} = \psi(x, w, z)$.

Now suppose that for all $(x, w) \in \mathscr{S}_{xw}$ and some t every $A \in \mathscr{D}_t(x, w)$ is identified and that $\mathscr{E}(x, w, A)$ is identified whenever $(x, w) \in \mathscr{S}_{xw}$ and $A \in \mathscr{D}_t(x, w)$. We establish that for all $(x, w) \in \mathscr{S}_{xw}$ every $A^* \in \mathscr{D}_{t+1}(x, w)$ is identified and that $\mathscr{E}(x, w, A^*)$ is identified whenever $(x, w) \in \mathscr{S}_{xw}$ and $A^* \in \mathscr{D}_{t+1}(x, w)$.

Pick an arbitrary $(x, w) \in \mathscr{S}_{xw}$. If $A^* \in \mathscr{D}_{t+1}(x, w)$, then by definition 1 one of four conditions is satisfied: (i) $A^* \in \mathscr{D}_t(x, w)$, (ii) $\exists A_1, A_2 \in \mathscr{D}_t(x, w)$ such that $A_1 \subset A_2$, $\mu(A_2 - A_1) > 0$, and $A^* = A_2 - A_1$, (iii) $\exists A_1, A_2 \in \mathscr{D}_t(x, w)$ such that $A_1 \cap A_2 = \emptyset$, $\mu(A_2 \cup A_1) > 0$, and $A^* = A_2 \cup A_1$, or (iv) $\exists(\tilde{x}, \tilde{w}) \in \mathscr{S}_{xw} : m(\tilde{x}, \tilde{w}) = m(x, w), \mathscr{D}_t(x, w) \cap \mathscr{D}_t(\tilde{x}, \tilde{w}) \neq \emptyset, A^* \in \mathscr{D}_t(\tilde{x}, \tilde{w})$. Case (i) is trivial. For (ii) (and similarly (iii)), A^* is identified because A_1, A_2 are identified. Now $\mathscr{E}(x, w, A^*)$ is identified because

$$\mathscr{E}(x, w, A^*) = \frac{\mathscr{E}(x, w, A_2)\mu(A_2) - \mathscr{E}(x, w, A_1)\mu(A_1)}{\mu(A_2 - A_1)}, \tag{6}$$

where $\mathscr{E}(x, w, A_1)$ and $\mathscr{E}(x, w, A_2)$ are identified since $A_1, A_2 \in \mathscr{D}_t(x, w)$. Finally, (iv). For each pair $(\tilde{x}, \tilde{w}) \in \mathscr{S}_{xw}$, check whether there exists an $\breve{A} \in \mathscr{D}_t(x, w) \cap \mathscr{D}_t(\tilde{x}, \tilde{w}) \neq \emptyset$, which can be done because $\mathscr{D}_t(x, w)$ and $\mathscr{D}_t(\tilde{x}, \tilde{w})$ are identified. Since the identification of $\mathscr{E}(x, w, \breve{A})$ and $\mathscr{E}(\tilde{x}, \tilde{w}, \breve{A})$ follows from the fact that $\breve{A} \in \mathscr{D}_t(x, w)$ and $\breve{A} \in \mathscr{D}_t(\tilde{x}, \tilde{w})$, respectively, it is possible to check whether they are equal which by assumption D occurs if and only if $m(x, w) = m(\tilde{x}, \tilde{w})$. Therefore, for such $\tilde{x}, \tilde{w}$ we have $\mathscr{E}(x, w, A^*) = \mathscr{E}(\tilde{x}, \tilde{w}, A^*)$ for any $A^* \subset \mathscr{U}$, noting that $\mathscr{E}(\tilde{x}, \tilde{w}, A^*)$ is identified when $A^* \in \mathscr{D}_t(\tilde{x}, \tilde{w})$. □

We are now in a position to state our main result, which requires one final assumption, which will be discussed following theorem 1.

Assumption E. $\mathscr{U} \in \mathscr{D}(x^*, w^*)$.

Theorem 1. *Suppose that assumptions A to E are satisfied. Then $\delta(x^*, w^*)$ is identified.*

Proof. Since by (5), $\delta(x^*, w^*) = \mathscr{E}(x^*, w^*, \mathscr{U})$, the stated result follows immediately from lemma 1. □

Assumption E is a restriction on the support of $(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z})$, which is satisfied whenever the support conditions of Vytlacil and Yıldız (2007) are satisfied in the case of binary $\boldsymbol{x}$. However, assumption E is weaker than the support conditions of Vytlacil and Yıldız (2007). Examples are provided in the following section.

4. Examples and sufficient conditions

4.1. **A simple example.** The idea of using a collection of sets that is generated from a simpler collection constitutes the core of theorem 1. We illustrate this idea by considering a simple example with binary endogenous variables and make a direct comparison with the VY approach. We start with a technical comparison of the two methods. At the end of the example we assign meaning to the model variables to facilitate intuitive understanding.

Suppose that $\boldsymbol{x}$ is binary (i.e. $J = 1$) and write $\eta = \eta_1$. Suppose that $\mathscr{S}_{\boldsymbol{wz}} = \{(0,0), (0,1), (1,0)\}$ and that there are no support restrictions between $(\boldsymbol{w}, \boldsymbol{z})$ and $\boldsymbol{x}$. Suppose further that $m(0,0) = m(0,1) = m(1,0) \neq m(1,1)$ and that $0 < \eta(0) < \eta(1) < 1$; note that η is directly identified from the data but that m is not. Consider the identification of $\delta(0,1) = \mathbb{E}[g\{m(0,1), \boldsymbol{u}\} | \boldsymbol{w} = 1]$.

The first step of both our and VY's approach is to write

$$\delta(0,1) = \mathscr{E}\{0, 1, (0, \eta(0)]\}\eta(0) + \mathscr{E}\{0, 1, (\eta(0), 1]\}\{1 - \eta(0)\}. \tag{7}$$

The first term on the right hand side of (7) is clearly identified because $\mathscr{E}\{0, 1, (0, \eta(0)]\} = \mathbb{E}\big[g\{m(0,1), \boldsymbol{u}\} | \boldsymbol{v} \in \big(0, \eta(0)\big]\big] = \mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = 0, \boldsymbol{w} = 1, \boldsymbol{z} = 0)$. The difference between our and VY's approaches is in identifying $\mathscr{E}\{0, 1, (\eta(0), 1]\}$.

VY use the idea of imputing $m(0,1)$ from $m(1,0)$, which is inspired by the fact that conditioning on $\boldsymbol{v} \in \big(\eta(0), 1\big]$ and $\boldsymbol{z} = 0$ is equivalent to conditioning on $\boldsymbol{x} = 1$ and $\boldsymbol{z} = 0$. Specifically, if one knew a priori that $m(0,1) = m(1,0)$, then it would follow that $\mathscr{E}\{0, 1, (\eta(0), 1]\} = \mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = 1, \boldsymbol{w} = 0, \boldsymbol{z} = 0)$. With this idea, VY provide a set of support restrictions that are useful to check whether $m(0,1) = m(1,0)$. In this example, the VY support restrictions require $(1,1) \in \mathscr{S}_{\boldsymbol{wz}}$, in which case one can use

$$\begin{aligned}\mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = 0, \boldsymbol{w} = 1, \boldsymbol{z} = 1)\eta(1) - \mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = 0, \boldsymbol{w} = 1, \boldsymbol{z} = 0)\eta(0)& \\ = \mathbb{E}\big[g\{m(0,1), \boldsymbol{u}\} | \eta(0) < \boldsymbol{v} \leq \eta(1)\big]\{\eta(1) - \eta(0)\}&\end{aligned} \tag{8}$$

and

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=1,\boldsymbol{w}=0,\boldsymbol{z}=0)\{1-\eta(0)\}-\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=1,\boldsymbol{w}=0,\boldsymbol{z}=1)\{1-\eta(1)\}$$
$$=\mathbb{E}\big[g\{m(1,0),\boldsymbol{u}\}|\eta(0)<\boldsymbol{v}\leq\eta(1)\big]\{\eta(1)-\eta(0)\}, \quad (9)$$

where by assumption D, (8) equals (9) if and only if $m(1,0)=m(0,1)$.

But in our example the support of $(\boldsymbol{w},\boldsymbol{z})$ does not contain $(1,1)$, so the direct matching strategy cannot be used. Instead, our procedure allows one to combine $m(0,1)=m(0,0)$ and $m(0,0)=m(1,0)$ to match $m(1,0)$ with $m(0,1)$. We now describe the details of this procedure.

We first match $m(0,1)$ with $m(0,0)$ by using assumption D; $m(0,1)=m(0,0)$ if and only if

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=0,\boldsymbol{w}=1,\boldsymbol{z}=0)=\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=0,\boldsymbol{w}=0,\boldsymbol{z}=0).$$

Likewise, we match $m(0,0)$ with $m(1,0)$ by equating (9) with

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=0,\boldsymbol{w}=0,\boldsymbol{z}=1)\eta(1)-\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=0,\boldsymbol{w}=0,\boldsymbol{z}=0)\eta(0)$$
$$=\mathbb{E}\big[g\{m(0,0),\boldsymbol{u}\}|\eta(0)<\boldsymbol{v}\leq\eta(1)\big]\{\eta(1)-\eta(0)\}. \quad (10)$$

Now since we know that $m(0,1)=m(0,0)=m(1,0)$, we conclude that

$$\mathscr{E}\{0,1,(\eta(0),1]\}=\mathscr{E}(0,0,(\eta(0),1]\}=\mathscr{E}\{1,0,(\eta(0),1]\}=\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}=1,\boldsymbol{w}=0,\boldsymbol{z}=0).$$

The above discussion is a simple illustration of theorem 1 and the conclusion of identification can be summarized by the fact that $\mathscr{D}(0,1)$ contains $\mathscr{U}$; see assumption E. Using the notation of definition 1,

$$\begin{aligned}\mathscr{D}_0(0,0)&=\big\{(0,\eta(0)],(0,\eta(1)]\big\}, & \mathscr{D}_0(0,1)&=\big\{(0,\eta(0)]\big\},\\ \mathscr{D}_0(1,0)&=\big\{(\eta(0),1],(\eta(1),1]\big\}, & \mathscr{D}_0(1,1)&=\big\{(\eta(0),1]\big\}.\end{aligned}$$

Note that by condition (ii) of definition 1, it follows that

$$\mathscr{D}_1(0,0)\supseteq\big\{(0,\eta(0)],(0,\eta(1)],(\eta(0),\eta(1)]\big\},$$

and that

$$\mathscr{D}_1(1,0)\supseteq\big\{(\eta(0),1],(\eta(1),1],(\eta(0),\eta(1)]\big\}.$$

Further, because $(0,\eta(0)]\in\mathscr{D}_0(0,1)\cap\mathscr{D}_0(0,0)$ and $m(0,0)=m(0,1)$ and since $(\eta(0),\eta(1)]\in\mathscr{D}_1(0,0)\cap\mathscr{D}_1(1,0)$ and $m(0,0)=m(1,0)$, we know that $\mathscr{D}_2(0,1)$ contains the union of $\mathscr{D}_0(0,1)$, $\mathscr{D}_1(0,0)$, and $\mathscr{D}_1(1,0)$ by condition (iv) of definition 1; i.e.

$$\mathscr{D}_2(0,1) \supseteq \mathscr{D}_0(0,1) \cup \mathscr{D}_1(0,0) \cup \mathscr{D}_1(1,0) \supseteq \{ (0,\eta(0)], (0,\eta(1)], (\eta(0),\eta(1)], (\eta(0),1], (\eta(1),1] \}.$$

Finally, condition (iii) of definition 1 implies that $\mathscr{U} = (0,\eta(0)] \cup (\eta(0),1] \in \mathscr{D}(0,1)$. So assumption E is satisfied and $\delta(0,1)$ is identified.

It is instructive to give labels to the variables. Suppose that $\boldsymbol{y}$ ='earnings,' $\boldsymbol{x}$ =college education' (yes=1), $\boldsymbol{w}$ ='poor background' (yes=1), $\boldsymbol{z}$ ='suburban' (yes=1, 'urban'=0).[1] Suppose further that we are looking for the counterfactual level of earnings for someone without a college education who comes from a poor background. With VY we would find a w such that $m(1,w) = m(0,1)$ by comparing the difference in average earnings between noncollege–educated individuals from a poor background from the suburbs and those from an urban area to the difference in average earnings between college–educated individuals with wealthy–background w from the suburbs and those from an urban area. The problem is that in the discussion above there are no poor individuals in the suburbs and the procedure breaks down.

With the new procedure we first note that noncollege–educated individuals from poor backgrounds have the same potential as noncollege–educated individuals from wealthy backgrounds before comparing noncollege–educated individuals from wealthy backgrounds to college–educated individuals from wealthy backgrounds, i.e. we use $m(0,1) = m(0,0)$ and $m(0,0) = m(1,0)$ to conclude that $m(0,1) = m(1,0)$ instead of trying to match $m(0,1)$ to $m(1,0)$ directly.

As mentioned just before (5), unlike VY we allow $\boldsymbol{z}$ to contain $\boldsymbol{w}$; the above example illustrates why. With VY one needs two different w–values for the same z–value, which is impossible if $\boldsymbol{z}$ contains $\boldsymbol{w}$. We, instead, can conclude that $m(0,0) = m(1,0)$ in the above example by exploiting variation in z–values only, which implies that $\delta(1,0)$ is identified. So whether $\boldsymbol{w}$ and $\boldsymbol{z}$ are entirely different, as in the example, or we use the pair $(\boldsymbol{w},\boldsymbol{z})$ in lieu of $\boldsymbol{z}$ is immaterial for the identification of $\delta(1,0)$.

4.2. **A more realistic example.** The example discussed in the previous subsection is not particularly interesting because there is only a minimal amount of endogeneity and it relies heavily on $\boldsymbol{z}$ being binary. The example in this subsection addresses both of these shortcomings.

Suppose that $\boldsymbol{x}$ is binary–valued, that $\mathscr{S}_{\boldsymbol{wz}} = \{(w,z) \in \mathbb{R}^3 : 0 < w_1, w_2 \leq 2, w_1 < z \leq w_1 + 1/2\}$, and that $\mathscr{S}_{\boldsymbol{xwz}} = \mathscr{S}_{\boldsymbol{x}} \times \mathscr{S}_{\boldsymbol{wz}}$. Suppose further that $p = p_1$ is strictly monotonic, that $m(x,w) = w_1 + w_2 - 1/2 - 3x/2$ and that $x^* = w_1^* = w_2^* = 1$. The situation is depicted in figure 1.

[1]These labels have the unrealistic implication that someone's background affects earnings directly without affecting education, but this is immaterial for the intuition provided here.

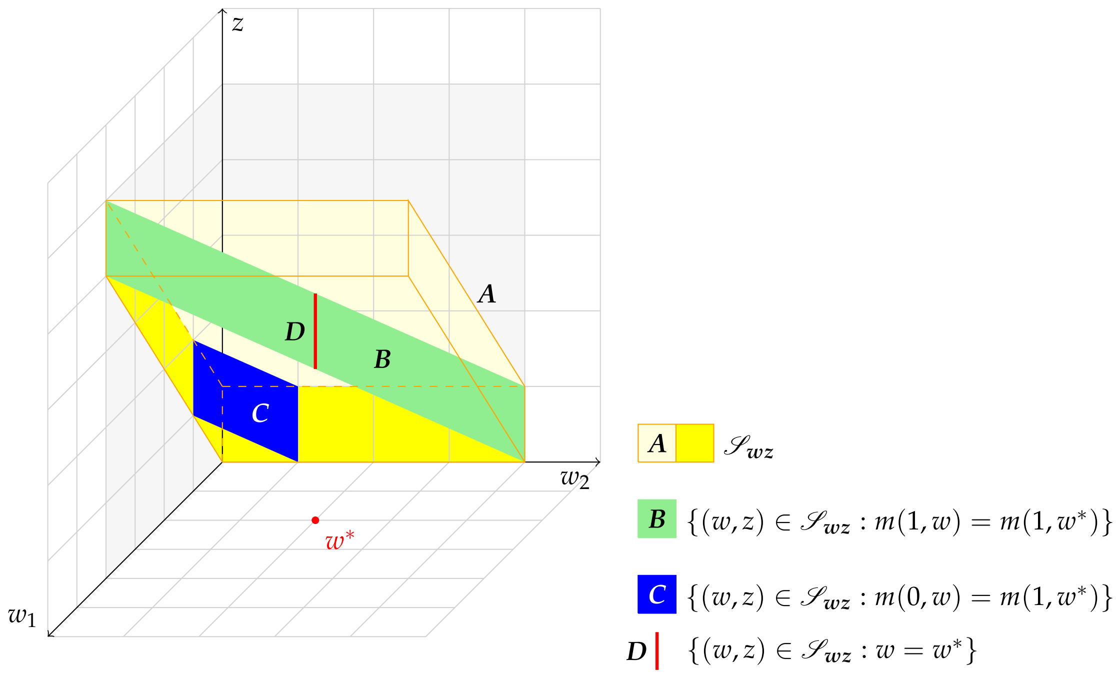

FIGURE 1. Difference in support conditions (section 4.2)

Please refer to the legend in figure 1 for the meaning of A, B, C, D. With VY one would need to find values $z_1 < z_2$ in the range indicated by D which are also in the range of values indicated by C; no such values exist.

Using theorem 1, however, all that needs to be done is to find values $z_1 < z_2$ in the range indicated by B which are also in the range of values indicated by C; for small w_1 many such values exist.

Looking at it from a different perspective, note that $m(1, w^*) = 0$ and that, given the shape of the support of $(\boldsymbol{w}, \boldsymbol{z})$, $\boldsymbol{w} = w^*$ implies that $\boldsymbol{z} > 1$. Note further that $m(0, w) = 0$ requires $w_1 \leq 1/2$ and hence $z \leq 1$. There is a contradiction.

With theorem 1 one can instead use the fact that for $\tilde{w} = (3/4, 5/4)$, we have $m(1, w^*) = m(1, \tilde{w})$ and that the support of $\boldsymbol{z}$ given $\boldsymbol{w} = \tilde{w}$ is $(3/4, 5/4]$, which overlaps with the support of $\boldsymbol{z}$ given $\boldsymbol{w} = w^*$, i.e. $(1, 3/2]$. But note that $(3/4, 5/4]$ also overlaps with the support of $\boldsymbol{z}$ given $\boldsymbol{w} = (1/2, 0)$, i.e. $(1/2, 1]$. So we first uncover that $m(1, w^*) = m(1, \tilde{w})$ before finding that $m(1, \tilde{w}) = m(0, w)$ for e.g. $w = (1/2, 0)$.

4.3. **Sufficient conditions.** The two examples above show that assumption E can be satisfied without satisfying the support restrictions used by VY. In this subsection we consider the binary endogenous regressor case with $h(z, v) = \mathbb{1}\{v > \eta(z)\}$ for which we provide restrictions on the support of $\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}$ that are sufficient for assumption E but are weaker than those in VY. We start with some strong alternative sufficient conditions which are simple to interpret before pointing out ways of relaxing and combining them. Let $\tilde{x}^* = 1 - x^*$.

We distinguish between four matching methods, including the one used in VY: VY, *between treatment chaining* (BTC), *bridging by chaining* (BBC), and *within treatment rollover* (WTR). The conditions are described in assumption F and illustrated in figure 2, where we included a special case of WTR (SWTR) for the purpose of illustration. Let $\mathscr{K}(x) = \{w : m(x, w) = m(x^*, w^*)\}$ and let $\mathscr{R}(\bar{x})$ denote the support of $(\boldsymbol{w}, \boldsymbol{z})$ conditional on $\boldsymbol{w} \in \mathscr{K}(\bar{x})$, i.e. $\mathscr{R}(\bar{x}) = \{(w, z) \in \mathscr{S}_{\boldsymbol{wz}} : m(\bar{x}, w) = m(x^*, w^*)\}$.

By the independence of $\boldsymbol{z}$ and $\boldsymbol{v}$ the conditional support of $(\boldsymbol{w}, \boldsymbol{z})$ given $\boldsymbol{x}$ depends on $\boldsymbol{x}$ if and only if there exists some $\bar{z}$ in the support of $\boldsymbol{z}$ such that $\eta(\bar{z}) = 0$. Relying on the existence of such $\bar{z}$ is essentially an identification–at–infinity type approach, which we aim to avoid in this paper. So, in the discussion below we take the support of $(\boldsymbol{w}, \boldsymbol{z})$ to be independent of $\boldsymbol{x}$ and use $\mathscr{S}_{\boldsymbol{z}}(w)$ to denote the conditional support of $\boldsymbol{z}$ given $\boldsymbol{w} = w$. We further shall take $\eta(z) = z$; otherwise $\boldsymbol{z}$ can be replaced by $\eta(\boldsymbol{z})$ in what follows.

Assumption F. *At least one of the conditions listed below is satisfied, where w, z–values are not necessarily all distinct.*

- **VY:** *There exist some $\tilde{w}^* \in \mathscr{K}(\tilde{x}^*)$ and some $z_0, z_1 \in \mathscr{S}_{\boldsymbol{z}}(w^*) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*)$ for which $z_0 < z_1$;*
- **BTC:** *There exist some $\tilde{w}^* \in \mathscr{K}(x^*) \cap \mathscr{K}(\tilde{x}^*)$ and some $z_0, z_1 \in \mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*)$ for which $z_0 < z_1$ and $\mathscr{S}_{\boldsymbol{z}}(w^*) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*) \neq \varnothing$;*
- **BBC:** *There exist some $w \in \mathscr{K}(x^*)$, $\tilde{w}^* \in \mathscr{K}(\tilde{x}^*)$, and $z_0, z_1 \in \mathscr{S}_{\boldsymbol{z}}(w) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*)$ for which $z_0 < z_1$ and $\mathscr{S}_{\boldsymbol{z}}(w^*) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*) \neq \varnothing$;*
- **WTR:** *There exist some $w_0, w_1 \in \mathscr{K}(x^*)$, $\tilde{w}^*, \tilde{w}_0, \tilde{w}_1 \in \mathscr{K}(\tilde{x}^*)$ such that for $j = 0, 1$ there exist $z_j \in \mathscr{S}_{\boldsymbol{z}}(w_j) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}_j)$ for which $z_0 < z_1$, $\mathscr{S}_{\boldsymbol{z}}(w^*) \cap \mathscr{S}_{\boldsymbol{z}}(w_j) \neq \varnothing$, and $\mathscr{S}_{\boldsymbol{z}}(\tilde{w}^*) \cap \mathscr{S}_{\boldsymbol{z}}(\tilde{w}_j) \neq \varnothing$.*

In comparing the conditions in assumption F it is helpful to bear in mind that $\tilde{w}^*$ is chosen, but w^* is fixed. If $\mathscr{S}_{\boldsymbol{z}}(w)$ does not depend on w, which is the case when the support of $(\boldsymbol{w}, \boldsymbol{z})$ is the Cartesian product of the marginal supports, then there is no difference among the conditions in assumption F; otherwise, the conditions differ.

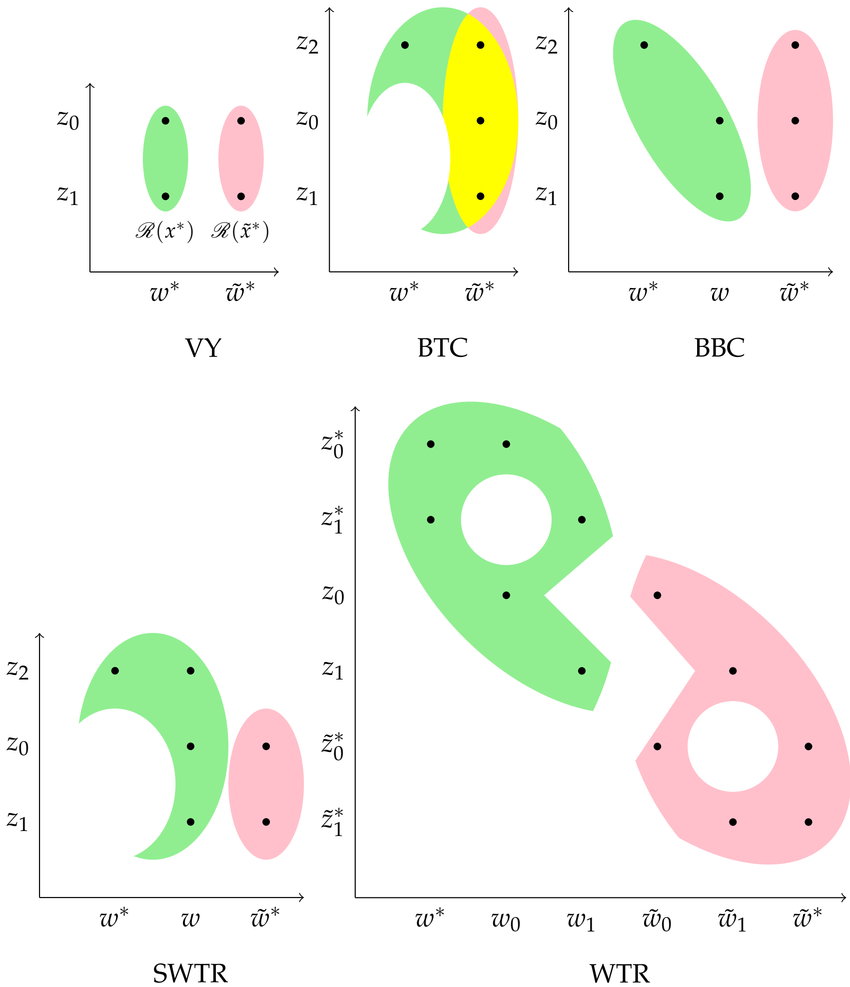

FIGURE 2. Assumption F illustrated

It is apparent from figure 2 that SWTR is a special case of WTR with $w_0 = w_1 = w \neq w^*$ and $\tilde{w}_0 = \tilde{w}_1 = \tilde{w}^*$. The dots in each of the shaded areas in figure 2 correspond to elements of the sets $\mathscr{R}(x^*)$ and $\mathscr{R}(\tilde{x}^*)$. The values of $z_2, z_0^*, z_1^*, \tilde{z}_0^*, \tilde{z}_1^*$ are such that $z_2 \in \mathscr{S}_z(w^*) \cap \mathscr{S}_z(\tilde{w}^*)$, as required by BTC and BBC in assumption F, and for $j = 0, 1$, $z_j^* \in \mathscr{S}_z(w^*) \cap \mathscr{S}_z(w_j)$ and $\tilde{z}_j^* \in \mathscr{S}_z(\tilde{w}^*) \cap \mathscr{S}_z(\tilde{w}_j)$.

With VY there must be $\tilde{w}^*, z_0, z_1$ such that *(i)* $m(\tilde{x}^*, \tilde{w}^*) = m(x^*, w^*)$, *(ii)* $z_0 < z_1$, and *(iii)* (w^*, z_0), (w^*, z_1), $(\tilde{w}^*, z_0)$, and $(\tilde{w}^*, z_1)$ all belong to $\mathscr{S}_{wz}$. Of the remaining conditions, BBC and SWTR (and hence WTR) allow VY as a special case. For instance, with BBC if $w = w^*$ then BBC reduces to VY and z_2 becomes redundant. However, BTC, BBC, and WTR are not generally nested; among the three conditions, BTC is the least interesting because it requires the existence of a w–value for which $m(0, w) = m(1, w)$, and only with BTC must $\mathscr{R}(x^*)$ and $\mathscr{R}(\tilde{x}^*)$ have elements in common.

The other extensions to VY exploit the existence of one or more w–values such that $m(x^*, w) = m(x^*, w^*)$ or $m(\tilde{x}^*, w) = m(x^*, w^*)$. Especially when $\boldsymbol{w}$ is vector–valued the existence of such values would be a common occurrence.

With SWTR, one can infer that $m(x^*, w^*) = m(x^*, w)$ by verifying that $\mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w^*, \boldsymbol{z} = z_2) = \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w, \boldsymbol{z} = z_2)$. Equality of $m(x^*, w_2)$ to $m(\tilde{x}^*, \tilde{w}^*)$ then follows using a VY step. WTR entails a repetition of similar steps. Indeed, analogous to SWTR one can infer that $m(x^*, w^*) = m(x^*, w_0) = m(x^*, w_1)$ and that $m(\tilde{w}^*, \tilde{x}^*) = m(\tilde{w}^*, \tilde{w}_0) = m(\tilde{w}^*, \tilde{w}_1)$, from which one can obtain

$$\begin{cases} \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w_0, \boldsymbol{z} = z_0) & = \theta\{m(x^*, w^*), V(x^*, z_0)\}, \\ \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w_1, \boldsymbol{z} = z_1) & = \theta\{m(x^*, w^*), V(x^*, z_1)\}, \\ \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = \tilde{x}^*, \boldsymbol{w} = \tilde{w}_0, \boldsymbol{z} = z_0) & = \theta\{m(\tilde{x}^*, \tilde{w}^*), V(\tilde{x}^*, z_0)\}, \\ \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = \tilde{x}^*, \boldsymbol{w} = \tilde{w}_1, \boldsymbol{z} = z_1) & = \theta\{m(\tilde{x}^*, \tilde{w}^*), V(\tilde{x}^*, z_1)\}, \end{cases}$$

such that from

$$\begin{aligned} &\mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w_0, \boldsymbol{z} = z_0)\mathbb{P}(\boldsymbol{x} = x^*|\boldsymbol{z} = z_0) - \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w_1, \boldsymbol{z} = z_1)\mathbb{P}(\boldsymbol{x} = x^*|\boldsymbol{z} = z_1) \\ &= \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = \tilde{x}^*, \boldsymbol{w} = \tilde{w}_1, \boldsymbol{z} = z_1)\mathbb{P}(\boldsymbol{x} = \tilde{x}^*|\boldsymbol{z} = z_1) - \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = \tilde{x}^*, \boldsymbol{w} = \tilde{w}_0, \boldsymbol{z} = z_0)\mathbb{P}(\boldsymbol{x} = \tilde{x}^*|\boldsymbol{z} = z_0), \end{aligned}$$

it can be concluded that $m(x^*, w^*) = m(\tilde{x}^*, \tilde{w}^*)$.

The Pacman–shaped conditional supports $\mathscr{R}(x^*)$ and $\mathscr{R}(\tilde{x}^*)$ in figure 2 may give the impression that WTR is extreme but WTR is useful with more natural–looking conditional supports if $\boldsymbol{w}$ is vector–valued.

BBC is perhaps the most interesting, albeit not necessarily the most useful, extension. Using VY matching (with $w \in \mathscr{K}(x^*)$ and $\tilde{w}^* \in \mathscr{K}(\tilde{x}^*)$), it can be established that $\delta(\tilde{x}^*, \tilde{w}^*) = \mathbb{E}g\{m(\tilde{x}^*, \tilde{w}^*), \boldsymbol{u}\}$ is identified. Then,

$$\frac{\delta(\tilde{x}^*, \tilde{w}^*) - \mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = \tilde{x}^*, \boldsymbol{w} = \tilde{w}^*, \boldsymbol{z} = z_2)\mathbb{P}(\boldsymbol{x} = \tilde{x}^*|\boldsymbol{z} = z_2)}{\mathbb{P}(\boldsymbol{x} = x^*|\boldsymbol{z} = z_2)} = \mathbb{E}\left[g\{m(\tilde{x}^*, \tilde{w}^*), \boldsymbol{u}\}|\boldsymbol{v} \in V(x^*, z_2)\right],$$

which equals $\mathbb{E}(\boldsymbol{y}|\boldsymbol{x} = x^*, \boldsymbol{w} = w^*, \boldsymbol{z} = z_2)$ if and only if $m(\tilde{x}^*, \tilde{w}^*) = m(x^*, w^*)$.

Assumption F is stronger than assumption E since the techniques described in assumption F can be combined and iterated ad nauseam. This is especially relevant if the objective is estimation rather than identification since being able to match with multiple w–values in $\mathscr{K}(x^*)$ and $\mathscr{K}(\tilde{x}^*)$

can be preferable because *(i)* the $\tilde{w}^*$–value found by VY may lie in a sparse data region and *(ii)* conditioning on a greater set generally results in more efficient estimates. Using multiple w–values when the VY conditions are satisfied can be interpreted as using overidentifying restrictions. If the set of w–values that satisfy assumption F is a continuum then the local parameter $\delta(x^*, w^*)$ can typically be estimated at a better convergence rate. At an extreme, if the set of such w–values has positive measure then $\sqrt{n}$–consistent estimation of the local parameter is possible.

We finish this section by formally showing that the conditions in assumption F are indeed sufficient, which is straightforward as the proof of theorem 2 demonstrates.

Theorem 2. *Assumption F implies assumption E.*

Proof. Suppose without loss of generality that $x^* = 0$.

For WTR, note that $\mathscr{D}_0(0, w^*) = \{(0, z_0^*], (0, z_1^*]\}$, $\mathscr{D}_0(0, w_0) = \{(0, z_0^*], (0, z_0]\}$, and $\mathscr{D}_0(0, w_1) = \{(0, z_1^*], (0, z_1]\}$. By definition 1 (ii) and (iv), $\mathscr{D}_1(0, w^*) \supseteq \{(0, z_0], (0, z_1]\}$, such that by definition 1 (ii), $\mathscr{D}_2(0, w^*) \supseteq \{(z_0, z_1], (0, z_0]\}$. Likewise $\mathscr{D}_2(1, \tilde{w}^*) \supseteq \{(z_0, z_1], (z_0, 1]\}$. Hence, again by definition 1 (iv), $\mathscr{D}_3(0, w^*) \ni \mathscr{U}$.

VY is a special case of SWTR, which is a special case of WTR. The proof of BTC is similar to but simpler than that of WTR.

For BBC note that $\mathscr{D}_0(0, w) = \{(0, z_0], (0, z_1]\}$, such that by definition 1 (ii), $\mathscr{D}_1(0, w) = \{(0, z_0], (0, z_1], (z_0, z_1]\}$. Since $\mathscr{D}_0(1, \tilde{w}^*) = \{(z_0, 1], (z_1, 1], (z_2, 1]\}$ definition 1 (iv) implies that $\mathscr{D}_2(1, \tilde{w}^*) \supseteq \{(z_2, 1], \mathscr{U}\}$, which yields $\mathscr{D}_3(1, \tilde{w}^*) \supseteq \{(0, z_2], \mathscr{U}\}$. Hence by definition 1 (iv), $\mathscr{D}_4(0, w^*) \ni \mathscr{U}$. □

5. Estimation

We now describe an estimation approach based on the WTR and BBC identification methods of assumption F.

Let $\boldsymbol{\eta} = \eta(\boldsymbol{z})$ and $\phi(x, w, \eta) = \mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = x, \boldsymbol{w} = w, \boldsymbol{\eta} = \eta)$. For ease of exposition, assume that the support $\mathscr{S}_{\boldsymbol{w}\boldsymbol{\eta}}$ of $(\boldsymbol{w}, \boldsymbol{\eta})$ is known. We first start with the local parameter $\delta(0, w^*)$, after which we consider the estimation of the global average $\delta(0) = \int_{\mathscr{S}_{\boldsymbol{w}}} \delta(0, w) dF_{\boldsymbol{w}}(w)$, where $F_{\boldsymbol{w}}$ is the distribution function of $\boldsymbol{w}$; we will assume that $\delta(0, w)$ is identified for all $w \in \mathscr{S}_{\boldsymbol{w}}$. When the support of $(\boldsymbol{w}, \boldsymbol{\eta})$ is the Cartesian product of the marginal supports, the estimation procedure below for the global parameter $\delta(0)$ is equivalent to the one discussed in VY. We implicitly assume continuity of the distribution of $\boldsymbol{\eta}$ given $\boldsymbol{w} = w$ and the existence of any derivatives used.

Let $\mathscr{S}_{\boldsymbol{\eta}}(w), \mathscr{S}_{\boldsymbol{w}}(\eta)$ denote the conditional supports of $\boldsymbol{\eta}$ given $\boldsymbol{w} = w$ and $\boldsymbol{w}$ given $\boldsymbol{\eta} = \eta$, respectively. Let further int denote 'interior' and

$$M(\bar{x}, \varpi) = \{w \in \mathscr{S}_{\boldsymbol{w}} : m(\bar{x}, w) = m(\bar{x}, \varpi),\ \operatorname{int} \mathscr{S}_{\boldsymbol{\eta}}(w) \cap \operatorname{int} \mathscr{S}_{\boldsymbol{\eta}}(\varpi) \neq \emptyset\}.$$

A natural estimate of $M(\bar{x}, \varpi)$ for a random sample $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i}) : i = 1, \ldots, n\}$ is

$$\hat{\boldsymbol{M}}(\bar{x}, \varpi) = \Big\{w \in \mathscr{S}_{\boldsymbol{w}} : \sum_{i=1}^{n} \{\hat{\boldsymbol{\phi}}(\bar{x}, \varpi, \boldsymbol{\eta_i}) - \hat{\boldsymbol{\phi}}(\bar{x}, w, \boldsymbol{\eta_i})\}^2 \mathbb{1}_{\boldsymbol{i}}(\varpi) \mathbb{1}_{\boldsymbol{i}}(w) \leq \epsilon_n \hat{\boldsymbol{N}}(w, \varpi) \text{ and } \hat{\boldsymbol{N}}(w, \varpi) \geq n\epsilon_n^{\circ}\Big\}, \quad (11)$$

where $\mathbb{1}_{\boldsymbol{i}}(w) = \mathbb{1}\{w \in \mathscr{S}_{\boldsymbol{w}}(\boldsymbol{\eta_i})\}$, $\hat{\boldsymbol{\phi}}$ is an estimate of ϕ, $\{\epsilon_n\}, \{\epsilon_n^{\circ}\}$ are vanishing sequences of input parameters for which $\{n\epsilon_n\}$ and $\{n\epsilon_n^{\circ}\}$ diverge, $\hat{\boldsymbol{N}}(w, \varpi) = \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{i}}(w) \mathbb{1}_{\boldsymbol{i}}(\varpi)$, and $\boldsymbol{\eta_i} = \eta(\boldsymbol{z_i})$, where η can be replaced by $\hat{\boldsymbol{\eta}}$ if the function η is unknown. The condition $\hat{\boldsymbol{N}}(w, \varpi) \geq n\epsilon_n^{\circ}$ in (11) is used to trim out w–values for which $\operatorname{int} \mathscr{S}_{\boldsymbol{\eta}}(w) \cap \mathscr{S}_{\boldsymbol{\eta}}(\varpi)$ is small for the sample size.

Now let

$$q(x, w, \eta) = \mathbb{E}\big[g\{m(x, w), \boldsymbol{u}\} | \boldsymbol{v} = \eta\big] = \begin{cases} \phi(0, w, \eta) + \eta \phi_\eta(0, w, \eta), & x = 0, \\ \phi(1, w, \eta) - (1 - \eta) \phi_\eta(1, w, \eta), & x = 1, \end{cases}$$

where ϕ_η denotes the partial derivative of ϕ with respect to η. By assumption D, we can achieve the matching of $m(0, w)$ with $m(1, \varpi)$ by using the q function; i.e. $m(0, w) = m(1, \varpi)$ if and only if $q(0, w, \eta) = q(1, \tilde{\varpi}, \eta)$ for any η with $(w, \eta), (\varpi, \eta) \in \mathscr{S}_{\boldsymbol{w\eta}}$. An estimate $\hat{\boldsymbol{q}}$ of q can be constructed using a (differentiable) estimate $\hat{\boldsymbol{\phi}}$.

Then

$$M^*(\bar{x}, \varpi) = \{w \in \mathscr{S}_{\boldsymbol{w}} : \exists \tilde{\varpi} \in M(1 - \bar{x}, \varpi) \text{ s.t. } m(\bar{x}, w) = m(1 - \bar{x}, \tilde{\varpi}),\ \operatorname{int} \mathscr{S}_{\boldsymbol{\eta}}(w) \cap \operatorname{int} \mathscr{S}_{\boldsymbol{\eta}}(\tilde{\varpi}) \neq \emptyset\},$$

can be estimated using

$$\hat{\boldsymbol{M}}^*(\bar{x}, \varpi) = \left\{w \in \mathscr{S}_{\boldsymbol{w}} : \min_{\substack{\tilde{\varpi} \in \hat{\boldsymbol{M}}(1 - \bar{x}, \varpi) \\ \hat{\boldsymbol{N}}(w, \tilde{\varpi}) \geq \epsilon_n^{\bullet}}} \frac{1}{\hat{\boldsymbol{N}}(w, \tilde{\varpi})} \sum_{i=1}^{n} \{\hat{\boldsymbol{q}}(\bar{x}, w, \boldsymbol{\eta_i}) - \hat{\boldsymbol{q}}(1 - \bar{x}, \tilde{\varpi}, \boldsymbol{\eta_i})\}^2 \mathbb{1}_{\boldsymbol{i}}(w) \mathbb{1}_{\boldsymbol{i}}(\tilde{\varpi}) \leq \epsilon_n^*\right\},$$

where $\{\epsilon_n^*\}$ and $\{\epsilon_n^{\bullet}\}$ have properties similar to those of $\{\epsilon_n\}$ and $\{\epsilon_n^{\circ}\}$. We again use trimming.

Finally, noting that

$$\delta(0, w^*) = \mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y} | \boldsymbol{w} \in M(0, w^*)\} + \mathbb{E}\{\boldsymbol{xy} | \boldsymbol{w} \in M^*(1, w^*)\}, \quad (12)$$

an estimate $\hat{\delta}(0, w^*)$ can be obtained by replacing the conditional expectations in (12) by estimates thereof using $\hat{\boldsymbol{M}}(0, w^*)$ and $\hat{\boldsymbol{M}}^*(1, w^*)$ in lieu of $M(0, w^*)$ and $M^*(1, w^*)$, respectively.

$M(0, w^*)$ and $M^*(1, w^*)$ are generally larger than $\{w^*\}$ and $\{\tilde{w}^*\}$, respectively. Therefore, estimates of the conditional expectations in (12) are typically more accurate than those of $\mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} = w^*\}$ and $\mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} = \tilde{w}^*\}$ for $\tilde{w}^* \in M^*(1, w^*)$. As mentioned earlier, in the extreme case in which $M(0, w^*)$ and $M^*(1, w^*)$ are sets of positive measure, $\delta(0, x^*)$ can be estimated at the parametric rate, although the estimation error resulting from the use of $\hat{\boldsymbol{M}}(0, w^*)$ and $\hat{\boldsymbol{M}}^*(1, w^*)$ may then dominate.

Estimation of the global parameter $\delta(0)$ can be based on the integral form of (12):

$$\delta(0) = \int_{\mathscr{S}_w} \mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} \in M(0, w)\}dF_{\boldsymbol{w}}(w) + \int_{\mathscr{S}_w} \mathbb{E}\{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w} \in M^*(1, w)\}dF_{\boldsymbol{w}}(w). \qquad (13)$$

where the first right hand side term is in fact equal to $\mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}\}$, because $\mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} \in M(0, w)\} = \mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} = w\}$. Thus $\delta(0)$ can be estimated by using the conditional expectations in (13), replacing $M(0, w), M^*(1, w), F_{\boldsymbol{w}}$ with $\hat{\boldsymbol{M}}(0, w), \hat{\boldsymbol{M}}^*(1, w)$, and $\hat{\boldsymbol{F}}_{\boldsymbol{w}}$, respectively.

If the support of $(\boldsymbol{w}_i, \boldsymbol{\eta}_i)$ is the Cartesian product of the marginal supports then the above–described procedure is equivalent to the estimation method sketched in VY. To see this, note that $\cup_{w \in \mathscr{S}_w} M(0, w) = \mathscr{S}_{\boldsymbol{w}}$, that then the elements of $\{M(0, w) : w \in \mathscr{S}_{\boldsymbol{w}}\}$ are either identical or disjoint, and that $M^*(1, w)$ is the set of all $\tilde{w}$–values for which $q(1, \tilde{w}, \eta) = q(0, w, \eta)$ for all η, which is what is used in VY.

However, this equivalence no longer holds when $\mathscr{S}_{\boldsymbol{\eta}}(w)$ depends on w and the new procedure generally yields more efficient estimates. Some weighting of observations may be necessary in the estimation of $\mathbb{E}\{(1 - \boldsymbol{x})\boldsymbol{y}|\boldsymbol{w} \in M(0, w)\}$ and $\mathbb{E}\{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w} \in M^*(1, w)\}$ to obtain optimal efficiency for the same reason that the generalized least squares (GLS) estimator is more efficient than the ordinary least squares (OLS) estimator in a standard linear regression context if there is heteroskedasticity. We finish our paper with an example producing a modest efficiency gain for an estimator $\hat{\delta}(0)$ of $\delta(0)$, focusing on the estimation of $\mathbb{E}\{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w} \in M^*(1, w)\}$; note that it is straightforward to construct examples generating arbitrarily large efficiency gains.

Example 1. Suppose that $\boldsymbol{w}$ consists of two elements with support $\{0, 1\}^2$, that all four outcomes have equal probability of occurring, that $\mathscr{S}_{\boldsymbol{\eta}}(0, 1) \cap \mathscr{S}_{\boldsymbol{\eta}}(1, 0) = \mathscr{S}_{\boldsymbol{\eta}}(0, 0) \cap \mathscr{S}_{\boldsymbol{\eta}}(1, 1) = \emptyset$, but that none of the other conditional support intersections are empty. Suppose further that $m(0, w) = 1 - \max(w_1, w_2)$ and $m(1, w) = w_2$. Finally, suppose that it is known that $\mathbb{V}(\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w} = w)$ is independent of w to avoid weighting issues and that the match identities are known, also. Then the following w–values are used for each indicated value of w^*:

w^*	VY	New
(0,0)	(0,1)	(0,1) and (1,1)
(0,1)	(0,0)	(0,0) and (1,0)
(1,0)	(0,0) and (1,0)	(0,0) and (1,0)
(1,1)	(1,0)	(0,0) and (1,0)

For the purpose of estimating the local parameters, VY and the new procedure are equivalent for $w^* = (1,0)$ but for the other w^*–values the relative efficiency is 0.5 in favor of the new procedure. Because aggregation reduces variance, the relative efficieny of the estimates of the global parameters is $10/11$ in favor of the new procedure since the two estimators can be expressed as

$$(3\hat{\boldsymbol{\lambda}}_{\mathbf{00}} + 2\hat{\boldsymbol{\lambda}}_{\mathbf{01}} + 3\hat{\boldsymbol{\lambda}}_{\mathbf{10}})/2, \quad \text{and} \quad (3\hat{\boldsymbol{\lambda}}_{\mathbf{00}} + \hat{\boldsymbol{\lambda}}_{\mathbf{01}} + 3\hat{\boldsymbol{\lambda}}_{\mathbf{10}} + \hat{\boldsymbol{\lambda}}_{\mathbf{11}})/2,$$

respectively, where $\hat{\boldsymbol{\lambda}}_{\boldsymbol{js}} = n^{-1}\sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{w_i} = (j,s)\}\boldsymbol{x_i}\boldsymbol{y_i}$ for $j,s = 0,1$, which all have the same variance; $(3^2 + 1^2 + 3^2 + 1^2)/(3^2 + 2^2 + 3^2) = 10/11$. □

References

Chamberlain, G. (1986): "Asymptotic efficiency in semi–parametric models with censoring," *Journal of Econometrics*, 32, 189–218.

Chiburis, R. (2010): "Semiparametric bounds on treatment effects," *Journal of Econometrics*.

Heckman, J. (1990): "Selectivity bias: new developments," *American Economic Review*, 80(2), 313–318.

Heckman, J. J. (1978): "Dummy endogenous variables in a simultaneous equation system," *Econometrica*, 46(6), 931–959.

Jun, S. J., J. Pinkse, and H. Xu (2009): "Tighter bounds in triangular models," Discussion paper, The Pennsylvania State University.

Jun, S. J., J. Pinkse, H. Xu, and N. Yildiz (2010): "Identification of treatment effects in a triangular system of equations," Discussion paper, The Pennsylvania State University.

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

Vytlacil, E., and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75(1), 757–779.