A Vector Monotonicity Assumption for Multiple Instruments

Leonard Goff*

May 22, 2020. Click here for most recent version.

Abstract

When a researcher wishes to use multiple instrumental variables for a single binary treatment, the familiar LATE monotonicity assumption can become restrictive: it requires that all units share a common direction of response even when different instruments are shifted in "opposing" directions. What I call *vector monotonicity*, by contrast, simply restricts treatment status to be monotonic in each instrument separately. I show that in a setting with a binary treatment and multiple binary instruments, a class of causal parameters is point identified under vector monotonicity, including the average treatment effect among units that are responsive to any particular subset of the instruments. I propose a simple "2SLS-like" estimator for the family of identified treatment effect parameters, and show how one can accommodate discrete instruments with finite support more generally. An empirical application revisits the labor market returns to college education.

1 Introduction

The local average treatment effects (LATE) framework introduced by Imbens and Angrist (1994) allows for causal inference with arbitrary heterogeneity in treatment effects, but in doing so imposes an important form of homogeneity on selection behavior. This homogeneity comes through the LATE *monotonicity* assumption, which is often quite natural to make when the researcher has a single instrumental variable at their disposal. However with multiple instruments, this traditional monotonicity assumption can become hard to justify – a point recently emphasized by Mogstad et al. (2019).

A natural question then is whether causal effects are still identified when monotonicity holds on an instrument-by-instrument basis, what I call *vector monotonicity*. Vector monotonicity (VM) captures the notion that each instrument has an impact on treatment uptake in a direction that is common across units (and typically known ex-ante by the researcher). For example, two instruments for college enrollment might be: i) proximity to a college; and ii) affordability of nearby colleges. It is reasonable to assume that each instrument induces some individuals towards going to college, while discouraging none. This contrasts with traditional LATE monotonicity, which as I describe below requires

*I am grateful to Simon Lee, Josh Angrist and Bernard Salanié for patient and insightful feedback on this project, as well as Isaiah Andrews, Jushan Bai, Junlong Feng, Peter Hull, Jack Light, José Luis Montiel Olea, Suresh Naidu, Serena Ng, Vitor Possebom, and Alex Torgovitsky for helpful comments and discussion. I also thank attendees of the Columbia econometrics colloquium, the 2019 Young Economists Symposium, and the 2019 Empirics and Methods in Economics Conference for their feedback. Any errors or other shortcomings are my own.

that either proximity or affordability effectively dominates in selection behavior for all units.

In this paper I provide a simple approach to estimating causal effects under vector monotonicity. I first show that in a setting with a binary treatment and finite discrete instruments that satisfy VM, average treatment effects can be point identified for subgroups of the population that satisfy a certain condition. The condition is satisfied by, for example, the set of all units that are responsive in any way to the collection of instruments, or more generally those units that respond to a fixed subset of the instruments (e.g. just one of them). These subgroups can further be chosen to additionally condition on treatment status. The result is established for binary instruments, and generalized by showing how discrete instruments can be re-expressed as a larger number of binary instruments that also satisfy vector monotonicity. I propose a simple two-step estimator for the corresponding causal parameters. This estimator performs well in simulations against the typical approach of estimating two-stage-least-squares (2SLS). I then apply the estimator to a study investigating the returns to schooling.

To appreciate the sense in which traditional LATE monotonicity is restrictive with multiple instruments, consider the two instruments for college mentioned above, with each coded as a binary variable (far/close and cheap/expensive). LATE monotonicity says that for any pair of points in the joint support of the two instruments, we can choose an assignment of labels z, z' in such a way that all units who would take treatment when the instruments take value z, would also take treatment at z'. This implies that either all units who would go to college when it is far but cheap would also go to college if it was close and expensive, or vice versa. We would generally expect this implication to fail if individuals are heterogeneous in how much each instrument matters to them: for example, if some students are primarily sensitive to distance and others are primarily sensitive to tuition. Vector monotonicity instead says something quite natural in this context: proximity to a college weakly encourages college attendance, regardless of price, and lower tuition weakly encourages college attendance, regardless of distance.

In a recent paper, Mogstad, Torgovitsky and Walters (2019) (henceforth MTW) underline the above difficulty for LATE monotonicity with multiple instruments, and set the stage for potential paths forward. MTW mention VM as as a possible alternative, but focus their identification analysis on a weaker assumption of *partial monotonicity* (PM), which allows the direction of "compliance" for each instrument to depend on the values of the others (see Section 3 for an explicit comparison). MTW characterize what the instruments can say under PM about a wide range of target causal parameters that are typically only bounded by IV methods, such as policy relevant treatment effects (Heckman and Vytlacil 2005). By contrast, I maintain the stronger assumption of VM and characterize a class of causal parameters that are then *point* identified without auxiliary assumptions. I show that VM differs from PM by adding to it a testable condition and that this restriction has additional identification power: the full class of causal parameters

identified under VM cannot generally be point identified under PM alone.

The estimator proposed in this paper can be seen as an alternative to two-stage-least-squares (2SLS), which has been a common method to make use of multiple instruments in applied work. 2SLS is known to provide a convex combination of local average treatment effects under the standard LATE assumptions provided that the first stage recovers the propensity score function, but this implication does not hold generally under VM or PM. MTW derive additional testable conditions which are sufficient for the 2SLS estimand to deliver a convex combination of treatment effects under PM, though the number of conditions to be verified generally grows combinatorially with the number of instruments. In the Supplemental Material, I consider two special cases in which *linear* 2SLS will uncover averages of causal effects under VM with binary instruments. A sufficient condition for one of the special cases – that the instruments are independent – is straightforward to test empirically. The other special case assumes that each unit is responsive to the value of one instrument only, and is quite restrictive. My main identification result eliminates the need to rely on such additional assumptions.

A large literature has considered extensions to the basic LATE model of Imbens and Angrist (1994), but has typically not emphasized the distinction between separate instruments, when more than one is available. Natural analogs of LATE monotonicity have been studied for treatments that are discrete (Angrist and Imbens, 1995), continuous (Angrist et al., 2000), or unordered (Heckman and Pinto, 2018). Other papers have considered identification under various violations of LATE monotonicity. In the case of a binary treatment, Gautier and Hoderlein (2011) and Lewbel and Yang (2016) consider models with particular additivity structures, while Chaisemartin (2017) shows that a weaker notion than monotonicity can be sufficient to give a causal interpretation to LATE estimands. Lee and Salanié (2018) obtain identification in a setting with multi-valued treatment and possible two-way flows between values of continuous instruments, generalizing results from the local instrumental variables approach of Heckman and Vytlacil (2005). LATE monotonicity is generally not assumed by triangular models with continuous endogenous treatment variables (e.g. Imbens and Newey 2009, Torgovitsky 2015, D'Haultfœuille and Février 2015, Gunsilius 2019), which typically rely on monotonicity in unobserved heterogeneity and/or continuity assumptions for identification.

In Section 2 I discuss the basic setup and definitions. I compare vector monotonicity to the traditional monotonicity assumption and MTW's proposal of partial monotonicity, and discuss examples in the context of a simple choice model. In Section 3, I show that like conventional monotonicity, VM partitions the population into well-defined "compliance groups" that can coexist alongside one another. I characterize these groups in a setting with any number of binary instruments. In Section 4 I use this taxonomy to demonstrate identification of a family of causal parameters. Section 5 proposes an estimator, considers its asymptotic properties, and shows simulation evidence on its performance. Section 6 reports results from an application to the labor market returns to schooling. In appen-

dices, I consider a generalization of the identification result that relaxes an assumption of rectangular support among the instruments, consider identification with covariates, and provide simulation results on the performance of the estimator proposed herein. In online Supplemental Material, I also consider some special cases in which linear 2SLS identifies a convex combination of treatment effects under VM, and provide additional examples pertaining to the main text, including a second empirical application to the labor supply effects of family size.

2 Setup

Here I fix notation and formalize the basic setup in which a researcher has multiple instrumental variables for a single binary treatment. Within this framework, I contrast the three alternative notions of monotonicity, emphasizing the point that vector monotonicity is a natural assumption to make in many empirical contexts.

Consider a setting with a binary treatment variable D, scalar outcome variable Y, and vector $Z = (Z_1 \dots Z_J)$ of J instrumental variables that can take values in set $\mathcal{Z} \subseteq (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$, where $\mathcal{Z}_j$ is the set of values that instrument Z_j can take.[1]

Definition (potential outcomes and treatments). *Let $D_i(z)$ denote the treatment status of unit i when their vector of instrumental variables takes value $z \in \mathcal{Z}$, and $Y_i(d, z)$ the realization of the outcome variable that would occur with treatment status $d \in \{0, 1\}$ and instrument value $z \in \mathcal{Z}$.*

Let G_i be a random variable defined via a one-to-one correspondence with the function $\{D_i(z)\}_{z \in \mathcal{Z}}$, having support $\mathcal{G}$. For example, with a single binary instrument $G_i =$ *"always-taker"* indicates that $D_i(0) = D_i(1) = 1$. G_i can be thought of as unit i's "compliance group", characterized by a complete mapping $D_i : \mathcal{Z} \to \{0, 1\}$ between points in $\mathcal{Z}$ and values of treatment. Unit i's compliance group G_i characterizes their counterfactual selection behavior across all realizations of the instrument values. Given these definitions, we can think of a set of *valid* instrumental variables as satisfying the following assumption:

Assumption 1 (exclusion and independence). *a) $Y_i(d, z) = Y_i(d)$ for all $z' \in \mathcal{Z}, d \in \{0, 1\}$; and b)*

$$(Y_i(1), Y_i(0), G_i) \perp (Z_{1i}, \dots, Z_{Ji})$$

The first part of Assumption 1 states that the instruments are "excludable" from the outcome function in the sense that potential outcomes do not depend on them once treatment status is fixed. The second part of Assumption 1 states that the instruments are independent of potential outcomes and potential treatments. In practice, it is common

[1] $\mathcal{Z}$ may be a strict subset when certain combinations of instrument values are ruled out on conceptual grounds, e.g. Z_1 indicates a mothers' first two births being girls and Z_2 indicates them both being boys. $\mathcal{Z}$ is allowed to differ from $supp(Z_i)$, since the former may contain values that are conceptually possible but take zero probability in a particular population.

to maintain a version of this independence assumption that holds only conditional on a set of observed covariates. For ease of exposition, I implicitly condition on any such covariates throughout, then consider incorporating them explicitly in Appendix B.

2.1 Notions of monotonicity

It is well-known that when treatment effects are heterogeneous, Assumption 1 alone is not sufficient for instrument variation to identify treatment effects. The seminal LATE model of Imbens and Angrist (1994) introduces the additional assumption of monotonicity:

Assumption IAM (traditional monotonicity). *For all* $z, z' \in \mathcal{Z}$*:* $P(D_i(z) \geq D_i(z')) = 1$ *or* $P(D_i(z) \leq D_i(z')) = 1$.

I follow the terminology of MTW and henceforth refer to this as Assumption IAM, or "Imbens and Angrist monotonicity". As pointed out by Heckman et al. (2006), IAM can be thought of as a "uniformity" assumption: it states that flows of selection into treatment between z in z' move only in one direction, whichever direction that is.

The proposed assumption of *vector monotonicity* captures "monotonicity" as the notion that increasing the value of any instrument weakly encourages (or discourages) all units to take treatment, regardless of the values of the other instruments:

Assumption 2 (vector monotonicity). *There exists an ordering* $\geq_j$ *on* $\mathcal{Z}_j$ *for each* $j \in \{1 \dots J\}$ *such that for all* $z, z' \in \mathcal{Z}$*, if* $z \geq z'$ *component-wise according to the* $\{\geq_j\}$*, then* $D_i(z) \geq D_i(z')$ *with probability one.*

Vector monotonicity is referred to as "actual monotonicity" by Mogstad et al. (2019), when each $\geq_j$ is the standard ordering on real numbers. Mountjoy (2018) imposes a version of VM in a case with a multivalued treatment and continuous instruments.

The partial monotonicity assumption introduced by MTW is weaker than both IAM and VM. Let (z_j, z_{-j}) denote a vector composed of $z_j \in \mathcal{Z}_j$ and $z_{-j} \in \mathcal{Z}_{-j}$, where $\mathcal{Z}_{-j}$ indicates the set of values that the vector of all instruments but Z_j can take.

Assumption PM (partial monotonicity). *For each* $j \in \{1 \dots J\}$*,* $z_j, z'_j \in \mathcal{Z}_j$*, and* $z_{-j} \in \mathcal{Z}_{-j}$ *such that* $(z_j, z_{-j}) \in \mathcal{Z}$ *and* $(z'_j, z_{-j}) \in \mathcal{Z}$*, either* $D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})$ *with probability one or* $D_i(z_j, z_{-j}) \leq D_i(z'_j, z_{-j})$ *with probability one.*

Note that under partial monotonicity, there will be a weak ordering on the points in $\mathcal{Z}_j$, for any fixed choice of j and z_{-j}. The crucial restriction made by vector monotonicity beyond partial monotonicity is that under VM, this ordering must be *the same* across all values of $z_{-j} \in \mathcal{Z}_{-j}$ for a given j. Partial monotonicity could for example capture a situation in which college proximity encourages attendance when nearby colleges are cheap but discourages attendance when they are expensive – while VM could not.

An alternative characterization of VM makes this relationship to PM more explicit. Call $\mathcal{Z}$ *connected* when for any two $z, z' \in \mathcal{Z}$ there exists a sequence of vectors $z_1, \dots, z_m$

with $z_1 = z$, $z_m = z'$ and each z_m and z_{m-1} differing on only one component, and such that $z_m \in \mathcal{Z}$ for all m.[2]

Proposition 1. *Let $\mathcal{Z}$ be connected. Then Assumption 2 holds iff for each $j \in \{1 \dots J\}$ there is an ordering $\geq_j$ on $\mathcal{Z}_j$ such that $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$ when $z_j \geq_j z'_j$, for all $z_{-j} \in \mathcal{Z}_{-j}$ such that both $(z_j, z_{-j}) \in \mathcal{Z}$ and $(z'_j, z_{-j}) \in \mathcal{Z}$.*

Proof. See Appendix D. □

The additional restriction made by VM over PM is empirically testable, by inspecting the propensity score function:

Proposition 2. *Suppose PM and Assumption 1 hold, and $\mathcal{Z}$ is connected. Then VM holds if and only if $E[D_i|Z_i = z]$ is component-wise monotonic in z, for some fixed ordering $\succeq_j$ on each $\mathcal{Z}_j$.*

Proof. See Appendix D. □

Note that if Assumption 1 holds conditional on covariates X_i, Proposition 2 also need only hold with respect to the *conditional* propensity score $E[D_i|Z_i = z, X_i = x]$ (see Section 6).

Since IAM implies PM, it follows as a corollary to Proposition 2 that if IAM and Assumption 1 hold and $E[D_i|Z_i = z]$ is component-wise monotonic in z, then VM holds. This establishes that after a researcher has verified that the propensity score function is monotonic, VM becomes a strictly weaker assumption than IAM. The relationship among Assumptions IAM, VM and PM is depicted graphically in Figure 1.

Examples of the points (a)-(e) in Figure 1 can be made more concrete by considering a case with two binary instruments $\mathcal{Z} = \{0, 1\} \times \{0, 1\}$, with an explicit selection model of the form:

$$D_i(z_1, z_2) = \mathbb{1}(\beta_{0i} + \beta_{1i} z_1 + \beta_{2i} z_2 + \beta_{3i} z_1 z_2 \geq 0) \tag{1}$$

where $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})' \perp Z_i$ (Assumption 1). Given the binary treatments, this model is general enough to capture all possible selection functions $D_i(z)$.

With no further restrictions on the joint distribution of the coefficients β_i, selection according to Eq. (1) could easily violate even partial monotonicity (e), for instance if β follows a uniform distribution over $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$. A simple model such as a) that satisfies all of VM, IAM, and PM can be constructed by setting $\beta_{3i} = 0$ with probability one and making the restriction that $\beta_{1i} = \beta_1$ and $\beta_{2i} = \beta_2$ are homogenous across i. This could capture a utility maximization model in which individuals trade off an incentive $\beta_1 z_1 + \beta_2 z_2$ produced by the instruments against a net cost $-\beta_{0i}$ of treatment that is heterogenous across individuals. Without loss, assume that $0 \leq \beta_2 \leq \beta_1$. Then:

$$D_i(0,0) \leq D_i(0,1) \leq D_i(1,0) \leq D_i(1,1) \quad \text{with prob. one} \tag{a}$$

[2] This rules out cases where $\mathcal{Z}$ is disjoint with respect to such chains of single-instrument switches, for example in a case of two binary instruments if $\mathcal{Z}$ consists only of the points $(0,0)$ and $(1,1)$. With this $\mathcal{Z}$, PM and VM are both vacuous.

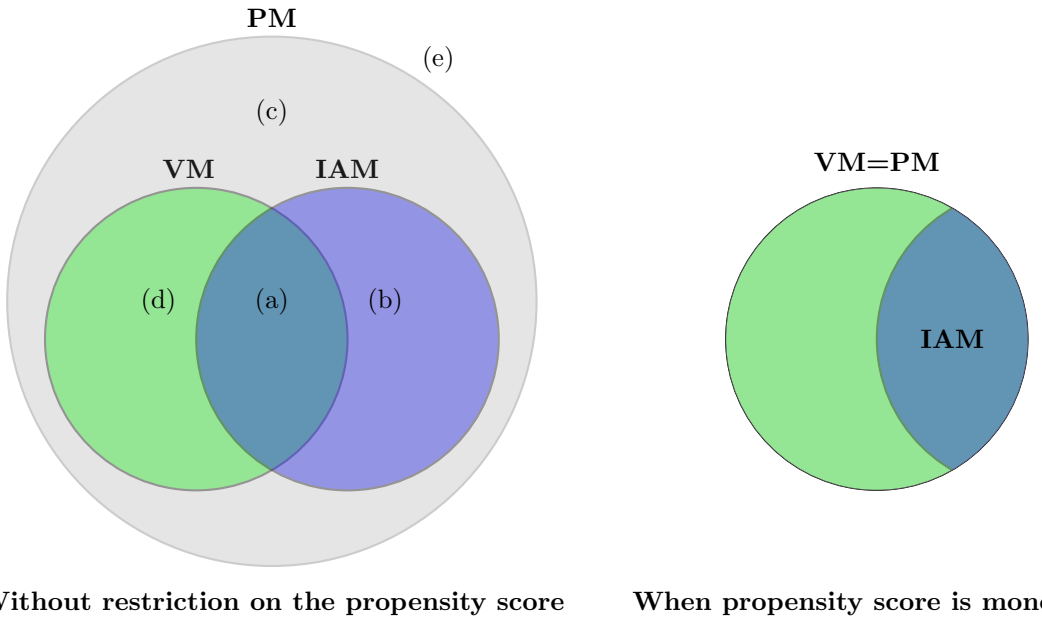

Figure 1: Left panel shows ex-ante comparison of Imbens & Angrist monotonicity (IAM), vector monotonicity (VM), and partial monotonicity (PM) before the propensity score function is known. Right panel depicts the relationship when the propensity score is component-wise monotonic: PM and VM become identical, with IAM a special case. Examples for points (a)-(e) are discussed in the text.

Model d), breaking IAM but not VM, can now be constructed by allowing heterogeneity in β_1 and β_2, but restricting each to remain positive with probability one. In this case, we have both that

$$\begin{aligned} D_i(0,0) \leq D_i(0,1) \leq D_i(1,1) \quad & \text{with prob. one} \\ D_i(0,0) \leq D_i(1,0) \leq D_i(1,1) \quad & \text{with prob. one} \end{aligned} \tag{d}$$

but there is no restriction on the relationship between $D_i(1,0)$ and $D_i(0,1)$ – there can be some units for which the former is one and the latter zero, and others for whom the former is zero and the latter one.

Example b) in which VM but not IAM is violated can be constructed from our example of a) by allowing a homogenous interaction $-\beta_1 \leq \beta_3 \leq -\beta_2 \leq 0$ between the instruments. This interaction is strong enough that now

$$D_i(0,0) \leq D_i(0,1) \leq D_i(1,1) \leq D_i(1,0) \quad \text{with prob. one} \tag{b}$$

An example of c), which violates both VM and IAM but not PM, is given by MTW: set $\beta_{2i} = 1$, $\beta_{3i} \leq -1$, and $0 \leq \beta_{1i} \leq -\beta_{3i}$ each with probability one.

Now consider the above cases in the returns to schooling example, with "cheap" and "close" the 1 states of Z_1 and Z_2, respectively. In a utility maximization model β_{0i} might denote the net benefit of attending college when it is far and expensive. If college then became either cheap or close, this might be expected to only increase the net benefit of college, incenting some individuals into enrolling while discouraging none. This motivates making the restrictions $\beta_{1i} \geq 0$ and $\beta_{2i} \geq 0$. If we then imagine changing to $(cheap, close)$

from either $(expensive, close)$ or $(cheap, far)$, it's reasonable to again assume that all students would move weakly towards college, unless there are individuals for whom the interaction coefficient β_{3i} is sufficiently strong and negative. It is possible to imagine scenarios in which this happens: for example, suppose there exist students who do not want to live with their parents during college, and feel that they will have to if attending a college near their parents' home. Accordingly, these students only opt out of college if there is a cheap option close to where they grew up. Note that in this case, PM would then require that there be no individuals that go to college only if it is both cheap and close. The Supplemental Material provides a taxonomy of such cases that break VM but not PM, as point c) does, with two binary instruments.

Finally, I note that a sufficient condition for the restriction from PM to VM is the existence of individuals that are sensitive to that instrument alone. For example, suppose Alice only cares about proximity, and Bob only cares about tuition, with:

$$D_{alice}(z_1, z_2) = \mathbb{1}(z_2 = close) \quad \text{and} \quad D_{bob}(z_1, z_2) = \mathbb{1}(z_1 = cheap)$$

Partial monotonicity then requires that the directions of "compliance" that Alice and Bob exhibit (lower distance and lower tuition, respectively) hold (weakly) for all other units in the population, which then implies VM.[3]

3 Characterizing compliance behavior under vector monotonicity

In this section I show that the assumption of vector monotonicity partitions the population of interest into a set of well-defined "compliance groups", denoted as realizations of the random variable G_i. These groups generalize the familiar taxonomy of always-takers, never-takers, and compliers from the case of a single binary instrument. Providing a characterization of the groups will be necessary to state the main identification result in Section 4.

3.1 With two binary instruments

To establish intuition for the running example of the returns to schooling, we first turn to the simplest case of two binary instruments.

Normalize the "1" state for each instrument to be the direction in which potential treatments are increasing. MTW have shown that VM then separates the population into the six compliance groups described in Table 1, and have introduced names for each group. A Z_1 complier, for example, is treated if and only if college is cheap, regardless of whether it is close. A Z_2 complier, in our example, would be treated if and only if college is close, regardless of whether it is cheap. A reluctant complier is "reluctant" in the sense

[3] That is, $D_{alice}(1, z_2) > D_{alice}(0, z_2)$ for all $z_2 \in \mathcal{Z}_2$ implies through PM that $P(D_i(1, z_2) \geq D_i(0, z_2)) = 1$ for all $z_2 \in \mathcal{Z}_2$, and similarly Bob implies that $P(D_i(z_1, 1) \geq D_i(z_1, 0)) = 1$ for all $z_1 \in \mathcal{Z}_1$.

that they require college to be both cheap and close to attend, while an eager complier receives treatment so long as college is cheap or close. Never and always takers are defined in the same way as they are under IAM: $\max_{z\in\mathcal{Z}} D_i(z) = 0$ and $\min_{z\in\mathcal{Z}} D_i(z) = 1$, respectively.

Name	$\mathbf{D_i(0,0)}$	$\mathbf{D_i(0,1)}$	$\mathbf{D_i(1,0)}$	$\mathbf{D_i(1,1)}$
never takers	N	N	N	N
always takers	T	T	T	T
Z_1 compliers	N	N	T	T
Z_2 compliers	N	T	N	T
eager compliers	N	T	T	T
reluctant compliers	N	N	N	T

Table 1: The six compliance groups under VM with two binary instruments.

The sizes of the six groups in Table 1 are not generally point identified, however two of them are: from the equality $P(z) := E[D_i|Z_i = z] = \sum_g D_g(z)$, we can deduce that $p_{n.t} = 1 - P(1,1)$ and $p_{a.t.} = P(0,0)$. For the others, we can identify certain linear combinations of the group occupancies, e.g. $P(1,0) - P(0,0) = p_{Z_1} + p_{eager}$, $P(0,1) - P(0,0) = p_{Z_2} + p_{eager}$, and $P(1,1) - P(0,1) = p_{Z_1} + p_{reluctant}$ (the other pairwise propensity score differences can be written as linear combinations of these three). This allows us to bound each of the four remaining group sizes, given that each must be positive. For example, $\{P(1,0) - P(0,0)\} - \{P(1,1) - P(0,1)\} \leq p_{eager} \leq \min\{P(0,1) - P(0,0), P(1,0) - P(0,0)\}$.

3.2 With multiple binary instruments

Now we see how the two-instrument case generalizes to a case where the researcher has any number of binary instruments. While the overall number of compliance groups explodes combinatorially, we can still keep track of the various groups in a systematic way.

Let there be J binary instruments $Z_1 \dots Z_J$. I focus on the baseline case in which $\mathcal{Z} = \{0,1\}^J$ (see Supplemental Material Section B.7 for some alternatives). We wish to characterize the subset of the 2^{2^J} possible mappings between vectors of instrument values and treatment that satisfy VM, where we continue to normalize the "1" state for each Z_j to be the direction in which potential treatments are weakly increasing.[4] The number of such compliance groups G_i as a function of J is equal to the number of isotone boolean functions on J variables, known to follow the so-called Dedekind sequence:

$$3, 6, 20, 168, 7581, 7828354 \dots$$

While an analytical expression for the Dedekind numbers was derived by Kisielewicz

[4]This "up" value for each instrument will be taken in our results to be known ex ante. In practice, this might follow from a maintained natural hypothesis, such as that lower price encourages rather than discourages college attendance. However, the directions could also be determined empirically under Assumption 2 from the propensity score function (see Proposition 2).

(1988), only the first eight have been calculated numerically due to the computational burden of evaluating it.[5]

I now explicitly characterize the compliance groups that satisfy VM, for any value of J. As before, there will always be a never-takers group for which $D_i(z) = 0$ for all values $z \in \mathcal{Z}$. Apart from these never-takers, each compliance group satisfying vector monotonicity can be associated with the various sets of instruments that are sufficient for the unit to take treatment, when all of the instruments in a set take a value of one. For example, in a setting with three instruments, one compliance group would be units that take treatment if either $Z_1 = 1$, or if $Z_2 = Z_3 = 1$. By vector monotonicity, then, any unit in this group must also take treatment if $Z_1 = Z_2 = Z_3 = 1$. However, another group of units might take treatment only if $Z_1 = Z_2 = Z_3 = 1$. This group is more "reluctant" than the former. The group of always-takers are the least "reluctant": they require no instruments to equal one in order for them to take treatment.

By the above logic, we see that compliance groups g under VM (aside from never takers) map one-to-one with families (i.e. sets) F of subsets $S \in \{0,1\}^J$ of the instruments, with the property that no element S of the family is a subset of any of the others. Such a family F of subsets is referred to as a *Sperner family* (see e.g. Kleitman and Milner 1973).

Definition (compliance group for a Sperner family). *For any Sperner family F on the set of instrument labels $\{1 \dots J\}$, let $g(F)$ denote the compliance group such that when $G_i = g(F)$, $D_i(z) = \mathbb{1}\left((\prod_{j \in S} z_j) = 1 \text{ for at least one } S \text{ in } F\right)$.*

Similarly, denote the Sperner family associated with a compliance group g as $F(g)$. Note that the reason we need only consider Sperner families is that other families of sets of the instruments would be redundant under VM – if any set S in F is a subset of another S' in F, S' can be dropped without affecting the implied selection function $D_i(z)$. Each set S in a Sperner family F represents a minimal set of instruments for which the following is true: if all instruments in S take a value of one, a unit in the compliance group $g(F)$ corresponding to F will take treatment.

In the simplest example of the above, when $J = 1$, vector monotonicity coincides with PM and IAM, and the Sperner families corresponding to this single instrument are simply the null set and the singleton $\{1\}$: corresponding to always-takers and compliers, respectively. Together with never-takers (which do not have a Sperner family representation), we have the familiar three groups from LATE analysis with a single binary instrument.

For $J = 2$, the five groups (aside from never takers) described in the previous section

[5] While the Dedekind numbers clearly explode quite rapidly, they do so much more slowly than the total number 2^{2^J} of boolean functions of J variables. For example while $3/4 = 75\%$ of conceivable compliance groups for $J = 1$ satisfy VM, only $20/256 \approx 7.8\%$ do for $J = 3$, and just $7581/4294967296 \approx 1.7 * 10^{-4}$ do for $J = 5$. Thus the "bite" of VM is increasing with J, in the sense that it rules out a larger and larger fraction of conceivable selection patterns.

map to Sperner families as follows:

$\mathbf{F}$	**name of** $\mathbf{G_F}$
$\emptyset$	"always takers"
$\{1\}$	"Z_1 compliers"
$\{2\}$	"Z_2 compliers"
$\{1\},\{2\}$	"eager compliers"
$\{1,2\}$	"reluctant compliers"

The rapidly expanding richness of selection behavior compatible with VM can be seen with $J = 3$, where there are 19 Sperner families, indicated within bold brackets:

$$\mathbf{\{}\emptyset\mathbf{\}},\mathbf{\{}1\mathbf{\}},\mathbf{\{}2\mathbf{\}},\mathbf{\{}3\mathbf{\}},$$

$$\mathbf{\{}1,2\mathbf{\}},\mathbf{\{}1,3\mathbf{\}},\mathbf{\{}2,3\mathbf{\}},\mathbf{\{}1,2,3\mathbf{\}},$$

$$\mathbf{\{}\{1\},\{2\}\mathbf{\}},\mathbf{\{}\{2\},\{3\}\mathbf{\}},\mathbf{\{}\{1\},\{3\}\mathbf{\}},\mathbf{\{}\{1\},\{2\},\{3\}\mathbf{\}},$$

$$\mathbf{\{}\{1,2\},\{3\}\mathbf{\}},\mathbf{\{}\{1,3\},\{2\}\mathbf{\}},\mathbf{\{}\{2,3\},\{1\}\mathbf{\}},$$

$$\mathbf{\{}\{1,2\},\{1,3\}\mathbf{\}},\mathbf{\{}\{1,2\},\{2,3\}\mathbf{\}},\mathbf{\{}\{1,3\},\{2,3\}\mathbf{\}},$$

$$\mathbf{\{}\{1,2\},\{1,3\},\{2,3\}\mathbf{\}}$$

For instance, an individual with G_i corresponding to $\mathbf{\{}\{1,2\},\{1,3\},\{2,3\}\mathbf{\}}$ takes treatment so long as any two instruments take the one value. Note that the number of Sperner families with $J = 3$ is one less than the third Dedekind number, which is 20. This is true for any J: absent any additional restrictions $\mathcal{D}_J - 1$ compliance groups are associated with Sperner families (the last group being never-takers), where $\mathcal{D}_J$ is the J^{th} number in the Dedekind sequence.

A central feature of the identification analysis will be that the selection functions corresponding to the various compliance groups are not all linearly independent from one another. Only 2^J such functions can be independent (though $\mathcal{D}_J$ is strictly larger for $J > 1$), since any function of binary variables can be written as a polynomial in them. Let $\mathcal{G}^c := \mathcal{G}/\{a.t., n.t.\}$ denote the set of $\mathcal{D}_J - 2$ groups compatible with Assumption VM that are not never-takers or always takers. The groups in $\mathcal{G}^c$ can be thought of as generalized "compliers": units that would vary treatment uptake in *some* way across instrument values.

Let $D_g(z)$ denote the common selection function for all units having $G_i = g$. A natural basis for the $\{D_g(z)\}_{g\in\mathcal{G}^c}$ under VM is formed by considering functions that are products over subsets of the instruments $z_S := \prod_{j\in S} z_j$, where $S \subseteq \{1 \dots J\}, S \neq \emptyset$.[6] For a given set S, z_S yields the selection function $D_{g(S)}(z)$ of the compliance group $g(S)$ corresponding

[6] Note that a similar construction plays a central role in Lee and Salanié, 2018.

to the Sperner family consisting of the single set S. I refer to such compliance groups $g(S)$ as *simple*. For $J = 2$, the selection functions for the simple compliance groups are:

$$D_{Z_1}(z) = z_1 \qquad D_{Z_2}(z) = z_2 \qquad D_{reluctant}(z) = z_1 z_2$$

The selection function for the remaining group, eager compliers, can be obtained as:

$$D_{eager}(z) = z_1 + z_2 - z_1 z_2 = D_{Z_1}(z) + D_{Z_2}(z) - D_{reluctant}(z)$$

We can express this linear dependency by the matrix M_J in the system:

$$\begin{pmatrix} D_{Z_1}(z) \\ D_{Z_2}(z) \\ D_{reluctant}(z) \\ D_{eager}(z) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & -1 \end{pmatrix}}_{:=M_2} \begin{pmatrix} D_{Z_1}(z) \\ D_{Z_2}(z) \\ D_{reluctant}(z) \end{pmatrix}$$

For general J, we define the matrix M_J from the analogous system of equations:

$$\{D_{g(F)}(z)\}_{F:\, g(F)\in\mathcal{G}^c} = M_J \{D_{g(S)}(z)\}_{S\subseteq\{1\dots J\}, S\neq\emptyset}$$

for all $z \in \mathcal{Z}$. The rows of matrix M_J are indexed by Sperner families (corresponding to the groups in $\mathcal{G}^c$), and the columns by the simple Sperner families for non-null S. The entries of M_J are given by the following expression:[7]

Proposition 3. $[M_J]_{F,S'} = \sum_{f\in s(F,S')} (-1)^{|f|+1}$ *where* $s(F, S') := \left\{ f \subseteq F : \left(\bigcup_{S\in f} S\right) = S' \right\}$.

Proof. See Appendix D. □

3.3 Vector monotonicity with discrete instruments

More generally, when the researcher has discrete instrumental variables that satisfy vector monotonicity, they can be re-expressed as a larger number of binary instruments in a way that preserves vector monotonicity. By introducing a single binary instrument for every value of each discrete instrument, the analysis can be extended to this much more general setting:

Proposition 4. *Let Z_1 be a discrete variable with M ordered points of support $z_1 < z_2 < \cdots < z_M$, and $Z_2 \dots Z_J$ be other instrumental variables. Let $\tilde{Z}_{mi} := \mathbb{1}(Z_{1i} \geq z_m)$. If the vector $Z = (Z_1, \dots Z_J)$ satisfies Assumption VM on a connected $\mathcal{Z}$ then so does the vector $(\tilde{Z}_2, \dots, \tilde{Z}_M, Z_2, \dots Z_J)$.*

Proof. See Appendix D. □

Applying Proposition 4 iteratively offers a fairly general recipe for mapping the instruments available in a given empirical setting into the framework of binary instruments.

[7] The matrix M_3, which has $\mathcal{D}_3 - 2 = 18$ rows and $2^3 - 1 = 7$ columns is given explicitly in the Supplemental Material.

Provided that the researcher is willing to approximate continuous instruments by finite discrete instruments, this can be applied to continuous instruments as well.

Note that the mapping in Proposition 3.3 introduces restrictions on $\mathcal{Z}$ for the resulting binary instruments, since for example we could not have both $\tilde{Z}_{2i} = 1$ and $\tilde{Z}_{1i} = 0$. As a result, not all of the compliance groups introduced in Section 3.2 are necessary to account for, since possible patterns of instrument variation pool some into equivalent groups. While in the next section I assume full binary instrument support for the baseline results, Appendix A provides the necessary generalizations to make use of Proposition 4.

4 Parameters of interest and identification

In this section I define and characterize a class of causal parameters, then show that they can be point identified under vector monotonicity. This section maintains a setup of J binary instruments with $\mathcal{Z} = \{0,1\}^J$ unless otherwise specified.

4.1 A class of conditional average treatment effects

I consider as parameters of interest conditional average treatment effects:

$$\Delta_c := E[Y_i(1) - Y_i(0)|C_i = 1]$$

where $C_i = c(G_i, Z_i)$ is a function $c : \mathcal{G} \times \mathcal{Z} \to \{0,1\}$ of individual i's compliance group and their realization of the instruments. Intuitively, the event $C_i = 1$ will indicate that unit i belongs to a certain subgroup of "compliers". Treatment effect parameters of this form are familiar in the IV literature: for instance local average treatment effects (Imbens and Angrist, 1994), and marginal treatment effects (Heckman and Vytlacil, 2005).

The main result in Section 4.3 is that identification is possible under VM for certain choices of the function $c(g,z)$. This applies the spirit of LATE analysis under IAM to VM, where the set of identified Δ_c will take on a richer structure. To motivate the necessary restriction on c, we can begin by considering what we could identify from a general type of "2SLS-like" estimand, in which a single scalar instrument $h(Z_i)$ is constructed from the vector of instruments Z_i based on some function h, and then used in a simple linear IV regression. Let $\Delta_g := E[Y_i(1) - Y_i(0)|G_i = g]$. Then:

Lemma 1. *Under Assumption 1 (exclusion and independence):*

$$\frac{Cov(Y_i, h(Z_i))}{Cov(D_i, h(Z_i))} = \sum_{g \in \mathcal{G}} \frac{P(G_i = g) \cdot Cov(D_g(Z_i), h(Z_i))}{\sum_{g' \in \mathcal{G}} P(G_i = g') \cdot Cov(D_{g'}(Z_i), h(Z_i))} \cdot \Delta_g$$

Proof. See Theorem 1, or D.3 in Supplemental Material for direct proof of this form. □

For a given choice of the function h, the weight placed on each compliance group g is governed by the covariance between $D_g(Z_i)$ and $h(Z_i)$.[8] Importantly, since the covariance

[8] Special cases include two stage least squares: $h(z) = E[D_i|Z_i = z]$, and Wald estimands: $h(z) = \frac{\mathbb{1}(Z_i = z)}{P(Z_i = z)} - \frac{\mathbb{1}(Z_i = z')}{P(Z_i = z')}$.

operator is linear, any perfect linear dependencies between the functions $D_g(z)$ across various g translate into necessary linear relationships between the weights on the corresponding Δ_g that can be so attained. Furthermore, this estimand puts zero weight on always-takers and never-takers, since there can be no variation in $D_g(Z_i)$ for these groups.

Note that like the estimand in Lemma 1, Δ_c may also be written as a convex combination of group-specific average treatment effects, where the weight that Δ_c places on compliance group g is proportional to the quantity $E[c(g, Z_i)] = P(C_i = 1|G_i = g)$. In particular, by Assumption 1 and the law of iterated expectations:

$$\Delta_c = \sum_{g \in \mathcal{G}} \left\{ \frac{P(G_i = g)E[c(g, Z_i)]}{E[c(G_i, Z_i)]} \right\} \cdot \Delta_g \tag{2}$$

where the weights on the Δ_g (in curly braces) sum to one by the law of total probability. Comparing Eq. (2) with the RHS of Lemma 1, we can see that Δ_c will be estimable if the function h can be chosen such that $Cov(D_g(Z_i), h(Z_i)) = E[c(g, Z_i)]$ for all $g \in \mathcal{G}^c$ (recall that $\mathcal{G}^c$ is the set of VM compliance groups besides always- and never-takers). Given the linear dependencies examined in Section 3.2, this imposes a set of linear restrictions among the weights in Δ_c. In particular, it requires that the vector of $E[c(g(F), Z_i)]$ for all Sperner families F belongs to the column-space of the matrix M_J. This is guaranteed under what I call "Property M":

Definition (Property M). *The function $c(g, z)$ satisfies Property M if for all $z \in \mathcal{Z}$: $c(a.t., z) = c(n.t., z) = 0$ and the remaining $g \in \mathcal{G}^c$ satisfy*

$$c(g, z) = \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot c(g(S), z)$$

I'll also say that Δ_c "satisfies Property M" if its underlying function $c(g, z)$ does.

The first requirement in Property M of zero weight on always-takers or never-takers is familiar from analysis based on IAM.[9] The second part of Property M however is new, and is a result of there being under VM more compliance groups in $\mathcal{G}^c$ than there are independent pairs of points in the support of the instruments. Under IAM with J binary instruments both are generally equal to $2^J - 1$, and it is possible to identify the average treatment effect within any single such compliance group (and hence also obtain any desired convex combination of the Δ_g). However, under VM the corresponding choice $c(g, z) = \mathbb{1}(g = g')$ fails to satisfy Property M, and we will not be able to identify the Δ_g individually in general.

While Property M is somewhat abstract, the following result shows that it is equivalent to $c(g, z)$ being equal to a linear combination of selection functions:

[9] Note that $E[c(g, Z_i)] = 0$ would also be necessary for any additional groups g for whom, given the distribution of Z_i, there is no actual variation in treatment status. In the baseline analysis, such additional groups will be ruled out by Assumption 3.

Proposition 5. *A function $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ satisfies Property M if and only if*

$$c(g, z) = \sum_{k=1}^{K} D_g(u_k(z)) - D_g(l_k(z))$$

for some $K \leq J/2$, where $u_k(\cdot)$ and $l_k(\cdot)$ are functions $\mathcal{Z} \to \mathcal{Z}$ such that $P(u_k(Z_i) \geq l_k(Z_i)) = 1$ for all $k = 1 \dots K$ and $P(l_k(Z_i) \geq u_{k+1}(Z_i)) = 1$ for all $k = 1 \dots K - 1$.

Proof. See Appendix D. □

Note: In the proof of Proposition 5, I also show that under IAM instead of VM, any function $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ for which $E[c(n.t., Z_i)] = E[c(a.t., Z_i)] = 0$ for all distributions of Z_i also admits this form, with $K \leq 2^J/2$. Such parameters are generally identified under IAM given the function c by very similar logic to that I present for VM. However, since these Δ_g can be written as convex combinations of LATEs (apply the law of iterated expectations to Equation 3), demonstrating identification is not novel in the IAM case.

Proposition 5 yields a natural interpretation of parameters Δ_c that satisfy Property M, which is that they can be written as

$$\Delta_c = E\left[Y_i(1) - Y_i(0) \middle| \bigcup_{k=1}^{K} \{D_i(u_k(Z_i)) > D_i(l_k(Z_i))\} \right] \tag{3}$$

for functions u_k and l_k having the properties stated in Proposition 5. This expression is obtained by substituting $C_i = c(G_i, Z_i)$, and noting that $\sum_{k=1}^{K} D_i(u_k(Z_i)) - D_i(l_k(Z_i))$ equals one if and only if $D_i(u_k(Z_i)) > D_i(l_k(Z_i))$ for some k. From Equation 3 we see that the types of complier groups that identified parameters can condition on are groups of individuals that respond to one of some number of unambiguous instrument shifts: that is, transitions $l_k(z) \to u_k(z)$ that do not induce two-way flows. This feature is common to both IAM and VM.

4.2 Examples from the family of identified parameters

While the form Equation (2) is very much analogous to LATE parameters under IAM, the natural structure of VM allows us to see that it nests several causal parameters with economically interesting interpretations. Table 2 presents some leading examples of Δ_c that satisfy Property M, as can be seen by applying Proposition 5. All of the cases presented in Table 2 admit the form of Equation (3) with a single term ($K = 1$), given in the third column.

The first item in Table 2 I call the "all compliers LATE" (ACL), which is the average treatment effect among all units who are not always-takers or never-takers. This is the largest subgroup of the population for which treatment effects can be generally identified from instrument variation.[10] With two instruments, the ACL is the average treatment effect among units who are Z_1, Z_2, eager or reluctant compliers. In the returns to schooling

[10] We may of course still be able to say something about treatment effects for never-takers and always-takers given additional restrictions (see e.g. Section 4.4 for bounds on the unconditional ATE when potential outcomes are bounded).

Parameter	$\mathbf{c(g, z)}$	Proposition 5 form
ACL	$\mathbb{1}(g \in \mathcal{G}^c)$	$D_g(1,1,\ldots,1) - D_g(0,0,\ldots,0)$
$SLATE_{\mathcal{J}}$	$D_g((1\ldots1), z_{-\mathcal{J}}) - D_g((0\ldots0), z_{-\mathcal{J}}))$	"
$SLATT_{\mathcal{J}}$	$D_g(z) \cdot (D_g((1\ldots1), z_{-\mathcal{J}}) - D_g((0\ldots0), z_{-\mathcal{J}})))$	$D_g(z) - D_g((0\ldots0), z_{-\mathcal{J}})$
$SLATU_{\mathcal{J}}$	$(1 - D_g(z)) \cdot (D_g((1\ldots1), z_{-\mathcal{J}}) - D_g((0\ldots0), z_{-\mathcal{J}})))$	$D_g((1\ldots1), z_{-\mathcal{J}}) - D_g(z)$
$MTE_j(z^*_{-j})$	$D_g(1, z^*_{-j}) - D_g(0, z^*_{-j}))$	"

Table 2: Leading parameters of interest satisfying Property M, including: the All Compliers LATE, set LATEs, set LATEs on the treated, and set LATEs on the untreated.

example, we can equivalently describe it as the average treatment effect among individuals who would go to college were it close and cheap, but would not were it far and expensive.

On the other end of the spectrum, the final row of Table 2 gives the most disaggregated type of parameter satisfying Property M, $MTE_j(z^*_{-j})$: the average treatment effect among individuals that move into treatment when a single instrument j is shifted from zero to one, while the other instrument values are held fixed at some vector of values z^*_{-j}. An example is the average treatment effect among individuals who go to college if it is close and cheap, but not if it is far and cheap. Ultimately, all Δ_c satisfying Property M can be written as convex combinations of such *marginal treatment effects* though the number could be quite large (see Supplemental Material Section B.6 for an explicit expression). The remaining parameters in Table 2 constitute a middle ground between the granular $MTEs$ and the very broad averaging of the ACL.

For example, the ACL is a special case of what I call a *set local average treatment effect*, or $SLATE_{\mathcal{J}}$, which captures the average treatment effect among units that move into treatment when all instruments in some fixed set $\mathcal{J}$ are changed from 0 to 1, with the other instruments not in $\mathcal{J}$ fixed at their realized values. The ACL takes this set to be all of the instruments: $\mathcal{J} = \{1, 2, \ldots J\}$. When $\mathcal{J}$ contains just one instrument index, SLATE recovers treatment effects among those who "comply" with that single instrument. For example, $SLATE_{\{2\}}$ is the average treatment effect among individuals who don't go to college if it is far, but do if it is close. This parameter may be of interest to policymakers considering whether to expand a community college to a new campus. The group of individuals included in $SLATE_{\{2\}}$ are Z_2 compliers, eager compliers with high tuition rates ($Z_{1i} = 0$), and reluctant compliers with low tuition rates ($Z_{1i} = 1$).[11]

For a discrete instrumental variable mapped to multiple binary instruments by Proposition 4, the LATE among units moved into treatment between any two of its values will also be an example of a SLATE. For example, if Z_1 has support $z_1 < z_2 < z_3 < z_4$, the average treatment effect among individuals for which $D_i(z_4, Z_{-1,i}) > D_i(z_2, Z_{-1,i})$ corresponds to $SLATE_{\mathcal{J}}$ with $\mathcal{J} = \{\tilde{Z}_3, \tilde{Z}_4\}$. SLATE allows the practitioner to flexibly

[11] Note that a single-instrument SLATE like $SLATE_{\{2\}}$ does not generally correspond to using Z_2 alone as an instrument, since this latter estimand does not control for variation in Z_1 that is correlated with Z_2. If on the other hand the instruments are independent of one another, using 2SLS may be justified, as I show in Supplemental Material Section A.

condition upon compliance with respect to individual or joint variation in the instruments.

The treatment effect parameters $SLATT_{\mathcal{J}}$ and $SLATU_{\mathcal{J}}$ in the final two rows of Table 2 are similar to $SLATE_{\mathcal{J}}$ but additionally condition on units' realized treatment status. For example $SLATT_{\{1,2\}}$ with our two instruments averages over individuals who do go to college, but wouldn't have were it far and expensive.[12] In Section 4.4, SLATT and SLATU are also used to construct bounds on the average treatment effect among the treated/untreated, when potential outcomes are bounded.

To construct some further examples of identified parameters from the ones mentioned in Table 2, one could make use of a closure property of the set of Δ_c that satisfy Property M. Let $\mathcal{C}$ denote the set of $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ that satisfy Property M, and let $c_a(g, z)$ and $c_b(g, z)$ be two functions in $\mathcal{C}$. Then it is straightforward to show that $c_a(g, z) - c_b(g, z) \in \mathcal{C}$ if and only if $c_b(g, z) \leq c_a(g, z)$ for all $z \in \mathcal{Z}, g \in \mathcal{G}^c$.[13] We can use this observation to generate parameters that condition on the "complement" of the complier group for Δ_{c_b} within the larger complier group for Δ_{c_a}. For example, with $J = 2$:

$$E[\Delta_i | G_i \in \mathcal{G}^c - \{D_i(1, Z_{2i}) - D_i(0, Z_{2i})\}]$$

yields the average treatment effect among individuals who are counted in the ACL but not in $SLATE_{\{1\}}$. These individuals would not respond to a counterfactual reduction in college tuition alone, but would respond if both instruments were shifted in concert.

4.3 Identification

This section presents the main identification theorem, and lays out some extensions. The main result is that causal parameters that satisfy Property M are identified under VM with binary instruments, provided the instruments provide sufficient independent variation in treatment uptake. The latter requirement holds when the binary instruments have full (rectangular) support:

Assumption 3 (full support). $P(Z_i = z) > 0$ *for all* $z \in \{0, 1\}^J$

Assumption 3 is stronger than is necessary but simplifies presentation – Appendix A presents a generalization.

Lemma 2 below gives an alternative expression of Assumption 3 that is useful for writing the function $h(z)$ explicitly. For an arbitrary ordering of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$, define the random vector $\Gamma_i = (Z_{S_1 i} \dots Z_{S_k i})$ from products of the Z_{ji} within each subset. Let Σ be the covariance matrix of Γ_i.

Lemma 2. *Assumption 3 holds if and only if* Σ *has full rank.*

Proof. See Appendix D. □

[12] Note that with a single binary instrument, $SLATT_{\{1\}}$ coincides with $ACL = SLATE_{\{1\}}$, as $E[Y_i(1) - Y_i(0) | D_i = 1, G_i = \textit{complier}] = E[Y_i(1) - Y_i(0) | Z_i = 1, \textit{complier}] = E[Y_i(1) - Y_i(0) | \textit{complier}]$, using Assumption 1. However, when the group $\mathcal{G}^c$ consists of more than one group, the "all-compliers" version of $SLATT$ generally differs from ACL.

[13] This follows from linearity and the definition of Property M, while $c_b(g, z) \leq c_a(g, z)$ is necessary for the image of the new function to remain $\{0, 1\}$.

We may now state the main result:

Theorem 1. *Under Assumptions 1-3 (independence & exclusion, VM, and full support), for any c satisfying Property M and any measurable function $f(Y)$ for each $d \in \{0,1\}$:*

$$E[f(Y_i(d))|C_i = 1] = (-1)^{d+1}\frac{E[h(Z_i)\mathbb{1}(D_i = d)f(Y_i)]}{E[h(Z_i)D_i]},$$

provided that $P(C_i = 1) > 0$, where $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - E[\Gamma_i])$ and

$$\lambda = (E[c(g(S_1), Z_i)], \dots E[c(g(S_k), Z_i)])'$$

Proof. See Appendix D. □

It follows immediately from Theorem 1 that $\Delta_c = E[Y_i(1) - Y_i(0)|C_i = 1]$ satisfying Property M are identified as:

$$\Delta_c = E[h(Z_i)Y_i]/E[h(Z_i)D_i]$$

Note that as the numerator of Δ_c depends on Z_i and Y_i only and the denominator depends on Z_i and D_i only, identification of Δ_c would hold in a "split-sample" setting where Y_i and D_i are not linked in the same dataset.

We can also re-express the empirical estimand for Δ_c delivered by Theorem 1 in a more illuminating form, directly in terms of conditional expectation functions of each of Y_i and D_i on the instruments:

Corollary 1. *Under the Assumptions of Theorem 1:*

$$\Delta_c = \frac{\sum_{S\subseteq\{1\dots J\},S\neq\emptyset}\lambda_S\sum_{z\in\mathcal{Z}}A_{S,z}E[Y_i|Z_i = z]}{\sum_{S\subseteq\{1\dots J\},S\neq\emptyset}\lambda_S\sum_{z\in\mathcal{Z}}A_{S,z}E[D_i|Z_i = z]}$$

where λ_S is as defined in Theorem 1 and $A_{S,z} = \sum_{\substack{f\subseteq z_0\\(z_1\cup f)=S}}(-1)^{|f|}$, with (z_1, z_0) a partition of the indices $j \in \{1 \dots J\}$ that take a value of zero or one in z, respectively.

Proof. See Appendix D. The proof gives the explicit form of A for $J = 2$. □

4.3.1 Discussion and extensions to Theorem 1

In this section I point out a few consequences and extensions of Theorem 1.

1) The size of the relevant complier sub-population is identified: The argument used in Theorem 1 can be leveraged to show that the proportion of relevant "compliers" associated with any causal parameter satisfying Property M is also identified, and is the denominator of the associated estimand:

Corollary 2 to Theorem 1. *Make Assumptions 1-3. For any c that satisfies Property M, $P(C_i = 1)$ is identified as $E[h(Z_i)D_i]$, where $h(z)$ is as given in Theorem 1.*

Proof. See Appendix D. □

2) Property M as a necessary condition. Property M was introduced in this section as part of a set of sufficient conditions for identification of Δ_c. One can show that, loosely speaking, any identified Δ_c must satisfy Property M. In this sense, Property M is also a necessary condition for identification. The simplest form of this result I express in terms of so-called "IV-like estimands" introduced by Mogstad et al. (2018), which are any cross moment $E[s(D_i, Z_i)Y_i]$ between Y_i and a function of treatment and instruments. Let $\mathcal{P}_{DZ}$ denote the joint distribution of D and Z, which is identified. Then:

Proposition 6. *Suppose Δ_c is identified by a finite set of IV-like estimands and $\mathcal{P}_{DZ}$, provided that Assumptions 1-3 hold and $P(C_i = 1) > 0$. Then Δ_c satisfies Property M.*

Proof. See Appendix D. □

The result can be strengthened given regularity conditions on the support of potential outcomes:

Proposition 7. *Suppose that the support of each potential outcome conditional on $G_i = g$ is independent of all $g \in \mathcal{G}^c$ for which $P(G_i = g) > 0$, and that the density (or p.m.f.) of each $Y_i(d)$ is uniformly bounded and separated from zero over that support, conditional on each such $G_i = g$. Then if Δ_c is point identified from the distribution of (Y_i, D_i, Z_i) whenever Assumptions 1-3 hold and $P(C_i = 1) > 0$, Δ_c must satisfy Property M.*

Proof. See Supplemental Material Section D.2. □

3) PM alone does not lead to identification. We can demonstrate that the assumption of vector monotonicity does have identifying power in Theorem 1, above and beyond that of partial monotonicity. For the $J = 2$ case, it is possible to see by explicit enumeration of the possible compliance groups that Theorem 1 cannot hold under PM only:

Proposition 8. *When $J = 2$, if PM holds but neither VM nor IAM hold, the ACL is not point identified from knowledge of any set of IV-like estimands and $\mathcal{P}_{DZ}$.*

Proof. See Appendix D. □

4) Linear dependency among the instruments: Assumption 3 is stronger than is strictly necessary for identification, since linear dependencies between products of the instruments may not pose a problem if the corresponding "weights" in Δ_c do not need be tuned independently from one another. In Appendix A, I give a version of Assumption 3 and generalization of the identification theorem that can accommodate instrument support restrictions and/or non-rectangular $\mathcal{Z}$ (for instance after applying Proposition 4).

5) Other treatment effect parameters. By choosing $f(Y) = \mathbb{1}(Y \leq y)$ in Theorem 1 for some value y in the support of Y_i, we can identify the CDF of each potential outcome at y conditional on $C_i = 1$ as: $F_{Y(d)|C=1}(y) = (-1)^{d+1} \frac{E[h(Z_i)\mathbb{1}(D_i=d)\mathbb{1}(Y_i \leq y)]}{E[h(Z_i)D_i]}$ (note that unlike identification of Δ_c this requires observing (Y_i, Z_i, D_i) all in the same sample). This

allows for the identification of $C_i = 1$ conditional quantile treatment effects, bounds on the distribution of treatment effects (Fan and Park, 2010), or distributional treatment effects: $F_{Y(1)|C=1}(y) - F_{Y(0)|C=1}(y)$ as $\frac{E[h(Z_i)\mathbb{1}(Y_i \leq y)]}{E[h(Z_i)D_i]}$.

6) Covariates. If Assumption 1 holds only conditional on a set of covariates X, and Assumption 3 also holds conditionally, then Theorem 1 can be taken to hold within a covariate cell $X_i = x$. In Appendix B, we describe how covariates can be accommodated nonparametrically.

Identification of the ACL from a single Wald ratio

In the case of the ACL, the RHS of Theorem 1 can be simplified to show that the ACL is in fact equivalent to the following Wald ratio:

$$\rho_{\bar{Z},\underline{Z}} := \frac{E[Y_i|Z_i = \bar{Z}] - E[Y_i|Z_i = \underline{Z}]}{E[D_i|Z_i = \bar{Z}] - E[D_i|Z_i = \underline{Z}]} \tag{4}$$

where $\bar{Z} = (1, 1, \ldots 1)'$ and $\underline{Z} = (0, 0, \ldots 0)'$. That $\rho_{\bar{Z},\underline{Z}}$ and $E[h(Z_i)Y_i]/E[h(Z_i)D_i]$ are equivalent is not obvious from the form of function $h(\cdot)$ given in Theorem 1, but the equality can be shown by applying Corollary 1 (see proof of Proposition 9 for details).

Alternatively, the equivalence between $\rho_{\bar{Z},\underline{Z}}$ and ACL can be shown directly. The result actually holds generally under either VM or IAM, with instruments having finite support. In general, let $\underline{Z}$ and $\bar{Z}$ denote the "lowest" and "highest" instrument values for which Z_i has support, as ranked by the propensity score:

Proposition 9. *Let Assumption 1 and VM or IAM hold with finite $|\mathcal{Z}|$. Then $ACL = \rho_{\bar{Z},\underline{Z}}$, where $\bar{Z} = argmax_{z\in\mathcal{Z}} E[D_i|Z_i = z]$ and $\underline{Z} = argmin_{z\in\mathcal{Z}} E[D_i|Z_i = z]$, with the generalized definition $ACL = E[\Delta_i|G_i \in \mathcal{G}^c]$ for $\mathcal{G}^c := \{g \in \mathcal{G} : E[D_g(Z_i)] \in (0, 1)\}$.*

Proof. See Appendix D. □

Proposition 9 shows that ACL is identified by a remarkably simple population quantity: one can restrict the population to $Z_i \in \{\underline{Z}, \bar{Z}\}$ and use $\mathbb{1}(Z_i = \bar{Z})$ as a single instrument. However, Theorem 1 yields identification of a much larger class of parameters than ACL alone, which do not appear to have single Wald counterparts. Furthermore, as we will see in Section 5, Theorem 1 suggests a means of improving estimation of the ACL. In particular, when the number of sample observations in $\underline{Z}$ and $\bar{Z}$ is not large, the Wald ratio $\rho_{\bar{Z},\underline{Z}}$ may be difficult to estimate precisely, and the sample analog of Eq. (4) can be expected to perform poorly. We will see that a regularization procedure based on the expression for Δ_c from Theorem 1 can be helpful in such cases.

4.4 Identified sets for ATE, ATT, and ATU

One drawback of the identification results in the preceding section is that since the parameters Δ_c satisfying Property M exclude never-takers and always-takers by assumption,

their definition depends upon the set of instruments available. This is not ideal unless the compliant subpopulation is directly of interest, for example when the policy-maker is interested in the effect of manipulating the instruments themselves.

When Y_i has bounded support, the parameters identified by Theorem 1 can be used to generate sharp worst-case bounds in the spirit of Manski (1990) for the unconditional average treatment effect (ATE), average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU). Suppose that $Y_i(d) \in [\underline{Y}, \bar{Y}]$ with probability one, for each $d \in \{0, 1\}$. Then

$$ATE \in \left[(1 - p_a - p_n)ACL - (p_a + p_n)(\bar{Y} - \underline{Y}), (1 - p_a - p_n)ACL + (p_a + p_n)(\bar{Y} - \underline{Y})\right]$$

where under Assumptions 1 and 3: $p_a = E[D_i|Z_i = \underline{Z}]$, $p_n = E[1 - D_i|Z_i = \bar{Z}]$, which are point identified. This can be seen by noting that $ATE := E[Y_i(1) - Y_i(0)] = p_a\Delta_a + p_n\Delta_n + (1 - p_t - p_a)ACL$ and that $\Delta_a, \Delta_n \in [-(\bar{Y} - \underline{Y}), +(\bar{Y} - \underline{Y})]$.

Note that under the bounded support condition the ATE can be partially identified whenever its conditional analog is identified for *some* subgroup of the population, and the size of that subgroup is also identified. Using variation in all of the instruments, as the ACL does, for the conditioning event leads to the narrowest possible such bounds.

We can place similar identified bounds on the ATT. Using that $P(G_i = a.t.|D_i = 1) = \frac{P(G_i=a.t.,D_i=1)}{P(D_i=1)} = \frac{p_a}{P(D_i=1)}$ and $P(G_i = n.t.|D_i = 1) = 0$, we have that

$$ATT \in \left[\left(1 - \frac{p_a}{E[D_i]}\right) SLATT_{\{1...J\}} - \frac{p_a}{E[D_i]}(\bar{Y} - \underline{Y}), \left(1 - \frac{p_a}{E[D_i]}\right) SLATT_{\{1...J\}} + \frac{p_a}{E[D_i]}(\bar{Y} - \underline{Y})\right]$$

and similarly for ATU, with $SLATU$ replacing $SLATT$, p_n replacing p_a, and $1 - E[D_i]$ replacinging $E[D_i]$.

5 Estimation

This section proposes a two-step estimator for the family of identified causal parameters introduced in Section 4. The finite sample performance of the estimator will be an important consideration, and I suggest a data-driven regularization procedure designed to improve performance in small samples. In Appendix C I provide simulation evidence that the regularized estimator performs reasonably well and improves upon existing ones.

5.1 A two-step estimator for vector monotonicity with binary instruments

Theorem 1 establishes that conditional average treatment effects Δ_c satisfying Property M are equal to a ratio of two population expectations – thus a natural plug-in estimator simply replaces these with their sample counterparts, provided $h(Z_i)$ is a strong enough instrument to avoid any weak identification issues.

Following $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - E[\Gamma_i])$ from Theorem 1, define $\hat{H}_i = n\tilde{\Gamma}(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\hat{\lambda}$, where $\tilde{\Gamma}$ is a $n \times k$ design matrix composed of empirically de-meaned Γ_i as rows, where recall that $\Gamma_i = (Z_{S_1 i} \dots Z_{S_k i})$ for an arbitrary ordering of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$. Thus $\tilde{\Gamma}_{il} = Z_{S_l i} - \hat{E}[Z_{S_j i}]$ for the l^{th} subset $S_j \subseteq \{1 \dots J\}$. $\hat{\lambda}$ is a sample estimator of $\lambda = (E[c(g(S_1), Z_i)], \dots E[c(g(S_k), Z_i)])'$, given explicitly below for our leading examples.

Given $\hat{H}_i$ as defined above, consider the plug-in estimator $\hat{\rho} = (\hat{H}'D)^{-1}(\hat{H}'Y)$, where Y and D are $n \times 1$ vectors of observations of Y_i and D_i, respectively. Noticing that for any vector $V \in \mathbb{R}^n$, $(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\tilde{\Gamma}'V$ is the sample linear projection coefficient vector of V on the de-meaned columns Γ_s, we can re-express it by the Frisch-Waugh-Lovell theorem as $(0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'V$ where $\Gamma = [1, \Gamma_1, \dots \Gamma_{|\mathcal{F}|}]$ adds a column of ones, and skips the demeaning of each Γ_s. The estimator can now be written as $\hat{\rho} = \hat{\rho}(\hat{\lambda})$, where

$$\hat{\rho}(\lambda) = \left((0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'D\right)^{-1}(0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'Y \tag{5}$$

In this section I assume existence of $(\Gamma'\Gamma)^{-1}$ in finite sample, and note that its population analog exists as a consequence of Assumption 3. When Assumption 3 does not hold but identification is still possible (see Appendix A), $\hat{H}_i$ may be defined in the same way from a subvector of Γ_i for $S \in \mathcal{F}$ that has a full-rank variance matrix can be used instead. For example, when using construction of Proposition 4 that maps discrete to binary instruments, $\mathcal{F}$ includes all products of the final binary instruments that do not contain distinct $\tilde{Z}$ from the same original discrete instrument. In all cases, let $\mathcal{F}$ index the elements of Γ_i, where $\mathcal{F} = \{S \subseteq \{1, 2, \dots J\}, S \neq \emptyset\}$ in the baseline setting.

Comparison with 2SLS: Note that the estimator $\hat{\rho}(\lambda)$ in Equation 5 is very similar in form to a "fully-saturated" 2SLS estimator that includes an indicator for each value of $Z_i \in \mathcal{Z}$ in the first stage. Indeed, that estimator is $\hat{\rho}_{2sls} = \left(D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'D\right)^{-1} D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'Y$.[14] The key difference is that rather than aggregating over linear projection coefficients $(\Gamma'\Gamma)^{-1}\Gamma'V$ for $V \in \{D, Y\}$ using the weights $D'\Gamma$ (which are governed asymptotically by the statistical distribution of D_i and Z_i), $\hat{\rho}(\lambda)$ uses weights $(0, \lambda')$, chosen to match the desired parameter of interest. Relative to 2SLS, $\hat{\rho}(\lambda)$ can be thought of as sacrificing some statistical efficiency in order to guarantee that it recovers a well-defined causal parameter under VM. In Section 5.2 I discuss regaining some of that lost efficiency through regularization, which is borne out in the simulation in Appendix C.

Under regularity conditions (see Theorem 2), we will have that for any $\hat{\lambda} \xrightarrow{p} \lambda \in \mathbb{R}^{|\mathcal{F}|}$:

$$\hat{\rho}(\hat{\lambda}) \xrightarrow{p} \sum_{g \in \mathcal{G}^c} \frac{P(G_i = g)[M_J\lambda]_g}{\sum_{g' \in \mathcal{G}^c} P(G_i = g')[M_J\lambda]_{g'}} \cdot \Delta_g$$

Matching the RHS of the above to particular estimands Δ_c that satisfy Property M is achieved by choosing $\hat{\lambda}$. Table 3 gives natural sample estimators for ACL, SLATE,

[14] The proof of Corollary 1 gives the basis transformation from a design matrix of indicators to Γ, which cancels in $\hat{\rho}_{2sls}$.

SLATT and SLATU that are consistent. Note that in the case of the ACL $\hat{\lambda}$ does not depend on the data and thus no "first-step" is necessary in estimation.

Parameter	Estimator $\hat{\lambda}$ of population λ
ACL	$(1,1,\dots 1)'$
$SLATE_{\mathcal{J}}$	$\hat{\lambda}_S^{SLATE_{\mathcal{J}}} = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S-\mathcal{J},i} = 1\right)$
$SLATT_{\mathcal{J}}$	$\hat{\lambda}_S^{SLATT_{\mathcal{J}}} = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S,i} = 1\right)$
$SLATU_{\mathcal{J}}$	$\hat{\lambda}_S^{SLATU_{\mathcal{J}}} = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S-\mathcal{J},i}(1 - Z_{\mathcal{J},i}) = 1\right)$

Table 3: Estimators $\hat{\lambda}$ for the leading parameters of interest. $S - \mathcal{J}$ denotes the set difference $\{j : j \in S, j \notin \mathcal{J}\}$.

5.2 Regularization of the estimator

In this section I propose a regularization procedure for the estimator, to improve its performance in small samples. This is of particular interest with regards to the All Compliers LATE, for which the simple plug-in estimator is seen to only use data at two points in $\mathcal{Z}$. The regularization procedure helps the estimator to make more efficient use of a full dataset by incorporating the observations at other instrument values.

From Proposition 9 there is a natural alternative Wald estimator of the ACL:

$$\hat{\rho}_{\bar{Z},\underline{Z}} := \frac{\hat{E}[Y_i|Z_i = \bar{Z}] - \hat{E}[Y_i|Z_i = \underline{z}]}{\hat{E}[D_i|Z_i = \bar{Z}] - \hat{E}[D_i|Z_i = \underline{z}]} \tag{6}$$

where recall that under Assumption 3 $\bar{Z} = (111\dots 1)'$ or $\underline{Z} = (000\dots 0)'$. It can be shown that $\hat{\rho}_{\bar{Z},\underline{Z}}$ and $\hat{\rho}((1,1,\dots 1)')$ in Equation 5 are numerically equivalent in finite sample.[15] In situations where there is non-zero but small support on the points $\bar{Z}$ and $\underline{Z}$, we may thus expect that $\hat{\rho}((1,1,\dots 1)')$ may perform quite poorly as an estimator of ACL in small samples, since it effectively ignores all of the data for which $Z_i \notin \{\underline{Z}, \bar{Z}\}$. This issue is mentioned by Frölich 2007 in the context of IAM, in which case $\hat{\rho}_{\bar{Z},\underline{Z}}$ is also consistent for the ACL with finite $\mathcal{Z}$ (see Proposition 9).

One way to see this issue is as a collinearity problem: when there are few observations in the points $\bar{Z}$ and $\underline{Z}$, the $n \times |\mathcal{F}|$ design matrix Γ will have singular values that are close to zero (to see this, note that $\Gamma'\Gamma = A'^{-1} n \cdot diag\{\hat{P}(Z_i = z)\}A^{-1}$). This observation suggests that the issue might be mitigated by employing a ridge-type shrinkage estimator (see e.g. Hoerl and Kennard, 1970). Accordingly, we allow a sequence of regularization parameters α_n:

$$\hat{\rho}(\hat{\lambda}, \alpha) = \left((0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1}\Gamma' D\right)^{-1} (0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1}\Gamma' Y \tag{7}$$

The estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ with a choice of $\alpha > 0$ establishes a floor on the singular values of the matrix Γ.

[15] To see this, note that the vector H of H_i solves the system of equations $\Gamma' H_i = (1\dots 1)'$. Among vectors that are in the column space of Γ, H is the unique such solution, given that the design matrix Γ has full column rank. One can readily verify that $\Gamma' H = (1,1,\dots 1)$ with the choice $H_i = \frac{\mathbb{1}(Z_i=(1\dots 1))}{\hat{P}(Z_i=(0\dots 0))} - \frac{\mathbb{1}(Z_i=(0\dots 0))}{\hat{P}(Z_i=(0\dots 0))}$, and that this $H = \Gamma\eta$ with $\eta = (1/\hat{P}(Z_i = (1\dots 1)), 0, \dots 0, -1/\hat{P}(Z_i = (0\dots 0)))$. In general, the empirical counterpart of any Wald ratio ρ_{zw} will be obtained by $\hat{\rho}(\lambda)$ with the choice $\lambda_S = z_S - w_S$, where $z_S := \prod_{j\in S} z_j$ and similarly for w.

In the case of the ACL, Corollary 1 can be leveraged to show that $\alpha > 0$ allows the estimator to make use of the full support of Z_i, rather than just the two points $\bar{Z}$ and $\underline{Z}$. But ridge regression comes at the expense of some bias. Proposition 10 below yields a means of navigating this tradeoff to choose α in practice. In particular, I propose choosing α to minimize a feasible estimator of the conditional MSE $E[(\hat{\rho}(\lambda, \alpha) - \Delta_c)^2 | Z_1 \dots Z_n]$.

Proposition 10. *Under the assumptions of Theorem 1, $E[(\hat{\rho}(\lambda, \alpha) - \Delta_c)^2 | Z_1 \dots Z_n]$ is, up to second order in estimation error and a positive constant of proportionality:*

$$\tilde{\lambda}'(\Gamma'\Gamma + \alpha I)^{-1} \left\{ \Gamma'(\Omega_Y + \Delta_c^2 \Omega_D - 2\Delta_c \Omega_{YD})\Gamma \right. \\ \left. + \alpha^2(\beta_Y \beta_Y' + \Delta_c^2 \beta_D \beta_D' - 2\Delta_c \beta_Y \beta_D') \right\} (\Gamma'\Gamma + \alpha I)^{-1} \tilde{\lambda} \quad (8)$$

where $\tilde{\lambda} := (0, \lambda')'$, $\beta_Y := E[\Gamma_i \Gamma_i']^{-1} E[\Gamma_i Y_i]$, $\beta_D := E[\Gamma_i \Gamma_i']^{-1} E[\Gamma_i D_i]$, and $\Omega_{VW} = E[(V - \beta_V \Gamma)(W - \beta_W \Gamma)' | \Gamma]$ for $V, W \in \{Y, D\}$, and all expectations are assumed to exist.

Furthermore, if $\hat{\alpha}_{mse}$ is chosen as the smallest positive local minimizer of the following estimate of the above:

$$\hat{M}(\alpha) := (0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1} \left\{ n\hat{\Pi} + \alpha^2(\hat{\beta}\hat{\beta}') \right\} (\Gamma'\Gamma + \alpha I)^{-1}(0, \hat{\lambda}')'$$

with $\hat{\beta}_V := (\Gamma'\Gamma)^{-1}\Gamma' V$ for each $V \in \{Y, D\}$, $\hat{\Pi} := \frac{1}{n}\sum_i (Y_i - \hat{\beta}_Y \Gamma_i - \frac{(0,\hat{\lambda}')\hat{\beta}_Y}{(0,\hat{\lambda}')\hat{\beta}_D}(D_i - \hat{\beta}_D \Gamma_i))^2 \Gamma_i \Gamma_i'$ and $\hat{\beta} := \hat{\beta}_Y - \frac{(0,\hat{\lambda}')\hat{\beta}_Y}{(0,\hat{\lambda}')\hat{\beta}_D}\hat{\beta}_D$ then

$$\hat{\alpha}_{mse}/\sqrt{n} \xrightarrow{p} 0$$

provided that $\hat{\lambda} \xrightarrow{p} \lambda$, $(0, \lambda')\Sigma^{-1}(\beta_Y + \Delta_c \beta_D) \neq 0$.

Proof. See Appendix D. □

The proposed data-driven choice $\hat{\alpha}_{mse}$ estimates the unknown quantities in Eq. (8) based on an initial guess of $\alpha = 0$, and then minimizes with respect to α. This can be seen as a "one-step" version of a more general iterative algorithm in which a value α_t is used to compute the function $\hat{M}(\alpha)$, which is then minimized to find α_{t+1} and so on until convergence. I implement the single-step version in Appendix C, and find that it indeed improves estimation error considerably for the simulation DGPs considered.

The reason that my proposed rule evaluates $\hat{\alpha}_{mse}$ as a local minimizer of $\hat{M}(\alpha)$ rather than a global minimzer, is that the function $\hat{M}(\alpha)$ is always positive but approaches zero as $\alpha \to \infty$. This stands in contrast with the standard case of ridge regression in which regularization bias always grows with α, eventually dominating any efficiency gains from increasing it further. In the present case, the vector $\hat{\beta}$ as defined above and $(0, \hat{\lambda}')'$ are orthogonal (in sample as well as in the population limit), and thus the "(squared) bias" term vanishes as $\alpha \to \infty$, along with the variance of the regularized estimator (this is roughly analogous to ridge regularizing a vector of regression coefficients when their true values are all zero). Nevertheless, the function $\hat{M}(\alpha)$ does have a well-defined local minimum that achieves a lower value than $\hat{M}(0)$ at some strictly positive α (see Appendix

D for details), and this local minimum is shown to provide a helpful guide to choosing α in the simulations of Appendix C. Note that the condition $(0, \lambda')\Sigma^{-1}(\beta_Y + \Delta_c\beta_D) \neq 0$ in Proposition 10 rules out a knife-edge case in which the Hessian of $\hat{M}(\alpha)$ is zero when the other arguments of $\hat{M}$ are evaluated at their probability limits.

5.3 Asymptotic distribution

Consistency and asymptotic normality of the estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ follows in a straightforward way from the results thus far. In particular, with $\alpha = 0$ the asymptotic variance can be computed as a special case of Theorem 3 in Imbens and Angrist (1994). In our setting, we can view estimation of $h(z)$ as a parametric problem $h(z) = g(z, \theta)$ where the parameter vector θ is the mean and variance of Γ_i, along with the vector λ:

$$\theta = (\mu_\Gamma, \Sigma, \lambda)' = (\{\mu_{\Gamma,l}\}_l, \{\Sigma_{lm}\}_{l\leq m}, \{\lambda\}_l)' \text{ with } l, m \in \{1 \dots |\mathcal{F}|\}$$

Then $\hat{\rho}(\lambda, \alpha) = \widehat{Cov}(g(Z_i, \hat{\theta}), Y_i)/\widehat{Cov}(g(Z_i, \hat{\theta}), D_i)$, where $\hat{\theta}$ solves a set of moment conditions $\sum_{i=1}^N \psi(Z_i, \hat{\theta}) = 0$ given explicitly in the theorem below.

Theorem 2 below allows $\alpha_n > 0$ provided that the sequence converges in probability to zero at a sufficient rate. By Proposition 10, we obtain this rate for the "one-step" minimizer of the feasible MSE estimate given in Eq (8).

Theorem 2. *Under the Assumptions of Theorem 1, if $\alpha_n = o_p(\sqrt{n})$ then*

$$\sqrt{n}(\hat{\rho}(\hat{\lambda}, \alpha_n) - \Delta_c) \xrightarrow{d} N(0, V)$$

where $V = \mathbf{e_1}'\Pi^{-1}\Omega(\Pi')^{-1}\mathbf{e_1}$ (i.e. the top-left element of $\Pi^{-1}\Omega(\Pi')^{-1}$) with:

$$\Omega = \begin{pmatrix} -E[D_i g(Z_i, \theta)] & -E[g(Z_i, \theta)] & E[U_i d_\theta g(Z_i, \theta)] \\ -E[D_i] & -1 & 0 \\ 0 & 0 & E[d_{\theta'}\psi(Z_i, \theta)] \end{pmatrix}$$

$$\Pi = \begin{pmatrix} E[g(Z_i, \theta)^2] & E[g(Z_i, \theta)U_i] & E[g(Z_i, \theta)\psi(Z_i, \theta)]' \\ E[g(Z_i, \theta)U_i] & E[U_i^2] & E[U_i\psi(Z_i, \theta)]' \\ E[g(Z_i, \theta)U_i\psi(Z_i, \theta)] & E[U_i\psi(Z_i, \theta)] & E[\psi(Z_i, \theta)\psi(Z_i, \theta)'] \end{pmatrix}$$

so long as Ω and Π are finite and Π has full rank, with the definitions:

$$U_i := Y_i - E[Y_i] - \Delta_c(D_i - E[D_i])$$

$$\theta = (\mu_\Gamma, \Sigma, \lambda)' = (\{\mu_{\Gamma,l}\}_l, \{\Sigma_{lm}\}_{l\leq m}, \{\lambda\}_l)'$$

$$g(z, \theta) = \lambda'\Sigma^{-1}(\Gamma(Z_i) - \mu_\Gamma)$$

$$\psi(Z_i, \theta) = ((\Gamma(Z_i) - \mu_\Gamma)', \{(\Gamma_l(Z_i) - \mu_{\Gamma,l})(\Gamma_m(Z_i) - \mu_{\Gamma,m}) - \Sigma_{lm}\}_{l\leq m}, \{c_l(Z_i) - \lambda_l\}_l)'$$

Here $\Gamma(Z_i) = (\Gamma_1(Z_i) \dots \Gamma_{|\mathcal{F}|}(Z_i))'$ where $\Gamma(Z_i)_l = Z_{S_l,i}$ for some arbitrary ordering S_l of the sets in $\mathcal{F}$, and $c_l(z) = c(g(S_l), z)$ (and thus $P(C_i = 1|G_i = g(S_l)) = E[c_l(Z_i)]$).

Proof. See Appendix D. □

6 Revisiting the returns to college

In this section I apply the results developed thus far to study the labor market returns to college. In the past, this literature has based IV methods on either an assumption of homogenous treatment effects, or the traditional IAM notion of monotonicity. Using the methods developed in this paper valid under VM, I find estimates broadly in line with existing ones. This puts previous findings onto firmer conceptual ground, as VM is likely to be a more plausible assumption than IAM in this context. A second empirical application to the effects of children on labor supply is reported in Supplemental Material Section C.2.

I use the dataset from Carneiro, Heckman and Vytlacil (2011) (henceforth CHV) constructed from the 1979 National Longitudinal Survey of Youth. The sample consists of 1,747 white males in the U.S., first interviewed in 1979 at ages that ranged from 14 to 22, and then again annually. The outcome of interest Y_i is the log of individual i's wage in 1991, and treatment $D_i = 1$ indicates i attended at least some college. As in CHV, log wages are divided by four so that treatment effects are expressed in roughly per-year equivalents.

I consider two binary instruments defined as follows: $Z_{2i} = 1$ indicates the presence of a public college in i's county of residence at age 14, while $Z_{1i} = 1$ indicates that average tuition rates local to i's residence around age 17 falls below the 90% sample percentile (see Carneiro et al. 2011 for details of the underlying tuition variable). This cutoff corresponds to about \$3,100 in 1993 dollars, and falls to the right of the modal mass of the distribution (figure in Supplemental Material Section C.1). I choose this weak definition of "cheap" in order to provide a comparatively sharp first stage, which allows for a meaningful discussion of the compliance group sizes despite a general lack of statistical power owing to the small sample size. Two alternative definitions of the tuition instrument are considered in the Supplemental Material, including treating it as a discrete rather than binary variable.

Distribution of the instruments

		Z_2		
		far	**close**	
Z_1	**expensive**	135	38	*173*
	cheap	695	879	*1,574*
		830	*917*	

Table 4: Cross-tabulation of the instruments. Total $N = 1,747$. It is clear that the sample mass is not evenly distributed across the four cells of $\mathcal{Z}$, and that the instruments are positively correlated.

While VM is a plausible assumption given the two available instruments, Assumption 1 is likely to be confounded by unobserved heterogeneity that is correlated with location during teenage years. I apply the common strategy of assuming a conditional version of

instrument validity, that

$$\{(Y_i(1), Y_i(0), Z_i) \perp Z_i\} \,|X_i \tag{9}$$

where X_i is a vector of observed covariates unaffected by treatment. As the primary focus of this paper is on identification and not estimation with covariates, I use just one control variable X_i, each individual's schooling-corrected Armed Forces Qualification Test score (see CHV for details).

Unconditional propensity scores

		Z_2	
		far	**close**
Z_1	**expensive**	0.437	0.500
	cheap	0.427	0.557

Mean fitted propensity scores

		Z_2	
		far	**close**
Z_1	**expensive**	0.426	0.441
	cheap	0.448	0.545

Table 5: Raw sample means $\hat{E}[D_i|Z_i]$ (left) and fitted propensity scores by linear regression (right) for Setting A. Fitted propensity scores are evaluated at the sample mean of X_i–see text for details. $N = 1,747$.

The value of conditioning on X can already be seen in Table 5, which reports the propensity score across the four points in $\mathcal{Z}$. While Equation 9 and VM imply that the conditional propensity score function $E[D_i|Z_i = z, X_i = x]$ should be monotonic componentwise in z (for any realization x), such monotonicity need not hold for the unconditional propensity score $E[D_i|Z_i = z]$. The latter of these is shown in the left panel of Table 5, which can be seen to be decreasing between the points $(far, expensive)$ and $(far, cheap)$. By contrast, the propensity scores reported in the right panel are increasing as one moves to the right or down in the table, as expected under VM.[16] The values in the right panel are estimated from a linear regression of D_i on X_i and indicators for the values of Z_i, and predictions are evaluated at the mean of X_i.

While the sizes of six compliance groups that can exist under VM with two binary instruments are not point identified, the right panel of Table 5 does offer some evidence on their sizes. Firstly, the share of always-takers is estimated to be about 43% of the population, while the share of never-takers is $1 - 0.55 = 45\%$. The remaining 12% are generalized "compliers" consisting of the tuition (Z_1), proximity (Z_2), eager and reluctant compliers. From the table we see that $P(D_i(expensive, close, x) > D_i(expensive, far, x)) \approx 1.5\%$, $P(D_i(cheap, far, x) > D_i(expensive, far, x)) \approx 2.2\%$, and $P(D_i(cheap, close, x) > D_i(expensive, close, x)) \approx 10.4\%$. This implies that between 8.2% and 9.7% of the population are reluctant compliers, the majority of all individuals in $\mathcal{G}^c$. We can also deduce that between 0.7% and 2.2% are tuition compliers, while no more than 1.5% each are proximity compliers or eager compliers (see Section 3).

We now turn to treatment effect estimates. Appendix B discusses nonparametric identification of causal parameters Δ_c with Equation 9 replacing the unconditional As-

[16] That propensity score monotonicity is an empirical prediction of VM coupled with conditional independence, suggests that VM could be used to test possible sets of control variables. I leave a formal investigation of this idea to future work.

sumption 1. Consider first the ACL, for which the results are particularly simple by virtue of its equivalence to a Wald ratio. By the results in Appendix B along with Corollary 1, it follows that:[17]

$$ACL = \frac{E\left\{E[Y_i|Z_i = (1,1), X_i] - E[Y_i|Z_i = (0,0), X_i]\right\}}{E\left\{E[D_i|Z_i = (1,1), X_i] - E[D_i|Z_i = (0,0), X_i]\right\}}$$

In a baseline specification, I assume that the expectations $E[Y_i|Z_i = z, X_i = x]$ are linear in x and additively separable between z and x. However, there is no reason to expect such additivity given the underlying model; I relax this assumption by reporting specifications that add interactions between the instrument indicators and X, or with a quadratic in X. Estimators are constructed from OLS coefficients in the natural way, with terms that include X evaluated at their sample means. As the asymptotic result from Theorem 2 based on Imbens and Angrist (1994) does not cover the case with covariates, I compute delta-method standard errors from the system of two regression equations (one for D_i and one for Y_i), allowing for heteroscedasticity and cross-correlation between the equations.

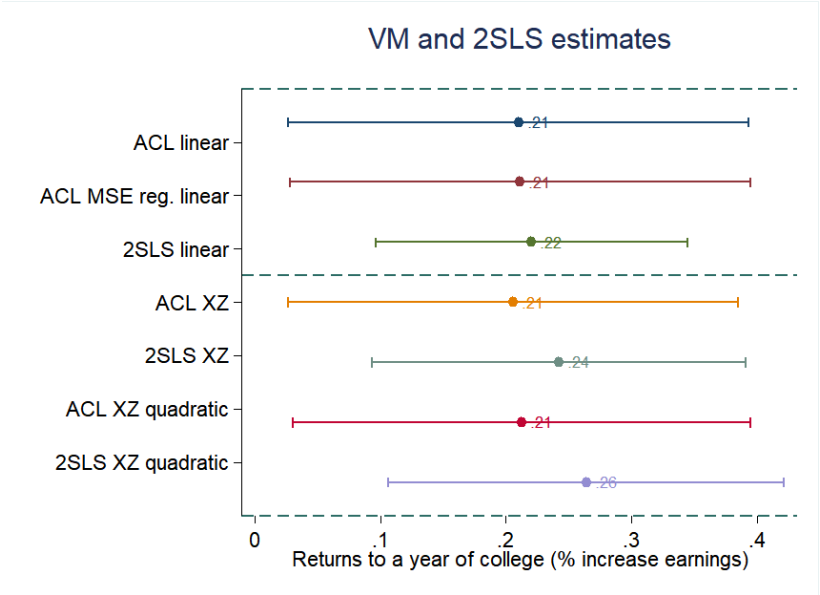

Figure 2: Estimates of the ACL. For each of the three specifications, the figure reports the ACL estimate and the analagous fully-saturated 2SLS model. Bars indicate 95% confidence intervals.

Figure 2 reports estimates of the ACL alongside the analagous 2SLS estimands for each of the three specifications. Results are fairly robust across specifications, so we focus here on the simplest case with a linear in AFQT and no interactions or regularization. The point estimate of 0.21 indicates that having attended a year of college increases wages in 1991 by roughly 21%. This estimate is in line with the marginal treatment effect function based on Assumption IAM estimated by CHV: which ranges between 0.4 and 0 with a mean of about 0.2. The 2SLS estimate also comports well at 0.22. However, the ACL standard errors are somewhat larger, a reflection of the fact that 2SLS weighs across the groups to minimize variance rather than pin down a specific target parameter.

[17] This type of estimand has also been considered by Frölich, 2007.

In the linear case, I also report a regularized version of the ACL estimator using the approximate MSE minimizing choice of α, with standard errors based on the unregularized estimator (see Proposition 10 and Theorem 2).[18] In this case the point estimate is essentially the same, as the optimal α is estimated to be quite small.

6.1 Capturing Heterogeneity

While the broad averaging implemented by the All Compliers LATE may be desirable from the perspective of learning about the population as a whole, the ACL can mask heterogeneity in treatment effects that is of economic or policy interest. One way to investigate such heterogeneity is to make use of other parameters Δ_c identified by Theorem 1 that capture average causal effects among smaller groups of compliers. The first two rows of Figure 3 report $SLATT_{\mathcal{J}}$ and $SLATU_{\mathcal{J}}$ for $\mathcal{J}$ containing both instruments, which are analogous to the ACL but additionally condition on the event $D_i = 1$ or $D_i = 0$, respectively. This does not reveal much heterogeneity: the point estimates are quite close to the ACL at 0.20 and 0.22.

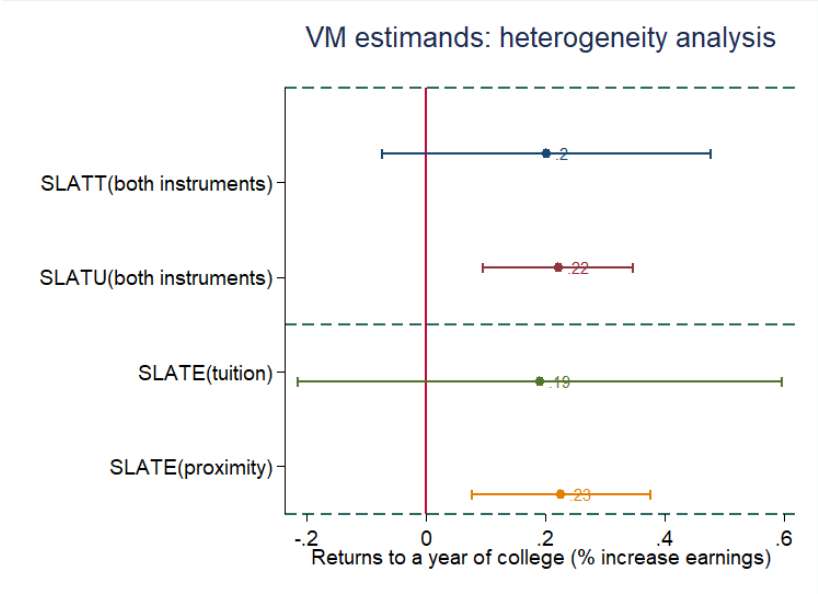

Figure 3: Additional VM estimands applied to Setting A. Bars indicate 95% confidence intervals.

The second two rows of Figure 3 report $SLATE_{\{1\}}$ and $SLATE_{\{2\}}$ respectively, the average treatment effect among students who are responsive to changes in the tuition instrument alone, and the average treatment effect among students who are responsive to changes in the proximity instrument alone. Neither are statistically different from the estimated ACL. While the point estimates for these single instrument LATEs may be policy relevant (see Section 4.1), comparing them is not ideal from the perspective of uncovering heterogeneity since the two parameters do not condition on disjoint groups (e.g. eager compliers for which $Z_i = (far, expensive)$ would respond to a shift in either the proximity or the tuition instrument alone). The ideal would be to compute the

[18] All variables are first linearly residualized with respect to X_i. Then $\hat{M}(\alpha)$ is minimized with respect to α.

average treatment effect Δ_g separately for each of the compliance groups g in $\mathcal{G}^c$. While this is possible under IAM, it is not under VM as the function $c(G_i, Z_i) = \mathbb{1}(G_i = g)$ does not Satisfy Property M.

Nevertheless, the limited complexity of the $J = 2$ case does allow us to speak meaningfully of heterogeneity in the Δ_g, because the data provide nearly enough distinct moments to identify them. Suppose that one group-specific average treatment effect – say say $\Delta_{tuition}$ – was known by external means, along with the associated proportion $p_{tuition}$ of such tuition-only compliers in the population. Then, the other three treatment effects Δ_g and associated group sizes p_g for $g \in \{proximity, eager, reluctant\}$ could be backed-out from three identified Δ_c (such as Wald ratios) and their associated $P(C_i = 1)$. Supplemental Material Section B.3 gives details of the calculation.

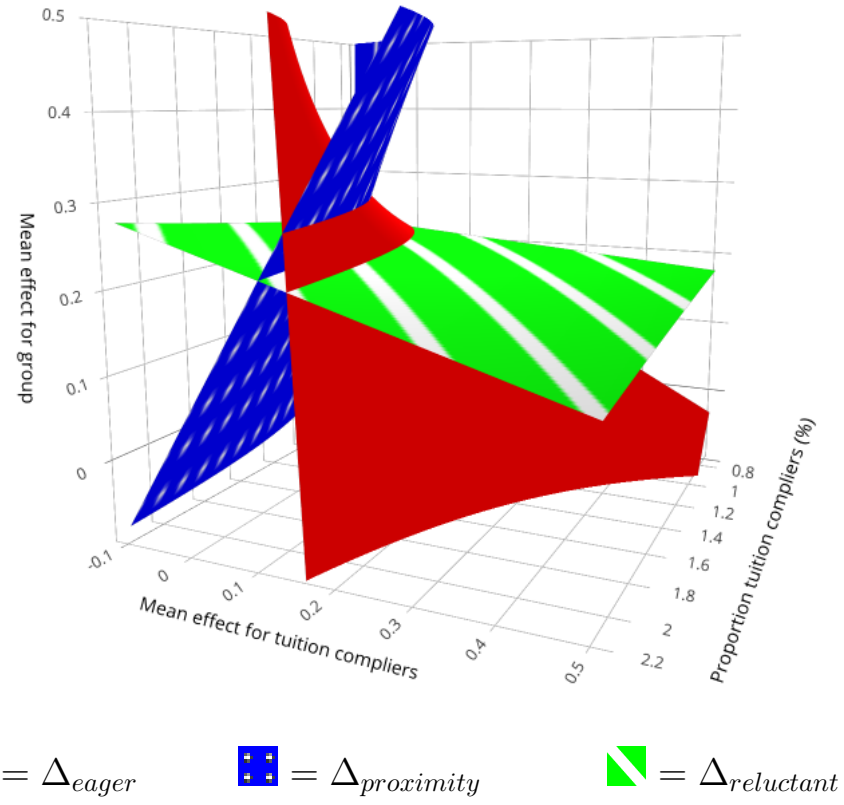

Figure 4: Group specific treatment effect point estimates as a function of the proportion of tuition compliers $p_{tuition}$ and their mean treatment effect $\Delta_{tuition}$. Treatment effect values are clipped to the range $[-0.1, 0.5]$.

Figure 4 reports the results as $p_{tuition}$ is varied across the range of possible values it can take: 0.7% to 2.2%, and $\Delta_{tuition}$ is varied across a range of "reasonable" treatment effect values: -0.1 to 0.5. The figure indicates that treatment effects must in fact be quite heterogeneous. For Δ_{eager}, $\Delta_{proximity}$ and $\Delta_{reluctant}$ to all take similar values to one another, we must have $\Delta_{tuition} \approx 0.1$, with the other three about 0.2 or greater. If $\Delta_{tuition}$ is much below 0.1, then Δ_{eager} is much larger than $\Delta_{proximity}$, and if $\Delta_{tuition}$ is much more than 0.1 then $\Delta_{proximity}$ is much larger than Δ_{eager}, which must then be negative. Even the general ranking of the Δ_g is thus unidentified without further information. Nonetheless,

the implied values of $\Delta_{reluctant}$ are comparatively stable, ranging from about 0.12 to 0.28 across the domain of the figure. Since these reluctant compliers make up the bulk of the complier population, this can explain the relative stability of the reported Δ_c across alternative choices of c, despite the underlying heterogeneity.

7 Conclusion

I have investigated an assumption I call vector monotonicity for cases in which a researcher has multiple instrumental variables for a single binary treatment, extending related results in Mogstad et al. (2019) and Mountjoy (2018). This paper has demonstrated that a class of interpretable causal parameters are generally point identified under vector monotonicity, and can be estimated by a regularized "2SLS-like" estimator. Simulation evidence (see Appendix C) suggests that the regularization can substantially improve small sample performance in some settings.

In an application to the labor market returns to college education, I find that estimates based on vector monotonicity comport well with existing estimates that have assumed classical IAM monotonicity. This is reassuring for the empirical setting considered; however simulation results in Appendix C underline the danger that the common 2SLS estimator may concentrate around a point outside the convex hull of treatment effects in the population, when vector monotonicity holds instead of IAM.

References

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish". *Review of Economic Studies* 67 (3), pp. 499–527.

Angrist, J. D. and Imbens, G. W. (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity". *Journal of the American Statistical Association* 90 (430), p. 431.

Carneiro, P., Heckman, J. J. and Vytlacil, E. J. (2011). "Estimating marginal returns to education". *American Economic Review* 101 (6), pp. 2754–2781.

Chaisemartin, C. de (2017). "Tolerating defiance? Local average treatment effects without monotonicity". *Quantitative Economics* 8 (2), pp. 367–396.

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

Fan, Y. and Park, S. S. (2010). "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference". *Econometric Theory* 26 (3), pp. 931–951.

Frölich, M. (2007). "Nonparametric IV estimation of local average treatment effects with covariates". *Journal of Econometrics* 139 (1), pp. 35–75.

Gautier, E. and Hoderlein, S. (2011). "A triangular treatment effect model with random coefficients in the selection equation", pp. 1–22. arXiv: 1109.0362.

Gunsilius, F. F. (2019). "Nonparametric point-identification of multivariate models with binary instruments". *Working Paper*, pp. 1–47.

Heckman, J. J. and Pinto, R. (2018). "Unordered Monotonicity". *Econometrica* 86 (1).

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006). "Understanding What Instrumental Variables Estimate in Models with Essential Heterogeneity". *The Review of Economics and Statistics* 88 (3).

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

Hoerl, A. and Kennard, R. (1970). "Ridge regression : Biased estimation for nonorthogonal problems". *Technometrics* 42 (1), pp. 80–86.

Imbens, B. G. W. and Newey, W. K. (2009). "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity". *Econometrica* 77 (5), pp. 1481–1512.

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

Kisielewicz, A. (1988). "A solution of Dedekind's problem on the number of isotone Boolean functions". *Journal fur die reine und angewandte Mathematik* 386.

Kleitman, D. J. and Milner, E. C. (1973). "On the average size of the sets in a Sperner family". *Discrete Mathematics* 6 (2), pp. 141–147.

Lee, S. and Salanié, B. (2018). "Identifying Effects of Multivalued Treatments". *Econometrica* 86 (6), pp. 1939–1963.

Lewbel, A. and Yang, T. T. (2016). "Identifying the average treatment effect in ordered treatment models without unconfoundedness". *Journal of Econometrics* 195 (1), pp. 1–22.

Manski, C. F. (1990). "Nonparametric Bounds on Treatment Effects". *The American Economic Review* 80 (2), pp. 829–823.

Mogstad, M., Santos, A. and Torgovitsky, A. (2018). "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters". *Econometrica* 86 (5), pp. 1589–1619.

Mogstad, M., Torgovitsky, A. and Walters, C. (2019). "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions". *Working Paper*.

Mountjoy, J. (2018). "Community Colleges and Upward Mobility". *Working Paper*, pp. 1–83.

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

Yin, J. et al. (2010). "Nonparametric covariance model". *Statistica Sinica* 20 (1), pp. 469–479.

Appendices

A Identification result with instrument degeneracy

This section provides an extension of Theorem 1 for cases when the support $\mathcal{Z}$ of the instruments is not rectangular (i.e. $supp(Z_i) \neq (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$), and there may be perfect linear dependencies between the instruments (of the form that would arise from the mapping from discrete to binary instruments presented in Section 3.3).

A weaker version of Assumption 3 is comprised of the following two conditions, with the definition that $Z_{\emptyset i}$ is a degenerate random variable that takes the value of one with probability one:

Assumption 3a* (existence of instruments). *There exists a family $\mathcal{F}$ of subsets of the instruments $S \subseteq \{1 \dots J\}$, where $\emptyset \in \mathcal{F}$ and $|\mathcal{F}| > 1$, such that random variables Z_{Si} for all $S \in \mathcal{F}$ are linearly independent, i.e. $P\left(\sum_{S \in \mathcal{F}} \omega_S Z_{Si} = 0\right) < 1$ for all vectors $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$.*

Assumption 3b* (non-degenerate subsets generate the compliance groups). *There exists a family $\mathcal{F}$ satisfying Assumption 3a*, such that for any $S \notin \mathcal{F}$, $g(F) \notin \mathcal{G}$ for all Sperner families that F that contain S.*

Assumption 3a* is in itself very weak, requiring only that there exists some product of the instruments that has strictly positive variance. Assumption 3b* is much more restrictive: it says that all compliance groups aside from never-takers can be generated from members of a family of linearly independent subsets of the instruments.

The construction in Proposition 4 mapping discrete instruments to binary instruments yields a case where Assumption 3* will hold, given rectangular support of the original discrete instruments.

Proposition. *Let each Z_j have M_j ordered points of support $z_1^j < z_2^j \cdots < z_{M_j}^j$ and let $\tilde{Z}_m^j = \mathbb{1}(Z_{ji} \geq z_m^j)$. If $P(Z_i = z) > 0$ for $z \in (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$, then Assumption 3* holds with $\mathcal{F}$ the family of all subsets of $\mathcal{M} := \{\tilde{Z}_m^j\}_{\substack{j \in \{1 \ldots J\} \\ m=2 \ldots m_j}}$ containing at most one Z_m^j for any given $j \in \{1 \ldots J\}$.*

Proof. See Appendix D. □

The above proposition allows us to make use of Assumption 3* in cases where discrete instruments are mapped to binary instruments via Proposition 4. To illustrate, consider a case with a single discrete instrument Z_1 having three levels $z_1 < z_2 < z_3$ and instruments $2 - J$ binary. Proposition 4 shows that if $Z_1 \ldots Z_J$ satisfies VM then so does the set of $J + 1$ instruments $\tilde{Z}_2, \tilde{Z}_3, Z_2, \ldots Z_J$ where $\tilde{Z}_2 = \mathbb{1}(Z_1 \geq z_2)$ and $\tilde{Z}_3 = \mathbb{1}(Z_1 \geq z_3)$. In this case there are 2^{J-1} "redundant" simple compliance groups vis-a-vis Assumption 3, since for any $S \subseteq \{2 \ldots J\}$: $\tilde{Z}_{2i}\tilde{Z}_{3i}Z_{Si} = \tilde{Z}_{3i}Z_{Si}$.

In this example, the vector Γ_i would contain all non-null subsets of $\{\tilde{Z}_2, \tilde{Z}_3, Z_2, \ldots Z_J\}$ that do not contain both of $\tilde{Z}_2$ and $\tilde{Z}_3$. In general, $\mathcal{F}$ can be constructed by considering all subsets of the instruments, and for each subset considering all possible assignments of a value to each instrument, with one fixed value for each instrument omitted from consideration throughout. Provided rectangular support on the original instruments, Assumption 3* then follows with this choice of $\mathcal{F}$, for which a generalized version of Theorem 1 can be stated:

Theorem 1*. *The results of Theorem 1 holds under Assumption 3* replacing Assumption 3, where now $\Gamma_i := \{Z_{Si}\}_{S \in \mathcal{F}, S \neq \emptyset}$, $\lambda := \{E[c(S), Z_i)]\}_{S \in \mathcal{F}, S \neq \emptyset}$ and again $h(Z_i) = \lambda' \Sigma^{-1}(\Gamma_i - E[\Gamma_i])$ with $\Sigma := Var(\Gamma_i)$, for any family $\mathcal{F}$ satisfying Assumption 3*.*

Proof. Identical to that of Theorem 1, except as noted therein. □

Theorem 1* may also be useful in other cases in which the practitioner has auxiliary knowledge that some of the compliance groups are not present in the population. In such cases, Assumption 3* may hold even without rectangular support among the instruments.

B Identification with covariates

This section discusses how one can accommodate, in a nonparametric way, covariates that need to be conditioned on for the instruments to be valid. In practice, it is often easier to justify a conditional version of Assumption 1:

$$\{(Y_i(1), Y_i(0), G_i) \perp Z_i\} \,|X_i$$

where X are a set of observed covariates unaffected by treatment. In this section I discuss identification and considerations for estimation in such a setting. I maintain that vector monotonicity continues to hold for a set of binary instruments, as VM is expressed in Assumption 2. This implies that the direction of "compliance" is the same regardless of X_i, since the condition in Assumption 2 holds with probability one. If Assumption 3 and Property M each hold conditional on $X_i = x$, then Theorem 1 implies that we can identify $\Delta_c(x) := E[\Delta_i | C_i = 1, X_i = x]$ for Δ_c satisfying Property M, from the distribution of $(Y_i, Z_i, D_i)|X_i = x$. In particular, the function $h(z)$ from Theorem 1 will now depend on the conditioning value of X_i:

$$h(Z_i, x) = \lambda(x)' Var(\Gamma_i | X_i = x)^{-1} (\Gamma_i - E[\Gamma_i | X_i = x])$$

for each $x \in \mathbb{X}$, where recall that Γ_i is a vector of products Γ_{Si} of Z_{ji} within subsets of the instruments, where S indexes such subsets. Here we define $\lambda(x)_S = E[c(g(S), Z_i)|X_i = x]$ – which is identified – for each simple compliance group $g(S)$. Under these assumptions, we have that $\Delta_c(x) = E[h(Z_i, x)Y_i|X_i = x]/E[h(Z_i, x)D_i|X_i = x]$.

If the support of X_i corresponds to a small number of "covariate-cells", it might be feasible to repeat the entire estimation on fixed-covariate subsamples, to estimate $\Delta_c(x)$ for each $x \in \mathbb{X}$. If the number of groups is large, or if X_i includes continuous variables, estimation of $\Delta_c(x)$ could still in principle be implemented by nonparametric regression of each component of Γ_i on X_i as well as nonparametrically estimating the conditional variance-covariance matrix $Var(\Gamma_i|X_i = x)$ (Yin et al. (2010) describe a kernel-based method for this). The vector $\lambda(x)$ can also be computed via nonparametric regression.

Furthermore, when the object of interest is simply the unconditional version of Δ_c, the conditional quantities become nuisance parameters. Notably, they can be integrated over separately in the numerator and the denominator of the empirical estimand. To see that this, write:

$$\begin{aligned}
\Delta_c = E[\Delta_i | C_i = 1] &= \int dF_{X|C}(x|1)\Delta_c(x) \\
&= \int dF_{X|C}(x|1)\frac{E[h(Z_i, x)Y_i|X_i = x]}{E[h(Z_i, x)Y_i|X_i = x]} = \int dF_{X|C}(x|1)\frac{E[h(Z_i, x)Y_i|X_i = x]}{P(C_i = 1|X_i = x)} \\
&= \frac{1}{P(C_i = 1)} \int dF_X(x) E[h(Z_i, X_i)Y_i|X_i = x] = \frac{E[h(Z_i, X_i)Y_i]}{E[h(Z_i, X_i)D_i]}
\end{aligned}$$

where we have used Bayes' rule and that $P(C_i = 1|X_i = x) = E[h(Z_i, x)D_i|X_i = x]$ (and hence $P(C_i = 1) = E[h(Z_i, X_i)D_i]$ as well). This provides a VM analog to a similar result that holds under IAM. In that context, Frölich (2007) shows that this fact can deliver $\sqrt{n}$-consistency of a nonparametric analog of the Wald ratio.

Note that by the conditional version of Corollary 1 we have that:

$$\Delta_c = \frac{E[\lambda(X_i)' A \{E[Y_i|Z_i = z, X_i]\}]}{E[\lambda(X_i)' A \{E[D_i|Z_i = z, X_i]\}]}$$

where the indices of $\lambda_S(x)$ range over $S \subseteq \{1 \dots J\}, S \neq \emptyset$ and $\{\cdot\}$ indicate vector representations of functions over $z \in \mathcal{Z}$. If the CEFs of Y and D happen to both be

separable between Z and X, i.e $E[Y_i|Z_i = z, X_i = x] = y(z) + w(x)$ and $E[Y_i|Z_i = z, X_i = x] = d(z) + v(x)$, then the expression simplifies:

$$\Delta_c = \frac{E[\lambda(X_i)'A\{y(z)\} + w(X_i)\lambda(X_i)'A\mathbf{1}]}{E[\lambda(X_i)'A\{d(z)\} + v(X_i)\lambda(X_i)'A\mathbf{1}]} = \frac{E[\lambda(X_i)'A\{y(z)\}]}{E[\lambda(X_i)'A\{d(z)\}]}$$

where $\mathbf{1}$ is a vector of ones and we have used that $\lambda(x)'A\mathbf{1} = 0$ for any x. This follows from the definition of the entries: $A_{S,z} = \sum_{\substack{f \subseteq z_0 \\ (z_1 \cup f) = S}} (-1)^{|f|}$ where z_0 is the set of components of z that are equal to zero. For any $S \neq \emptyset$,the identity $\sum_{f \subseteq S}(-1)^{|f|} = 0$ implies that $[A\mathbf{1}]_S = \sum_{z_1 \subseteq S} \sum_{f \subseteq (S - z_1)} (-1)^{|f|} = 0$. The first component of $A\mathbf{1}$, corresponding to $S = \emptyset$, does not contribute since the first component of $\lambda(x)$ is always zero, by construction.

Now, since each $\lambda_S(x)$ is defined as $E[C_i = 1|G_i = g(S), X_i = x]$, its expectation delivers the unconditional analog: $\lambda_S := E[C_i = 1|G_i = g(S)] = E[\lambda(X_i)_S]$. Thus we can write $\Delta_c = \frac{\lambda' A\{y(z)\}}{\lambda' A\{d(z)\}}$. This shows that in this separable case the estimand that identifies Δ_c is essentially unchanged from the baseline case without covariates, aside from the need to control semi-parametrically for X_i to obtain the functions $y(z)$ and $d(z)$. The estimates reported in Figure 3 use this result, with $w(x)$ and $v(x)$ taken to be linear.

C Simulation study

This appendix reports a Monte Carlo experiment in which the regularized estimator proposed in Section 5 is compared against its unregularized version and 2SLS. I proceed in two steps. In a first simulation involving three binary instruments, I demonstrate the practical importance of regularization. A second simulation with two binary instruments highlights the potential dangers of using 2SLS.

C.1 Three instrument DGP

We first let $J = 3$, and put equal weight $P(G_i = g) = .05$ over each of the 20 compliance groups. To introduce endogeneity, I let $Y_i(0) = G_i \cdot U_i$ where the G_i are numbered arbitrarily from one to 20 and $U_i \sim Unif[0, 1]$. The treatment effect within each group g is chosen to be constant and equal to g, so that

$$Y_i(1) = Y_i(0) + G_i + V_i$$

with $V_i \sim Unif[0, 1]$. With this setup, $ACL = 10$.

For the joint distribution of the instruments, I consider two alternatives, meant to capture different extremes regarding statistical dependence among the instruments:

1. (Z_{1i}, Z_{2i}, Z_{3i}) generated as uncorrelated coin tosses
2. (1) followed by the following transformation: if $Z_{2i} = 1$ set $Z_{3i} = 0$ with probability 95%

I let the sample size be $n = 1000$, and perform one thousand simulations. Our primary goal is to compare the estimator $\hat{\rho}(1, 1, \ldots, 1, \alpha)$, where α chosen by the feasible approximate MSE minimizing procedure described in Section 5, to the simple Wald estimator of ACL $(\hat{E}[Y_i|Z_i = (111)] - \hat{E}[Y_i|Z_i = (000)])/(\hat{E}[D_i|Z_i = (111)] - \hat{E}[D_i|Z_i = (000)])$, which is equal to $\hat{\rho}(1 \ldots 1, \alpha = 0)$. I also benchmark both estimators against fully saturated 2SLS. I stress that 2SLS is not generally consistent for the ACL (or any convex combination of treatment effects) under vector monotonicity. Nevertheless, given the popularity of 2SLS and its desirable properties under traditional LATE monotonicity, it is important to know if and when the proposed estimator $\hat{\rho}(\lambda, \alpha)$ outperforms 2SLS in practice.

Figure 5 shows the results for the first DGP, where the Z_j are independent Bernoulli random variables with mean 1/2. We see that with the good overlap of the points $\bar{Z} = (1, 1, 1)$ and $\bar{Z} = (0, 0, 0)$ (which are each equal to 1/8), the Wald estimator performs well. For this DGP, the procedure to choose $\hat{\alpha}_{mse}$, minimizing MSE, results in small values with high probability. Hence the regularized estimator $\hat{\rho}((1, 1, \ldots 1)', \hat{\alpha}_{mse})$ according to Proposition 10 is very close to the Wald estimator (recall that they are identical when $\alpha = 0$). However, my estimator does deliver a slightly smaller RMSE, as expected, at the cost of some bias. Fully saturated 2SLS happens to also perform well for this DGP.

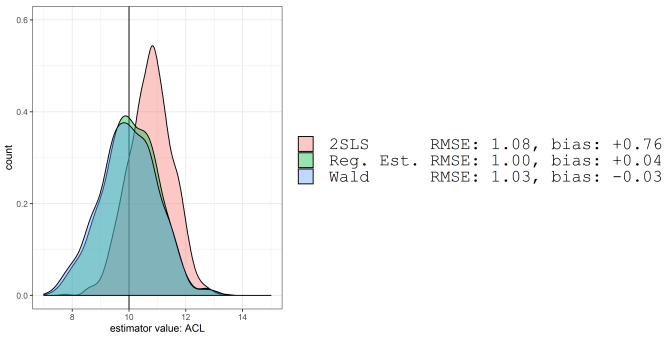

Figure 5: Monte Carlo distributions of estimators, for the first DGP (Z uncorrelated coin tosses) with three binary instruments. "Reg. Est." indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

Figure 6 shows the results for the second DGP, where I modify the joint distribution of (Z_1, Z_2, Z_3) to impose $E(Z_{3i}|Z_{2i} = 1) = 0.05$. In this case, the Wald estimator performs comparatively poorly. We see that regularizing the estimator to use the full sample rather than just the points $\bar{Z} = (1, 1, 1)$ and $\bar{Z} = (0, 0, 0)$ can help considerably.

C.2 Two instrument DGP

Note that in both Figures 5 and 6, fully saturated 2SLS (regression on the propensity score) performs well, in the latter case actually outperforming both of the alternative estimators. This is despite the fact that it is not consistent for the ACL, and is in general not even guaranteed to be consistent for Δ_c for any choice of the function $c(g, z)$. To demonstrate that 2SLS can in practice perform very poorly under vector monotonicity, I below report results from an additional simulation in which $J = 2$.

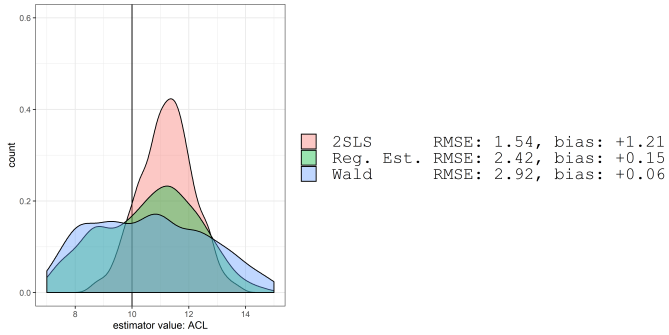

Figure 6: Monte Carlo distributions of estimators, for the first DGP ($P(Z_{3i}|Z_{2i}=1)=0.05$) with three binary instruments. "Reg. Est." indicates $\hat{\rho}(1,\ldots,1,\hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

For this simulation, the DGP is as follows. Among the six possible compliance groups under vector monotonicity, I give units a 90% chance of being Z_1 complier and a 10% chance of Z_2 complier. The treatment effect is set to 2 for Z_1 compliers, and -8 for Z_2 compliers, resulting in a ACL of unity. I generate negatively correlated binary instruments (with correlation of about $-.1$) from a multivariate normal. In particular, with

$$\begin{pmatrix} Z_1^* \\ Z_2^* \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & -.8 \\ -.8 & 1 \end{bmatrix}\right]$$

I set $Z_{1i}=1$ when Z_{1i}^* is over its median and $Z_{2i}=1$ when Z_{2i}^* is over its median. I again let the sample size be $n=1000$, and perform a thousand simulations.

Figure 7 shows that in this case, 2SLS is indeed outside of the convex hull of treatment effects, despite having high precision. The proposed regularized estimator clearly outperforms both of the alternatives for this DGP.

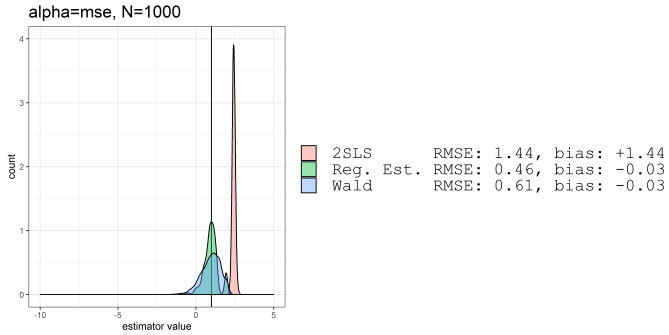

Figure 7: Monte Carlo distributions of estimators, for the two-instrument DGP. "Reg. Est." indicates $\hat{\rho}(1,\ldots,1,\hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

D Proofs

This section provides proofs for the formal results presented in the body of the paper.

D.1 Proof of Proposition 1

To simplify notation take each ordering $\geq_j$ to be the ordering on the natural numbers $\geq$, without loss. The two versions of VM are:

Assumption VM (vector monotonicity). *For $z, z' \in \mathcal{Z}$, if $z \geq z'$ component-wise, then $P(D_i(z) \geq D_i(z')) = 1$*

Assumption VM' (alternative characterization). $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$ *when $z_j \geq z'_j$ and both (z_j, z_{-j}) and $(z'_j, z_{-j}) \in \mathcal{Z}$*

The claim is that $VM \iff VM'$.

- **VM** $\implies$ **VM'** : immediate, since $(z_j, z_{-j}) \geq (z'_j, z_{-j})$ in a vector sense when $z_j \geq z'_j$
- **VM'** $\implies$ **VM** : consider $z, z' \in \mathcal{Z}$ such that $z \geq z'$ in a vector sense, i.e. $z_j \geq z'_j$ for all $j \in \{1 \dots J\}$. Then by VM' and connectedness of $\mathcal{Z}$, then for some ordering of the instrument labels $1 \dots J$:

$$P\left(D_i\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i\begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix}\right) = 1 \quad P\left(D_i\begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i\begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z_J \end{pmatrix}\right) = 1 \quad etc\dots$$

and thus:

$$P\left(D_i\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i\begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i\begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z_J \end{pmatrix} \geq \dots \geq D_i\begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_J \end{pmatrix}\right) = 1$$

D.2 Proof of Proposition 2

Let $P(z) := E[D_i | Z_i = z]$ be the propensity score function. By the law of iterated expectations and Assumption 1:

$$P(z) = \sum_{g \in \mathcal{G}} P(G_i = g | Z_i = z) E[D_i(Z_i) | G_i = g, Z_i = z] = \sum_{g \in \mathcal{G}} P(G_i = g) D_g(z)$$

By VM, $D_g(z)$ is component-wise monotonic for any g in the support of G_i. As a convex combination of component-wise monotonic functions, $P(z)$ will thus also be component-wise monotonic.

In the other direction, note that by PM if $P(z_j, z_{-j}) > P(z'_j, z_{-j})$, then we must have that $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$. Thus component-wise monotonicity of $P(z)$ with

respect to some collection of orderings $\{\geq_j\}_{j\in\{1\dots J\}}$ implies $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$ for all choices of $j \in \{1\dots J\}$, $z_j \geq_j z'_j$ and $z_{-j} \in \mathcal{Z}_{-j}$. This is the equivalent form of VM stated in Proposition 1.

D.3 Proof of Proposition 4

Let $\tilde{\mathcal{Z}}$ be the set of possible values for the new set of instruments $(\tilde{Z}_2, \dots \tilde{Z}_m, Z_{-1})$. Since $P(\tilde{Z}_{mi} = 0 \,\&\, \tilde{Z}_{ni} = 1) = 0$ for any $m > n$, we can take $\tilde{\mathcal{Z}}$ to only consist of cases where for all m: $\tilde{Z}_{-m}$ is composed of all zeros for the first $m-1$ entries, and then ones for $m+1\dots M$. Note that fixing Z_1 is equivalent to fixing $\tilde{Z}_2 \dots \tilde{Z}_M$.

If $\mathcal{Z}$ is connected, then the $\tilde{\mathcal{Z}}$ given above is connected. Then, by Proposition 1, we simply need to show that for any $Z_{-1} = (Z_2, \dots, Z_J)$ and $\tilde{Z}_{-m} = (\tilde{Z}_2, \dots, \tilde{Z}_m, \tilde{Z}_{m+1}, \dots, \tilde{Z}_M)$ such that $(0, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$ and $(1, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$:

$$D_i(1, \tilde{Z}_{-m}; Z_{-1}) \geq D_i(0, \tilde{Z}_{-m}; Z_{-1})$$

where the notation $D_i(a, b; c)$ is understood as $D_i(d, c)$ where d is the value of Z_1 corresponding to $\tilde{Z}$ with value a for $\tilde{Z}_m$ and b for $\tilde{Z}_{-m}$. For any $\tilde{Z}_{-m}$ satisfying $(0, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$ and $(1, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$, switching $\tilde{Z}_m$ from zero to ones corresponds to switching Z_1 from value z_{m-1} to value z_m. Since

$$D_i(1, \tilde{Z}_{-m}; Z_{-1}) - D_i(0, \tilde{Z}_{-m}; Z_{-1}) = D_i(z_m, Z_{-1}) - D_i(z_{m-1}, Z_{-1}) \geq 0$$

by vector monotonicity on the original vector $(Z_1 \dots Z_J)$, the result now follows.

D.4 Proof of Proposition 3

For any fixed z, write the condition $D_{g(F)}(z) = 1$ as

$$\left\{D_{g(F)}(z) = 1\right\} \iff \left\{\bigcup_{S\in F} \left\{D_{g(S)}(z) = 1\right\}\right\} \iff \text{not } \left\{\bigcap_{S\in F} \left\{D_{g(S)}(z) = 0\right\}\right\}$$

which can be written as

$$D_g(z) = 1 - \prod_{S\in F} \left(1 - D_{g(S)}(z)\right) = \sum_{f\subseteq F: f\neq\emptyset} (-1)^{|f|+1} \prod_{S\in F} D_{g(S)}(z)$$

Let $\mathbf{z}(z) = \{j \in \{1\dots J\} : z_j = 1\}$ represent z as the subset of instrument indices for which the associated instrument takes the value of one. Then, using that for a simple

compliance group $D_{g(S)}(z) = \mathbb{1}(S \subseteq \mathbf{z}(z))$:

$$\begin{aligned}
D_g(z) &= \sum_{f \subseteq F: f \neq \emptyset} (-1)^{|f|+1} \prod_{s \in F} D_{g(S)}(z) \\
&= \sum_{f \subseteq F: f \neq \emptyset} (-1)^{|f|+1} \cdot D_{g\left(\left(\bigcup_{S \in f} S\right)\right)}(z) \\
&= \sum_{f \subseteq F: f \neq \emptyset} (-1)^{|f|+1} \cdot \mathbb{1}\left(\left(\bigcup_{S \in f} S\right) \subseteq \mathbf{z}(z)\right) \\
&= \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) \subseteq \mathbf{z}(z)}} (-1)^{|f|+1} = \sum_{S' \subseteq \mathbf{z}(z)} \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \\
&= \sum_{S' \subseteq \{1 \ldots J\}} \mathbb{1}\left(S' \subseteq \mathbf{z}(z)\right) \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \\
&= \sum_{S' \subseteq \{1 \ldots J\}} \left[\sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] D_{g(S')}(z) = \sum_{\emptyset \subset S' \subseteq \{1 \ldots J\}} \left[\sum_{\substack{f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] D_{g(S')}(z)
\end{aligned}$$

Thus, letting $s(F, S') := \left\{ f \subseteq F : \left(\bigcup_{S \in f} S\right) = S' \right\}$, we have $D_{g(F)}(z) = \sum_{S'} [M_J]_{F,S'} D_{g(S)}(z)$, where the sum ranges over non-null subsets of the instruments $\emptyset \subset S' \subseteq \{1 \ldots J\}$ and $[M_J]_{F,S'} = \sum_{f \in s(F,S')} (-1)^{|f|+1}$.

D.5 Proof of Proposition 5

D.5.1 VM case

The if direction is most straightforward. From Proposition 3 we have that for any $z \in \mathcal{Z}$ and $g \in \mathcal{G}^c$:

$$D_g(z) = \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot D_{g(S)}(z)$$

Thus, for any such $c(g, z)$:

$$\begin{aligned}
c(g,z) &= \sum_{k=1}^{K} \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot D_{g(S)}(h_k(z))) - \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot D_{g(S)}(l_k(z))) \\
&= \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot \left\{ \sum_{k=1}^{K} D_{g(S)}(h_k(z))) - D_{g(S)}(l_k(z))) \right\} \\
&= \sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot c(g(S), z)
\end{aligned}$$

for any $z \in \mathcal{Z}$. To finish verifying Property M, we need only observe that $c(a.t., z) = c(n.t., z) = 0$ for all z since $D_g(h_k(z)) = D_g(l_k(z))$ for any h_k, l_k when $g \in \{a.t., n.t.\}$.

Now we turn to the other implication of the Proposition, that any c satisfying Property M has a representation like the above. For shorthand, let $c^{-1}(z)$ indicate the family of $S \subseteq \{1 \dots J\}$ such that $c(g(S), z) = 1$. The following Lemma establishes that the family $c^{-1}(z)$ and its complement are each closed under unions:

Lemma. *Let c be a function from $\mathcal{G} \times \mathcal{Z}$ to $\{0, 1\}$ satisfies Property M. If $A \in c^{-1}(z)$ and $B \in c^{-1}(z)$, then $A \cup B \in c^{-1}(z)$, and if $A \notin c^{-1}(z)$ and $B \notin c^{-1}(z)$, then $A \cup B \notin c^{-1}(z)$.*

Proof. If the sets A and B are nested, then the result follows trivially. Now suppose neither set contains the other, and consider the Sperner family $A \sqcup B$ constructed of the two sets A and B. By Property M and using Proposition 3:

$$\begin{aligned} c(g(A \sqcup B), z) &= \sum_{\emptyset \subset S' \subseteq \{1 \dots J\}} \left[\sum_{\substack{f \subseteq \{A, B\}: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] c\left(\bigcup_{S \in f} S, z \right) \\ &= \sum_{\emptyset \subset f \subseteq \{A, B\}} c\left(\bigcup_{S \in f} S, z \right) \\ &= c(g(A), z) + c(g(B), z) - c(g(A \cup B), z) \end{aligned}$$

In the first case, if both A and B are in $c^{-1}(z)$, then we must have $c(g(A \cup B), z) = 1$ to prevent $c(g(A \sqcup B), z)$ from evaluating to 2, which contradicts the assumption that c takes values in $\{0, 1\}$. In the second case, when both $c(g(A), z)$ and $c(g(B), z)$ are zero, we must have $c(g(A \cup B), z) = 1$ to prevent $c(g(A \sqcup B), z)$ from evaluating to -1. □

As a consequence of the Lemma, since $c^{-1}(z)$ is a finite set, there exists a member $S_1(z)$ of $c^{-1}(z)$ that satisfies $S_1(z) = \bigcup_{S \in c^{-1}(z)} S$ (similarly, there exists a $S_0(z) = \bigcup_{S \notin c^{-1}(z)} S$ with $S_0(z) \notin c^{-1}(z)$). All members of the family $c^{-1}(z)$ are subsets of $S_1(z)$, and all $S \subseteq \{1 \dots J\}$ that are not in $c^{-1}(z)$ are subsets of $S_0(z)$.

Let z take some fixed value, and beginning with the set $S_1 = S_1(z)$, define a sequence of sets $\{S_1, S_2, S_3, \dots\}$ as follows:

$$S_{2k} = \bigcup_{\substack{S' \subseteq S_{2k-1}: \\ S' \notin c^{-1}(z)}} S' \qquad \text{and} \qquad S_{2k+1} = \bigcup_{\substack{S' \subseteq S_{2k}: \\ S' \in c^{-1}(z)}} S'$$

where we take $\bigcup_{S' \in \emptyset} S'$ to evaluate to the empty set. This sequence provides a charac-

terization of the family $c^{-1}(z)$ as follows. For any $\emptyset \subset S \subseteq \{1 \dots J\}$:

$$\begin{aligned}
c(g(S), z) &= \mathbb{1}(S \in c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1 : S \in c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1) - \mathbb{1}(S \subseteq S_1 : S \notin c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1) - \big(\mathbb{1}(S \subseteq S_2) - \mathbb{1}(S \subseteq S_2 : S \in c^{-1}(z))\big) \\
&= \mathbb{1}(S \subseteq S_1) - \mathbb{1}(S \subseteq S_2) + \big(\mathbb{1}(S \subseteq S_3) - \mathbb{1}(S \subseteq S_3 : S \notin c^{-1}(z))\big) \\
&= \dots \\
&= \sum_{n=1}^{N} (-1)^{n+1} \cdot \mathbb{1}(S \subseteq S_n) + (-1)^N \cdot \begin{cases} \mathbb{1}(S \subseteq S_N : S \in c^{-1}(z)) & \text{if } N \text{ even} \\ \mathbb{1}(S \subseteq S_N : S \notin c^{-1}(z)) & \text{if } N \text{ odd} \end{cases}
\end{aligned}$$

for any natural number N.

Think of the power set of S_1 as a "first-order" approximation to the family $c^{-1}(z)$. However, in most cases this family is too large, as there will be subsets of S_1 that are not found in $c^{-1}(z)$. Define S_2 to be the union of all such offending sets. The power set of S_2 now provides a possible "overestimate" of the family of offending sets (since they are all in 2^{S_2}) and hence removing all subsets of S_2 as a correction to be applied to 2^{S_1} as an estimate of $c^{-2}(z)$ will overcompensate: we will have removed some sets which are indeed in $c^{-1}(z)$. We thus define S_3 analagously, whose power set provides an approximation to the error in S_2 as an approximation to the error in S_1, and so on.

Does this process of over-correction eventually terminate, so that the final remainder term is zero? Note that for any n: $S_n \subseteq S_{n-1}$. If $S_n = S_{n-1} \neq \emptyset$, then we have a fixed point S where $\bigcup_{S' \subseteq S : S' \in c^{-1}(z)} S' = \bigcup_{S' \subseteq S : S' \notin c^{-1}(z)} S'$. But by the Lemma, this would imply that S is a member both of $\{S' \subseteq S : S' \in c^{-1}(z)\}$ and of $\{S' \subseteq S : S' \notin c^{-1}(z)\}$, and therefore that both $c(g(S), z) = 1$ and $c(g(S), z) = 0$, a contradiction. Thus, $S_n \subset S_{n-1}$, and $|S_n|$ is a decreasing sequence of non-negative integers that is strictly decreasing so long as $|S_n| > 0$. It must thus converge to zero in at most $|S_1|$ iterations, so that $S_n = \emptyset$ for all $n \geq |S_1|$.

Without loss, we can terminate the sequence on an even term, since $\mathbb{1}(S \subseteq \emptyset) = 0$ for any $S \supset \emptyset$. Let $2K$ denote the smallest even number such that $S_n = \emptyset$ for all $n > 2K$, for a fixed z. Thus, we have for any $\emptyset \subset S \subseteq \{1 \dots J\}$:

$$c(g(S), z) = \sum_{n=1}^{2K} (-1)^{n+1} \cdot D_{g(S)}(S_n) = \sum_{k=1}^{K} D_{g(S)}(S_{2k-1}) - D_{g(S)}(S_{2k})$$

where $2K \leq |S_1| \leq J$, and we have used that $D_{g(S)}(S') = \mathbb{1}(S \subset S')$ for any S'.

Now recall that we have left the dependence of each of the sets S_n (as well as the integer K) on z implicit, and have also adopted the notational convention of $D_g(S)$ as a shorthand for $D_g(z)$ where z is a point in $\mathcal{Z}$ that takes a value of one for exactly the instruments in the set S. To obtain the notation of the final result, define for each $k = 1 \dots K$ the point $u_k(z) \in \mathcal{Z}$ to have a value of one exactly for the elements in S_{2k-1}

for that value of z, and $l_k(z) \in \mathcal{Z}$ to have a value of one exactly for the elements in S_{2k} for that value of z. We may thus write, for any $\emptyset \subset S \subseteq \{1 \dots J\}$ and any $z \in \mathcal{Z}$:

$$c(g(S), z) = \sum_{k=1}^{K(z)} D_{g(S)}(u_k(z)) - D_{g(S)}(l_k(z)) = \sum_{k=1}^{K} D_{g(S)}(u_k(z)) - D_{g(S)}(l_k(z))$$

where we let K be the maximum of $K(z)$ over the finite set $\mathcal{Z}$, and we define $u_k(z)$ and $l_k(z)$ to each be a vector of zeros whenever $k > K(z)$. For each z, the relations $u_k(z) \succeq l_k(z)$ and $l_k(z) \succeq u_{k+1}(z)$ now follow from $S_n \subseteq S_{n+1}$.

Now we may apply Property M to construct $c(g, z)$ for any of the non-simple compliance groups as well. Recall that Property M says that $c(g(F), z) = \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot c(g(S), z)$ for all z, for any Sperner family F. Thus:

$$\begin{aligned} c(g(F), z) &= \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot \sum_{k=1}^{K} \{D_{g(S)}(u_k(z)) - D_{g(S)}(l_k(z))\} \\ &= \sum_{k=1}^{K} \{ \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot D_{g(S)}(u_k(z))\} - \{ \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot D_{g(S)}(l_k(z))\} \\ &= \sum_{k=1}^{K} D_{g(F)}(u_k(z)) - D_{g(F)}(l_k(z)) \end{aligned}$$

Finally, note that $D_g(u_k(z)) = D_g(l_k(z))$ for any $g \in \{a.t., n.t.\}$ so the following expression holds for all $g \in \mathcal{G}$:

$$c(g, z) = \sum_{k=1}^{K} D_g(u_k(z)) - D_g(l_k(z))$$

D.5.2 IAM case

Now I prove that representation from Proposition 5 also holds under IAM. Note that under IAM Property M places no restriction beyond $c(a.t., z) = c(n.t., z) = 0$ since there is no perfect linear dependency between the functions $D_g(z)$ to worry about. Under IAM, each $g \in \mathcal{G}^c$ can be associated with an integer $m = \in \{1, 2 \dots 2^J - 1\}$ and characterized directly as $\mathbb{1}(g = m) = D_g(z_{m+1}) - D_g(z'_m)$, where $z_1, z_2, \dots, z_{2^J}$ is any fixed ordering of the points that is weakly increasing according to the propensity score $E[D_i | Z_i = z_m]$. Thus, for any function $g : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ such that $c(a.t., z) = c(n.t., z) = 0$:

$$\begin{aligned} c(g, z) &= \sum_{m=1}^{2^J - 1} c(m, z) \cdot (D_g(z_{m+1}) - D_g(z'_m)) \\ &= \sum_{k=1}^{K} D_g(u_k(z)) - D_g(l_k(z)) \end{aligned}$$

with $K = 2^J - 1$ where for each z we let $l_k(z) = z_m$ and we let $u_k(z) = \begin{cases} z_k & \text{if } c(k, z) = 0 \\ z_{k+1} & \text{if } c(k, z) = 1 \end{cases}$ .Note that if any set of consecutive $c(k, z) = c(k + 1, z) \dots c(k + T, z)$ are all equal to

one, then one can drop $T-1$ of these terms as the inner terms will all cancel leaving $D_g(u_{k+T}(z)) - D_g(l_k(z))$. Thus we may take without loss $K \leq 2^J/2 = 2^{J-1}$ (corresponding to the case where $c(1,z)=1$, $c(2,z)=0$, $c(3,z)=1$ etc.).

D.6 Proof of Lemma 2

Any indicator $\mathbb{1}(Z_i = z)$ for a value $z \in \{0,1\}^J$ can be expanded out as a polynomial in the instrument indicators as $\mathbb{1}(Z_i = z) = \prod_{j \in z_1} Z_{ji} \prod_{j \in z_0}(1 - Z_{ji}) = \sum_{f \subseteq z_0}(-1)^{|f|} Z_{(z_1 \cup f),i}$, where (z_1, z_0) is a partition of the indices $j \in \{1 \dots J\}$ that take a value of zero or one in z, respectively. With $J=2$ for example,

$$((1-Z_{1i})(1-Z_{2i}), Z_{1i}(1-Z_{2i}), Z_{2i}(1-Z_{1i}), Z_{1i}Z_{2i}) = (1, Z_{1i}, Z_{2i}, Z_{1i}Z_{2i})A = (1, \Gamma_i')A$$

where $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$. Denote the random vector of such indicators $\mathfrak{Z}_i$. Then $(1, \Gamma_i')A = \mathfrak{Z}_i'$, with the matrix of coefficients $A_{S,z} = \sum_{\substack{f \subseteq z_0 \\ (z_1 \cup f) = S}} (-1)^{|f|}$. The matrix A so defined must be invertible, because any product of the instruments Z_{Si} for $S \subseteq \{1 \dots J\}$ can similarly be expressed as a linear combination of the components of $\mathfrak{Z}_i$, where we define $Z_{\emptyset i} = 1$. Specifically, $Z_{Si} = \sum_{z \in \mathcal{Z}} \mathbb{1}(\forall_{j \in S}, z_j = 1)\, \mathbb{1}(Z_i = z)$.

Consider the matrix

$$\Sigma^* := E[(1, \Gamma_i')'(1, \Gamma_i')] = A'^{-1} E[\mathfrak{Z}_i \mathfrak{Z}_i'] A^{-1} = A'^{-1} diag\{P(Z_i = z)\} A^{-1}$$

where $E[\mathfrak{Z}_i \mathfrak{Z}_i']$ is diagonal since the events that Z_i take on two different values are exclusive. Since A^{-1} exists, the rank of Σ^* must be equal to the rank of $diag\{P(Z_i = z)\}$, which is in turn equal to the cardinality of $\mathcal{Z}$. Assumption 3 thus holds if and only if Σ^* has full rank of 2^J. Note that although A^{-1} diagonalizes the matrix Σ^*, it does not provide its eigen-decomposition, as $A^{-1} = A'$ (A is not orthogonal).

Now we prove that Σ^* has full rank whenever Σ has full rank, and vice versa. Note that $\Sigma = Var(\Gamma_i)$ has full rank if and only if $\omega' E[(\Gamma_i - E\Gamma_i)(\Gamma_i - E\Gamma_i)]\omega = E[\omega'(\Gamma_i - E\Gamma_i)(\Gamma_i - E\Gamma_i)\omega] > 0$, i.e. $P(\omega'(\Gamma_i - E\Gamma_i) = 0) < 1$ for any $\omega \in \mathbb{R}^{2^J - 1}/\mathbf{0}$. Similarly Σ^* has full rank if $P((\omega_0, \omega)'((1, \Gamma_i) = 0) < 1$ for any $\omega_0 \in \mathbb{R}, \omega \in \mathbb{R}^{2^J-1}$ where (ω_0, ω) is not the zero vector in $\mathbb{R}^{2^J}$. But if for some ω, $\omega'(\Gamma_i - E\Gamma_i) = 0$ w.p.1., then we also have $(\omega_0, \omega)'(1, \Gamma_i) = 0$ w.p.1. by choosing $\omega_0 = -\omega' E[\Gamma_i]$. In the other direction, note that $(\omega_0, \omega)'(1, \Gamma_i) = 0$ w.p.1. implies that $\omega'\Gamma_i = -\omega_0$ and hence $\omega'(\Gamma_i - E\Gamma_i) = -\omega_0 - \omega' E\Gamma_i = -\omega_0 - E[\omega'\Gamma_i] = -\omega_0 + \omega_0 = 0$.

D.7 Proof of the Appendix A Proposition

Introduce the notation that $\sqcup$ indicates inclusion of a new set among a family of sets (while $\cup$ continues to indicate taking the union of elements across sets).

For any $S \subseteq \mathcal{M}$ that contains both Z_m^j and $Z_{m'}^j$ for some j and $m < m'$, $g(F \sqcup S)$ and $g(F \sqcup S/\{Z_m^j\})$ generate the same selection behavior for any Sperner family F on all of $\mathcal{Z}$ (this can be seen by mapping the implied selection behavior back to the original discrete instrument Z_j). Thus, we can take $\mathcal{G}$ to exclude such S without loss of generality.

Now, consider the family $\mathcal{F}$ of all $S \subset \mathcal{M}$ that contain at most one Z_m^j for any given j. By the above, this choice of $\mathcal{F}$ satisfies Assumption 3b*. Suppose it did not satisfy Assumption 3a*. Then, there would need to exist a non-zero vector ω such that $P\left(\sum_{S\in\mathcal{F}} \omega_S Z_{Si} = 0\right) = 1$ with $Z_{Si} := \prod_{(j,m)\in S} \tilde{Z}_m^j$. This would imply non-invertibility of $\Sigma^* := E[(1,\Gamma_i)(1,\Gamma_i)']$, where $\Gamma_i := \{Z_{Si}\}_{S\in\mathcal{F},S\neq\emptyset}$ by the same argument as in the proof of Lemma 2 (Γ_i and a vector of indicators for all $z \in \mathcal{Z}$ are each related by an invertible linear map), which in turn contradicts the assumption of full support. Note that invertibility of Σ^* is again equivalent to invertibility of $Var(\Gamma_i)$ as before.

D.8 Proof of Theorem 1

We first note that any measurable function $f(Y)$ preserves Assumption 1, that is

$$(f(Y_i(1)), f(Y_i(0)), G_i) \perp Z_i$$

and Assumptions 2-3 are unaffected by such a transformation to the outcome variable. Thus, we continue without loss with Y_i, $Y_i(1)$ and $Y_i(0)$ possibly redefined as $f(Y_i)$, $f(Y_i(1))$ and $f(Y_i(0))$ respectively.

Note that the function $h(\cdot)$ given in Theorem 1 has the property that $E[h(Z_i)] = 0$, for any distribution of the instruments. Consider the quantity $E[Y_i D_i h(Z_i)]$ for a function h having this property. By the law of iterated expectations, and the independence assumption:

$$\begin{aligned}
E[Y_i D_i h(Z_i)] &= \sum_g P(G_i = g) E[Y_i D_i h(Z_i) | G_i = g] \\
&= \sum_g P(G_i = g) E[Y_i(1) D_g(Z_i) h(Z_i) | G_i = g] \\
&= \sum_g P(G_i = g) E[Y_i(1) | G_i = g] E[D_g(Z_i) h(Z_i)] \qquad (10)
\end{aligned}$$

where $D_g(z)$ denotes the selection function for compliance group g. Similarly,

$$\begin{aligned}
E[Y_i(1 - D_i) h(Z_i)] &= \sum_g P(G_i = g) E[Y_i(0)(1 - D_i) h(Z_i) | G_i = g] \\
&= \sum_g P(G_i = g) \{E[Y_i(0) | G_i = g] E[h(Z_i)] \\
&\qquad - E[Y_i(0) | G_i = g] E[D_g(Z_i) h(Z_i)]\} \\
&= \sum_g -P(G_i = g) E[Y_i(0) | G_i = g] E[D_g(Z_i) h(Z_i)] \qquad (11)
\end{aligned}$$

where we have used that $Z_i \perp (Y_i(0), Z_i)$ and $E[h(Z_i)] = 0$.

Combining these two results:

$$E[Y_i h(Z_i)] = E[Y_i D_i h(Z_i)] + E[Y_i(1-D_i)h(Z_i)] = \sum_g P(G_i = g) E[D_g(Z_i)h(Z_i)]\Delta_g \quad (12)$$

where $\Delta_g := E[Y_i(1) - Y_i(0)|G_i = g]$. By the law of iterated expectations, we also have that

$$E[D_i h(Z_i)] = \sum_g P(G_i = g) E[D_g(Z_i)h(Z_i)] \quad (13)$$

Note that in all of Eqs (10), (11) and (12), the weighing over various groups g is governed by the quantity $E[D_g(Z_i)h(Z_i)]$. It can be seen that never takers and always takers receive no weight, since $E[D_{n.t}(Z_i)h(Z_i)] = E[0] = 0$ and since $E[D_{a.t}(Z_i)h(Z_i)] = E[h(Z_i)] = 0$.

Let $\mathcal{F}$ denote the set of non-empty subsets of the instrument indices: $\mathcal{F} := \{S \subseteq \{1, 2, \dots J\}, S \neq \emptyset\}$, and recall that these correspond each to a simple compliance group $g(S)$, where $D_{g(S)}(Z_i) = Z_{Si}$. I first show that for any $\lambda \in \mathbb{R}^{|\mathcal{F}|}$, Assumption 3 allows us to define an $h(Z_i)$ such that $E[D_{g(S)}(Z_i)h(Z_i)] = E[Z_{Si}h(Z_i)] = \lambda_S$. Note that since $E[h(Z_i)] = 0$, this is the same as tuning each covariance $Cov(Z_{Si}, h(Z_i))$ to λ_S (c.f. Lemma 1). In particular, consider the choice $h(Z_i) = (\Gamma_i - E[\Gamma_i])'\Sigma^{-1}\lambda$, where recall that Γ_i is a vector of Z_{Si} for each $S \in \mathcal{F}$.

$$\begin{aligned}(E[h(Z_i)_i, \Gamma_{i1}], E[h(Z_i), \Gamma_{i2}], \dots, E[h(Z_i), \Gamma_{ik}])' &= E[(\Gamma_i - E[\Gamma_i])h(Z_i)] \\ &= E[(\Gamma_i - E[\Gamma_i])(\Gamma_i - E[\Gamma_i])']\Sigma^{-1}\lambda \\ &= \Sigma\Sigma^{-1}\lambda = \lambda\end{aligned}$$

We can understand the algebra of this result as follows. Let $V = span(\{Z_{Si} - E[Z_{Si}]\}_{S\in\mathcal{F}})$. V is a subspace of the vector space $\mathcal{V}$ of random variables on $\mathcal{Z}$, with the zero vector being a degenerate random variable equal to zero. Since the matrix Σ is positive semidefinite by construction, Assumption 3 is equivalent to the statement that for all $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$, $\omega' E[(\Gamma_i - E[\Gamma_i])(\Gamma_i - E[\Gamma_i])']\omega = E[|\omega'(\Gamma_i - E[\Gamma_i])|^2] > 0$: i.e. $P\left(\sum_{S\in\mathcal{F}} \omega_S(Z_{Si} - E[Z_{Si}])) = 0\right) < 1$ for all $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$. In other words, the random variables $(Z_{Si} - E[Z_{Si}])$ for $S \in \mathcal{F}$ are linearly independent, and hence form a basis of V. Since V is finite dimensional, there exists an orthonormal basis of random vectors of the same cardinality, $|\mathcal{F}|$, where orthonormality is defined with respect to the expectation inner product: $\langle A, B\rangle := E[A_i B_i]$. It is this orthogonalized version of the Z_{Si} that affords the ability to separately tune each of the $E[h(Z_i)Z_{Si}]$ to the desired value λ_S, without disrupting the others.

Expanding out the denominator of Eq (2) for parameters Δ_c:

$$\Delta_c = \sum_{g\in\mathcal{G}} \left\{ \frac{P(G_i = g)P(C_i = 1|G_i = g)}{P(C_i = 1)} \right\} \cdot \Delta_g = \frac{\sum_{g\in\mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g) \cdot \Delta_g}{\sum_{g\in\mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g)}$$

Comparing with Eqs (12) and (13), the equality $\Delta_c = E[h(Z_i)Y_i]/E[D_i h(Z_i)]$ follows (provided that $P(C_i = 1) > 0$) if the coefficients match. That is: $E[D_g(Z_i)h(Z_i)] =$

$P(C_i = 1|G_i = g)$, for all $g \in \mathcal{G}^c$. By the above, this is guaranteed under Property M if we choose $\lambda_S = P(C_i = 1|G_i = g(S)) = E[c(g(S), Z_i)]$, since the quantity $E[D_g(Z_i)h(Z_i)]$ appearing in Eq (12) is linear in $D_g(Z_i)$. The same logic follows for causal parameters of the form $E[Y_i(d)|C_i = 1]$ for $d \in \{0, 1\}$, using Eqs (10) and (11), and an analagous expression to Eq (2) i.e.

$$\begin{aligned} E[Y_i(d)|C_i = 1] &= \sum_{g \in \mathcal{G}} P(G_i = g|C_i = 1)E[Y_i(d)|G_i = g, c(g, Z_i) = 1] \\ &= P(C_i = 1)^{-1} \sum_{g \in \mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g)E[Y_i(d)|G_i = g] \end{aligned}$$

by independence. Note that the quantity λ_S for each S can be computed from the observed distribution of Z.

To replace Assumption 3 with Assumption 3* from Appendix A, simply replace $\mathcal{F}$ as defined here with a maximal $\mathcal{F}$ from Assumption 3a*.

D.9 Proof of Corollary 1 to Theorem 1

The proof of Lemma 2 shows that $(1, \Gamma_i')A$ is a vector of indicators $\mathfrak{Z}_i'$ for values of Z, where A is the matrix with entries given in Corollary 1, which is invertible, and $\mathfrak{Z}_i$ is a vector of indicators $\mathbb{1}(Z_i = z)$ for each of the values $z \in \mathcal{Z}$. We can thus write $h(Z_i)$ from Theorem 1 as

$$\begin{aligned} h(Z_i) &= \lambda'\Sigma^{-1}(\Gamma_i - E[\Gamma_i]) = (0, \lambda')E[(1, \Gamma_i')'(1, \Gamma_i')]^{-1}(1, \Gamma_i')' \\ &= (0, \lambda')E[A'^{-1}A'(1, \Gamma_i')'(1, \Gamma_i')AA^{-1}]^{-1}A'^{-1}\mathfrak{Z}_i \\ &= (0, \lambda')AE[\mathfrak{Z}_i\mathfrak{Z}_i']\mathfrak{Z}_i \end{aligned}$$

This is useful because $E[\mathfrak{Z}_i\mathfrak{Z}_i']$ is diagonal, since the events that Z_i take on two different values are exclusive: $E[\mathfrak{Z}_i\mathfrak{Z}_i'] = diag\{P(Z_i = z)\}_{z \in \mathcal{Z}}$.

Now, for $V \in \{Y, D\}$, $E[h(z)V_i] = (0, \lambda')Adiag\{P(Z_i = z)\}_{z \in \mathcal{Z}}^{-1}\{E[\mathbb{1}(Z_i = z)V_i]\}_{z \in \mathcal{Z}} = (0, \lambda')A\{E[V_i|Z_i = z]\}_{z \in \mathcal{Z}}$. Thus $(0, \lambda')A$ describes the coefficients in an expansion of $E[h(z)V_i]$ into CEFs of V_i across the support of Z_i.

D.10 Proof of Corollary 2 to Theorem 1

Using independence and Property M:

$$\begin{aligned}
E[h(Z_i)D_i] &= \sum_g P(G_i = g)E[h(Z_i)D_g(Z_i)] \\
&= \sum_g P(G_i = g)E\left[h(Z_i)\left\{\sum_S [M_J]_{F(g),S} D_{g(s)}(Z_i)\right\}\right]) \\
&= \sum_g P(G_i = g)\sum_S [M_J]_{F(g),S} P(C_i = 1|D_{g(s)}(Z_i)) \\
&= \sum_g P(G_i = g)P(C_i = 1|G_i = g) \\
&= P(C_i = 1)
\end{aligned}$$

D.11 An Equivalence Result

The proofs of Proposition 6 and 8 will make use of the following equivalence result:

Proposition 11. *Let the support $\mathcal{Z}$ of the instruments be discrete and finite. Fix a function $c(g, z)$. Let $\mathcal{P}_{DZ}$ denote the joint distribution of D_i and Z_i. Then the following are equivalent:*

1. *Δ_c is (point) identified by $\mathcal{P}_{DZ}$ and $\{\beta_s\}_{s\in\mathcal{S}}$, for some finite set $\mathcal{S}$ of known or identified (from $\mathcal{P}_{DZ}$) measurable functions $s(d, z)$, and $\beta_s := E[s(D_i, Z_i)Y_i]$*
2. *$\Delta_c = \beta_s$ for a single such $s(d, z)$*
3. *$\Delta_c = E[t(D_i, Z_i, Y_i)]$ with $t(d, z, y)$ a known or identified (from $\mathcal{P}_{DZ}$) measurable function*
4. *Δ_c is identified from the set of CEFs $\{E[Y_i|D_i = d, Z_i = z]\}$ for $d \in \{0, 1\}$, $z \in \mathcal{Z}$ along with the joint distribution $\mathcal{P}_{DZ}$*

Proof. See Supplemental Material Section D.1. □

In saying that a parameter θ is *identified* by some set of empirical estimands, I mean that the set of values of θ that are compatible with the empirical estimands is a singleton, regardless of the distribution of the latent variables $(G_i, Y_i(1), Y_i(0))$ – for all $\mathcal{P}_{DZ}$ within some class (note that the marginal distribution of G_i must also be compatible with $\mathcal{P}_{DZ}$). For example, by writing the estimand of Theorem 1 $\sum_{z\in\mathcal{Z}} \frac{P(Z_i=z)h(z;\mathcal{P}_{DZ})}{E[h(Z_i;\mathcal{P}_{DZ})D_i]} \cdot E[Y_i|Z_i = z]$, where we make explicit that the function h depends on $\mathcal{P}_{DZ}$, it is clear that for any Δ_c satisfying Property M and under Assumptions 1-2, Δ_c is identified in the sense of item 4., for all $\mathcal{P}_{DZ}$ with the properties: i) the marginal distribution of Z_i satisfies Assumption 3; and ii) $E[h(Z_i; \mathcal{P}_{DZ})D_i] > 0$.

D.12 Proof of Proposition 6

By Proposition 11, we know that if Δ_c is identified from a finite set of IV-like estimands and $\mathcal{P}_{DZ}$, it can be written as a single one: $\Delta_c = \beta_s$ with $s(d,z)$ an identified functional of $\mathcal{P}_{\mathcal{DZ}}$. Now, using that $Y_i = Y_i(0) + D_i\Delta_i$ where $\Delta_i := Y_i(1) - Y_i(0)$:

$$\begin{aligned}
\Delta_c = \beta_s &= \{E[s(D_i, Z_i)Y_i(0)] + E[s(D_i, Z_i)D_i\Delta_i]\} \\
&= \sum_g P(G_i = g)\left\{E[s(D_g(Z_i), Z_i)Y_i(0)|G_i = g] + E[s(D_g(Z_i), Z_i)D_g(Z_i)\Delta_i|G_i = g]\right\} \\
&= \sum_g P(G_i = g)\left(\cancel{E[s(D_g(Z_i), Z_i)]}\right) E[Y_i(0)|G_i = g] \\
&\qquad + \sum_g P(G_i = g)\left(E[s(D_g(Z_i), Z_i)D_g(Z_i)]\right) E[\Delta_i|G_i = g] \\
&= \sum_g P(G_i = g)\left(E[s(1, Z_i)D_g(Z_i)]\right)\Delta_g
\end{aligned}$$

where we've used independence, and that the crossed out term must be equal to zero for every g by the assumption that $\beta_s = \Delta_c$ for every joint distribution of compliance groups and potential outcomes compatible with $\mathcal{P}_{DZ}$ in some class (it is always possible to translate the support of the distribution of $Y_i(0)$ and $Y_i(1)$ by the same constant without affecting Δ_i). Finally, $s(D_g(Z_i), Z_i)D_g(Z_i) = s(1, Z_i)D_g(Z_i)$ with probability one, establishing the final equality.

Recall that from Equation (2) that Δ_c can also be written as a weighted average of group-specific average treatment effects $\Delta_g = E[Y_i(1) - Y_i(0)|G_i = g]$ as:

$$\Delta_c = \frac{1}{P(C_i = 1)} \sum_g P(G_i = g)E[c(g, Z_i)] \cdot \Delta_g$$

Since $\beta_s = \Delta_c$ holds for any vector of $\{\Delta_g\}$ across all of the g for which $P(G_i = g) > 0$ is compatible with $\mathcal{P}_{DZ}$, we can match coefficients within this group to establish that $E[c(g, Z_i)] = P(C_i = 1)E[s(1, Z_i)D_g(Z_i)]$. This set of weights satisfies Property M, since for any $g \in \mathcal{G}^c$:

$$\begin{aligned}
E[c(g, Z_i)] &= P(C_i = 1)E[s(1, Z_i)\sum_S [M_J]_{F(g),S} D_{g(S)}(Z_i)] \\
&= \sum_S [M_J]_{F(g),S}\left(P(C_i = 1)E[s(1, Z_i)D_{g(S)}(Z_i)]\right) \\
&= \sum_S [M_J]_{F(g),S} \cdot E[c(Z_i, g(S))]
\end{aligned}$$

If this holds for any distribution of Z_i satisfying Assumption 3, then we must have $c(g,z) = \sum_S [M_J]_{F(g),S} \cdot c(g(S), z)$ for all $z \in \mathcal{Z}$. To see this, consider a sequence of distributions for Z_i that converges point-wise to a degenerate distribution at any single point z, but satisfies Assumption 3 for each term in the sequence. Applying the dominated convergence theorem to $E[c(g, Z_i)] - \sum_S [M_J]_{F(g),S} \cdot E[c(g(S), Z_i)] = 0$ along this sequence, we have that $c(g,z) = \sum_S [M_J]_{F(g),S} \cdot c(g(S), z)$. We can apply a similar argument to

establish that $c(a.t., z) = c(n.t., z) = 0$ for all $z \in \mathcal{Z}$ given that $E[c(g, Z_i)] = P(C_i = 1)E[s(1, Z_i)D_g(Z_i)]$ and $E[s(1, Z_i)] = 0$.

D.13 Proof of Proposition 8

In Supplemental Material Section B.1, I show that with two binary instruments, if PM holds but not VM or IAM, then $\mathcal{G}$ consists of seven compliance groups, whose definitions are given in Supplemental Material. We suppose that all 7 groups are possibly present, and the practitioner has knowledge of $E[Y_i|D_i = d, Z_i = z]$ for all eight combinations of (d, z), as well as the joint distribution of D_i and Z_i. This is equivalent to knowledge of $E[Y_i D_i|Z_i = z]$ and $E[Y_i(1-D_i)|Z_i = z]$ for all $z \in \mathcal{Z}$ and the joint distribution of (D_i, Z_i). Point identification from these moments is in turn equivalent to point identification from a finite set of IV-like estimands, by Proposition 11.

Using Supplemental Material Table 2, these eight moments can be written in matrix form as

$$\left(\begin{array}{c} E[Y_iD_i|Z_i=(0,0)] \\ E[Y_iD_i|Z_i=(0,1)] \\ E[Y_iD_i|Z_i=(1,0)] \\ E[Y_iD_i|Z_i=(1,1)] \\ \hline E[Y_i(1-D_i)|Z_i=(0,0)] \\ E[Y_i(1-D_i)|Z_i=(0,1)] \\ E[Y_i(1-D_i)|Z_i=(1,0)] \\ E[Y_i(1-D_i)|Z_i=(1,1)] \end{array}\right) = \left[\begin{array}{ccccccc|ccccccc} 0&0&1&0&0&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&0&1&0&0&0&0&0&0&0&0 \\ 1&1&1&0&1&1&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&1&0&0&0&0&0&0&0&0 \\ \hline 0&0&0&0&0&0&0&1&1&0&1&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&1 \\ 0&0&0&0&0&0&0&1&0&1&1&1&0&1 \end{array}\right] \left(\begin{array}{c} p_{odd} \cdot E[Y_i(1)|G_i = odd] \\ p_{eager} \cdot E[Y_i(1)|G_i = eager] \\ p_{reluct.} \cdot E[Y_i(1)|G_i = reluct.] \\ p_1 \cdot E[Y_i(1)|G_i = 1only] \\ p_2 \cdot E[Y_i(1)|G_i = 2only] \\ p_a \cdot E[Y_i(1)|G_i = a.t.] \\ p_n \cdot E[Y_i(1)|G_i = n.t.] \\ \hline p_{odd} \cdot E[Y_i(0)|G_i = odd] \\ p_{eager} \cdot E[Y_i(0)|G_i = eager] \\ p_{reluct.} \cdot E[Y_i(0)|G_i = reluct.] \\ p_1 \cdot E[Y_i(0)|G_i = 1only] \\ p_2 \cdot E[Y_i(0)|G_i = 2only] \\ p_a \cdot E[Y_i(0)|G_i = a.t.] \\ p_n \cdot E[Y_i(0)|G_i = n.t.] \end{array}\right),$$

for some labelling of the instrument values, where the groups "reluctant defiers" and "odd compliers" are defined in the Supplemental Material. If this equation is written as $b = Ax$, where b is the 8×1 vector of identified quantities, and x the 14×1 unknown vector of potential outcome moments (note the matrix A here is not the same as the matrix A defined in Corollary 1), then ACL can be written as

$$ACL = \frac{1}{1 - p_a - p_n} \cdot \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \end{pmatrix}'}_{:=\lambda} x \qquad (14)$$

ACL is identified only if the vector λ is in the row space of matrix A (the column space of A'), which follows from the proof of $\mathbf{4} \to \mathbf{2}$ in Proposition 11. This can be readily verified not to hold, since

$$A'(AA')^{-1}A\lambda \approx \begin{pmatrix} 1.45 & .82 & .82 & .73 & .73 & .18 & 0 & -1.45 & -.73 & -.73 & -.82 & -.82 & 0 \end{pmatrix}$$

where $A'(AA')^{-1}A$ is the orthogonal projector into the row space of A (which has full row rank). Since the RHS of the above is not equal to λ (given explicitly in Eq. 14), λ is not in the row space of A.

D.14 Proof of Proposition 9

Define $A = \{g \in \mathcal{G} : E[D_g(Z_i)] = 1\}$ and $N = \{g \in \mathcal{G} : E[D_g(Z_i)] = 0\}$. Then

$$E[Y_i|Z_i = \bar{Z}] = p_n E[Y_i(0)|G_i \in N] + p_a E[Y_i(1)|G_i \in A] + P(G_i \in \mathcal{G}^c)E[Y_i(1)|G_i \in \mathcal{G}^c]$$

and

$$E[Y_i|Z_i = \underline{Z}] = p_n E[Y_i(0)|G_i \in N] + p_a E[Y_i(1)|G_i \in A] + P(G_i \in \mathcal{G}^c)E[Y_i(0)|G_i \in \mathcal{G}^c]$$

where $p_a = P(E[D_{G_i}(Z_i)] = 1) = E[D_i|Z_i = \underline{Z}]$, $p_n = P(E[D_{G_i}(Z_i)] = 0) = E[1-D_i|Z_i = \bar{Z}]$, and note that $P(G_i \in \mathcal{G}^c) = 1 - p_a - p_n$. Combining, we have that $E[Y_i|Z_i = \bar{Z}] - E[Y_i|Z_i = \underline{Z}] = P(G_i \in \mathcal{G}^c)ACL$, and that $E[D_i|Z_i = \bar{Z}] - E[D_i|Z_i = \underline{Z}] = P(G_i \in \mathcal{G}^c)$. It follows now that $ACL = \rho_{\bar{Z},\underline{Z}}$. The result can be seen as following from the fact that under either VM or IAM: for any point $z \in \mathcal{Z}$, $D_i(\underline{Z}) \leq D_i(z) \leq D_i(\bar{Z})$ with probability one.

To see that $\rho_{\bar{Z},\underline{Z}}$ is equal to the ACL as expressed by Corollary 1, note that each entry of the vector $(0, 1, \ldots, 1)A$ is a column sum of the matrix A where we omit the first entry. The identity $\sum_{f \subseteq X}(-1)^{|f|} = 0$ for any finite set X will be useful in what follows (the number of "even" and "odd" terms in a binomial expansion are equal). The first entry of $(0, 1, \ldots, 1)A$ is equal to the first entry of $(1, 1, \ldots, 1)A$, minus A_{11} which is equal to one. Thus, we have $-1 + \sum_{f \subseteq \{1 \ldots J\}}(-1)^{|f|} = -1$, by the identity. For any column corresponding to a non-empty z_0, the first entry on that column is equal to zero, and thus the corresponding entry in $(0, 1, \ldots, 1)A$ is equal to the full column sum, which is equal to zero by the identity. The last column of A is always equal to a column of zeros followed by a single one, and hence the final component of $(0, 1, \ldots, 1)A$ is equal to positive one.

D.15 Proof of Proposition 10

Write the parameter of interest Δ_c as θ_Y/θ_D, where for $V \in \{Y, D\}$, $\theta_V = \tilde{\lambda}'\beta_V$ with $\beta_V := E[\Gamma_i\Gamma_i']^{-1}E[\Gamma_i' V_i]$ and $\tilde{\lambda} = (0, \lambda')'$. Denote the estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ as $\hat{\Delta}_c$ for shorthand. It takes the form $\hat{\Delta}_c = \hat{\theta}_Y/\hat{\theta}_D$, where $\hat{\theta}_V := (0, \hat{\lambda}')'(\Gamma'\Gamma + K)^{-1}\Gamma' V$, and $K = \alpha I$. I keep the notation in terms of K as the first part of the argument below will go through with any diagonal matrix of positive entries, allowing a different regularization parameter corresponding to each singular vector of $\Gamma'\Gamma$. Write each $\hat{\theta}_V := (0, \hat{\lambda}')'\hat{\beta}_V^*$ where $\hat{\beta}_V^*$ is the ridge-regression estimate of β_V, and let $\hat{\beta}_V = (\Gamma'\Gamma)^{-1}\Gamma' V$ be the unregularized regression coefficient estimator.

Consider the conditional MSE $M = E[(\hat{\Delta}_c - \Delta_c)^2|\Gamma]$. It can be rearranged as:

$$\begin{aligned} M &= E\left[\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D} - \frac{\theta_Y}{\theta_D}\right)^2 \middle| \Gamma\right] = \frac{1}{\theta_D^2}E\left[\left((\hat{\theta}_Y - \theta_Y) - \hat{\Delta}_c(\hat{\theta}_D - \theta_D)\right)^2 \middle| \Gamma\right] \\ &= \frac{1}{\theta_D^2}E\left[(\hat{\theta}_Y - \theta_Y)^2 + \hat{\Delta}_c^2(\hat{\theta}_D - \theta_D)^2 - 2\hat{\Delta}_c(\hat{\theta}_Y - \theta_Y)(\hat{\theta}_D - \theta_D)\middle| \Gamma\right] \end{aligned} \tag{15}$$

For any $V, W \in \{Y, D\}$, and $m \geq 1$:

$$E\left[(\hat{\Delta}_c)^m(\hat{\theta}_V - \theta_V)(\hat{\theta}_W - \theta_W)\middle|\Gamma\right] = E\left[(\hat{\Delta}_c)^m(0,\hat{\lambda})'(\hat{\beta}_V^* - \beta_V)(\hat{\beta}_W^* - \beta_W)'(0,\hat{\lambda})'\middle|\Gamma\right]$$
$$= (\Delta_c)^m\tilde{\lambda}'E\left[(\hat{\beta}_V^* - \beta_V)(\hat{\beta}_W^* - \beta_W)'\middle|\Gamma\right]\tilde{\lambda} + R_n^m$$

where the first term in the above is viewed as an approximation that ignores terms that are of third or higher order in estimation errors. The asymptotic rate at which the approximation error captured by the R_n^m converges to zero is considered explicitly at the end of this section.

Let $Z = (\Gamma'\Gamma + K)^{-1}\Gamma'\Gamma$ and notice that $\hat{\beta}_V^* = Z\hat{\beta}_V$. Using that $E[\hat{\beta}_V|\Gamma] = \beta_V$ (as Γ_i includes all products of the instruments the CEF must be linear) for $V \in \{Y, D\}$:

$$E\left[(\hat{\beta}_V^* - \beta_V)(\hat{\beta}_W^* - \beta_W)'\middle|\Gamma\right] = ZE\left[(\hat{\beta}_V - \beta_V)(\hat{\beta}_W - \beta_W)'\middle|\Gamma\right]Z' + (Z-I)\beta_V\beta_W'(Z-I)'$$
$$= (\Gamma'\Gamma + K)^{-1}(\Gamma'\Omega_{VW}\Gamma + K\beta_V\beta_W'K)(\Gamma'\Gamma + K)^{-1}$$

where we define the $n \times 1$ vector $U_V = V - \Gamma\beta_V$ and $\Omega_{VW} = E[U_V U_W'|\Gamma]$. Thus, total conditional MSE is, by Equation (15):

$$M \approx \frac{1}{\theta_D^2}\tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\left\{\Gamma'(\Omega_Y + \Delta_c^2\Omega_D - 2\Delta_c\Omega_{YD})\Gamma\right.$$
$$\left. + K(\beta_Y\beta_Y' + \Delta_c^2\beta_D\beta_D' - 2\Delta_c\beta_Y\beta_D')K\right\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda}$$

This development follows and generalizes that of Hoerl and Kennard (1970), who consider MSE optimal regularization via ridge regression for estimating a single regression vector, under homoskedasticity. Our case targets the ratio $\hat{\theta}_Y/\hat{\theta}_D$ rather than a vector of regression coefficients, and also allows for heteroskedasticity.

We now prove that $\alpha/\sqrt{n} \xrightarrow{p} 0$ if α is chosen to minimize the following "single-step" estimator of the MSE (ignoring the positive factor of θ_D^{-2} that does not depend on K):

$$\hat{M} := \tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\left\{\Gamma'\left(\hat{\Omega}_Y + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\Omega}_D - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\Omega}_{YD}\right)\Gamma + \right.$$
$$\left. K\left(\hat{\beta}_Y\hat{\beta}_Y' + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\beta}_D\hat{\beta}_D' - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_Y\hat{\beta}_D'\right)K\right\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda}$$

where $\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)$ is the un-regularized estimator of Δ_c. The problem can be re parameterized as a choice of $b := \alpha/n$, where

$$\hat{M}(b) := \tilde{\lambda}'\left(\frac{\Gamma'\Gamma}{n} + bI\right)^{-1}\left\{\frac{1}{n}\frac{\Gamma'\left(\hat{\Omega}_Y + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\Omega}_D - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\Omega}_{YD}\right)\Gamma}{n} + \right.$$
$$\left. b^2\left(\hat{\beta}_Y - \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_D\right)\left(\hat{\beta}_Y - \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_D\right)'\right\}\left(\frac{\Gamma'\Gamma}{n} + bI\right)^{-1}\tilde{\lambda}$$
$$:= m(b, \hat{\Pi}, \hat{\beta}, \hat{\Sigma}, \hat{\lambda})$$

where $\hat{\Pi} := \frac{1}{n}\sum_i (\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D \hat{U}_{Di})^2 \Gamma_i \Gamma_i'$, $\hat{\beta} := (\hat{\beta}_Y - \hat{\theta}_Y/\hat{\theta}_D \hat{\beta}_D)$, and $\hat{\Sigma}^* := \frac{1}{n}\sum_i \Gamma_i \Gamma_i'$. Note that $\hat{\beta} \xrightarrow{p} \beta := \beta_Y - \Delta_c \beta_D$, $\hat{\Sigma}^* \xrightarrow{p} \Sigma^* := E[(1, \Gamma_i')'(1, \Gamma_i')]$, $\sqrt{n}\left(\hat{\Pi} - \Pi\right) \xrightarrow{d} N(0, V)$ for some V provided that the variance of $(\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D \hat{U}_{Di})^2 \Gamma_i \Gamma_i'$ exists, where $\Pi := E[(\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D U_{Di})^2 \Gamma_i \Gamma_i']$. The function m is

$$m(b, \Pi/n, \beta, \Sigma^*, \lambda) = (0, \lambda') \left(\Sigma^* + bI\right)^{-1} \left\{\Pi/n + b^2 \beta\beta'\right\} \left(\Sigma^* + bI\right)^{-1} (0, \lambda')'$$

We wish to show that $\sqrt{n}b = \alpha/\sqrt{n} \xrightarrow{p} 0$, when b is chosen as the smallest positive minimizer of $m(\cdot, \hat{\Pi}/n, \hat{\beta}, \hat{\Sigma}, \hat{\lambda})$. The strategy will be to show that $nb \xrightarrow{p} X$ where X is a finite degenerate random variable. Since Π and $\beta\beta'$ are positive definite, it is clear that $m(b, \Pi/n, \beta, \Sigma^*, \lambda)$ is weakly positive for any choice of b. Further, $m(b, \Pi/n, \beta, \Sigma^*, \lambda)$ is typically strictly positive at $b = 0$, and it can also be seen that $\lim_{b\to\infty} m(b, \Pi/n, \beta, \Sigma^*, \lambda) = 0$ (see Section 5.2 for discussion). However, m is generally not monotonically decreasing in between, as we shall see below.

Observe that $b = 0$ minimizes $m(b, \mathbf{0}, \beta, \Sigma^*, \lambda)$ with respect to b regardless of the values of β, Σ^*, λ, where $\mathbf{0}$ is a $k \times k$ matrix of zeros (the dimension of Π), since $m(\cdot)$ is always positive and when its second argument vanishes can be made equal to zero by choosing $b = 0$. Furthermore, $b = 0$ is a local minimizer when $\Pi/n = \mathbf{0}$, since m_b vanishes when evaluated at $(0, \mathbf{0}, \beta, \Sigma^*, \lambda)$–see below, while the second derivative of m with respect to b, evaluated at $(0, \mathbf{0}, \beta, \Sigma^*, \lambda)$, is equal to

$$(0, \lambda')\Sigma^{*-1}\beta\beta'\Sigma^{*-1}\lambda = \left((0, \lambda')\Sigma^{*-1}\beta\right)^2$$

up to a strictly positive constant. We have assumed that the quantity in parenthesis is non-zero. By the implicit function theorem, there then exists a unique function $g(\Pi/n; \beta, \Sigma^*, \lambda)$ such that $g(\mathbf{0}; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$ and $m_b(g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}), \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$, in a neighborhood $\mathcal{N}$ of the probability limits $(\mathbf{0}, \beta, \Sigma^*, \lambda)$ of $(\hat{\Pi}/n, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$, and this function is continuously differentiable with respect to all parameters, (including, in particular, the elements of Π). Since the second derivative of m is strictly positive at $(0, \mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ and continuous with respect to all arguments, $\mathcal{N}$ can furthermore be chosen such that the critical point at $(g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}), \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ is always a local minimum within $\mathcal{N}$.

Since for any realization of $\hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}$:

$$m_b(0, \mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 2\tilde{\lambda}'(\hat{\Sigma}^* + bI)^{-1} \left\{bI - b^2(\hat{\Sigma}^* + bI)^{-1}\right\} \hat{\beta}\hat{\beta}'(\hat{\Sigma}^* + bI)^{-1}\tilde{\lambda}\Big|_{b=0} = 0$$

we see that m has a critical point at $b = 0$ for values $(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ of the other arguments. By uniqueness of the function $g(\Pi/n; \beta, \Sigma^*, \lambda)$, this implies then that $g(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$. By the mean value theorem, we can write

$$\begin{aligned} g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) &= g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) - g(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) \\ &= \frac{\partial}{\partial x} g(vec(cn^{-1}\hat{\Pi}); \hat{\beta}, \hat{\Sigma}^*, \hat{\lambda}) \cdot \frac{vec(\hat{\Pi})}{n} \end{aligned}$$

for some $c \in [0, 1]$, where $vec(\Pi)$ denotes the vectorization x of the matrix Π, and we let $\frac{\partial}{\partial x} g(x; \beta, \Sigma^*, \lambda)$ denote a gradient of g with respect to that vector (recall that existence

of the derivative is a consequence of the implicit function theorem). By continuity of $\frac{\partial}{\partial x}g(x;\beta,\Sigma^*,\lambda)$ and the continuous mapping theorem then,

$$n \cdot g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda}) \xrightarrow{p} \frac{\partial}{\partial x}g(\mathbf{0},\beta,\Sigma^*,\lambda)vec(\Pi) \tag{16}$$

which is a finite scalar.

To complete the proof, we now simply note that with probability approaching unity, $(\hat{\Pi}/n,\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})$ is within the neighborhood $\mathcal{N}$, and thus if b is chosen as the smallest positive local minimizer of $m(b,\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})$ we have that $b = g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})$. We have now established the result, since for any $B > 0$:

$$\begin{aligned}
P(|\alpha/\sqrt{n}| > B) &\leq P(|\alpha/\sqrt{n}| > B \text{ and } b = g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})) + P(b \neq g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})) \\
&= P(|n \cdot g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})| > \sqrt{n}B) + P(b \neq g(\hat{\Pi}/n;\hat{\beta},\hat{\Sigma}^*,\tilde{\lambda})) \\
&\xrightarrow{n} 0 + 0
\end{aligned}$$

Finally, I consider the error involved in the approximation made to Equation (15). Write this as:

$$\begin{aligned}
R_n &:= R_n^m + R_n^m = \\
&= \frac{1}{\theta_D^2}\tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\Big\{(\hat{\Delta}_c^2 - \Delta_c^2)(\Gamma'\Omega_D\Gamma + K\beta_D\beta_D'K) \\
&\qquad\qquad -2(\hat{\Delta}_c - \Delta_c)(\Gamma'\Omega_{YD}\Gamma + K\beta_Y\beta_D'K)\Big\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda} \\
&= \frac{1}{\theta_D^2 \cdot n^{3/2}} \cdot \tilde{\lambda}'\left(\frac{\Gamma'\Gamma}{n} + \frac{K}{n}\right)^{-1}\left\{\sqrt{n}(\hat{\Delta}_c^2 - \Delta_c^2)\left(\frac{\Gamma'\Omega_D\Gamma}{n} + \frac{K}{\sqrt{n}}\beta_D\beta_D'\frac{K}{\sqrt{n}}\right)\right. \\
&\qquad\qquad \left. -2\sqrt{n}(\hat{\Delta}_c - \Delta_c)\left(\frac{\Gamma'\Omega_{YD}\Gamma}{n} + \frac{K}{\sqrt{n}}\beta_Y\beta_D'\frac{K}{\sqrt{n}}\right)\right\}\left(\frac{\Gamma'\Gamma}{n} + \frac{K}{n}\right)^{-1}\tilde{\lambda}
\end{aligned}$$

Provided that $\alpha/\sqrt{n} \xrightarrow{p} 0$ as above, we will show in Theorem 2 that $\hat{\Delta}_c$ is $\sqrt{n}$-consistent for Δ_c. In this case, the approximation error term is $O_p(n^{-3/2})$.

D.16 Proof of Theorem 2

When $\alpha_n = 0$, the result follows from Theorem 3 of Imbens and Angrist (1994). To see that $o_p(\sqrt{n})$ regularization has no asymptotic effect, note that

$$\begin{aligned}
(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}\Gamma'Y &= (0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma + \alpha I - \alpha I)(\Gamma'\Gamma)^{-1}\Gamma'Y \\
&= (0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'Y - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'Y
\end{aligned}$$

and similarly for D, thus:

$$\begin{aligned}
\rho(\hat{\lambda},\alpha) &= \frac{(0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'Y - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'Y}{(0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'D - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'D} \\
&= \frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'Y}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'D} \\
&= \frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} + \frac{\alpha}{n} \cdot \frac{(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\left\{\frac{1}{n}\Gamma'D \cdot \frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} - \frac{1}{n}\Gamma'Y\right\}}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'D}
\end{aligned}$$

and thus the asymptotic distribution of $\sqrt{n}(\hat{\rho}(\hat{\lambda}, 0) - \Delta_c)$ is the same as that of $\sqrt{n}\left(\frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} - \Delta_c\right)$, provided that $\alpha_n/\sqrt{n} \xrightarrow{p} 0$ (in which case the second term above is $o_p(n^{-1/2})$).