Supplemental Material for "A Vector Monotonicity Assumption for Multiple Instruments"

Leonard Goff

Last updated: October 8, 2019

Contents

A PM without VM, with two binary instruments

Suppose there are two binary instruments, and that PM holds but not VM. For VM to be violated, there must be a "flip" in which value of one of the instruments –say Z_2– that is the "pro-treatment" state, depending on the value of the other instrument. In other words, for some choice of which instrument is called Z_2, and some choice of labeling for the "0" and "1" values of each instrument, we have that:

$$P(D_i(0,1) \geq D_i(0,0)) = 1 \text{ and } P(D_i(1,1) \leq D_i(1,0)) = 1$$

with both

$$P(D_i(0,1) > D_i(0,0) > 0 \text{ and } P(D_i(1,1) < D_i(1,0)) > 0$$

This is without loss of generality, given the choice to arbitrarily assign the labels $0, 1$.

Now consider the set of possible compliance groups that satisfy PM but not VM, denoted as $\mathcal{G}^{PM-VM}$. Any compliance group $g \in \mathcal{G}^{PM-VM}$ must then be either a complier, always-taker, or never-taker with respect to Z_2, when $Z_1 = 0$. Similarly, any compliance group g must then be either a "defier", always-taker, or never-taker with respect to Z_2 when $Z_1 = 1$. The set of pairs (g_0, g_1), where $g_0 \in \{c, a, n\}$ and $g_1 \in \{d, a, n\}$ exhausts the possible compliance groups, since knowing g_0 and g_1 pins down $D_i(z_1, z_2)$ for all four values of z_1, z_2. This generates a set of 9 possible compliance groups: However, all nine of the above compliance groups cannot coexist at the same time. This is because comparisons across values of Z_1, with Z_2 held fixed, can violate partial monotonicity (if for example both odd compliers and Z_1 defiers both exist in the population). This is a consequence of what Mogstad et al. (2019) call *logical consistency*, applied to the first stage.

name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
odd compliers	N	T	T	N
eager compliers	N	T	T	T
1-only	N	T	N	N
reluctant defiers	T	T	T	N
always takers	T	T	T	T
Z_1 defiers	T	T	N	N
2-only	N	N	T	N
Z_1 compliers	N	N	T	T
never takers	N	N	N	N

Table 1: Rows are possible compliance groups in the set $\mathcal{G}^{PM-VM}$. T and N indicate treatment, or non-treatment, respectively. Not all of these groups can coexist in the population without violating PM.

The possibilities cleanly separate into two cases, depending on whether there are "odd compliers" in the population. If there are, then there can be no Z_1 compliers or Z_1 defiers in the population. This leaves seven groups, depicted in Table 2.

If, on the other hand, $P(G_i = \text{odd complier}) = 0$, then there can be either Z_1 compliers, or Z_1 defiers, but not both. This creates a second type of case. Supposing that $P(G_i = Z_1 \text{ complier}) > 0$, there can be no Z_1 defiers, 1-only units, or reluctant defiers. This leaves five possible groups, depicted in Table 3.

The remaining $P(G_i = Z_1 \text{ defier}) > 0$ case is symmetric with respect to Table 3, up to a relabeling of "0" and "1" for Z_1: in addition to Z_1 defiers, there can be reluctant defiers, 1-only units, always takers and never takers.

Case 1 and Case 2 have very different implications for identification. In Case 2, the

group name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
odd compliers	N	T	T	N
eager compliers	N	T	T	T
reluctant defiers	T	T	T	N
1-only	N	T	N	N
2-only	N	N	T	N
always takers	T	T	T	T
never takers	N	N	N	N

Table 2: Case 1, when $P(G_i = \text{odd complier}) > 0$.

group name	$\mathbf{Z_1 = 0, Z_2 = 0}$	$\mathbf{Z_1 = 0, Z_2 = 1}$	$\mathbf{Z_1 = 1, Z_2 = 0}$	$\mathbf{Z_1 = 1, Z_2 = 1}$
eager compliers	N	T	T	T
Z_1 compliers	N	N	T	T
2-only	N	N	T	N
always takers	T	T	T	T
never takers	N	N	N	N

Table 3: Case 2, when $P(G_i = \text{odd complier}) = 0$ and $P(G_i = Z_1 \text{ complier}) > 0$.

group-specific average treatment effects Δ_g are identified for all groups aside from always takers and never takers. However, it can be readily verified that Assumption IAM holds in Case 2.

However, in Case 1, the Δ_g for $g \notin \{a.t., n.t.\}$ are not identified. With three linearly independent Wald ratios, we only have three equations for five unknowns. This is also generally true under VM, that the Δ_g are not separately identified. However, here we can also show that the Wald estimands do not even identify the Big LATE. This is easiest to see by assuming that the 1-only and 2-only groups are not present. Even then we cannot identify the Big LATE. With this restriction, we would still have $E[Y_i|Z_i = (0,1)] - E[Y_i|Z_i = (0,0)] = p_{odd}\Delta_{odd} + p_{eager}\Delta_{eager}$, $E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (1,1)] = p_{odd}\Delta_{odd} + p_{reluct.}\Delta_{reluct.}$, and $E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (0,0)] = E[Y_i|Z_i = (0,1)] - E[Y_i|Z_i = (0,0)]$. Thus the third equation gives no further information beyond the first. The observable conditional means of Y_i are compatible with any numerical value of the BLATE, which is equal to $p_{odd}\Delta_{odd} + p_{eager}\Delta_{eager} + p_{reluct.}\Delta_{reluct.}$.

B Results about linear 2SLS

The standard method of combining multiple instruments in applied work is to use some variant of the two-stage least squares (2SLS) estimator. In the language of Lemma 2, this corresponds to either letting $h(z)$ be a linear projection of a treatment indicator on the instruments (what we'll call "linear 2SLS"), and letting $h(z)$ equal the propensity score (what we call "regression on the propensity score", also referred to as *fully-saturated*

2SLS).

A known result from Angrist and Imbens (1994) is that under conventional IAM monotonicity, regressing Y on the propensity score function recovers a convex combination of group-specific average treatment effects. In Section C.4, we provide a novel demonstration of this, starting from Lemma 2. However, this result does not extend to vector monotonicity, a property that arises from the fat that VM allows two-way flows between some pairs of points in $\mathcal{Z}$.

In this section we demonstrate two special cases of vector monotonicity in which linear 2SLS with binary instruments recovers a convex combination of causal effects. The first special case is one in which each unit is responsive to at most *one* of the instruments:

Assumption 4 (separable compliance). *For each unit i (i.e. with probability one), there exists a $j \in \{1 \dots J\}$ such that $D_i(z, z_{-j}) = D_i(z, z'_{-j})$ for all z_{-j}, z'_{-j} in the support of Z_{-j} and $z \in \{0,1\}$, i.e. treatment assignment only depends on the value of Z_j.*

In the language of Section 3, this corresponds to a case where all units are Z_j "compliers" for some j (equivalently, all compliance groups correspond to Sperner families $\{j\}$ for some j).

In the following theorem, we will also use a slight strengthening of the notion of vector monotonicity:

Assumption 2* (vector monotonicity with aligned covariances). *With the Z_j normalized such that $C(D_i, Z_{ij}) \geq 0$, for each $j \in \{1 \dots J\}$ and $z_{-j} \in \{0,1\}^{J-1}$, we have that $D_i(1, z_{-j}) \geq D_i(0, z_{-j})$*

Assumption 2* is stronger than Assumption 2, because when some of the instruments are negatively correlated it is possible that the unconditional covariances are negative, even with $Z_j i = 1$ corresponding to the "pro-treatment" state for instrument j.[1]

Our result is then:

Theorem SM1. *Assumptions 1, 2* and 4:*

$$\rho_{2sls} = \sum_{j=1}^{J} w_j \cdot E[Y_i(1) - Y_i(0) | D_i^j(1) > D_i^j(0)]$$

where the weights w_j are positive and sum to one: $w_j = \frac{P(D_i^j(1) > D_i^j(0)) C(D_i, Z_{ji})}{\sum_{j=1}^{J} P(D_i^j(1) > D_i^j(0)) C(D_i, Z_{ji})}$.

Proof. See Appendix F. □

Discussion: Linear 2SLS always identifies a sum of single instrument IV estimators $\rho_j := \frac{C(Y_i, Z_{ji})}{C(D_i, Z_{ji})}$, with weights π_j that add to one (but may be negative). Separable monotonicity allows linear 2SLS to identify a convex combination of LATEs despite the fact

[1] Note that under the construction in Section 3.3 from discrete to binary instruments, the resulting vector of binary instruments will satisfy Assumption 2* so long as if the CEF $E[D|Z_1 = z_m]$ is monotonic in m.

that even under separable compliance, ρ_j need not put positive weight on all compliance groups when the intruments are correlated (this problem has been pointed out by Heckman 2010). The proof of Theorem SM1 reveals that the 2SLS weights π_j are such that the overall weight for each compliance group ends up being positive, despite the fact that each ρ_j is a linear combination of SLATE's (defined in Section 4.1) that generally places some negative weights.

Theorem SM1 also extends to the estimator defined by regression on the propensity score, because under VM and separable compliance it turns out that the propensity score function must be linear, and hence consistently estimated by the first stage of standard 2SLS:

Corollary to Theorem SM1. *Under Assumptions 1, 2*, and 4:*

$$\frac{C(Y_i, P(Z_i))}{V(P(Z_i))} = \rho_{2sls},$$

where $P(Z_i) := E[D_i|Z_i]$.

Proof. By Assumption 1:

$$E[D_i|Z_i] = \sum_g P(G_i = g)D_g(Z_i) = p_{a.t.} + \sum_{j=1}^{J} p_{Z_j} Z_{ji}$$

Since the propensity score is linear in the Z_{ki}, it coincides with the linear projection function used by 2SLS. Now Apply Theorem SM1. □

Our second special case in which linear 2SLS can be justified in a context with VM is when the instruments are independent of one another, or slightly more generally, are what we call "unentangled" in selection:

Assumption 5 (instruments *unentangled* in selection). *For* $j \in \{1 \dots J\}$*:*

$$(D_i(0, Z_{-j,i}), D_i(1, Z_{-j,i})) \perp Z_{ji}$$

With these assumptions:

Lemma 1. *Under Assumptions 1, 2 and 5:*

$$\rho_j = E[Y_i(1) - Y_i(0)|D_i^j(1) > D_i^j(0)]$$

Proof. See Appendix F (note that the proof makes use of Assumption 2*, but this is implied by Assumption 2 when Assumption 5 holds). □

Now we can state our result:

Theorem SM2 (2SLS with unentangled binary instruments). *Under Assumptions 1, 2*, and 5, the two stage least squares estimand is*

$$\rho_{2sls} = \sum_{j=1}^{J} w_j E[Y_i(1) - Y_i(0)|D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i})]$$

where the coefficients w_j *are positive and sum to unity.*

Proof. See Appendix F. $\square$

Although made implicitly in some applied work, the assumption that the instruments are unentangled is very strong. While it appears to be weaker than full independence of the instruments, it is hard to articulate concisely a case in which it holds without independence. For instance, it would be violated in a simple latent index model with homogeneous coefficients:

$$D_i = \mathbb{1}(\alpha + \beta_1 Z_{1i} + \beta_2 Z_{2i} \geq \nu_i) \tag{1}$$

where $\beta_j > 0$ and $(Z_1, Z_2) \perp \nu$, but $Z_1 \not\perp Z_2$. This selection model satisfies both monotonicity and vector monotonicity.

C Various examples from the main text

C.1 The identifying power of Wald ratios with two binary instruments

For example, with two binary instruments, there are five unique pairs (z, w) that are ordered in a vector sense. Table 4 describes these in terms of the compliance groups introduced in Section 3.1. Corresponding to each row in 4 is a LATE that can be identified via a Wald ratio ρ_{zw}, in a case of two binary instruments under VM. For instance, from the first row, $\rho_{(1,0),(0,0)}$ is:

z	**w**	$\mathbf{D_i(z) > D_i(w) \iff G_i = \ldots}$
(1,0)	(0,0)	Z_1 complier or eager complier
(0,1)	(0,0)	Z_2 complier or eager complier
(1,1)	(0,1)	Z_1 complier or reluctant complier
(1,1)	(1,0)	Z_2 complier or reluctant complier
(1,1)	(0,0)	any $g \in \mathcal{G}^c$

Table 4: Third column indicates which compliance groups G_i lead to $D_i(z) > D_i(w)$ with the indicated z,w. For each row in this Table, a Wald ratio ρ_{zw} identifies a LATE under Assumption VM. In the two binary instrument vase under VM: $\mathcal{G}^c = \{Z_1, Z_2, or, and\}$.

$$E[Y_i(1)-Y_i(0)|G_i \in \{Z_1 \text{ comp., reluctant}\}] = \frac{p_{Z_1}}{p_{Z_1} + p_{reluctant}}\Delta_{Z_1} + \frac{p_{reluctant}}{p_{Z_1} + p_{reluctant}}\Delta_{reluctant}$$

Recall that the number of compliance groups under VM scales with the so-called Dedekind numbers $\mathcal{D}_J$, which grow much faster than $(2^J \times (2^J - 1))/2$, the number of unique Wald estimands. Furthermore, the ρ_{zw} are not even all linearly-independent: in the example of two binary instruments only three of the five are.

Thus, we can see that it will in general be hopeless to identify Δ_g for each compliance group $g \in \mathcal{G}^c$ separately, because we will lack an order condition on the number of pairs (z, w) in which there is variation in treatment take-up.[2] The exception is the familiar

[2] Mountjoy (2018) alludes to this point in the context of a closely related monotonicty assumption.

case of $J = 1$, where $(2^J \times (2^J - 1))/2 = \mathcal{D}_J - 2 = 1$ and IAM and VM are equivalent.[3] Furthermore, for $J > 1$ under VM, we also can't separately identify the occupancy of the compliance groups $p_g = P(G_i = g)$, aside from never-takers and always-takers, without further assumptions. An interesting problem is whether for such identification it is sufficient to assume a linear single index model underlying selection.[4]

C.2 Example of ρ_h for Theorem 1

In a case with two binary instruments, suppose we are interested in $SLATE_1$, the average treatment effect among units for whom $D_i(1, Z_{2i}) > D_i(0, Z_{2i})$. This event is equivalent to the event that i is a Z_1 complier, or i is an *and*-complier and $Z_{2i} = 1$, or i is an *or*-complier and $Z_{2i} = 0$. If one forms the linear combination:

$$\begin{aligned}
&P(Z_{2i} = 1) \left(E[Y_i|Z_i = (1,1)] - E[Y_i|Z_i = (0,1)]\right) + P(Z_{2i} = 0) \left(E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (0,0)]\right) \\
&\qquad = P(Z_{2i} = 1) \left(p_{Z_1} + p_{and}\right) \left[\frac{p_{Z_1}}{p_{Z_1} + p_{and}} \Delta_{Z_1} + \frac{p_{and}}{p_{Z_1} + p_{and}} \Delta_{and}\right] \\
&\qquad + P(Z_{2i} = 0) \left(p_{Z_1} + p_{or}\right) \left[\frac{p_{Z_1}}{p_{Z_1} + p_{or}} \Delta_{Z_1} + \frac{p_{or}}{p_{Z_1} + p_{or}} \Delta_{or}\right] \\
&\qquad = P(Z_{2i} = 1)(p_{Z_1}\Delta_{Z_1} + p_{and}\Delta_{and}) + P(Z_{2i} = 0)(p_{Z_1}\Delta_{Z_1} + p_{or}\Delta_{or}) \\
&\qquad = p_{Z_1} \cdot \Delta_{Z_1} + P(Z_{2i} = 1)p_{and} \cdot \Delta_{and} + P(Z_{2i} = 0)p_{or} \cdot \Delta_{or}
\end{aligned}$$

Similarly

$$\begin{aligned}
&P(Z_{2i} = 1) \left(E[D_i|Z_i = (1,1)] - E[D_i|Z_i = (0,1)]\right) + P(Z_{2i} = 0) \left(E[D_i|Z_i = (1,0)] - E[D_i|Z_i = (0,0)]\right) \\
&\qquad = p_{Z_1} + P(Z_{2i} = 1)p_{and} + P(Z_{2i} = 0)p_{or} \\
&\qquad = P\left(D_i(1, Z_{2i}) > D_i(0, Z_{2i})\right)
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{P(Z_{2i} = 1) \left(E[Y_i|Z_i = (1,1)] - E[Y_i|Z_i = (0,1)]\right) + P(Z_{2i} = 0) \left(E[Y_i|Z_i = (1,0)] - E[Y_i|Z_i = (0,0)]\right)}{P(Z_{2i} = 1) \left(E[D_i|Z_i = (1,1)] - E[D_i|Z_i = (0,1)]\right) + P(Z_{2i} = 0) \left(E[D_i|Z_i = (1,0)] - E[D_i|Z_i = (0,0)]\right)} \\
&= \frac{p_{Z_1}}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{Z_1} + \frac{P(Z_{2i} = 1)p_{and}}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{and} + \frac{P(Z_{2i} = 0)p_{or}}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{or} \\
&= \frac{P(G_i = Z_1)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{Z_1} + \frac{P(Z_{2i} = 1 \& G_i = and)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{and} + \frac{P(Z_{2i} = 0 \& G_i = or)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{or} \\
&= \frac{P(D_i(1, Z_{2i}) > D_i(0, Z_{2i}) \& G_i = Z_1)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{Z_1} + \frac{P(D_i(1, Z_{2i}) > D_i(0, Z_{2i}) \& G_i = and)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \Delta_{and} + \frac{P(D_i(1, Z_{2i}) > D_i(0, Z_{2i}) \& G_i = or)}{D_i(1, Z_{2i}) > D_i(0, Z_{2i})} \\
&= P(G_i = Z_1 | D_i(1, Z_{2i}) > D_i(0, Z_{2i}))\Delta_{Z_1} + P(G_i = and | D_i(1, Z_{2i}) > D_i(0, Z_{2i}))\Delta_{and} + P(G_i = or | D_i(1, Z_{2i}) > D_i(0, Z_{2i}))\Delta_{or} \\
&= E[Y_i(1) - Y_i(0) | D_i(1, Z_{2i}) > D_i(0, Z_{2i})] = SLATE_1
\end{aligned}$$

To see that this same particular combination of causal effects is operationalized by the 2SLS-like estimator ρ_h suppose we choose $h(z)$ such that $C(Z_{1i}, H_i) = 1$, $C(Z_{2i}, H_i) = 0$,

[3] With IAM, we can identify Δ_g for all $g \in \mathcal{G}^c$ that occur with positive probability when the instruments have full support; the number of linearly-independent Walds is and $|\mathcal{G}^c|$ are both equal to $2^J - 1$.

[4] With continuous instruments and an assumption of no never-takers, Theorem 1 of Ichimura and Thompson (1998) would imply identification of F_β up to a scale normalization, in a model where $D_i = \mathbb{1}(Z_i'\beta_i \geq \beta_{0i})$, with $\beta_i \perp Z_i$ (the VM positivity restriction: $P(\beta_{ji} \geq 0) = 1$ for $j > 0$, is not necessary here for identification). Given the marginal distributions of β_i and Z_i, one could compute p_g for all $g \in \mathcal{G}$.

and $C(Z_{1i}Z_{2i}, H_i) = P(Z_{2i} = 1)$. Then, by Lemma 2:

$$\begin{aligned}
\rho_h &= \frac{p_{Z_1}C(Z_{1i}, H_i)\Delta_{Z_1} + p_{Z_1}C(Z_{1i}, H_i)\Delta_{Z_2} + p_{and}C(Z_{1i}Z_{2i}, H_i)\Delta_{and} + p_{or}C(Z_{1i} + Z_{2i} - Z_{1i}Z_{2i}, H_i)\Delta_{or}}{p_{Z_1}C(Z_{1i}, H_i) + p_{Z_1}C(Z_{1i}, H_i) + p_{and}C(Z_{1i}Z_{2i}, H_i) + p_{or}C(Z_{1i} + Z_{2i} - Z_{1i}Z_{2i}, H_i)} \\
&= \frac{p_{Z_1}\Delta_{Z_1} + p_{and}P(Z_{2i} = 1)\Delta_{and} + p_{or}(1 - P(Z_{2i} = 1))\Delta_{or}}{p_{Z_1} + p_{and}P(Z_{2i} = 1) + p_{or}(1 - P(Z_{2i} = 1))} \\
&= \frac{{}_{Z_1} \cdot \Delta_{Z_1} + P(Z_{2i} = 1)p_{and} \cdot \Delta_{and} + P(Z_{2i} = 0)p_{or} \cdot \Delta_{or}}{P\left(D_i(1, Z_{2i}) > D_i(0, Z_{2i})\right)}
\end{aligned}$$

C.3 The matrix M_J for $J = 3$

	{1}	{2}	{3}	{1,2}	{1,3}	{2,3}	{1,2,3}
{1}	1						
{2}		1					
{3}			1				
{1,2}				1			
{1,3}					1		
{2,3}						1	
{1,2,3}							1
{1},{2}	1	1		-1			
{2},{3}		1	1			-1	
{1},{3}	1		1		-1		
{1},{2},{3}	1	1	1	-1	-1	-1	1
{1,2},{3}			1	1			-1
{1,3},{2}		1			1		-1
{2,3},{1}	1					1	-1
{1,2},{1,3}				1	1		-1
{1,2},{2,3}				1		1	-1
{1,3},{2,3}					1	1	-1
{1,2},{1,3},{2,3}				1	1	1	-2

Table 5: The matrix M_3 defined in Section 4. Empty cells indicate a zero.

C.4 Special cases of Lemma 2 under IAM

When we have a single binary instrument, consider the function $h(Z_i) = Z_i$. Under monotonicity, the compliance groups are $\mathcal{G} = \{never - taker, always - taker, complier\}$. The functions $D_{a.t.}(Z_i)$ and $D_{n.t.}(Z_i)$ are constants, so only the complier term contributes. Since $D_{complier}(Z_i) = Z_i$, we have that $C(Y_i, h(Z_i)) = P_{complier}\alpha_{complier}$ and $C(D_i, h(Z_i)) = P_{complier}$, justifying the traditional Wald IV estimator for $\alpha_{complier}$.

In the case of a vector instrument with finite support and with monotonicity, there is a well-defined ordering $z_1 \dots z_{\mathcal{M}}$ of points in $\mathcal{Z}$ (where $\mathcal{M} = |\mathcal{Z}|$) such that $P(D_i(z_m) \geq D_i(z_{m-1})) = 1$ (this ordering may be non-unique if there are no "compliers" between some pairs of points in $\mathcal{Z}$). If we choose $h(Z_i) = P(Z_i)$, the propensity score function,

ρ_h is equal to the "regression on the propensity score" estimator $C(Y_i, P(Z_i))/V(P(Z_i))$ (since $C(D_i, P(Z_i)) = V(P(Z_i))$). Recall that G_i maps one-to-one with as the smallest m for which $D_i(z_m) = 1$ (provided i is not a "never-taker"). Thus we can use the notation $D_m(z) := \mathbb{1}(z \succ z_m)$, indicating that z succeeds z_m in the sequence $z_1 \dots z_{\mathcal{M}}$. The weights are positive so long as for all m: $C(D_m(Z_i), P(Z_i)) \geq 0$, which occurs iff:

$$E[P(Z_i)|D_m(Z_i) = 1] - E[P(Z_i)|D_m(Z_i) = 0] = E[P(Z_i)|Z_i \succ z_m] - E[P(Z_i)|Z_i \preceq z_m] > 0$$

This inequality will hold for all m, since (using independence) $P(z_m)$ is monotonically increasing in m. This yields a novel demonstration of fact that the"regression on the propensity score" estimator identifies a convex combination of LATEs, as shown in Angrist and Imbens (1994).

C.5 Proof of Lemma 2

First, note that for any $z \in \mathcal{Z}$, since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$ we have that:

$$\begin{aligned} E[Y_i|Z_i = z, G_i = g] &= E[Y_i(0)|Z_i = z, G_i = g] + E[D_i(z)(Y_i(1) - Y_i(0))|Z_i = z, G_i = g] \\ &= E[Y_i(0)|G_i = g] + D_g(z) \cdot E[Y_i(1) - Y_i(0)|G_i = g] \end{aligned}$$

using independence, and thus for any $z, w \in \mathcal{Z}$:

$$E[Y_i|Z_i = z, G_i = g] - E[Y_i|Z_i = w, G_i = g] = (D_g(z) - D_g(w)) \cdot E[(Y_i(1) - Y_i(0))|G_i = g]$$

Now:

$$\begin{aligned} &C(Y_i, h(Z_i)) \\ &= E[C(Y_i, h(Z_i)|G_i)] \qquad \text{(since } C(A,B) = E[C(A,B|C)] + C(E[A|C], E[B|C])) \\ &= \sum_g P(G_i = g) C(Y_i, h(Z_i)|G_i = g) \\ &= \sum_g P(G_i = g) \sum_z h(z) C(Y_i, \mathbb{1}(Z_i = z)|G_i = g) \qquad \text{(since } h(Z_i) = \sum_z h(z)\mathbb{1}(Z_i = z)) \\ &= \sum_g P(G_i = g) \sum_z h(z) \left(E[Y_i \mathbb{1}(Z_i = z)|G_i = g] - E[Y_i|G_i = g] P(Z_i = z|G_i = g)\right) \\ &= \sum_g P(G_i = g) \sum_z h(z)\pi_z \sum_w \pi_w \left(E[Y_i|Z_i = z, G_i = g] - E[Y_i|Z_i = w, G_i = g]\right) \\ &= \sum_g P(G_i = g) \left\{ \sum_{z,w} h(z)\pi_z\pi_w \left(D_g(z) - D_g(w)\right) \right\} \Delta_g \end{aligned}$$

Furthermore

$$\begin{aligned} \sum_{z,w} h(z)\pi_z\pi_w \left(D_g(z) - D_g(w)\right) &= \sum_z h(z)\pi_z D_g(z) - \left(\sum_z h(z)\pi_z\right)\left(\sum_w \pi_w D_g(w)\right) \\ &= C(D_g(Z_i), h(Z_i)) \end{aligned}$$

An analagous sequence of steps shows that the denominator

$$C(D_i, h(Z_i)) = \sum_g P(G_i = g) C(D_g(Z_i), h(Z_i))$$

D Latent index representation theorems

A well-known result by Vytlacil (2002) shows that the traditional LATE framework with IAM is equivalent to an implicit latent index model in which selection behavior is described by a function of the form:

$$D_i(z) = \mathbb{1}(v(z) \geq U_i) \tag{2}$$

where $U_i \perp Z_i$ and the function $v(\cdot)$ is common to all agents. Without loss of generality, $\nu(z)$ can be taken to be the propensity score function $P(z) := E[D_i|Z_i = z]$ and U_i is distributed uniformly on the unit interval, conditional on any realization of Z_i.

An IV model that assumes either VM or PM will not generally admit of a representation of the form of Equation 2 with a scalar U_i, since either assumption allows two-way flows between some points in $\mathcal{Z}$. Equation 2 can never allow two-way flows $P(D_i(z) > D_i(z')) > 0$ and $P(D_i(z') > D_i(z)) > 0$, because this would imply that for some U_i both $\nu(z) \geq U_i > \nu(z')$ and $\nu(z) < U_i \leq \nu(z')$.

Mogstad et al. (2019) show that in the case of PM, since IAM is satisfied with respect to Z_j conditional on a realization of Z_j, there exists a set of J latent indices yielding redundant representations of treatment:

$$D_i = D_i^1(Z_{1i}) = D_i^2(Z_{2i}) = \dots D_i^J(Z_{Ji}) \tag{3}$$

where $D_i^j(z_i) = \mathbb{1}(P(z_j, z_{-j}) \geq U_{ji})$ and $U_i := (U_{1i}, U_{2i}, \dots, U_{Ji})$ is a set of J heterogeneity parameters with uniform marginal distributions, interpreted in terms of unit i's treatment responsiveness to each instrument j.

Since VM is a special case of Assumption PM, a representation of selection in terms of U_i and the equations $D_i^j(z_i) = \mathbb{1}(P(z_j, z_{-j}) \geq U_{ji})$ does exist under Assumption VM. This represents heterogeneity in terms of a J-dimensional vector of real numbers. However, an alternative representation under vector monotonicity is simply in terms of "compliance group" G_i, as a nonseparable threshold crossing model:

$$D_i(z) = \mathbb{1}(\mu_D(z, G_i) \geq 0) \tag{4}$$

where the function $\mu_D(z, G_i)$ is monotonic in each component of the vector z for all groups g. Existence of this representation is trivial because we can set $\mu_D(z, g) = D_i(z) - \epsilon$ for any i in group g, and ϵ is any small number in $(0, 1)$ (to turn the weak inequality into a strict one). When the instruments are discrete and finite, this allows heterogeneity to represents by a finite number of groups (as we show in the main text, G_i is a discrete random variable taking on $\mathcal{D}(J)+1$ values, where $\mathcal{D}(J)$ is the J^{th} number in the *Dedekind* sequence.

Equation (4) is a special case of the nonseparable threshold crossing models considered by Heckman and Vytlacil (2001). They discuss how in general, such models break the connection between the local instrumental variables (LIV) estimator, and marginal treatment effects, which could be defined in this case as $\Delta_g = E[Y_i(1) - Y_i(0)|G_i = g]$.

Indeed, we have seen that these parameters do not appear to be identified, even under VM. On the other hand, we have shown that weighted averages of the Δ_g which satisfy Property M *are* identified, though not by a LIV procedure.

E A linear special case of estimation with covariates

A special case in which we can manage covariates in a much simplified estimator is when $\Delta_g(x) = \Delta_g$, that is group-specific average treatment effects do not depend on X, and furthermore groups are independent of the covariates: $G_i \perp X_i$. Both conditions are strong. The latter has the testable implication that $P(C_i = 1|X_i = x)$ does not depend on x for any $c(\cdot)$ satisfying property M. The former condition describes a case in which, borrowing from the language of **?**, the local average response function $LARF_g(x,d) = E[Y_i(d)|X_i = x, G_i = g]$ is additively separable between x and d. In this case, we will be able to handle covariates by adding them as additional "regressors" in the numerator and denominator of ρ_h, when the relevant CEFs are linear (in this case no interactions between X_i and the Z_i are necessary in the first stage).

An alternative derivation of $\Delta_c = E[h(Z_i,X_i)Y_i]/E[h(Z_i,X_i)D_i]$ helps motivate this result. The numerator of ρ_h is, by the law of iterated expectations and the conditional independence assumption:

$$\begin{aligned}
E[Y_i h(Z_i,X_i)] &= \sum_g P(G_i = g)\int dF_{X|G}(x|g)E[Y_i h(Z_i,x)|G_i = g, X_i = x] \\
&= \sum_g P(G_i = g)\int dF_{X|G}(x|g)E[\{Y_i(0) + D_i\Delta_i\}\, h(Z_i,x)|G_i = g, X_i = x] \\
&= \sum_g P(G_i = g)\int dF_{X|G}(x|g)\Big\{E[Y_i(0)|G_i = g, X_i = x]\cancel{E[h(Z_i,x)|X_i = x]} \\
&\qquad\qquad +E[D_g(Z_i)\Delta_i h(Z_i,x)|G_i = g, X_i = x]\Big\} \\
&= \sum_g P(G_i = g)\int dF_{X|G}(x|g)E[D_g(Z_i)h(Z_i,x)|X_i = x]\cdot\Delta_g(x)
\end{aligned}$$

where $\Delta_g(x) := E[\Delta_i|G_i = g, X_i = x]$. By the same sequence of steps:

$$E[D_i h(Z_i,X_i)] = \sum_g P(G_i = g)\int dF_{X|G}(x|g)E[D_g(Z_i)h(Z_i,x)|X_i = x]$$

Analogously:

$$\Delta_c = \sum_g P(G_i = g)\int dF_{X|G}(x|g)\frac{E[c(g,Z_i)|X_i = x]}{E[c(G_i,Z_i)]}\cdot\Delta_g(x)$$

with $E[c(G_i,Z_i)] = \sum_g P(G_i = g)\int dF_{X|G}(x|g)E[c(g,Z_i)|X_i = x]$. Taken together, these expressions show that if it is possible to simultaneously choose h such that

1. $E[h(Z_i,x)|X_i = x] = 0$ for all $x \in \mathbb{X}$
2. $E[D_g(Z_i)h(Z_i,x)|X_i = x] = E[c(g,Z_i)|X_i = x]$ for all $g \in \mathcal{G}^c$, $x \in \mathbb{X}$

Note that condition 2. above requires that it be possible to tune the expectation $E[D_{g(S)}(Z_i)h(Z_i,x)|X_i =$ $x]$ for each of the simple compliance groups $g(S)$ to a desired value, conditional on each value of x (with the rest of the compliance groups then pinned down by Property M).

If one is willing to maintain the assumptions of $\Delta_g(x)=\Delta_g$ and $X_i \perp G_i$, we have that

$$\sum_g P(G_i=g)\Delta_g \int dF_{X|G}(x|g)E[D_g(Z_i)h(Z_i,x)|X_i=x] = \sum_g P(G_i=g)\Delta_g E[D_g(Z_i)h(Z_i,X_i)]$$

and similarly for Δ_c, so we simply need for $E[D_g(Z_i)h(Z_i,X_i)] = E[c(g(S),Z_i)]$ overall, rather than conditional on each value of x. In practice, this along with condition 1., can be achieved by residualizing both the numerator and denominator of our estimator with respect to X_i, in a conditional expectation sense.

To construct the estimator, we consider a vector $(X_{1i}\dots X_{Li})$ of functions of X_i that includes a constant and is rich enough that $E[\Gamma_{ji}|X_i]$ is linear in the X_{li} for each j (for example, if $(X_{1i}\dots X_{Li})$ is a "fully-saturated" set of group dummies). Let X be an $n\times L$ design matrix of observations of the X_{li}, and define $\mathcal{X}=[\Gamma,X]$. Now consider the estimator:

$$\hat{\rho} = ((\hat{\lambda},0\dots0)'\mathcal{X}^{\dagger}D)^{-1}(\hat{\lambda},0\dots0)'\mathcal{X}^{\dagger}Y = (\hat{\lambda}'(\Gamma'M_X\Gamma)^{-1}\Gamma'M_XD)^{-1}\hat{\lambda}'(\Gamma'M_X\Gamma)^{-1}\Gamma'M_XY$$

Here $(0\dots0)$ is a vector of L zeros, thus $\hat{\rho}$ corresponds to adding the X_j as regressors when calculating the linear projections of Y and D on the instrument products in Γ, but then ignoring the coefficients on the X. The second equality uses Frisch-Waugh-Lovell to rewrite this expression in terms of M_X, the an $n\times n$ projection matrix into the null space of of X. Recall that $\hat{\lambda}$ depends on the causal parameter of interest, and may itself depend on the data but is assumed to be consistent component-wise for $\lambda_j = P(C_i=1|G_i=g(S_j))$ for an arbitrary indexing S_j of the simple compliance groups. Here we also focused on the un-regularized case of estimation ($\alpha=0$) for simplicity.

The estimator above corresponds to setting $\hat{H}_i := \hat{h}(Z_i,X_i)$ to be the i^{th} component of the vector $n(\mathcal{X}')^{\dagger}(\hat{\lambda},0\dots0) = M_X\Gamma'(\Gamma M_X\Gamma')^{-1}\hat{\lambda}$. I now show that this choice of $\hat{h}$ satisfies the desired properties $E[h(Z_i,x)|X_i=x]=0$ and $E[\Gamma_{ji}h(Z_i,X_i)]=\lambda_j$, asymptotically. We assume that $\mathcal{X}$ has full column rank, so that $\mathcal{X}^{\dagger}=(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'$. Then we have:

$$\hat{E}[\Gamma_{ji}\hat{H}_i] = \frac{1}{n}[\Gamma'H]_j = [\Gamma'M_X\Gamma(\Gamma M_X\Gamma')^{-1}\hat{\lambda}]_j = \hat{\lambda}_j$$

verifying the sample analogue of the unconditional version of condition 2.

Now consider condition 1. Note that assuming finite second moments, the probability limit of numerator $\hat{\lambda}'(\Gamma'M_X\Gamma)^{-1}\Gamma'M_XY$ of $\hat{\rho}$ exists and can be written:

$$\begin{aligned}\lambda'E\left[(\Gamma_i - E[X_iX_i']^{-1}E[X_i\Gamma_i])(\Gamma_i - E[X_iX_i']^{-1}E[X_i\Gamma_i])'\right]^{-1}E\left[(\Gamma_i - E[X_iX_i']^{-1}E[X_i\Gamma_i])Y_i\right]\\ = \lambda'E\left[(\Gamma_i - E[\Gamma_i|X_i])(\Gamma_i - E[\Gamma_i|X_i])'\right]^{-1}E\left[(\Gamma_i - E[\Gamma_i|X_i])Y_i\right]\\ = E\left[\sum_j w_j(\Gamma_{ji} - E[\Gamma_{ji}|X_i])Y_i\right]\end{aligned}$$

where w_j is the j^{th} component of the vector $E\left[(\Gamma_i - E[\Gamma_i|X_i])(\Gamma_i - E[\Gamma_i|X_i])'\right]^{-1}\lambda$, and I have used the assumption that $E[\Gamma_{ji}|X_i]$ is linear in each component X_{ji}, and hence coincides with the corresponding linear projection. This expression reveals that the probability limit of $\hat{\rho}$ sets $h(Z_i, X_i) = \sum_j w_j(\Gamma_{ji} - E[\Gamma_{ji}|X_i])$, where note that the w_j are non-stochastic. Thus,

$$E[h(Z_i, X_i)|X_i = x] = \sum_j w_j E\left[\Gamma_{ji} - E[\Gamma_{ji}|X_i]|X_i = x\right] = 0$$

for any $x \in \mathbb{X}$, verifying condition 1.

F Supplemental Material Proofs

F.1 Proof of Theorem SM1

We start with the $J = 2$ case to build the intuition, and present the generalization afterwards. Simple algebra shows that the 2SLS estimand can be written

$$\rho_{2sls} = \frac{\pi_1 C(Y_i, Z_{1i}) + \pi_2 C(Y_i, Z_{2i})}{\pi_1 C(D_i, Z_{1i}) + \pi_2 C(D_i, Z_{2i})}$$

where π_1 and π_2 are the population regression coefficients from the first-stage regression of D on Z_1 and Z_2.

As we've already shown:

$$E[Y_i|Z_{1i} = 1] - E[Y_i|Z_{1i} = 0] = E[D_i(1, Z_{2i})(Y_i(1) - Y_i(0))|Z_{1i} = 1] - E[D_i(0, Z_{2i})(Y_i(1) - Y_i(0))|Z_{1i} = 0]$$

By separable monotonicity, we can divide all units into 4 groups: always-takers (a.t.), never-takers (n.t.), compliers for the first instrument (Z_1), and compliers for the second instrument (Z_2). Applying the law of total probability to the above expression, only the two complier groups contribute, since $D_i(1, Z_{2i}) = D_i(0, Z_{2i}) = 0$ for the never-takers and

$$E[Y_i(1) - Y_i(0)|Z_{1i} = 1, a.t.] = E[Y_i(1) - Y_i(0)|Z_{1i} = 0, a.t.] = E[Y_i(1) - Y_i(0)|a.t.]$$

by the independence assumption. Thus we have:

$$\begin{aligned} E[Y|Z_1 = 1] - E[Y_i|Z_1 = 0] &= p_{Z_1} E[Y(1) - Y(0)|Z_1 = 1, Z_1] \\ &\quad + p_{Z_2}\left(E[Z_2|Z_1 = 1, G = Z_2] - E[Z_2|Z_1 = 0, G = Z_2]\right) E[Y(1) - Y(0)|G = \\ &= p_{Z_1} E[Y(1) - Y(0)|G = Z_1] + p_{Z_2} \frac{C(Z_1, Z_2)}{V(Z_1)} E[Y(1) - Y(0)|G = Z_2] \end{aligned} \tag{5}$$

where we've used the independence assumption. The same steps lead to an analagous expression for Z_2. Now consider the regression coefficient π_1. It is:

$$\begin{aligned} \pi_1 &= \tfrac{1}{V(Z_1)(1-\rho_{12}^2)} \left[C(D, Z_1) - \tfrac{C(Z_1, Z_2)}{V(Z_2)} C(D, Z_2) \right] \\ &= \tfrac{1}{1-\rho_{12}^2} \left[\tfrac{C(D, Z_1)}{V(Z_1)} - \tfrac{C(Z_1, Z_2)}{V(Z_1)} \cdot \tfrac{C(D, Z_2)}{V(Z_2)} \right] \end{aligned}$$

where ρ_{12} is the Pearson correlation coeffient between Z_1 and Z_2, and we've simplified $C(Z_1, Z_1 - \frac{C(Z_1,Z_2)}{V(Z_2)} Z_2)$ to $V(Z_1)(1-\rho_{12}^2)$. By the same steps as those leading to Eq. (5):

$$\frac{C(D,Z_1)}{V(Z_1)} = E[D|Z_1=1] - E[D|Z_1=0] = p_{Z_1} + p_{Z_2}\frac{C(Z_{1i},Z_{2i})}{V(Z_{1i})}$$

and

$$\frac{C(D,Z_2)}{V(Z_2)} = E[D_i|Z_2=1] - E[D|Z_2=0] = p_{Z_2} + p_{Z_1}\frac{C(Z_{1i},Z_{2i})}{V(Z_{2i})}$$

Thus

$$\begin{aligned}\pi_1 &= \frac{1}{1-\rho_{12}^2}\left[p_{Z_1} + \cancel{p_{Z_2}\frac{C(Z_1,Z_2)}{V(Z_1)}} - \frac{C(Z_1,Z_2)}{V(Z_1)}\left(\cancel{p_{Z_2}} + p_{Z_1}\frac{C(Z_1,Z_2)}{V(Z_2)}\right)\right] \\ &= p_{Z_1}\frac{1-\rho_{12}^2}{1-\rho_{12}^2} = p_{Z_1}\end{aligned}$$

and similarly $\pi_2 = p_{Z_2}$. In other words, under separable monotonicity, the linear regression control in 2SLS is sufficient to isolate the compliers for each instrument (we shall see that this property also holds for $J > 2$).

The 2SLS estimator can now be written, using Equation (5):

$$\begin{aligned}\rho_{2sls} &= \frac{p_{Z_1}C(Y,Z_1) + p_{Z_2}C(Y,Z_2)}{p_{Z_1}C(D,Z_1) + p_{Z_2}C(D,Z_2)} \\ &= \frac{p_{Z_1}\left(p_{Z_1} + p_{Z_2}\frac{C(Z_1,Z_2)}{V(Z_1)}\right)}{p_{Z_1}C(D,Z_1) + p_{Z_2}C(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_1] \\ &\quad + \frac{p_{Z_2}\left(p_{Z_2} + p_{Z_1}\frac{C(Z_1,Z_2)}{V(Z_2)}\right)}{p_{Z_1}C(D,Z_1) + p_{Z_2}C(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_2] \\ &= \frac{p_{Z_1}C(D,Z_1)}{p_{Z_1}C(D,Z_1) + p_{Z_2}C(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_1] \\ &\quad + \frac{p_{Z_2}C(D,Z_2)}{p_{Z_1}C(D,Z_1) + p_{Z_2}C(D,Z_2)} \cdot E[Y(1)-Y(0)|G=Z_2]\end{aligned}$$

Since $C(D,Z_j) \geq 0$ by Assumption 2*, the weights are positive.

To show the $J > 2$ case, note that we now have $J+2$ disjoint groups: always-takers, never-takers and compliers for each instrument 1 to J. We use the notation Ck_i to indicate the event that $D_i^k(1, z_{-j}) > D_i^k(0, z_{-j})$ for all z_{-j} and hence $D^k(1) > D^k(0)$. Equation (5) now generalizes, by the law of iterated expections, to:

$$E[Y|Z_j=1] - E[Y|Z_j=0] = \sum_k p_{Ck}\left(E[Z_k|Z_j=1] - E[Z_k|Z_j=0]\right) E[Y(1)-Y(0)|Ck]$$

Similarly, we have that

$$E[D|Z_j=1] - E[D|Z_j=0] = \sum_k p_{Ck}\left(E[Z_k|Z_j=1] - E[Z_k|Z_j=0]\right) \tag{6}$$

This latter expression gives us the property that the multiple regression coefficient $\pi_j = p_{Cj}$ for all j. The reason is that the vector of regression coefficients π is the unique vector satisfying $\Sigma\pi = C$, where Σ is the $J \times J$ covariance matrix of the instruments and C is a vector of covariances between the treatment D and each instrument Z_j. This can be rewritten as for each j:

$$\sum_k C(Z_k, Z_j)\pi_k = C(D, Z_j)$$

Substituting in the guess that $\pi_k = p_{Ck}$ yields Equation (6). The 2SLS estimand is:

$$\begin{aligned}\rho_{2sls} &= \frac{\sum_j \pi_j C(Y, Z_j)}{\sum_j \pi_j C(D, Z_j)} = \frac{\sum_j p_{Cj} \sum_k p_{Ck} C(Z_j, Z_k) E[Y(1) - Y(0)|Ck]}{\sum_j p_{Cj} C(D, Z_j)} \\ &= \frac{\sum_k p_{Ck} \left(\sum_j p_{Cj} C(Z_j, Z_k)\right) E[Y(1) - Y(0)|Ck]}{\sum_j p_{Cj} C(D, Z_j)} \\ &= \frac{\sum_k p_{Ck} C(D, Z_k) E[Y(1) - Y(0)|Ck]}{\sum_j p_{Cj} C(D, Z_j)}\end{aligned}$$

where we've used that $C(Y, Z_j) = \sum_k p_{Ck} C(Z_j, Z_k) E[Y(1) - Y(0)|Ck]$ and $C(D, Z_j) = \sum_k p_{Ck} C(Z_j, Z_k)$. $C(D, Z_k)$ is positive for all k by Assumption 2*.

F.2 Proof of Lemma 1

Fix any $j \in \{1 \ldots J\}$.

$$\begin{aligned}E[Y_i|Z_{ji} = 1] - &E[Y_i|Z_{ji} = 0] \\ &= E[D_i(1, Z_{ji})(Y_i(1) - Y_i(0))|Z_{ji} = 1] - E[D_i(0, Z_{2i})(Y_i(1) - Y_i(0))|Z_{ji} = 0] \\ &= E[(D_i(1, Z_{2i}) - D_i(0, Z_{-j,i}))(Y_i(1) - Y_i(0))] \\ &= P(D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i}))E[Y_i(1) - Y_i(0)|D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i})]\end{aligned}$$

where we've used $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$ and Assumption 1 in the first step, and vector monotonicity in the second. Similarly $E[D_i|Z_{ji} = 1] - E[D_i|Z_{ji} = 0] = P(D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i}))$ and thus

$$\rho_j = E[Y_i(1) - Y_i(0)|D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i})]$$

F.3 Proof of Theorem SM2

Note that the 2SLS estimand can be written:

$$\rho_{2sls} = \frac{\sum_j \pi_j C(Y, Z_j)}{\sum_j \pi_j C(D, Z_j)} = \frac{\sum_j \pi_j C(D, Z_j)\rho_j}{\sum_j \pi_j C(D, Z_j)}$$

Given Lemma 1, it only remains to be shown that $\pi_j \geq 0$ for all j. As a vector:

$$\pi = \Sigma^{-1}C$$

where Σ is the $J \times J$ covariance matrix of the instruments and C is a vector of covariances between the treatment D and each instrument Z_j. A result known as Farkas' Lemma (see e.g. Gale et al. 1951) states the following: for matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$, exactly one of the following is true:

1. There exists an $x \in \mathbb{R}^n$ such that $Ax = b$ and $x \geq 0$
2. There exists a $y \in \mathbb{R}^m$ such that $A'y \geq 0$ and $b'y < 0$

where for any vector, the notation ≥ 0 indicates that each of its components is weakly positive, etc. Since Σ is an invertible, square, symmetric matrix, Farkas' Lemma is in our case equivalent to:

$$\pi \geq 0 \iff \left(\forall y \in \mathbb{R}^J : \Sigma y \geq 0 \implies C'y \geq 0\right)$$

Now note that element j of the vector Σy is equal to $C(Z_j, Z'y)$ and $C'y$ is equal to $C(Z'y, D)$ where $Z'y = \sum_{k=1}^{J} y_k Z_k$ is a linear combination of the instruments. Thus, what we wish to show is that for any $y \in \mathbb{R}^m$:

$$E[Z'y|Z_j = 1] \geq E[Z'y|Z_j = 0] \quad \forall j \implies E[Z'y|D = 1] \geq E[Z'y|D = 0]$$

In fact, given the strength of Assumption 5, the left-hand inequality holding for any single j will be sufficient. By the law of iterated expectations:

$$\begin{aligned}
&E[Z'y|D = 1] - E[Z'y|D = 0] \\
&= \sum_{z \in \{0,1\}} P(Z_j = z|D = 1)E[Z'y|D(z, Z_{-j}) = 1, Z_j = z] - P(Z_j = z|D = 0)E[Z'y|D(z, Z_{-j}) = 0, Z_j = \\
&= P(Z_j = 1|D = 1)\left\{E[Z'y|D(1, Z_{-j}) = 1, Z_j = 1] - E[Z'y|D(0, Z_{-j}) = 1, Z_j = 0]\right\} \\
&- P(Z_j = 1|D = 0)\left\{E[Z'y|D(1, Z_{-j}) = 0, Z_j = 1] - E[Z'y|D(0, Z_{-j}) = 0, Z_j = 0]\right\} \\
&+ E[Z'y|D(0, Z_{-j}) = 1, Z_j = 0] - E[Z'y|D(0, Z_{-j}) = 0, Z_j = 0]
\end{aligned}$$

By Assumption 2 and 1, for any $z, d \in \{0, 1\}$:

$$\begin{aligned}
E[Z'y|D(z, Z_{-j}) = d, Z_j = z] &= \sum_{z_{-j} \in \{0,1\}^{L-1}} (z, z_{-j})'y \cdot P(Z_{-j} = z_{-j}|D(z, z_{-j}) = d, Z_j = z) \\
&= \sum_{z_{-j} \in \{0,1\}^{L-1}} (z, z_{-j})'y \cdot P(Z_{-j} = z_{-j}|Z_j = z) \\
&= E[Z'y|Z_j = z]
\end{aligned}$$

and thus $E[Z'y|D = 1] - E[Z'y|D = 0]$ can be simplified to

$$\left(E[Z_j|D = 1] - E[Z_j|D = 0]\right)\left(E[Z'y|Z_j = 1] - E[Z'y|Z_j = 0]\right)$$

which is positive whenever $E[Z'y|Z_j = 1] - E[Z'y|Z_j = 0]$ is positive, since $C(D, Z_j) \geq 0$ by Assumption 2*. While we did not make Assumption 2* in the statement of this theorem, it is implied by Assumptions 2 and 5, since:

$$\begin{aligned}
C(D, Z_j) &= P(Z_j)(1 - P(Z_j))\left(E[D|Z_j = 1] - E[D|Z_j = 0]\right) \\
&= P(Z_j)(1 - P(Z_j))\left(E[D(1, Z_{-j})|Z_j = 1] - E[D(0, Z_{-j})|Z_j = 0]\right) \\
&= P(Z_j)(1 - P(Z_j))E[D(1, Z_{-j}) - D(0, Z_{-j})] \qquad =\geq 0
\end{aligned}$$

where the third equality follows by Assumption 1 (independence) and the final inequality follows by Assumption 2 and the law of iterated expectations (over Z_{-j}).