A Vector Monotonicity Assumption for Multiple Instruments

Leonard Goff*

October 10, 2019. Preliminary and incomplete.
Click here for most recent version.

Abstract

I study a monotonicity assumption that may be natural in cases where the researcher has more than one instrumental variable for a single binary treatment. In such cases, the traditional LATE monotonicity assumption may become restrictive, as it requires that all units share a common direction of response (either into or out of treatment) to any change in the values of the instruments as a vector: even when one instrument is made "lower" while another is made "higher". What I call *vector monotonicity*, by contrast, simply restricts treatment status to be monotonic in each instrument separately. This allows two-way flows between values of the instrument vector that are not ordered component-wise, and nests natural selection models, for example random coefficients models with positive coefficients. I show that in a setting with 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 (including as special cases all instruments or a single instrument). I also show how discrete instruments can be mapped into this framework more generally. I propose a regularized "2SLS-like" estimator for the class of treatment effect parameters, and provide simulation evidence on its performance. By highlighting the additional identification power of vector monotonicity, these results complement a recent analysis by Mogstad et al. (2019) that focuses on a weaker *partial monotonicity* assumption.

1 Introduction

The influential local average treatment effects (LATE) framework pioneered by Imbens and Angrist (1994) allows for causal inference with instrumental variables while allowing unrestricted heterogeneity in treatment effects. To acheive this, it imposes an important type of homogeneity across units in selection behavior, formalized by what has become known as the LATE *monotonicity* assumption. This condition says that for any pair of points in the (joint) support of the instruments, one of the points has the property that if any unit takes treatment when the instruments take that value (as a vector), they would also take treatment if their instrument vector took the other value.

This assumption is quite natural for a single instrumental variable for which there is a natural ordering among the values it can take, with the property that "higher" values of the instrument constitute a greater incentive towards treatment than "lower" values.

*I thank Simon Lee, Josh Angrist, Bernard Salanié, Suresh Naidu, Serena Ng, José Luis Montiel Olea, Junlong Feng and Vitor Possebom for helpful comments and discussion. I also thank attendees of the Columbia econometrics colloquium and the 2019 Young Economists Symposium for their feedback. Any errors or other shortcomings are my own.

For example, a lower distance to the nearest college can be though to induce individuals towards receiving a college degree. But as I describe below, the assumption becomes harder to justify when the researcher has multiple instrumental variables, thought to be jointly valid.

In this paper I propose what I call *vector monotonicity* as an assumption for cases in which the researcher has more than one instrument for a single binary treatment. Vector monotonicity (VM) simply assumes that potential treatments are always monotonic with respect to each instrument separately, regardless of the values that the other instruments take. This captures the intuitive notion that each instrument has an impact on treatment uptake that is of a known sign.

Through most of the analysis, I take the available instrumental variables to be represented by J binary instruments, and discuss in Section 3.3 how discrete instruments can be re-represented as binary instruments, while preserving vector monotonicity. My main contribution is to show that under vector monotonicity, a family of LATE parameters are point identified with such a set of valid binary instrumental variables. I furthermore propose a simple "2SLS-like" estimator for these parameters. The identified family includes naturally interpretable summary statistics of heterogeneous treatment effects, such as: i) the average treatment effect among units that are responsive to *any* combination of instrument values (what I call the "Big LATE"); ii) the average treatment effect among units that are responsive to any fixed subset of the instruments (including a single instrument); and iii) the Big LATE among the treated. When the outcome variable is bounded, I discuss how these parameters can be used to construct identified sets for parameters that do not depend on the instruments available, such as the average treatment effect and the average treatment on the treated.

As a motivating example, consider a researcher studying the returns to a college education. Suppose, inspired by Card (1995) and Kane and Rouse (1993), that two binary instruments are available: distance to the closest university (coded as "close" and "far") and local tuition rates (coded as "low" and "high").The traditional LATE monotonicity assumption 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. Otherwise, we would have a failure of the assumption between the instrument values $z = (far, cheap)$ and $z' = (close, expensive)$. We would generally expect such a violation if individuals are heterogeneous in how much each instrument matters to them: for example, if there are some students who care about distance only and others who care about tuition only. Vector monotonicity, on the other hand, says something quite natural: proximity to a college weakly encourages college attendance, regardless of the price, and lower tuition weakly encourages college attendance, regardless of distance. "How much" each matters can differ arbitrarily across individuals.[1]

[1] As a second example, consider evaluating the effects of a job search workshop that is advertized by mail to a random subset of individuals on the rolls for unemployment assistance. To increase uptake, program staff also visit some individuals' homes to encourage participation. Program staff visit individuals from the mailing list at random times during the day,

The above difficulties for traditional LATE monotonicity with multiple instruments have recently been highlighted by Mogstad et al. (2019) (henceforth MTW). MTW introduce VM, and then consider identification under a weaker assumption they call *partial monotonicity* (PM), which nests both VM and traditional LATE monotonicity as special cases. MTW derive testable conditions which under PM are sufficient for the standard two stage least squares (2SLS) estimand to deliver a convex combination of treatment effects in the population. However, the conditions may not hold, and they may be difficult to verify in practice as the number of them generally grows combinatorially with the number of instruments. MTW develop an alternative by considering what the instruments can say about a large class of specific target parameters, such as policy relevant treatment effects (Heckman and Vytlacil 2005), that are generally only bounded by IV methods. I complement this analysis by characterizing a class of interpretable causal parameters that are *point* identified maintaining the stronger assumption of VM. While VM is stronger, I argue that it is typically quite natural and is perhaps the canonical alternative to traditional LATE monotonicity within the class PM. I show that if a class of causal parameters that includes the Big LATE is generically identified by the various conditional means of the outcome on treatment and instruments, and PM holds, then either VM or traditional LATE monotonicity must too. Thus VM (like traditional LATE monotonicity) has particular identifying power beyond that of PM.

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 simple choice models. In Section 3, I show that like conventional monotonicity, VM separates the population into mutually compatible "compliance groups" based on each unit's selection behavior.[2] 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. An appendix discusses estimation with covariates. In the Supplemental Material, I also consider special cases in which 2SLS can be expected to uncover averages of causal effects under VM with binary instruments. However, these special cases are restrictive.

making contact with a random subset of them. The staff also visit a smaller number of unemployed invidivuals who were not sent a mailer. This leads to two binary instruments that are correlated with one another, but each uncorrelated with potential outcomes and potential treatments. Traditional LATE monotonicity would imply, for example, that anyone who enrolls in the program with an in-person visit but no mailer would also enroll if they had only received the mailer (or that the same applies in the other direction).

[2]These compliance groups would allow one to represent selection behavior in terms of a non-separable latent index model, as I discuss in the Supplemental Material. Heckman and Vytlacil (2001) point out that in such models the straightforward relationships between conventional IV estimands and their causal parameter counterparts breaks down. In this paper I show that useful identification is nevertheless possible when VM holds.

2 Setup

Consider a setting with a binary treatment variable D, an outcome variable Y, and a vector $Z = (Z_1 \dots Z_J)$ of J instrumental variables with support $\mathcal{Z} \subseteq (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$.

Definition (potential outcomes and treatments). *Let $D_i(z)$ denote the treatment status of unit i when their vector of instrumental variables takes value $z \in \mathcal{Z}$. Actual realized treatment is $D_i = D_i(Z_i)$, where Z_i is unit i's value of Z. Let $Y_i(d, z)$ denote the realization of the outcome variable that would occur with treatment status $d \in \{0, 1\}$ and instrument value $z \in \mathcal{Z}$. The observed value of Y is then $Y_i = Y_i(D_i, Z_i) = Y_i(D_i(Z_i), Z_i)$.*

Let G_i be a random variable defined via a one-to-one correspondence with the function $\{D_i(z)\}_{z \in \mathcal{Z}}$. This 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. 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 the 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 (selection/"compliance" behavior). In practice, it is common 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 C.

2.1 Notions of monotonicity

The second assumption of conventional LATE analysis is traditional monotonicity:

Assumption IAM (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 phrasing of MTW and henceforth call this Assumption IAM, or "Imbens and Angrist monotonicity", for its introduction in Imbens and Angrist (1994). 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 value of IAM for identification is demonstrated by Imbens and Angrist (1994), who show that Assumptions 1 and IAM are sufficient to identify a LATE among units

that change treatment status between z and z', where z, z' are any two points in $\mathcal{Z}$ such that $P(D_i(z) = 1) > P(D_i(z') = 1)$.

$$E[Y_i(1) - Y_i(0)|D_i(z) > D_i(z')] = \frac{E[Y_i|Z_i = z] - E[Y_i|Z_i = z']}{E[D_i|Z_i = z] - E[D_i|Z_i = z']}$$

However, as suggested in the introduction, and as has been compellingly argued by MTW, Assumption IAM can be restrictive when the researcher is making use of multiple instrumental variables.

My proposed assumption of *vector monotonicity* captures the notion that each instrument either encourages or discourages all units to take treatment, regardless of the values of the other instruments:

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

Let $\mathcal{Z}_{-j}$ represent the joint support of all the instruments aside from instrument j, and let (z_j, z_{-j}) denote a vector composed of $z_j \in \mathcal{Z}_j$ and $z_{-j} \in \mathcal{Z}_{-j}$. An alternative characterization of VM is given by the following Proposition:

Proposition 1. *Assumption 2 holds iff $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$ when $z_j \geq z'_j$, according to some ordering $\geq$ on $\mathcal{Z}_j$, for all choices of $j \in \{1 \dots J\}$, $(z_j, z'_j) \in Z_j$ and $z_{-j} \in \mathcal{Z}_{-j}$.*

Proof. See Appendix A. □

Vector monotonicity is referred to as "actual monotonicity" in Mogstad et al. (2019). Mountjoy (2018) makes a version of VM in a case with a multivalued treatment, and continuous instruments. Note that vector monotonicity has the testable implication that the propensity score function $P(z) := E[D_i|Z_i = z]$ should be weakly monotonically increasing for values of z that are ordered in a vector sense.

MTW also introduce the notion of partial monotonicity (PM), which nests both Assumptions 2 and IAM:

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

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 (this is most apparent in the characterization of vector

monotonicity given by Proposition 1).

The relationship between Assumptions IAM, VM and PM is depicted graphically in Figure 1, with examples for a case of two binary instruments $\mathcal{Z} = \{0,1\} \times \{0,1\}$.

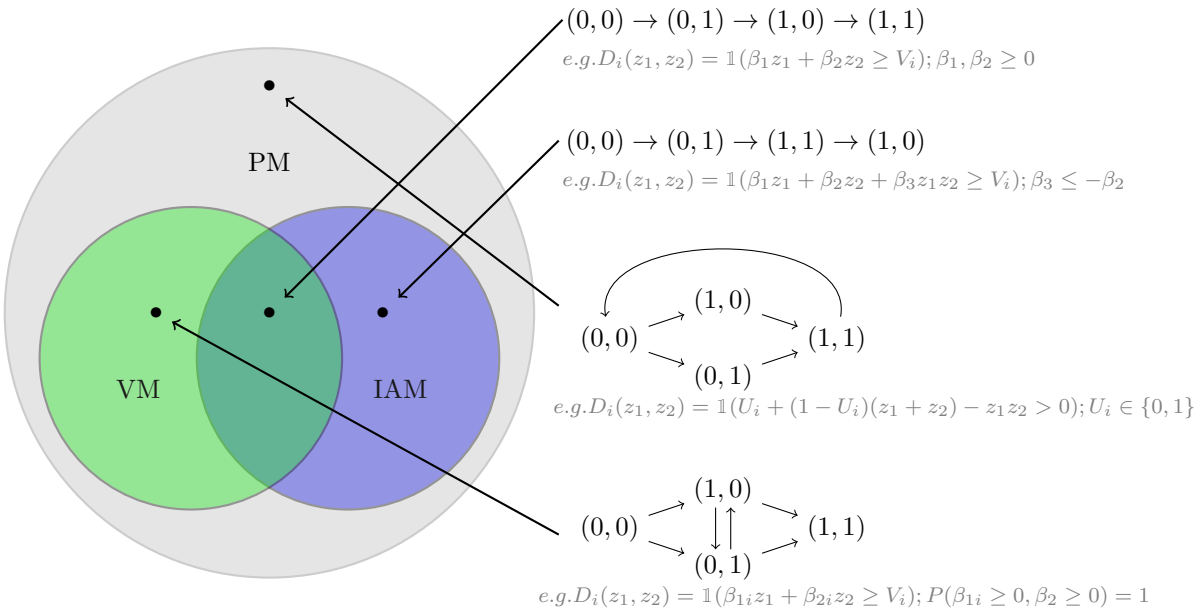

Figure 1: Comparison of Imbens & Angrist monotonicity (IAM), vector monotonicity(VM), and partial monotonicity (PM). Each dot represents an example in a case with two binary instruments ($\mathcal{Z} = \{0,1\} \times \{0,1\}$). An arrow from point z' to z indicates that $P(D_i(z) > D_i(z')) > 0$ is allowed in the example (for any z, z', if there is no path from z' to z, then $P(D_i(z) > D_i(z')) = 0$). Possible latent-index models underlying each example are given in small gray text (see Section 2.2).

Assumption IAM confers a well-defined ordering among points in the support of the vector of instruments according to the relation $P(D_i(z) \geq D_i(z')) = 1$. This relation may not be unique if $P(D_i(z) = D_i(z')) = 1$ for some pairs of points z, z', but can otherwise be detected empirically by ranking points in $\mathcal{Z}$ according to the propensity score. Under Assumption PM, such a "chain" exists along the support of Z_j, conditional on each fixed realization of Z_{-j}. Under VM, there is a partial order on $\mathcal{Z}$ defined by the property $P(D_i(z) \geq D_i(z')) = 1$: we can only make this claim for points for which $z \geq z'$ as a vector.

2.2 Examples in a latent-index selection model and the college decision

In this section I consider the latent-index models depicted in Figure 1 as examples of IAM, VM and PM, and discuss them in the context of the returns to schooling application. This is for exposition purposes only; I will not make use of them in the formal results. In the Supplemental Material, I discuss general latent-index representation existence results under VM, such as the well-known theorem by Vytlacil (2002) for IAM.

In a case with two instruments Z_1 and Z_2, the following linear selection model would satisfy IAM:

$$D_i(z_1, z_2) = \mathbb{1}(\beta_1 z_1 + \beta_2 z_2 \geq U_i) \tag{1}$$

Consider the returns to schooling example described in the introduction, with local tuition and distance to college each coded as binary variables for simplicity, with 1 corresponding to "cheap" (Z_1) and "close" (Z_2), respectively. Equation (1) in this case could reflect a model in which individuals differ in their perceived benefit $-U_i$ of attending college, but all place the same value on tuition and distance to college.

A latent-index model that satisfies VM, but not IAM, can be constructed by simply making the β's heterogeneous but having a common sign:

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

where $(\beta_{1i}, \beta_{2i}, U_i) \perp (Z_{1i}, Z_{2i})$, and $P(\beta_{1i} \geq 0, \beta_{2i} \geq 0) = 1$ (here we can also think of $-U_i$ as "β_0", a heterogeneous intercept term). With Z_1 and Z_2 binary, IAM will be violated if $\beta_{1i} < U_i \leq \beta_{2i}$ for some units but $\beta_{2i} < U_i \leq \beta_{1i}$ for some different units in the population. For example, suppose Alice and Bob are two students with the same U_i (e.g. they both place the same value on a college degree). Alice has greater means to afford tuition than Bob, but values living with her parents more than he does. Placing less value on tuition, Alice has a low β_{1i}, but a high β_{2i} because she wants to live at home. Bob has a higher β_{1i}, but doesn't mind living far from his parents and his β_{2i} is low. In this case, Alice and Bob might move in opposite directions between the instrument points $z = (cheap, far)$ and $z' = (expensive, close)$, violating IAM.

It is also possible to write down selection models that satisfy partial monotonicity but not VM or IAM, in the two instrument case. For example:

$$D_i(z_1, z_2) = \mathbb{1}(V_i + (1 - V_i)(z_1 + z_2) - z_1 z_2 > 0) \tag{3}$$

where $V_i \in \{0, 1\}$ with positive probability of each. The individuals with $V_i = 0$ take treatment if either of the instruments are set to one, while the $V_i = 1$ individuals take treatment unless both instruments are set to one. The two-way flow between $(0, 0)$ and $(1, 1)$ does not violate assumption PM, however, since these two points in $\mathcal{Z}$ are not related by changing the value of one instrument while keeping the other fixed.

In the returns to school example, this model could capture a situation in which the $V_i = 1$ units are students who really don't want to live with their parents during college, and feel that they will have to if attending a college near their parents' home. Accordingly, but perhaps counter-intuitively, these students only opt out of college if there is a cheap option close to where they grew up. The $V_i = 0$ group, on the other hand, exhibit more typical selection behavior, responding positively to both proximity and affordability of college. Assumption PM allows these two groups to coexist in the population. However, given the presence of both groups, PM requires that there be, among other groups, no individuals that go to college only if it is both cheap and close. Thus, while it is possible

to allow individuals with less obvious selection patterns (the $V_i = 1$ group) under PM, this can mean ruling out other individuals that exhibit quite natural varieties of selection behavior, which we may want to allow.

Furthermore, it's straightforward to see that if PM holds and there exist individuals that are sensitive to any one of the binary instruments alone, then VM must hold. For example, suppose Alice's selection behavior depends non-trivially only on the value of the distance instrument, and Bob's selection behavior is responsive only to the value of the tuition instrument, i.e.

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

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 vector monotonicity.[3]. For instance, the selection behavior described by $V_i = 1$ in Eq. (3) would violate PM given the presence of either Alice or Bob in the population.

3 Characterizing compliance behavior under vector monotonicity with binary instruments

Throughout this section we consider J binary instruments $Z_1 \dots Z_J$ that satisfy VM. We normalize the "1" state for each instrument to be the direction in which potential treatments are weakly increasing. Implicitly, this "up" value for each instrument will be taken in our results to be known ex ante. In practice, this may 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 if VM holds, by examining the propensity score function.

3.1 With two binary instruments

Consider the $J = 2$ case with two binary instruments Z_1 and Z_2. Each unit has 4 "potential treatments": counterfactual treatment values $D_i(z_1, z_2) \in \{0, 1\}$ given values of both of the instruments. With no monotonicity restrictions, there would be $2^4 = 16$ points in the support of G_i: each representing a distinct mapping from combinations of instrument values to treatment states. It turns out that 10 of these will violate VM, leaving 6. One can enumerate these explicitly by identifying the random variable G_i with the tuple $(D_i(0,0), D_i(0,1), D_i(1,0), D_i(1,1))$:

$$G_i \in \{NNNN, NNNT, \cancel{NNTN}, NNTT, \cancel{NTNN}, NTNT, \cancel{NTTN}, NTTT, \cancel{TNNN}, \cancel{TNNT}, \cancel{TNTN}, \cancel{TNTT}, \cancel{TTNN}, \cancel{TTNT}, \cancel{TTTN}, TTTT\}$$
$$= \{NNNN, NNNT, NNTT, NTNT, NTTT, TTTT\}$$

[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 $z_1 \in \mathcal{Z}_1$.

where N indicates no treatment, and T indicates treatment, and I have crossed out the tuples that violate VM.

I follow Mogstad et al. (2019) and refer to the six remaining groups as never-takers (NNNN), *eager* compliers (NTTT), Z_1 compliers (NNTT), Z_2 compliers (NTNT), *reluctant* compliers (NNNT), and always takers (TTTT). A Z_1 complier, for example, is treated if and only if $Z_1 = 1$, regardless of the value of Z_2. A Z_2 complier, for example, is treated if and only if $Z_2 = 1$, regardless of the value of Z_1. A reluctant complier is "reluctant" in the sense that they require both instruments equal to one to take treatment, while an eager complier receives treatment if either instrument is equal to one. Never and always takers are defined in the same way as they are under IAM.

In the language of a linear latent index model such as Eq (2), eager compliers would be units with $\min\{\beta_{1i}, \beta_{2i}\} \geq U_i$, while reluctant compliers would have

$$\max\{\beta_{1i}, \beta_{2i}\} < U_i \leq \beta_{1i} + \beta_{2i}$$

Z_1 compliers are those for whom $\beta_{1i} \geq U_i$ and $\beta_{2i} < U_i$, and Z_2 compliers have $\beta_{2i} \geq U_i$ but $\beta_{1i} < U_i$. Always takers have $U_i \leq 0$ and never takers have $\beta_{1i} + \beta_{2i} < U_i$.

3.2 With many binary instruments

Suppose now we have J binary instruments $Z_{1i} \dots Z_{Ji}$ (within this setup we can also accommodate discrete instruments – see Section 3.3). Now we wish to characterize the subset of the 2^{2^J} possible mappings between vectors of instrument values and treatment that satisfy VM. Given our normalization of $Z_j = 1$ as the "up" state of each instrument, 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, which are 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 (1988), only the first eight have been calculated numerically due to computational limitations. The Dedekind numbers clearly explode quite rapidly; however they do much more slowly than the total number 2^{2^J} of boolean functions of J variables, as shown in Table 1. 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.

We can now explicitly characterize the compliance groups that satisfy VM. For any value of J, there will be a never-takers group for which $D_i(z) = 0$ for all values $z \in \mathcal{Z}$. Apart from these never-takers, the compliance groups satisfying vector monotonicity can be associated with 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. The simplest such group is the group of always-takers, who require no instruments to equal one for the unit to take treatment. Another group, in a setting with three instruments, would be units that take

J	# possible	# vector-monotonic	% satisfying vector monotonicity
1	4	3	75%
2	16	6	37.5%
3	256	20	7.8%
4	65536	168	0.26%
5	4294967296	7581	0.000177%

Table 1: Number of possible compliance groups 2^{2^J} and number of compliance groups satisfying vector monotonicity, as a function of number J of binary instruments.

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 may take treatment only if $Z_1 = Z_2 = Z_3 = 1$.

By this 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). Each set S in a Sperner family F represents a minimal set of instruments for which the following is true: if all instrument numbers in S take a value of one, a unit in the compliance group $g(F)$ corresponding to F will take treatment. Since VM then implies that for any unit any superset $S' \supseteq S$, units in $g(F)$ will also take treatment if all instruments in S' take a value of one, we need only consider Sperner families defined on $\{1, 2, \ldots J\}$: the other families would be redundant. For notation, I refer to the compliance group associated with any Sperner family F as $g(F)$, and the Sperner family associated with any compliance group $g \in \mathcal{G}$ (under VM) as $F(g)$.

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 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"
$\{12\}$	"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:

$$\{\emptyset\}, \{1\}, \{2\}, \{3\},$$
$$\{12\}, \{13\}, \{23\}, \{123\},$$
$$\{\{1\}, \{2\}\}, \{\{2\}, \{3\}\}, \{\{1\}, \{3\}\}, \{\{1\}, \{2\}, \{3\}\},$$
$$\{\{12\}, \{3\}\}, \{\{13\}, \{2\}\}, \{\{23\}, \{1\}\},$$
$$\{\{12\}, \{13\}\}, \{\{12\}, \{23\}\}, \{\{13\}, \{23\}\},$$
$$\{\{12\}, \{13\}, \{23\}\}$$

For instance, an individual with G_i corresponding to $\{\{12\}, \{13\}, \{23\}\}$ takes treatment so long as any two instruments take the "on" value. Note that the number of Sperner families with $J = 3$ is one less than the third Dedekind number, which is 20. This true for any J: $|\mathcal{G}| = \mathcal{D}_J + 1$ under VM, where $\mathcal{D}_J$ is the J^{th} number in the Dedekind sequence.

3.3 Vector monotonicity with binary instruments from vector monotonicity with discrete instruments

The preceding section considered a case with any number of instruments J but where all instruments are binary. Below I note that this is without loss of generality in the sense that any set of discrete instruments with finite support satisfying vector monotonicity can be represented as a set of binary instruments that satisfy vector monotonicity:

Proposition 2. *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. If the vector $Z = (Z_1, \dots Z_J)$ satisfies Assumption VM then so does the vector $(\tilde{Z}_2, \dots, \tilde{Z}_M, Z_2, \dots Z_J)$ where $\tilde{Z}_{mi} := \mathbb{1}(Z_{1i} \geq z_m)$, between any two points in the support of $(\tilde{Z}_2, \dots, \tilde{Z}_M, Z_2, \dots Z_J)$ that occur with positive probability.*

Proof. See Appendix A. □

Proposition 2 offers a fairly general recipe for mapping the instruments available in a given empirical setting to our framework of binary instruments. This introduces support restrictions ($\mathcal{Z}$ is no longer rectangular), but the methods in the foregoing sections can still be applied to the transformed set of instruments expressed as binary variables. Provided that the researcher is willing to approximate continuous instruments by finite discrete instruments, this can be applied to continuous instruments. In practice, the number of instruments and number of levels per instrument may be limited by computational constraints, as will be discussed in Section 5.3 (though this problem is not worse than flexible 2SLS with discrete instruments).

4 Parameters of interest and identification

4.1 A class of conditional average treatment effects

We will consider as parameters of interest conditional average treatment effects:

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

where membership in the conditioning set $C_i = 1$ is determined by a unit's compliance group and realization of the instruments: $C_i = c(G_i, Z_i)$, where $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ is a function satisfying certain properties. Intuitively, $c(g, z) = 1$ will correspond to being a type of "complier" in the population. Without loss, this setup also nests functions of the triple (G_i, Z_i, D_i), since treatment status is a deterministic function of G_i and Z_i.

Using that $c(g, Z_i) \perp (Y_i(1), Y_i(0), G_i)$ by independence, and the law of iterated expectations, we can see that parameters of the form Δ_c are convex combinations of group-specific average treatment effects $\Delta_g := E[Y_i(1) - Y_i(0)|G_i = g]$:

$$\begin{aligned}\Delta_c &= \sum_{g \in \mathcal{G}} P(G_i = g|C_i = 1)E[Y_i(1) - Y_i(0)|G_i = g, C_i = 1] \\ &= \sum_{g \in \mathcal{G}} \left\{ \frac{P(G_i = g)P(C_i = 1|G_i = g)}{E[P(C_i = 1)]} \right\} \cdot \Delta_g \end{aligned} \tag{4}$$

where the weight for group g is proportional to $P(G_i = g)P(C_i = 1|G_i = g)$, and it can be readily verified that the weights sum to one by the law of total probability.[4]

Identification will require that Δ_c places no weight on always-takers or never-takers:

$$P(c(a.t., Z_i) = 1) = P(c(n.t., Z_i) = 1) = 0$$

These two groups can always be defined by the conditions $\{\max_{z \in \mathcal{Z}} D_i(z) = 0\}$ and $\{\min_{z \in \mathcal{Z}} D_i(z) = 1\}$, respectively, for instruments with finite support $\mathcal{Z}$. Absent additional assumptions, we cannot hope to learn about treatment effects among always-takers and never-takers, since the instruments do not provide any variation in treatment takeup for individuals in these groups.[5] Let $\mathcal{G}^c := \mathcal{G}/\{a.t., n.t.\}$ denote the remaining set of compliance groups compatible with Assumption VM.

4.2 Leading examples

My main identification result, Theorem 1, will be that Δ_c is identified for choices of the function $c(g, z)$ that satisfy a further property on how $E[c(g, Z_i)]$ is related across values

[4] It can be shown that if c is allowed to also depend on a random variable $\mathcal{E}_i$ that is independent of $(Y_i(0), Y_i(1), G_i, Z_i)$, for simplicity taken to be uniformly distributed on the unit interval, I can obtain any set of weights $\{w_g\}_{g \in \mathcal{G}^c}$ through a Δ_c by letting $c(g, z) = \mathbb{1}(\mathcal{E}_i \leq w_g)$ for each $g \in \mathcal{G}^c$. I omit this generality to keep things simple.

[5] Note that the same would also be true for any additional groups g for whom, given the distribution of Z_i, there is no actual variation in treatment status: i.e. $P(D_g(Z_i) = 1) = 0$ or $P(D_g(Z_i) = 1) = 1$, where $D_g(Z_i)$ is the selection function for compliance group g. In the baseline analysis, I will make assumptions that imply the set $\mathcal{N} := \{g \in \mathcal{G}/\{a.t., n.t.\}_{P(D_g(Z_i)=1) \in \{0,1\}}$ of such further groups is empty (see Assumption 3). In a more general setting, we would let $\mathcal{G}^c := \mathcal{G}/(\{a.t., n.t.\} \cup \mathcal{N})$ in the definitions below.

of $g \in \mathcal{G}^c$. This will include a number of easily interpretable parameters, and in this Section I give leading examples.

The Big LATE: Of particular interest is the "Big" LATE, which I define as the unweighted average treatment effect among all units whose G_i belongs to the set $\mathcal{G}^c$:

$$BLATE := E[Y_i(1) - Y_i(0)|G_i \in \mathcal{G}^c]$$

$BLATE$ is the ATE among all units that comply with *any* subset of the instruments. The Big LATE is "big" in the sense that it conditions on the largest subgroup of the population for which treatment effects can be point identified by instrumental variables methods. It corresponds to a choice of $c(g, z) = \mathbb{1}(g \in \mathcal{G}^c)$. In the returns to schooling example considered in the introduction, the BLATE is the average treatment effect among all individuals for whom $D_i(1, 1) > D_i(0, 0)$: that is, those who would go to college were it close and cheap, but would not were it far and expensive.

Big LATE on the treated: Another version of BLATE is the big local average treatment effect *on the treated* $BLATT := E[Y_i(1) - Y_i(0)|D_i = 1, G_i \in \mathcal{G}^c]$. Let $D_g(Z_i)$ be the selection function for compliance group g. The BLATT sets $c(g, z) = \mathbb{1}(g \in \mathcal{G}^c)\mathbb{1}(D_g(z) = 1)$. In the returns to schooling example, the BLATT would correspond to individuals who do go to college, but wouldn't have were it far and expensive.

Note that with a single binary instrument, $BLATT = BLATE$, because

$$E[Y_i(1)-Y_i(0)|D_i = 1, G_i = comp.] = E[Y_i(1)-Y_i(0)|Z_i = 1, comp.] = E[Y_i(1)-Y_i(0)|comp.]$$

where I have used the independence assumption. However, when the group $\mathcal{G}^c$ consists of more than one group, $BLATT$ generally differs from $BLATE$.

In Section 4.4 I show how the BLATE and BLATT lead to bounds on the ATE and ATT, respectively, when potential outcomes are bounded.

Set LATEs: Consider any (non-empty) subset $\mathcal{J} \subseteq \{1 \dots J\}$ of the instruments. Define a class of "set local average treatment effect" parameters $SLATE_{\mathcal{J}}$ that capture the average treatment effect among units that "comply" when all instruments in the set $\mathcal{J}$ are moved from 0 to 1, with the other instruments fixed at their realized values:

$$SLATE_{\mathcal{J}} = E\left[Y_i(1) - Y_i(0)\,|D_i((1 \dots 1), Z_{-\mathcal{J},i}) > D_i((0 \dots 0), Z_{-\mathcal{J},i})\right]$$

Here the notation $D_i((d \dots d), Z_{-\mathcal{J},i})$ indicates that $z_j = d$ for all $j \in \mathcal{J}$ and $z_j = Z_{ji}$ for all $j \notin \mathcal{J}$. $SLATE_{\mathcal{J}}$ sets $c(g, z) = \mathbb{1}\left(D_i((1 \dots 1), Z_{-\mathcal{J},i}) > D_i((0 \dots 0), Z_{-\mathcal{J},i})\right))$. The special case when $\mathcal{J} = \{j\}$, a singleton, yields a "Single instrument LATE": $E[Y_i(1) - Y_i(0)|D_i(1, Z_{-j,i}) > D_i(0, Z_{-j,i})]$.[6] The BLATE is also a special case of SLATE, with

[6]Note that the single-instrument LATE for instrument j is not generally identified by a "partial-Wald" estimand $C(Y, Z_j)/C(D, Z_j)$ when the instruments are not independent, a property pointed out by Heckman (2010). If on the other hand the instruments are independent of one another, using 2SLS may be justified, as we show in the Supplemental Material.

$\mathcal{J} = \{1, 2, \ldots J\}$ (all of the instruments).

In the returns to schooling example considered in the introduction, the only set LATEs are the BLATE $\mathcal{J} = \{1, 2\}$ and the LATE's corresponding to each instrument on it's own: for example $SLATE_{\{1\}}$ is the average treatment effect among individuals who don't go to college if it is far, but do if it is close. Note however that $SLATE_{\{1\}}$ does not generally correspond to using Z_1 alone as an instrument, since this latter estimand does not control for variation in Z_2 that is correlated Z_1 (see footnote 6).

For a discrete instrumental variable mapped to multiple binary instruments by Proposition 2, 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 $\mathcal{J} = \{\tilde{Z}_3, \tilde{Z}_4\}$.

4.3 Identification

Throughout this section, I will continue to focus on a setup with J binary instruments (see Proposition 2 for a discussion of mapping finite, discrete instruments to this setting).

In order to state the main identification result, it will be necessary to introduce some notation describing the way that the selection functions $D_g(Z_i)$ relate to one another across values of g. This will allow us to characterize the class of causal parameters that are point identified under VM, which includes the leading examples from the preceding section.

In particular, I begin with the observation that, under VM with binary instruments, the selection functions $D_g(z)$ associated with the various $g \in \mathcal{G}^c$ are not all linearly independent from one another. In fact, the vector space of functions $\{D_g(z)\}_{g \in \mathcal{G}^c}$ under VM maintains the same dimension as it does under IAM: $2^J - 1$ (this despite the fact that $\mathcal{G}^c$ is generally much larger under VM, growing with the Dedekind numbers $\mathcal{D}_J$). A natural basis for the $\{D_g(z)\}_{\mathcal{G}^c}$ under VM is formed by products of the instruments $Z_S := \prod_{j \in S} Z_j$, where $S \subseteq \{1 \ldots J\}, S \neq \emptyset$. I shall refer to the associated compliance groups $g(S)$ as *simple* compliance groups. The simple compliance groups correspond to Sperner families that consist of a single set: S.

For $J = 2$, this product basis consists of the three functions

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

while the remaining group in $\mathcal{G}^c$, eager compliers, can be obtained as:

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

In this $J = 2$ case, we can express the linear dependency by the the matrix M_J in the

system of equations:

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

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

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

for all $z \in \mathcal{Z}$, where $\{D_g(z)\}_{g\in\mathcal{G}^c}$ and $\{D_g(z)\}_{F(g)=\{S\}_{S\subseteq\{1\dots J\}},S\neq\emptyset}$ are understood as vectors in $\mathbb{R}^{|\mathcal{G}^c|}\mathbb{R}^{|2^J-1|}$, respectively, for each $z \in \mathcal{Z}$. The following Lemma gives the entries of M_J.

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

Proof. See Appendix A. □

The explicit form of M_J is useful for the proofs, but is not important from the standpoint of estimation. The matrix M_J for $J = 3$, which has $\mathcal{D}_3 - 2 = 18$ rows and $2^3 - 1 = 7$ columns is given explicitly in the Supplemental Material.

Now consider the causal parameter Δ_c for a given choice of c. I will say that Δ_c satisfies "Property M" if the conditional probabilities $P(C_i = 1|G_i = g) = E[c(g, Z_i)]$ for all $g \in \mathcal{G}^c$ combine linearly from the simple compliance groups in the same way as the $D_g(z)$ do; that is, according to the matrix M_J:

Definition (Property M). *The conditional average treatment effect Δ_c with binary random variable $C_i = c(G_i, Z_i)$ satisfies Property M if for all $g \in \mathcal{G}^c$:*

$$P(C_i = 1|G_i = g) = \sum_{S\subseteq\{1\dots J\},S\neq\emptyset} [M_J]_{F(g),S} \cdot P(C_i = 1|F(G_i) = S)$$

The following theorem establishes that causal parameters that satisfy Property M are identified under VM with binary instruments, so long as the instruments provide sufficient independent variation in treatment uptake. Define the vector $\Gamma_i = (Z_{S_1 i} \dots Z_{S_k i})$ for an arbitrary indexing of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$.[7] Let Σ be the variance-covariance matrix of Γ_i.

Assumption 3 (instrument non-degeneracy). Σ *has full rank.*

[7] When employing the mapping from discrete to binary instruments from Section 3.3, any S that contain more than one instrument $\tilde{Z}$ associated with the same original discrete instrument Z_j would be omitted from the definition of Γ_i. Their inclusion would necessarily violate Assumption 3, but as we show in Appendix B, this does not hamper identification.

Theorem 1. *Under Assumptions 1-3 (independence & exclusion, VM, and instrument non-degeneracy),* $\Delta_c = E[Y_i(1) - Y_i(0)|C_i = 1]$ *is identified for any* c *satisfying Property M as:*

$$\Delta_c = \frac{E[h(Z_i)Y_i]}{E[h(Z_i)D_i]}$$

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

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

Furthermore, BLATE, BLATT, and $SLATE_{\mathcal{J}}$ *all satisfy Property M.*

Proof. See Appendix A. □

Note that Theorem 1 expresses Δ_c in terms of the joint distributions $\mathcal{P}_{YZ}$ and $\mathcal{P}_{DZ}$: thus identification holds in a "split-sample" setting in which Y_i and D_i are not linked in the same dataset.

I now give some intuition for how Theorem 1 is established, before stating extensions.

Discussion of Theorem 1

To motivate Theorem 1, consider what is identified by a class of "2SLS-like" estimands where in a first stage, a single scalar instrument $h(Z_i)$ is constructed from the vector of instruments Z_i. Then, $h(Z_i)$ is used in simple linear IV regression: $\rho_h := \frac{C(Y_i,h(Z_i))}{C(D_i,h(Z_i))}$ where for random variables X and Y, $C(X, Y)$ denotes their covariance. Assuming only exclusion and independence (i.e. without any monotonicity assumptions) the following Lemma allows us to express ρ_h in terms of treatment effects:

Lemma 2. *Under Assumption 1 (exclusion and independence):*

$$\rho_h = \sum_g \frac{P(G_i = g)C(D_g(Z_i), h(Z_i))}{\sum_{g'} P(G_i = g')C(D_{g'}(Z_i), h(Z_i))} \cdot \Delta_g$$

Proof. See proof of Theorem 1, or Supplemental Material for proof as stated here. □

Note that the construction in Theorem 1 with $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - E[\Gamma_i])$ is an example of such an estimand ρ_h, since $E[h(Z_i)] = 0$ (and hence the covariance is equivalent to the expectation of the product).[8]

Comparing the form of ρ_h with expression (4) for Δ_c, it is evident that we will have $\rho_h = \Delta_c$ if the function $h(\cdot)$ can be chosen such that $C(D_g(Z_i), h(Z_i)) = P(C_i = 1|G_i = g)$ for each compliance group $g \in \mathcal{G}^c$. Since the covariance operator is linear, this equality will be guaranteed for a choice of c that satisfies Property M, so long as it holds for the simple compliance groups $g(S)$. Assumption 3 guarantees this is possible: with Σ

[8] Other examples of ρ_h include: i) 2SLS, in which h is equal to the linear projection of D_i on Z_i; ii) the "regression on the propensity score" estimand ("fully saturated 2SLS"), which sets $h(z) = P(z)$, with $P(z) = E[D_i|Z_i = z]$ the propensity score function; and iii) single Wald ratios $\rho_{zw} = \frac{E[Y_i|Z_i=z]-E[Y_i|Z_i=w]}{E[D_i|Z_i=z]-E[D_i|Z_i=w]}$, which set $h(Z_i) = \frac{\mathbb{1}(Z_i=z)}{P(Z_i=z)} - \frac{\mathbb{1}(Z_i=w)}{P(Z_i=w)}$. In general, the weights $C(D_g(Z_i), h(Z_i))$ are not gauranteed to be positive for an arbitrary choice of $h(\cdot)$.

invertible, any desired vector of the covariances $(D_g(Z_i), h(Z_i))$ for the simple compliance groups will be attainable by judicious choice of h (see Appendix A for further discussion).

Assumption 3 is in fact 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 need not be tuned independently. This is fortunate, since Assumption 3 rules out cases such as $(Z_{4i} = 1) \implies (Z_{3i} = 1)$ (with the implication that $P(Z_{3i}Z_{4i} = Z_{4i}) = 1$), which can be expected when using the construction in Proposition 2 to map discrete into binary instruments. In Appendix B, I give a version of Assumption 3 and generalization of the theorem below that can accommodate such instrument support restrictions. For ease of exposition, however, I here state the identification result under Assumption 3, which represents the simplest case.

What remains to be shown in Theorem 1 is that the parameters $BLATE$, $BLATT$ and $SLATE_{\mathcal{J}}$ all satisfy Property M. The reader is referred to Appendix A for details.

Extensions of Theorem 1

1) The size of the relevant sub-population is also 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 ρ_h:

Corollary to Theorem 1. *Make Assumptions 1-3. For any $C_i = c(G_i, Z_i)$ 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 A. □

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. I now show that any Δ_c that is identified by a broad set of empirical estimands must satisfy Property M. In this sense, Property M is also a *necessary* condition for identification. Specifically, we make the following definition:

Definition (expectation identification). *We say that a parameter θ is "expectation identified" if the set of values θ that are compatible with $E[Y_i|D_i = d, Z_i = z]$ for $d \in \{0, 1\}$, $z \in \mathcal{Z}$ and the joint distribution $\mathcal{P}_{DZ}$ of (D_i, Z_i) is a singleton, for all possible joint distributions of the latent variables $(G_i, Y_i(1), Y_i(0))$ (and hence values of θ).*

By writing $\rho_h = \sum_{z \in \mathcal{Z}} \frac{P(Z_i=z)h(z)}{E[h(Z_i)D_i]} \cdot E[Y_i|Z_i = z]$, it is clear that Δ_c is expectation identified under the assumptions of Theorem 1, since in that case $\Delta_c = \rho_h$ with $h(z)$ an identified function of z from $\mathcal{P}_{DZ}$. In Appendix Proposition 7, I show that expectation identification is equivalent to being point identified by a set of the so-called "IV-like estimands" introduced by Mogstad et al. (2018), as well as being equal to a single moment of a known or identified function of the i.i.d. triples (D_i, Z_i, Y_i).

Proposition 3. *Suppose Δ_c is expectation identified. Then Δ_c satisfies Property M.*

Proof. See Appendix A. □

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. Indeed, in the absence of traditional IAM, Theorem 1 cannot hold under PM alone, at least for all J:

Proposition 4. *When $J = 2$, if PM holds but neither VM nor IAM hold, BLATE is not expectation identified.*

Proof. See Appendix A. □

4) Identification of BLATE from a single Wald ratio: In the case of the Big LATE, identification can be demonstrated in a much simpler way than through Theorem 1. Under VM with finite discrete instruments, BLATE will always be equal to the Wald ratio $\rho_{\bar{Z},\underline{Z}} := \frac{E[Y_i|Z_i=z]-E[Y_i|Z_i=w]}{E[D_i|Z_i=z]-E[D_i|Z_i=w]}$ between $\underline{Z}$ and $\bar{Z}$, with $\underline{Z}$ and $\bar{Z}$ the "lowest" and "highest" instrument values, in a vector sense, for which Z_i has support. In the case at hand. with binary instruments having full support: $\bar{Z} = (1, 1, \dots 1)'$ and $\underline{Z} = (0, 0, \dots 0)'$. However, the result holds more generally with finite instrument support:

Proposition 5. *Let Assumption 1 and VM or IAM hold with finite $|\mathcal{Z}|$. Then $BLATE = \rho_{\bar{Z},\underline{Z}}$, 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 $BLATE = E[\Delta_i|G_i \in \mathcal{G}^c]$ for $\mathcal{G}^c := \{g \in \mathcal{G} : E[D_g(Z_i)] \in (0, 1)\}$.*

Proof. See Appendix A □

Thus, identification of BLATE is somewhat immediate: 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.[9] However, with Theorem 1 we are able to identify a much larger class of parameters than BLATE (for example BLATT and SLATEs do not appear to have single Wald counterparts). Furthermore, the expression for Δ_c in Theorem 1 suggests a facilitates improving estimation through a regularization procedure. In Section 5 I show that this regularization can be important when the number of observations in $\underline{Z}$ and $\bar{Z}$ is not large. The finite sample equivalence between sample analogs of $\rho_{\bar{Z},\underline{Z}}$ and ρ_h is also examined in Section 5.

5) 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 C, we show that covariates can be accommodated non-parametrically, or in a parsimonious way under additional assumptions in the Supplemental Material.

[9] Frölich 2007 has a discussion of using $\rho_{\bar{Z},\underline{Z}}$ in the case of IAM.

4.4 Identified sets for ATE and ATT

One drawback of the identification results in the preceding section is that since the class Δ_c explicitly excludes never-takers and always-takers, all such parameters depend 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 $BLATE$ and $BLATT$ can be used to generate sharp worst-case bounds in the spirit of Manski (1990) for the unconditional average treatment effect (ATE) and average treatment effect on the treated (ATT), respectively. Suppose that $Y_i(d) \in [\underline{Y}, \bar{Y}]$ with probability one, for each $d \in \{0, 1\}$. Then

$$ATE := E[Y_i(1) - Y_i(0)] = p_a \Delta_a + p_n \Delta_n + (1 - p_t - p_a) BLATE$$

where $p_a = E[D_i | Z_i = \underline{Z}]$, $p_n = E[1 - D_i | Z_i = \bar{Z}]$. Thus, under the support restriction that $Y_i(d) \in [\underline{Y}, \bar{Y}]$:

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

since $\Delta_a, \Delta_n \in [-(\bar{Y} - \underline{Y}), +(\bar{Y} - \underline{Y})]$. Note that the probabilities p_n and p_a are point identified.

We can place similar identified bounds on the ATT. Using that $P(G_i \in A | D_i = 1) = \frac{P(G_i \in A, D_i = 1)}{P(D_i = 1)} = \frac{p_a}{P(D_i = 1)}$ and $P(G_i \in N | D_i = 1) = 0$, I have that

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

Note that under the bounded support condition the ATE and ATT can be partially identified whenever their conditional analogs are identified for some subgroup of the population, where the size of that subgroup is also identified. The BLATE and BLATT are unique in providing the largest such subgroup that does not require making assumptions about treatment effects for never-takers and always-takers, thus providing the narrowest possible bounds under that constraint.

5 Estimation

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 certain 2SLS-like estimands ρ_h. As these are defined as the ratio of two covariances, a natural plug-in estimator simply replaces these with their sample counterparts, assuming $h(Z_i)$ is a strong enough instrument to avoid weak identification issues. In particular, following the choice $h(Z_i) = \lambda' \Sigma^{-1} (\Gamma_i - E[\Gamma_i])$, 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 de-meaned Γ_j as columns, where recall that $\Gamma_i = (Z_{S_1 i} \dots Z_{S_k i})$ for an arbitrary indexing of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$, thus $\tilde{\Gamma}_{ij} = Z_{S_j i} - \hat{E}[Z_{S_j i}]$ for the j^{th} subset $S_j \subseteq \{1 \dots J\}$. We assume existence of $n(\tilde{\Gamma}'\tilde{\Gamma})^{-1}$ in finite sample, and note that it is guaranteed in the asymptotic limit by Assumption 3. $\hat{\lambda}$ will be an 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.

Note that via Assumption 3 we've assumed that the Z_{Si} for all $S \subseteq \{1 \dots J\}$ are linearly independent, meaning that $\tilde{\Gamma}$ has full column rank. In the Supplemental Material, I show that we can handle the linear dependencies introduced by the construction of Proposition 2 that maps discrete to binary instruments. In practice, using this construction requires that for each discrete instrument mapped into several binary instruments, we drop from Γ_i any product of the final binary instruments that contain distinct binary instruments from the same original discrete instrument.

In general, let $\mathcal{F}$ denote the set of S represented in the vector Γ_i (i.e. $\mathcal{F} := \{S \subseteq \{1, 2, \dots J\}, S \neq \emptyset\}$) in the baseline case. Since each column of $\tilde{\Gamma}$ sums to zero, $\hat{C}(H_i, \tilde{\Gamma}_{ij}) = \hat{E}[H_i, \tilde{\Gamma}_{ij}]$ and:

$$(\hat{E}[\hat{H}_i, \tilde{\Gamma}_{i1}], \hat{E}[\hat{H}_i, \tilde{\Gamma}_{i2}], \dots, \hat{E}[\hat{H}_i, \tilde{\Gamma}_{i|\mathcal{F}|}])' = \frac{1}{n}\tilde{\Gamma}'\hat{H} = \tilde{\Gamma}'\tilde{\Gamma}(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\hat{\lambda} = \hat{\lambda},$$

i.e. $\hat{H}$ tunes the covariance between $\hat{H}_i$ and Z_{Si} to be exactly $\hat{\lambda}_S$ in finite sample.

Given $\hat{H}_i$ as defined above, consider the plug-in estimator

$$\hat{\rho} = (\hat{H}'D)^{-1}(\hat{H}'Y) = \left\{\hat{\lambda}'(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\tilde{\Gamma}'D\right\}^{-1}\hat{\lambda}'(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\tilde{\Gamma}'Y$$

where Y and D are $n \times 1$ vectors of observations of Y_i and D_i, respectively. Noticing that for any $V \in \mathbb{R}^n$, $(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\tilde{\Gamma}'V$ is the sample linear projection coefficient of V on the de-meaned Γ_j, we can re-express it as $(0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'V$ where $\Gamma = [1, \Gamma_1, \dots \Gamma_{|\mathcal{F}|}]$ skips the demeaning of each vector $\Gamma_S = (Z_{S1} \dots Z_{Sn})$ and adds a column of ones. The estimator can now be written as $\hat{\rho} = \hat{\rho}(\hat{\lambda})$, where

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

We note that the estimator $\hat{\rho}(\lambda)$ 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, such an estimator can be written as $\hat{\rho}_{2sls} = (D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'D)^{-1}D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'Y$. The key difference, then 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 the weights $(0, \lambda')$ chosen to match the desired parameter of interest.

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

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

Matching the RHS to particular estimands Δ_c that satisfy Property M is achieved by choosing $\hat{\lambda}$. We now consider the estimation of λ. For the BLATE:

$$\hat{\rho}((1,1,\dots 1)') \xrightarrow{p} \sum_{g \in \mathcal{G}^c} \frac{P(G_i = g)}{P(G_i \in \mathcal{G}^c)} \cdot \Delta_g = BLATE \tag{5}$$

and thus no first-step is required $\hat{\lambda}$ does not depend on the data.

For the BLATT (written here in the case of full instrument support):

$$\hat{\rho}(\hat{E}[Z_1], \hat{E}[Z_2], \dots, \hat{E}[Z_1 Z_2 \dots Z_J]) \xrightarrow{p} E[Y_i(1) - Y_i(0) | D_i = 1, G_i \in \mathcal{G}^c] = BLATT$$

and for $SLATE_j$:

$$\hat{\rho}(\hat{\lambda}_1^{SLATE_{\mathcal{J}}} \dots \hat{\lambda}_{|\mathcal{F}|}^{SLATE_{\mathcal{J}}}) \xrightarrow{p} E[Y_i(1) - Y_i(0) | D_i((1\dots 1), Z_{-\mathcal{J},i}) > D_i((0\dots 0), Z_{-\mathcal{J},i})] = SLATE_{\mathcal{J}}$$

where $\hat{\lambda}_S^{SLATE_{\mathcal{J}}} = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S-\mathcal{J},i} = 1\right)$, where $S - \mathcal{J}$ is the set difference $\{j : j \in S, j \notin \mathcal{J}\}$.

5.2 Regularization of the estimator

Recall that for the BLATE there is a natural alternative Wald estimator:

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

where $\bar{Z} = (111\dots 1)'$ or $\underline{Z} = (000\dots 0)'$. Though it is not obvious, it turns out that $\hat{\rho}_{\bar{Z},\underline{Z}}$ and $\hat{\rho}((1,1,\dots 1)')$ in Equation 5 are numerically equivalent. 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.

However, in situations where there is non-zero but small support on the points $\bar{Z}$ and $\underline{z}$, we can expect that $\hat{\rho}_{\bar{Z},\underline{Z}}$ may perform quite poorly as an estimator of BLATE in finite samples, since it effectively ignores all of the data for which $z \notin \{\underline{Z}, \bar{Z}\}$. 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 BLATE with finite $\mathcal{Z}$ (see Proposition 5).

In the case of VM, the equivalence between $\hat{\rho}_{\bar{Z},\underline{Z}}$ and $\hat{\rho}((1,1,\dots 1)')$ suggests a straightforward way to address this problem. When there are few observations in the points $\bar{Z}$ and $\underline{z}$, the finite sample matrix Γ will have singular values that are close to zero. To mitigate the effects of this, I allow a sequence of ridge regularization parameters α_n in the estimator, by making the replacement:

$$(\Gamma'\Gamma)^{-1}\Gamma' \to (\Gamma'\Gamma + \alpha_n I)^{-1}\Gamma'$$

Using this regularized version of $(\Gamma'\Gamma)^{-1}\Gamma'$ has the effect of establishing a floor on the singular values of Γ. In doing so, a choice of $\alpha > 0$ allows the estimator to use the full dataset, at the expense of some bias. Proposition 6 below yields a means of navigating this tradeoff to choose α in practice. In the simulations of Section 5.4, I show the practical importance of the issue, and how regularization helps.

With this generalization, I let the final class of estimators be:

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

with $\hat{\lambda}$ defined by the particular parameter of interest. Consistency and asymptotic normality of the estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ now follows in a straightforward way from the results thus far. In particular, the asymptotic variance can be computed as a special case of Theorem 3 in Imbens and Angrist (1994). In our case, 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 $\hat{\rho}(\lambda, \alpha) = \hat{C}(g(Z_i, \hat{\theta})Y_i)/\hat{C}(g(Z_i, \hat{\theta})D_i)$, 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. *Under Assumptions 1-3, 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 $P(C_i = 1|G_i = g(S_l)) = E[c_l(Z_i)]$).

Recall that under Assumption 3, $\mathcal{F}$ is the set of all non-empty subsets of instrument indices $\{1 \dots J\}$ (i.e. $\Gamma_i = (Z_{1i}, Z_{2i}, \dots, Z_{1i}Z_{2i} \dots Z_{Ji})$).

Proof. See Appendix A. □

As the expression for the asymptotic variance is messy, I leave its estimation by a plug-in estimator implicit.

5.3 Practical considerations

Computational cost: The estimator $\hat{\rho}(\lambda, \alpha)$ requires generalized inversion of an $n \times (|\mathcal{F}|+1)$ matrix Γ, where $|\mathcal{F}|$ is typically (and at most) $2^J - 1$, possibly preceded by computation of the λ_n. Aside from computation of the λ_n, the overall cost of computation is the same as would be required to calculate such "fully saturated" 2SLS via separate evaluation of the first stage and the reduced form.

Choice of α: 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 6. *Under the assumptions of Theorem 2, $E[(\hat{\rho}(\lambda, \alpha) - \Delta_c)^2 | Z_1 \dots Z_n]$ is, up to second order in estimation error and a positive constant of proportionality:*

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

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

Proof. See Appendix A □

The above expression can be consistently estimated by an iterative algorithm in which Δ, β_Y and β_D are estimated starting with $\alpha = 0$, and then the expression minimized over α to choose α for the the next iteration. This is implemented for the simulations in Section 5.4. Choosing α_n to minimize the expression in Proposition 6 ensures that although regularization $\alpha > 0$ introduces bias, it will decrease overall estimation error (as measured by the MSE), at least in large enough samples.[10]

5.4 Simulation evidence

In this section, I conduct an estimation study of $\hat{\rho}(\lambda_n, \alpha_n)$ for the estimation of BLATE. In particular, we want to compare it to the simple Wald estimator, given in Equation 6.

[10] In the proof of Proposition 6, I derive an expression for the conditional MSE of a more general type of ridge regularization, which may offer better performance if J is small enough to allow optimization over a $2^J - 1$ dimensional parameter space.

For this I 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 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: $BLATE = 10$.

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 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 5.3, to the simple Wald estimator of BLATE $(\hat{E}[Y_i|Z_i = (111)] - \hat{E}[Y_i|Z_i = (000)])/(\hat{E}[D_i|Z_i = (111)] - \hat{E}[D_i|Z_i = (000)])$, 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 BLATE (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 2 shows the results for the first DGP, where the Z_j are independent Bernoulli random variables with mean 1/2. We see that with given 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, converges to small values. Hence the "regularized SVD" estimator $\hat{\rho}((1, 1, \ldots 1)', \hat{\alpha}_{mse})$ is very close to the Wald estimator (recall that they are numerically 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 3 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. The first panel shows when the MSE is chosen by the iterative procedure to minimize MSE. Here, the improvement is appreciable, but not huge. In the second panel, α is arbitrarily set to 100. We see that this delivers a significantly lower RMSE than does the data-driven value of α. This underlines that the MSE-minimizing procedure does rely on a finite sample approximation, and is thus imperfect in delivering the optimal finite sample value of α.

We note that in both Figures 2 and 3, fully saturated 2SLS (regresion on the propensity score) performs well, in some cases actually outperforming both of the alternative

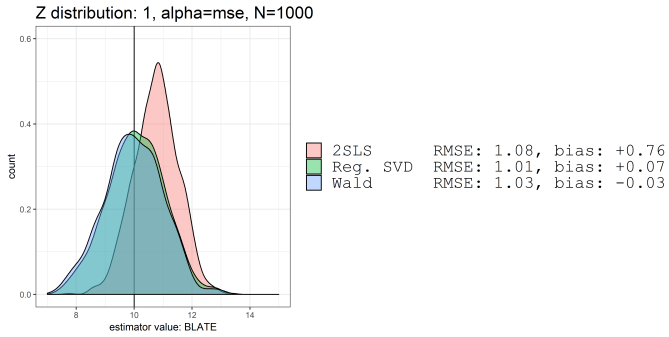

Figure 2: Monte Carlo distributions of estimators, for the first DGP (Z uncorrelated coin tosses) with three binary instruments. "Reg. SVD" indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of BLATE.

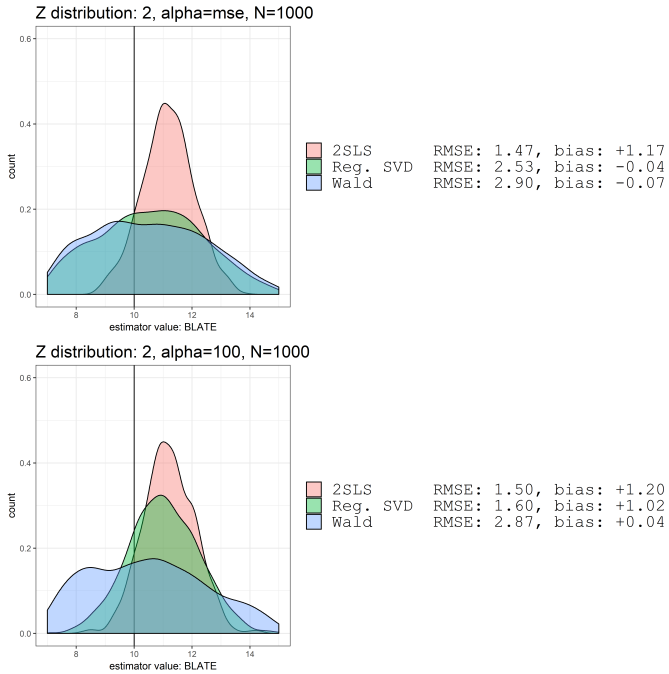

Figure 3: Monte Carlo distributions of estimators, for the first DGP ($P(Z_{3i}|Z_{2i} = 1) = 0.05$) with three binary instruments. "Reg. SVD" indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of BLATE.

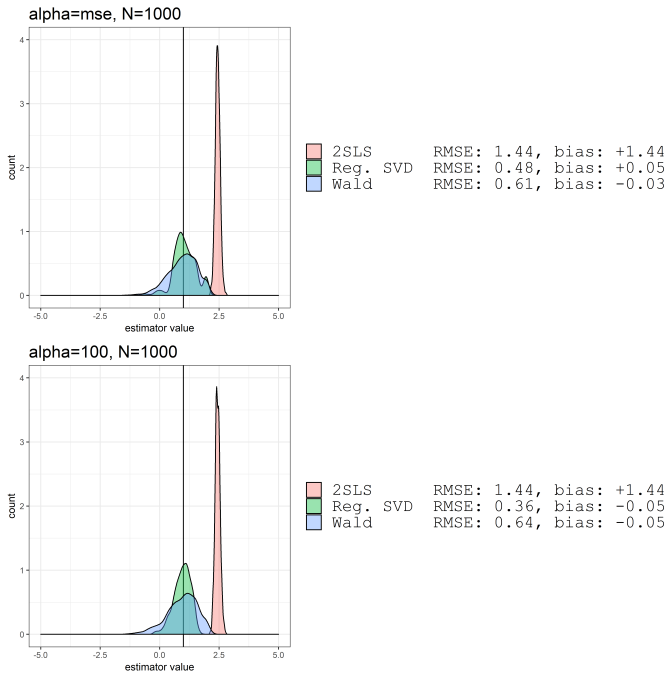

Figure 4: Monte Carlo distributions of estimators, for the two-instrument DGP. "Reg. SVD" indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of BLATE.

estimators. This is despite the fact that it is not consistent for the $BLATE$, 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 $J = 2$.

For this secondary 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 $BLATE$ of unity. I generate negatively correlated instruments (with correlation of about $-.1$) by slicing up 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 4 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.

6 Conclusion

In this paper I have considered a revised "vector monotonicity" assumption for cases when a researcher has multiple valid instrumental variables for the same binary treatment, a situation in which the traditional LATE monotonicity assumption may be hard to defend. This vector monotonicity assumption has also been recently discussed (as "Assumption AM") by Mogstad et al. (2019). I have focused on the implications of vector monotonicity for identification and estimation of causal parameters of interest, providing positive point identification results under VM.

When the researcher encodes their instrument vector as several binary instruments, I have shown that a class of weighted averages of causal effects are point identified under VM, and can be estimated by a 2SLS-like estimator that is consistent and asymptotically normal. Under support restrictions on the outcome variable, these identification results also yield sharp bounds on the average treatment effect and average treatment on the treated. In the Supplemental Appendix, I also consider auxillary conditions under which the common procedure of estimating 2SLS yields a convex combination of causal effects under vector monotonicity. These conditions are restrictive, so the identification results and proposed estimator from this paper may be of use when researchers beeive that VM is a good assumption for the context at hand.

References

David Card. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp*, 1995.

Markus Frölich. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75, 2007. ISSN 03044076. doi: 10.1016/j.jeconom.2006.06.004.

David; Gale, Harold; Kuhn, and Albert; Tucker. *Linear Programming And The Theory Of Games*. 1951.

James J Heckman. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, 48(2):356–398, 2010. ISSN 00220515. doi: 10.1257/jel.48.2.356.

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

James J. Heckman and Edward J Vytlacil. Local Instrumental Variables. *Nonlinear Statistical Modeling*, 2001.

James J. Heckman, Sergio Urzua, and Edward Vytlacil. Understanding What Instrumental Variables Estimate in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 3, 2006.

Arthur Hoerl and Robert Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 1970.

Hidehiko Ichimura and T. Scott Thompson. Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295, 1998.

Guido W Imbens and Joshua D Angrist. Identification and Estimation of Local Average Treatment Effects Published by : The Econometric Society Stable URL : https://www.jstor.org/stable/2951620 to Econometrica. 62(2):467–475, 1994.

By Thomas J Kane and Cecilia Elena Rouse. Labor-Market Returns to Two- and Four-Year College: Is a Credit a Credit and Do Degrees Matter? *NBER Working Paper 4268*, 1993.

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

D. J. Kleitman and E. C. Milner. On the average size of the sets in a Sperner family. *Discrete Mathematics*, 6(2):141–147, 1973. ISSN 0012365X. doi: 10.1016/0012-365X(73)90043-5.

Charles F Manski. Nonparametric Bounds on Treatment Effects. *The American Economic Review*, 80(2):829–823, 1990. ISSN 0002-8282. doi: 10.2307/2006592.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. 86(5):1589–1619, 2018. doi: 10.3982/ECTA15463.

Magne Mogstad, Alexander Torgovitsky, and Christopher Walters. Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions. *Working Paper*, page Working paper, 2019.

Jack Mountjoy. Community Colleges and Upward Mobility. *Mimeo*, pages 1–83, 2018.

Edward Vytlacil. Independence , Monotonicity , and Latent Index Models : An Equivalence Result. *Econometrica: Notes and Comments*, 70(1):331–341, 2002.

Jianxin Yin, Zhi Geng, Runze Li, and Hansheng Wang. Nonparametric covariance model. *Statistica Sinica*, 20(1):469–479, 2010. ISSN 10170405.

Appendices

A Proofs

A.1 Proof of Proposition 1

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). $P(D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})) = 1$ *for all $j \in \{1 \dots J\}$, $z_j \geq z'_j \in \mathcal{Z}_j$ and $z_{-j} \in \mathcal{Z}_{-j}$*

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':

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

and thus:

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

A.2 Proof of Proposition 2

Since fixing Z_1 is equivalent to fixing $\tilde{Z}_2 \dots \tilde{Z}_M$, I only 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 $(z, \tilde{Z}_{-m}, Z_{-1})$ has positive probability for both $z \in \{0, 1\}$:

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

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 the positive probability condition to hold, it must be the case that $\tilde{Z}_{-m}$ is composed of all zeros for the first $m-1$ entries, and then ones for $m+1 \dots M$ (since $P(\tilde{Z}_{mi} = 0 \,\&\, \tilde{Z}_{ni} = 1) = 0$ when $m < n$). Thus, for any such $\tilde{Z}_{-m}$, switching $\tilde{Z}_m$ from one to zero corresponds to switching Z_1 from value z_m to value z_{m-1}, and

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

by vector monotonicity on the original vector $(Z_1 \dots Z_J)$.

A.3 Proof of Lemma 1

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

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

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

As an alternative, to the above expression, one can also pin down the matrix M_J by "growing" it inductively. Begin with the first $2^J - 1$ rows of the matrix of M_J, for the simple compliance groups. This sub-matrix is equal to the identity matrix. Now consider any Sperner family F and a subset S of $\{1\dots J\}$ not contained in F. Now we add the set S to the family of sets F. To distinguish between adding sets to a Sperner family from adding elements to the sets within a Sperner family, I introduce the notation that $\sqcup$ indicates inclusion of a new set among a family (while $\cup$ indicates taking the union of elements across sets.) If $F' = F \sqcup S$ is also a Sperner family, the associated row f' in M_J

can be constructed as follows. Begin with the row f associated with F. Set the entry in the column associated with S to one. Then, for each set S' in F, subtract the entry in the column associated with S' from the entry in the column associated with the union of the set S' and S.

A.4 Proof of Theorem 1

A.4.1 First part: Assumption 3 yields ability to tune covariances

Consider the quantity $E[Y_i D_i h(Z_i)]$ for an arbitrary function $h(\cdot)$ such that $E[h(Z_i)] = 0$. Using the law of iterated expectations, and the independence assumption:

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

where $D_g(z)$ denotes the selection function for compliance group g. A similar set of steps shows that $E[Y_i(1 - D_i)h(Z_i)] = \sum_g -P(G_i = g)E[Y_i(0)|G_i = g]E[D_g(Z_i)h(Z_i)]$ (using that $E[h(Z_i)] = 0$), and thus, combining:

$$E[Y_i h(Z_i)] = E[Y_i D_i h(Z_i)] + E[Y_i(1 - D_i)h(Z_i)] = \sum_g E[D_g(Z_i)h(Z_i)]\Delta_g$$

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 $D_{g(S)}(Z_i) = Z_{Si}$. We first show that for any $\lambda \in \mathbb{R}^{|\mathcal{F}|}$, Assumption 3 allows us to define an $h(Z_i)$ such that $E[Z_{Si}h(Z_i)] = \lambda_S$. Note that since $E[h(Z_i) = 0]$, this is the same as tuning each covariance $C(Z_{Si}, h(Z_i))$ to λ_S (c.f. Lemma 2).

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.

Given the above, the equality $\Delta_c = \rho_h$ now follows from Property M if we choose $\lambda_S = P(C_i = 1|G_i = g(S)) = E[c(g(S), Z_i)]$, since $E[D_g(Z_i)h(Z_i)]$ is linear in $D_g(Z_i)$. Note that the quantity $E[c(g(S), Z_i)]$ for each S can be computed from the observed distribution of Z_i.

To replace Assumption 3 with Assumption 3* from Appendix B, simply replace $\mathcal{F}$ as defined here with a maximal $\mathcal{F}$ from Assumption 3a*. For any subsets S of the instruments that are not in the set $\mathcal{F}$, we appeal to Assumption 3b*.

A.4.2 Second part: BLATE, BLATT and SLATE satisfy Property M

Now I show that the BLATE, BLATT and SLATEs all satisfy Property M. Since the BLATE is a special case of SLATE, it doesn't require separate treatment. However, since the proof for BLATE is much more straightforward, I include it here nonetheless. Recall that the conditional average treatment effect $E[Y_i(1) - Y_i(0)|C_i = 1]$ with $C_i = c(G_i, Z_i)$ satisfies Property M if

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

A.4.3 BLATE ($\mathbf{c(g,z) = \mathbb{1}(g \in \mathcal{G}^c)}$)

Corollary to Lemma 1. *For any J, the sum of the entries along any row of M_J is one.*

Proof. By induction using Lemma 1. If the row associated with F sums to one, then so does the row associated with $F \sqcup S$, if it is a Sperner family. The row corresponding to any singleton Sperner family (a simple compliance group) has one 1 and the rest zeroes. □

Since $P(C_i = 1|G_i = g) = 1$ for the g associated with each simple compliance group, the vector $M_J(1\ldots1)'$ will yields a sum across each row of M_J. By the Corollary, this recovers that $P(C_i = 1|G_i = g)$ for all $g \in \mathcal{G}^c$.

A.4.4 BLATT ($\mathbf{c(g,z) = \mathbb{1}(g \in \mathcal{G}^c)\mathbb{1}(D_g(z) = 1)}$)

Note that: $P(C_i = 1|G_i = g(F)) = P(D_{g(F)}(Z_i) = 1)$. The "inclusion-exclusion principle" states that for any finite set of events $A_1, A_2, \ldots A_n$:

$$P\left(\bigcup_{i\in\{1\ldots n\}}\right) = \sum_{f\subseteq\{1\ldots n\}} (-1)^{|f|+1} P\left(\bigcap_{i\in f} A_i\right)$$

Using that

$$\left\{D_{g(F)}(z)=1\right\} \iff \left\{\bigcup_{S\in F}\left\{D_{g(S)}(z)=1\right\}\right\}$$

and applying this principle we have:

$$\begin{aligned}
P(C_i=1|G_i=g(F)) &= \sum_{f\subseteq F}(-1)^{|f|+1}P\left(\bigcap_{S\in f}\left\{D_{g(S)}(Z_i)=1\right\}\right)\\
&= \sum_{f\subseteq F}(-1)^{|f|+1}P\left(D_{g\left(\bigcup_{S\in f}\right)}(Z_i)=1\right)\\
&= \sum_{S'}\sum_{\substack{f\subseteq F\\ \left(\bigcup_{S\in f}\right)=S'}}(-1)^{|f|+1}P\left(D_{g(S')}(Z_i)=1\right)\\
&= \sum_{S'}\left[\sum_{\substack{f\subseteq F\\ \left(\bigcup_{S\in f}\right)=S'}}(-1)^{|f|+1}\right]P(C_i=1|G_i=g(S'))
\end{aligned}$$

matching M_J in Lemma 1.

A.4.5 SLATE $(\mathbf{c}(\mathbf{g},\mathbf{z})=\mathbb{1}\left(\mathbf{D_i}((\mathbf{1}\dots\mathbf{1}),\mathbf{Z}_{-\mathcal{J},\mathbf{i}})>\mathbf{D_i}((\mathbf{0}\dots\mathbf{0}),\mathbf{Z}_{-\mathcal{J},\mathbf{i}})\right))$

Call i a $\mathcal{J}$-complier if $\mathbb{1}\left(D_i((1\dots1),Z_{-\mathcal{J},i})>D_i((0\dots0),Z_{-\mathcal{J},i})\right)$. We can write this condition as:

$$\left(Z_{Si}=0\ \forall\ S\in F(G_i)\text{ for which }\mathcal{J}\cap S=\emptyset\right)\text{ and }\left(\exists\ S\in F(G_i)\text{ for which }\mathcal{J}\cap S\neq\emptyset\text{ such that }Z_{(S-\mathcal{J}),i}=1\right)$$

where $S\in F$ denotes that the set S is a member of the family of sets F, and $S-\mathcal{J}$ denotes the complement of $\mathcal{J}$ within S. Being a $\mathcal{J}$-complier requires two things. First, if $Z_{Si}=1$ for any $S\in F(G_i)$ that contains no elements in common with $\mathcal{J}$, then unit i will be an "always-taker" with respect to shifting the Z_j for all $j\in\mathcal{J}$ from zero to one. Second, to avoid being a "never-taker" with respect to this transition, it must be the case that for some $S\in F(G_i)$, $Z_{ji}=1$ for all j in the complement of $\mathcal{J}$ within S (and that this complement is not null). Otherwise having $Z_j=1$ for all $j\in\mathcal{J}$ would not be sufficient for the unit to take treatment. If i satisfies both of these conditions, they are a $\mathcal{J}$-complier, since then $(D_i((1\dots1),Z_{-\mathcal{J},i})=1$ and $D_i((0\dots0),Z_{-\mathcal{J},i}))=0$.

Fixing the set $\mathcal{J}$ and an F, let F_0 be the set of $S'\in F$ such that $\mathcal{J}\cap S'=\emptyset$, and F_1 the set of $S'\in F$ such that $\mathcal{J}\cap S'\neq\emptyset$. As a further shorthand, let E_{Fi} indicate the event $c(g(F),Z_i)=1$. In the foregoing, we shall suppress the i index throughout for functions of realized Z_i. In this notation we can write the above characterization of $\mathcal{J}$-compliers as:

$$E_F=\begin{cases}\bigcup_{S'\in F_1}\left(1=\left[\prod_{S''\in F_0}(1-Z_{S''})\right]Z_{S'-\mathcal{J}}\right) & \text{if } F_1\neq\emptyset\\ \emptyset & \text{if } F_1=\emptyset\end{cases} \tag{8}$$

By the inclusion-exclusion principle, and then by expanding out the product of sums, we can express the probability of E_F as:

$$\begin{aligned}
P(E_F) &= \sum_{f\subseteq F_1}(-1)^{|f|+1}\cdot P\left(\bigcap_{S'\in f}\left\{1=\left[\prod_{S''\in F_0}(1-Z_{S''})\right]Z_{S'-\mathcal{J}}\right\}\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\cdot P\left(\left\{\prod_{S'\in f}Z_{S'-\mathcal{J}}\right\}\left\{\prod_{S''\in F_0}(1-Z_{S''})\right\}=1\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot\mathbb{E}\left[\left\{\prod_{S'\in f}Z_{S'-\mathcal{J}}\right\}\left\{\prod_{S''\in f'}Z_{S''}\right\}\right]\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot P\left(Z_{(\bigcup_{\{S'\in f\}}S'-\mathcal{J})\cup(\bigcup_{\{S''\in f'\}}S'')}=1\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot P\left(Z_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}-\mathcal{J}}=1\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot P\left(E_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}}\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot P\left(Z_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}-\mathcal{J}}=1\right)\\
&= \sum_{\substack{f\subseteq F_1\\ f\neq\emptyset}}(-1)^{|f|+1}\sum_{f'\subseteq F_0:}(-1)^{|f'|}\cdot P\left(E_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}}\right)\\
&= \sum_{f\subseteq F_1}(-1)^{|f|+1}\sum_{f'\subseteq F_0}(-1)^{|f'|}\cdot P\left(E_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}}\right)
\end{aligned}$$

where in the last line we use that if $f=\emptyset$, $P\left(E_{\{(\bigcup_{\{S'\in f\}}S')\cup(\bigcup_{\{S''\in f'\}}S'')\}}\right)=0$ for any $f'\subseteq F_0$.

Expressing $P(E_F)$ as a sum over distinct events:

$$P(E_F)=\sum_{S'}\left(\sum_{\substack{f\subseteq F_1,f'\subseteq F_0\\ (\bigcup_{\{S''\in f\cup f'\}}S'')=S'}}(-1)^{|f|+|f'|+1}\right)P(E'_S)=\sum_{S'}\left(\sum_{\substack{f\subseteq F\\ (\bigcup_{\{S''\in f\}}S'')=S'}}(-1)^{|f|+1}\right)P(E'_S)$$

We can see from the above that $P(E_F)=\sum_{S'}M_{FS'}P(E_{S'})$ where

$$M_{FS'}:=\sum_{\substack{f\subseteq F\\ (\bigcup_{\{S''\in f\}}S'')=S'}}(-1)^{|f|+1}$$

matching M_J as given by Lemma 1. Since this is true for any F, we have confirmed that $SLATE$ satisfies Property M.

A.5 Proof of the Corollary to Theorem 1

Using independence and Property M:

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

A.6 An Equivalence Result

The proofs of Proposition 3 and 4 will make use of the following equivalence result regarding expectation identification:

Proposition 7. *Let the support $\mathcal{Z}$ of the instruments be discrete and finite. Fix a function $c(g, z)$. Then the following are equivalent:*

1. *Δ_c is expectation identified*
2. *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 measurable functions $s(d, z)$, and $\beta_s := E[s(D_i, Z_i)Y_i]$*
3. *$\Delta_c = \beta_s$ for a single such $s(d, z)$*
4. *$\Delta_c = E[t(D_i, Z_i, Y_i)]$ with $t(d, z, y)$ a known or identified measurable function, for all joint distributions of $(Y_i(1), Y_i(0), G_i)$ and distributions of Z_i within some fixed class $\mathcal{P}$*

Proof. We can show each of the following implications:

- **1 → 2** Let $\mathcal{S} = \{s(d, z) = \mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)\}_{d\in\{0,1\},z\in\mathcal{Z}}$. Then each β_s is equal to $P(D_i = d, Z_i = z)E[Y_i|D_i = d, Z_i = z]$ for some d, z. The coefficient is known from $\mathcal{P}_{DZ}$, thus 1. is a case of 2.
- **1 → 3** Write any $E[Y_i|D_i = d, Z_i = z] = E[Y_i(d)|D_i = d, Z_i = z] = P(D_i = d|Z_i = z)^{-1}E[Y_i(d)\mathbb{1}(D_i = d)|Z_i = z] = P(D_i = d|Z_i = z)^{-1}\sum_g P(G_i = g|Z_i = z)E[Y_i(d)\mathbb{1}(D_i = d)|G_i = g, Z_i = z] = P(D_i = d|Z_i = z)^{-1}\sum_{g:D_g(z)=d} P(G_i = g)E[Y_i(d)|G_i = g]$, where we have used independence.

  To eliminate the coefficient, simply write: $E[Y_i\mathbb{1}(D_i = d)|Z_i = z] = \sum_{g:D_g(z)=d} P(G_i = g)E[Y_i(d)|G_i = g]$. If we stack the unknown quantities $P(G_i = g)E[Y_i(d)|G_i = g]$ for all $g \in \mathcal{G}, d \in \{0, 1\}$ into a vector x, and the identified quantities $E[Y_i\mathbb{1}(D_i = d)|Z_i =$

$z]$ for all $d \in \{0,1\}, z \in \mathcal{Z}$ into a vector b, then we have a system of linear equations $Ax = b$, where A is a fixed matrix of entries of the form $[A]_{dz,g} = \mathbb{1}(D_g(z) = d)$.

Similarly, as we have seen Δ_c can be written as a linear combination of the components of the vector z. Specifically, from Equation(4):

$$\Delta_c = \sum_g \frac{E[c(g, Z_i)]}{E[c(G_i, Z_i)]} P(G_i = g) \{E[Y_i(1)|G_i = g] - E[Y_i(0)|G_i = g]\}$$

We can now write $\Delta_c = \theta_c' x$, where θ_c is the vector of coefficients $\pm\frac{E[c(g,Z_i)]}{E[c(G_i,Z_i)]}$ from the above equation.

The set of vectors x compatible with the set of identifying restrictions $Ax = b$ can be written as $\{A^\dagger b + (I - A^\dagger A)w\}$ for all arbitrary vectors $w \in \mathbb{R}^{2|\mathcal{G}|}$, where $A^\dagger$ is the Moore-Penrose pseudo-inverse of A. The corresponding set of values for Δ_c is $\{\theta_c' A^\dagger b + \theta_c'(I - A^\dagger A)w\}$. For this set to be a singleton for all w, we must either have $A^\dagger A = 0$ (i.e. A has full column rank), or the vector θ_c must lie in the row space of the matrix A, so that in either case $\theta_c'(I - A^\dagger A)$ is equal to the zero vector. If the set were not a singleton, then Δ_c would not be expectation identified, since an infinite collection of values of Δ_c would be compatible with the full set of restrictions $Ax = b$. Thus, by **1.**, we have that $\Delta_c = \theta_c' A^\dagger b$. This then implies **3.**, if we take $s(d, z) = \frac{P(D_i=d|Z_i=z)}{P(D_i=d,Z_i=z)} \cdot [\theta_c' A^\dagger]_{(d,z)}$, where $[\theta_c' A^\dagger]_{(d,z)}$ is the component of the vector $\theta_c' A^\dagger$ corresponding to the pair (d, z). Note that $A^\dagger$ is a known matrix (without looking at the data), and θ_c is a known function of the marginal distribution of Z_i, up to the factor $E[c(G_i, Z_i)]$, for a fixed function c.

It only remains to be shown that $E[c(G_i, Z_i)]$ is also identified under assumption of 1. For Δ_c to be pinned down for all joint distributions of $(G_i, Y_i(1), Y_i(0))$, it must be pinned down in the special case where each potential outcome distribution is a point mass: $Y_i(d) = d$. In this case each $E[Y_i|D_i = d, Z_i = z] = d$, and $\Delta_c = 1$. Thus, using our result above we have that $E[c(G_i, Z_i)] = E[\tilde{s}(d, z)D_i]$, where $\tilde{s}(d, z) := \frac{P(D_i=d|Z_i=z)}{P(D_i=d,Z_i=z)} \cdot [\tilde{\theta}_c' A^\dagger]_{(d,z)}$, where $\tilde{\theta}_c := E[c(G_i, Z_i)]\theta_c$. Unlike θ_c, $\tilde{\theta}_c$ is pinned down from knowledge of the function c and the marginal distribution of Z_i.

This last point generalizes the result of Corollary 4.3, that $E[c(G_i, Z_i)] = P(C_i = 1)$ is identified under the assumptions of Theorem 1, to hold whenever Δ_c is expectation identified (though it may not always hold in a "split-sample" sense). Note that for the BLATE under PM or IAM, we know that $E[c(G_i, Z_i)] = 1 - p_n - p_a$, and is thus easily computable from the identified proportions of never-takers and always-takers in the population.

- **3** $\to$ **1** Any β_s can be written: $\beta_s = \sum_{d,z} P(D_i = d, Z_i = z)s(d, z)E[Y_i|D_i = d, Z_i = z]$, and is thus pinned down by the CEFs $E[Y_i|D_i = d, Z_i = z]$, the joint distribution $\mathcal{P}_{DZ}$, and the known function s.

- **3** $\to$ **2** Immediate, since 3 is a special case of 2 with $\mathcal{S}$ a singleton

- **3** $\to$ **4** This is immediate, since $s(d,z)y$ is a possible function $t(d,z,y)$.
- **4** $\to$ **3** Consider a joint distribution F of fundamentals (potential outcomes, compliance groups, and instruments) and an alternative distribution F', where the potential outcomes are rescaled by a factor $b \in \mathbb{R}$: i.e. if $(Y_i(1), Y_i(0), G_i) \sim F$ then $(bY_i(1), bY_i(0), G_i) \sim F$. Let $\Delta_c(\cdot)$ denote the causal parameter Δ_c as a function of the joint distribution of $(Y_i(1), Y_i(0), G_i)$. Clearly $\Delta(F') = b\Delta_c(F)$. Note that if the distribution of Z_i is held fixed, the distribution of (Y_i, D_i, Z_i) under F' is the same as the distribution of (bY_i, D_i, Z_i) under F, since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$. Thus, by assumption $\beta = \Delta_c(F')$ when the observables are generated under F', I must have that $E[s(D_i, Z_i, bY_i)] = bE[s(D_i, Z_i, Y_i)]$. For this to be true for any distribution of (D_i, Z_i, Y_i), it must be that $s(d,z,by) = bs(d,z,y)$ for all d, z, y, b. Defining $s(d,z)$ as $s(d,z,1)$, I can then write $s(d,z,y)$ as $s(d,z)y$.[11]

□

A.7 Proof of Proposition 3

By Proposition 7, we know that if Δ_c is expectation identified, it can be written as $\Delta_c = \beta_s$, where $\beta_s = E[s(D_i, Z_i)Y_i]$ and $s(d,z)$ is an identified function of $\mathcal{P}_{\mathcal{DZ}}$. Now, using that $Y_i = Y_i(0) + D_i\Delta_i$ where $\Delta_i := Y_i(1) - Y_i(0)$:

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

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 group and potential outcomes (it is always possible to translate the support of the distribution of $Y_i(0)$ 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 (4) 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:

$$\Delta_c = \frac{1}{P(C_i = 1)} \sum_g P(G_i = g)P(C_i = 1|G_i = g) \cdot \Delta_g$$

[11] Note that a similar argument with an F' such that $(Y_i(1) + b, Y_i(0) + b, G_i, Z_i) \sim F$ reveals that the random variable $s(D_i, Z_i)$ must be mean zero.

Matching coefficients, this establishes that $P(C_i = 1|G_i = g) = E[s(1, Z_i)D_g(Z_i)]/P(C_i = 1)$, since $\beta_s = \Delta_c$ holds for any vector of $\{\Delta_g\}$. This set of weights satisfies Property M, since

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

A.8 Proof of Proposition 4

In the Supplemental Material, I show that with two binary instruments, if PM holds but not VM or IAM, then up to arbitrary labeling, $\mathcal{G}$ 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). Using Supplemental Material Table 2, these eight moments can be written in matrix form as

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

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, then BLATE can be written as

$$BLATE = \frac{1}{1 - p_a - p_n} \cdot \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \end{pmatrix}'}_{:=\lambda} x$$

BLATE 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 $\mathbf{1} \to \mathbf{3}$ in Proposition 7. This can be readily verified not to hold, since

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

where $A'(AA')^{-1}A$ is the orthogonal projector into the row space of A (which has full row rank).

A.9 Proof of Proposition 5

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 $E[Y_i|Z_i = \bar{Z}] - E[Y_i|Z_i = \underline{Z}] = P(G_i \in \mathcal{G}^c)BLATE$, 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 $BLATE = \rho_{\bar{Z},\underline{Z}}$. 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.

A.10 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:

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

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

A.11 Proof of Proposition 6

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 $\hat{\theta}_V := (0, \hat{\lambda}')'(\Gamma'\Gamma + K)^{-1}\Gamma' V$, where K is a diagonal matrix of positive entries. This generalizes from the $K = \alpha I$ case slightly to allow 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 a 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:

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

For any $V, W \in \{Y, D\}$, and $m \geq 0$:

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

where in the approximation I ignore terms that are of third or higher order in estimation error. Let $Z = (\Gamma'\Gamma + K)^{-1}\Gamma'\Gamma$ and notice that $\hat{\beta}_V^* = Z\hat{\beta}_V$. Using that $E[\hat{\beta}_V|\Gamma] = \beta_V$ 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 w 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:

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

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.

B Identification result with instrument degeneracy

A weaker version of Assumption 3 is comprised of the following two conditions:

Assumption 3a* (non-zero first stage). *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. $P\left(\sum_{S\subseteq\{1\dots J\}} \omega_S Z_{Si} = 0\right) < 1$ for all vectors $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$.*

Assumption 3b* (restriction on degenerate subsets). *For any of the $S \notin \mathcal{F}$ with $\mathcal{F}$ a maximal family from Assumption 3a*, write $Z_{Si} = \alpha_{S,\emptyset} + \sum_{S' \in \mathcal{F}} \alpha_{S,S'} Z_{S',i}$. Then the estimand Δ_c is such that $P(C_i = 1 | G_i = g(S)) = \sum_{S' \in \mathcal{F}} \alpha_{S,S'} P(C_i = 1 | G_i = g(S'))$.*

Assumption 3a* is in itself very weak: requiring only that there exists some product of the instruments that has strictly positive variance. Assumption 3b* however can be restrictive, when 3a* holds only for families $\mathcal{F}$ that do not contain all subsets of the instruments (Assumption 3). Assumption 3b* suggests that we can only identify the parameter Δ_c when $P(C_i = 1 | G_i = g)$ for redundant products of the instruments follows the same pattern of linear dependency as do the products of the instruments themselves. However, we can show a general case in which Assumption 3b* may not be as restrictive.

Proposition. *Suppose Assumption 3a* holds for a family $\mathcal{F}$, such that for any $S \notin \mathcal{F}$, $Z_{Si} = Z_{S'(S),i}$ with probability one for some "matching" subset $S'(S)$. Then, if $c(g(S), Z_i)$ can be written as a function of Z_{Si} only, Assumption 3b* holds.*

Proof. Since for any such S: $P(Z_{S'(S),i} = Z_{Si}) = 1$, it then follows that $c(g(S), Z_i) = c(g(S'(S)), Z_i)$ with probability one, and thus $P(C_i = 1 | G_i = g(S)) = P(C_i = 1 | G_i = g(S'(S)))$, which delivers Assumption 3b*. □

The first assumption of the Proposition captures cases where certain products of the instruments are truly redundant in the sense that they are exactly equal to a some single other product of the instruments, and not a linear combination of such products. The second part, that $c(g(S), Z_i)$ depends on Z_{Si} only, is satisfied by our leading examples BLATE, BLATT and SLATE.

The construction in Proposition 2 mapping discrete instruments to binary instruments yields a case where the above Proposition may be useful. Consider a case with a discrete instrument Z_1 with three levels $z_1 < z_2 < z_3$. Proposition 2 shows that if $Z_1 \dots Z_J$ satisfies VM then so does the set of $J + 1$ instruments $\tilde{Z}_2, \tilde{Z}_3, Z_2, \dots Z_J$ 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 redundant simple compliance groups, since for any $S \subseteq \{2 \dots J\}$: $\tilde{Z}_{2i} \tilde{Z}_{3i} Z_{Si} = \tilde{Z}_{3i} Z_{Si}$. However, if Assumption 3a* holds for the family $\mathcal{F}$ of all subsets of $\{\tilde{Z}_2, \tilde{Z}_3, Z_2, \dots Z_J\}$ that do not contain both of $\tilde{Z}_2$ and $\tilde{Z}_3$, the above Proposition will yield Assumption 3b*, for a class of causal parameters that includes BLATE, BLATT, and the SLATEs.

Theorem 1*. *Under Assumptions 1, 2, and 3*, the result of Theorem 1 follows for any Δ_c satisfying Property M provided that $P(C_i = 1) > 0$, where now $\Gamma_i = \{Z_{Si}\}_{S \in \mathcal{F}}$ and $h(Z_i) = \{Z_{Si}\}'_{S \in \mathcal{F}} \Sigma^{-1} (\Gamma_i - E[\Gamma_i])$.*

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

In the example above, the vector Γ_i would contain all non-null subsets of $\{\tilde{Z}_2, \tilde{Z}_3, Z_2, \dots Z_J\}$ that do not contain both of $\tilde{Z}_2$ and $\tilde{Z}_3$. In general, a maximal set $\mathcal{F}$ that satisfies Assumption 3b* 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.

C Identification and estimation with covariates

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 briefly consider identification and then 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:

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

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 will have that $\rho_h(x) = \Delta_c(x)$ where $\rho_h(x) = E[h(Z_i, x)Y_i|X_i = x]/E[h(Z_i, x)D_i|X_i = x]$.

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 non-parametric regression of each component of Γ_i on X_i as well as non-parametrically estimating the conditional variance-covariance matrix $\mathbb{V}[\Gamma_i|X_i = x]$ (Yin et al. (2010) describe a kernel-based method for this). The vector $\lambda(x)$ can also be computed via non-parametric regression.

However, when the object of interest is simply the unconditional version of Δ_c, the conditional quantities become nuisance parameters. It turns out that we can integrate over them separately in the numerator and the denominator in the estimand

$$\rho_h = \frac{E[h(Z_i, X_i)Y_i]}{E[h(Z_i, X_i)D_i]} \tag{9}$$

To see that this holds, write:

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

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.

In the Supplemental Material, I also work out a special case in which the X_i can simply be added as "regressors" in the numerator and denominator of Δ_c.