ESTIMATION IN AN INSTRUMENTAL VARIABLES MODEL WITH TREATMENT EFFECT HETEROGENEITY

Michal Kolesár[*]
Cowles Foundation
Yale University

Version 1.3.1, November 5, 2013[†]

Abstract

This paper analyzes estimators based on the classic linear instrumental variables model when the treatment effects are in fact heterogeneous, as in Imbens and Angrist (1994). I show that if the local average treatment effects vary, two-step instrumental variables estimators (tsiv), such as the two-stage least squares estimator (tsls) typically all estimate the same convex combination of them. In contrast, estimands of minimum distance estimators, such as the limited information maximum likelihood (liml) estimator, may be outside of the convex hull of the local average treatment effects, and may therefore not correspond to a causal effect. This result questions the standard recommendation to use liml when the number of instruments is large as a way of addressing the bias exhibited by tsls in these settings. Instead, I propose a new tsiv estimator, a version of the jackknife instrumental variables estimator (ujive). Unlike tsls or liml, ujive is consistent for a convex combination of local average treatment effects under many instrument asymptotics that also allow for many covariates and heteroscedasticity.

Keywords: Instrumental Variables, Local Average Treatment Effects, Limited Information Maximum Likelihood, Jackknife.

[*] Electronic correspondence: `mkolesar@princeton.edu`. I am deeply grateful to Gary Chamberlain and Guido Imbens for their constant guidance and support. I also thank Alberto Abadie, Nikhil Agarwal, Josh Angrist, Raj Chetty, Paul Goldsmith-Pinkham, Adam Guren, Pepe Montiel, Whitney Newey, Jim Stock, Fanyin Zheng, and seminar participants at Brown University, Harvard University, Michigan, MIT, Princeton University, University of Maryland, and Yale University for helpful comments and suggestions.

[†] Version hash df1f500, compiled November 5, 2013

1 Introduction

The classic linear instrumental variables model is commonly used to estimate treatment effects. When the individual treatment effect is independent of treatment status and covariates, estimators based on this model estimate the population average treatment effect (Heckman, 1997). However, since this assumption rules out selection into treatment based on anticipated gains from treatment, it is not very plausible in many empirical settings. It is therefore important to understand the properties of these estimators when the individual treatment effect is allowed to be correlated with treatment status.

The first contribution of this paper is to characterize the estimands of estimators based on the classic linear instrumental variables (IV) model when the treatment effects are unrestricted. I assume that the instruments satisfy the monotonicity condition of Imbens and Angrist (1994), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). I show that the two-stage least squares (TSLS) estimator, under some mild assumptions about the first stage, estimates a convex combination of these local average treatment effects, weighted over different pairs of instrument values and covariates. On the other hand, unless all LATES are the same, the estimand of the limited information maximum likelihood (LIML, Anderson and Rubin, 1949) depends on the covariance matrix of the reduced-form errors, and may lie outside the convex hull of the local average treatment effects. Therefore, the estimand may not correspond to a causal effect. Moreover, other estimators based on the classic linear IV model will, depending on how they are constructed, either estimate the same convex combination of LATES as TSLS, or else behave similarly to LIML.

In particular, estimators that behave like TSLS can be thought of as two-step estimators. In the first step, they construct a single instrument, a predictor of the treatment status based on the first-stage regression. In the second step, an instrumental variables estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect. I refer to these estimators as two-step instrumental variables estimators. In the limit under standard asymptotics, the exact way of constructing the single instrument does not matter; all two-step IV estimators converge to the same probability limit as the infeasible instrumental variables estimator that uses a population linear predictor of the treatment status as a single instrument. In turn, the probability limit of this IV estimator corresponds to a weighted average of LATES. The weights are non-negative if the single instrument itself satisfies monotonicity in that changing its value does not induce two-way flows in and out of treatment.

In contrast, estimators that behave like LIML are based on the property of the classic linear iv model that the coefficients on the instruments in the first-stage regression are proportional to the coefficients in the reduced-form outcome regression. These estimators, which I refer to as minimum distance estimators, minimize a minimum distance objective function that directly enforces this proportionality with respect to some weight matrix. The estimator of the treatment effect is given by the estimator of the constant of proportionality. Malinvaud (1966, Chapter 20) and Goldberger and Olkin (1971) show that liml can be thought of in this way, with the weight matrix depending on the covariance matrix of the reduced-form errors.

This approach yields a different estimand under treatment effect heterogeneity because imposing proportionality of the reduced-form coefficients implies that the treatment and the outcome are treated symmetrically. In particular, it requires that the estimand of the reverse two-stage least squares estimator (rtsls) be equal to the estimand of tsls. The rtsls estimator is obtained as the reciprocal of the tsls estimator in the instrumental variables model that swaps the treatment and the outcome. This requirement makes sense if the instrumental variables model is supposed to solve an errors-in-variables problem (Zellner, 1970), or an omitted variable bias (Chamberlain, 2007). However, in the context of estimating treatment effects, the reduced-form coefficients are no longer proportional to each other unless all lates are equal. Therefore, the tsls and rtsls estimands are in general different; the probability limit of the rtsls estimator is the same as that of an instrumental variables estimator that uses a linear predictor of the outcome based on the reduced-form outcome regression as an instrument. This instrument induces a different weighting scheme for the lates, and hence a different estimand, than using a linear predictor of the treatment status as an instrument. Unlike the tsls weights, these weights are proportional to the effect size, with the bigger lates receiving more weight.

There are two ways in which this difference between tsls and rtsls estimands can cause a minimum distance estimand to be outside the convex hull of lates. First, if some lates are negative, the rtsls estimand gives them a negative weight, so that the estimand may end up being outside the convex hull of the lates. Consequently, the minimum distance estimand, trying to equate rtsls with tsls, may end up being outside the convex hull. Second, even if the rtsls estimand is inside the convex hull, if the weight matrix that is used to equate rtsls with tsls is non-diagonal, as is the case with liml, the minimum distance estimand is not guaranteed to lie between the rtsls and tsls estimands.

It is easy to avoid these problems by simply avoiding liml and using tsls. However, when many instruments are used, tsls may be severely biased even in large samples (Bound, Jaeger and Baker, 1995), and it is inconsistent under the many instrument asymptotic sequence of Kunitomo (1980), Morimune (1983), and Bekker (1994). Therefore, when the number of instruments is large, the standard recommendation has been to use LIML, which is not only consistent under many instrument asymptotics, but also efficient among rotation invariant estimators and homoscedasticity (Chioda and Jansson, 2009; Anderson, Kunitomo and Matsushita, 2010). Recently, other estimators have been proposed that behave better than LIML under heteroscedasticity. Hausman, Newey, Woutersen, Chao and Swanson (2012) propose a Fuller (1977) type modification to a jackknife version of LIML (HLIM). Bekker and Crudu (2012) propose a similar estimator, which they call symmetric jackknife. However, all of these estimators are minimum distance estimators, and therefore not likely to work well under treatment effect heterogeneity.

The second contribution of this paper is to propose a new estimator in the two-step IV class, the unbiased jackknife IV estimator (UJIVE), that remains consistent for a convex combination of LATES even under many instrument asymptotics and heteroscedasticity. This estimator is similar to the jackknife instrumental variables estimator (JIVE, Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999) in that it also uses a "leave-one-out" jackknife-type predictor of the treatment in the first stage, but differs from JIVE in the way it deals with covariates. In particular, in constructing the single instrument in the two-step IV procedure, we need to partial out the effect of covariates on the treatment. Suppose, for example, that the instruments are classroom indicators, and the covariates are school indicators (school "fixed effects"). Then the JIVE estimate of the effect of covariates on the treatment status of individual i is given by an average treatment status of individuals in the same school as individual i. With a finite number of observations in each school, this estimate is noisy, and since it depends on the treatment status of individual i, the estimation error is correlated with the outcome. Therefore, the single constructed instrument is also correlated with the outcome, causing JIVE to be biased when the number of covariates is large (Ackerberg and Devereux, 2009). In contrast, the UJIVE estimate of the effect of the covariates is given by a sample average that excludes individual i, which guarantees that the prediction error will be uncorrelated with the outcome. As a result, unlike JIVE, UJIVE is consistent for a convex combination of LATES even when we let the number of covariates, in addition to the number of instruments, increase in proportion to the sample size, as in Anatolyev (2011) and Kolesár, Chetty, Friedman, Glaeser and Imbens (2011).

The estimand of two-step IV estimators can be seen as one way of summarizing the effect of the treatment on outcome. For particular policy questions, however, we might be interested in a weighting scheme that is different than the one used by these estimators. For this purpose, a number of alternative approaches, not based on the classic IV model, have been proposed in the literature. For example, Frölich (2007) derives a non-parametric estimator for the largest subpopulation of compliers for which a treatment effect can be identified. When the instrument is binary, Abadie (2003) works out a semi-parametric approach to approximating a treatment response function, and Hirano, Imbens, Rubin and Zhou (2000) and Yau and Little (2001) use a parametric approach to estimate a LATE that does not condition on covariates. Alternatively, instead of focusing on the LATEs, one might be interested in other features of the distribution of the potential outcomes, such as the average treatment effect. Although such features are not point-identified, Balke and Pearl (1997), Kitagawa (2009), and Machado, Shaikh and Vytlacil (2013) derive informative bounds for such parameters. To keep the paper focused, I do not try to compare the classic IV estimators with these alternative approaches.

The rest of the paper is organized as follows. In Section 2, I set up the problem of estimating causal effects in a potential outcomes framework. In Section 3, I review assumptions underlying the classic linear IV model, and I introduce the classes of two-step IV and minimum distance estimators that are based on this model. In Section 4, I introduce the local average treatment effects framework of Imbens and Angrist (1994). In Section 5, I derive the first main result of the paper, the estimands of two-step IV and minimum distance estimators under the LATE assumptions. In Section 6, I derive the second main result of the paper that UJIVE is consistent for a convex combination of LATEs under many instrument asymptotics. Section 7 illustrates the results from Sections 5 and 6 with a Monte Carlo experiment. Section 8 concludes. The Appendix contains proofs and an extension of the results to the case when the treatment is multi-valued.

2 Potential outcomes framework

Using a random sample of n individuals indexed by $i = 1, \dots, n$, we want to learn about the causal effect of a treatment $T_i \in \mathcal{T}$ on some outcome of interest. For clarity of exposition, I focus on the case when the treatment T_i is binary, so that $\mathcal{T} = \{0, 1\}$. I discuss the extension to multi-valued treatment in Appendix A. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes in the treated and untreated states. The treatment effect for individual i is then given by $\tau_i = Y_i(1) - Y_i(0)$.

The fundamental problem is that for each individual, we only observe the potential outcome corresponding to the observed treatment state, $Y_i = Y_i(T_i)$. Therefore, we cannot compute τ_i directly for any individual. Moreover, there is a concern that anticipated potential outcomes affect selection into treatment, so that comparing the average outcome of the subsample of individuals who are treated in our sample with those who are not is likely to lead to a biased estimate of the population average treatment effect $\mathbb{E}[\tau_i]$.

We do, however, observe instruments Q_i with support $\mathcal{Q}$ that help to identify average treatment effects for at least some subpopulations. For each possible realization $q \in \mathcal{Q}$, let $T_i(q)$ denote the potential treatment variable that equals one if individual i would receive treatment if their instrument value was changed to $Q_i = q$, and equals zero if they would not receive treatment. The observed treatment status is given by $T_i = T_i(Q_i)$; the other potential treatments are not observed.

We also observe a vector of covariates X_i with support $\mathcal{X}$. I include these covariates explicitly for two related reasons. First, in many empirical applications the identification assumptions that underlie the instrumental variables framework may only be plausible conditional on X_i. One simple approach in this case is to carry out the analysis separately for all values of the covariates. However, when the covariate set is detailed so that the support $\mathcal{X}$ is rich, this approach is unlikely to be satisfactory. Second, even when the identification assumptions are plausible unconditionally, inference without covariates might not be precise enough. A common solution to both of these problems in practice is to estimate a single model with covariates. It is therefore important to understand how the presence of covariates affects inference.

In summary, the observed data vector for each individual is given by (Y_i, T_i, Q_i, X_i).

Two important functions of the distribution of the observed data are given by the two regression functions

$$r(q, x) = \mathbb{E}[Y_i \mid Q_i = q, X_i = x], \tag{1}$$

$$p(q, x) = \mathbb{E}[T_i \mid Q_i = q, X_i = x]. \tag{2}$$

With binary treatment, $p(q, x)$ equals the conditional treatment probability, $\mathbb{P}(T_i = 1 \mid Q_i = q, X_i = x)$, also known as the propensity score. When viewed as a random variable, I will denote it by $P_i = p(Q_i, X_i)$. Similarly, $r(q, x)$ denotes the conditional expectation of the outcome, and I denote it by $R_i = r(Q_i, X_i)$ when viewed as a random variable. Without further assumptions, these regression function are not directly informative about the objects of interest—the treatment effects. They are therefore known as the reduced form equations.

In general, both reduced form equations will be non-linear. The linear IV estimators that I will consider are based on a linear approximation to the true non-linear reduced form

$$R_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] = Z_i'\pi_1 + W_i'\psi_1, \tag{3}$$

$$P_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] = Z_i'\pi_2 + W_i'\psi_2, \tag{4}$$

where $\mathbb{E}^*$ denotes population (minimum mean-squared-error) linear projection[1], and $Z_i = z(Q_i, X_i)$ and $W_i = w(X_i)$ are expansions of the original instruments and covariates, with $\dim(Z_i) = K$ and $\dim(W_i) = L$. I assume that W_i spans a column of ones. The estimators that I will consider will use these constructed instruments and covariates.

For example, in Angrist and Krueger (1991), the basic instruments Q_i were three quarter of birth indicators, and the constructed instruments Z_i were obtained by interacting Q_i with year of birth and state of birth indicators. A similar specification was used in Dobbie and Fryer (2011), who study the effect of Harlem Children Zone (HCZ) charters on educational outcomes. In particular, Dobbie and Fryer (2011) construct Z_i by interacting an indicator for living within HCZ, Q_i, with cohort, so that $Z_{i,\ell} = Q_i \mathbb{1}_{X_i=\ell}$, where ℓ indexes cohorts, $\ell \in \{1, \dots, L\}$. If we also set $W_{i\ell} = \mathbb{1}_{X_i=\ell}$, and cohort is the only covariate that we observe, then the linear approximation is exact, and $P_i = P_i^L$. With continuous instruments and covariates, we could use series expansions to construct Z_i and W_i. Of course, we can also simply set $z(Q_i, X_i) = Q_i$ and $w(X_i) = X_i$. I make the distinction between the original instruments and covariates, (Q_i, X_i), and the constructed ones, (Z_i, W_i), because it will matter for the estimands of these estimators under treatment effect heterogeneity how exactly the instruments were constructed.

I use matrix notation to help keep the definitions and results compact. I denote the n-component vector with ith element Y_i by $\mathbf{Y}$. Similarly, let $\mathbf{T}$, $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{P}$, $\mathbf{P}^L$, $\mathbf{R}$ and $\mathbf{R}^L$ denote vectors and matrices with rows T_i, W_i', Z_i', P_i, P_i^L, R_i and R_i^L. For any full-rank $n \times m$ matrix $\mathbf{A}$, let $\mathbf{H_A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ denote the associated $n \times n$ projection matrix (also known as the hat matrix), and let $\mathbf{D_A}$ be an $n \times n$ diagonal matrix with $(\mathbf{H_A})_{ii}$ on the diagonal. Let $\mathbf{I}_m$ denote the $m \times m$ identity matrix, and let $\mathbf{M_A} = \mathbf{I}_n - \mathbf{H_A}$ denote the annihilator matrix. Let $\mathbf{A}_\perp = \mathbf{M_W A}$ denote the residual from the sample projection of $\mathbf{A}$ onto $\mathbf{W}$, and let $\tilde{A}_i = A_i - \mathbb{E}^*[A_i \mid W_i]$ denote the residual from the population projection of A_i onto W_i. Also, let a.s. denote almost surely (i.e. with probability one).

[1] In other words, the linear projection of A_i onto B_i, $\mathbb{E}^*[A_i \mid B_i] = B_i'\gamma$, minimizes $\min_\gamma \mathbb{E}[(A_i - B_i'\gamma)^2]$. If the covariance matrix of B_i is non-singular so that $\mathbb{E}[B_i B_i']$ is invertible, then the solution is uniquely given by $\gamma = \mathbb{E}[B_i B_i']^{-1}\mathbb{E}[B_i A_i]$.

3 Classic linear IV model and estimators

The classic linear IV model is usually defined in terms of a structural equation (see, for example Wooldridge, 2002, Chapter 5)

$$Y_i = T_i\beta + W_i'\delta + \epsilon_i, \tag{5}$$

where the covariates W_i and the instruments Z_i are assumed to be uncorrelated with the structural error ϵ_i:

$$\mathbb{E}[\epsilon_i W_i] = 0, \qquad\qquad \mathbb{E}[\epsilon_i Z_i] = 0. \tag{6}$$

The second assumption is that the instruments are relevant in the sense that the coefficient π_2 in Equation (4) is non-zero. The parameter of interest is β, and it represents the causal effect of T_i on Y_i. Equations (5) and (6) can be compactly written as a moment condition

$$\mathbb{E}^*[Y_i - T_i\beta - W_i'\delta \mid Z_i, W_i] = 0. \tag{7}$$

In this section, I use the potential outcomes framework to formulate assumptions that deliver the moment condition (7) and that give β a direct causal interpretation as the population average treatment effect. This allows me to more easily link the classic IV model to the LATE framework of Imbens and Angrist (1994). Second, I define three classes of estimators of β: the class of two-step IV estimators (that includes the two-stage least squares estimator), the class of reverse two-step IV estimators (that includes the reverse two-stage least squares estimator), and finally the class of minimum-distance estimators (that includes LIML). This classification will be more useful when considering the behavior of the classic IV estimators under the LATE framework than the traditional division into estimators that fit into the k-class (Nagar, 1959; Theil, 1961, 1971), and estimators that do not.

3.1 Assumptions underlying the classic linear IV model

Interpreting β in the moment condition (7) as the population average treatment effect requires three assumptions that correspond to Assumptions IV, CTE and L below.

First, the instruments need to be valid in the sense that they only affect potential outcomes through their effect on the treatment. To formalize this notion, we need to include Q in the definition of potential outcomes. Let $Y_i(t, q)$ be the potential outcome when individual i receives treatment t and instrument q, so that the observed outcome is given by $Y_i = Y_i(T_i, Q_i)$.

Assumption IV.

(i) (Random assignment) $\{Y_i(t,q), T_i(q)\}_{t\in\mathcal{T},q\in\mathcal{Q}} \perp\!\!\!\perp Q_i \mid X_i$;

(ii) (Exclusion restriction) $\mathbb{P}(Y_i(t,q) = Y_i(t,q') \mid X_i) = 1$ for all (t,q,q') a.s.; and

(iii) (Relevance) The distribution of P_i^L conditional on X_i is non-degenerate with positive probability.

Part (i) requires that conditional on covariates, the instruments are as good as randomly assigned in the sense that they are independent of potential outcomes and potential treatments. Part (ii) requires that the instruments only affect outcomes through their effect on the treatment. This assumption justifies writing the potential outcomes as functions of the treatment only, so that $Y_i(t) = Y_i(t,q)$. Finally, Part (iii) is a rank condition—requires that the constructed instruments Z_i have a non-zero effect on the treatment, at least for some values of covariates; it ensures that the coefficient π_2 in Equation (4) is non-zero. Since $Z_i = z(Q_i, X_i)$, a necessary condition is that the original instruments Q_i have a non-zero effect on the treatment.

The second assumption restricts treatment effect heterogeneity:

Assumption CTE (Constant Average Treatment Effects). For all $(t_1,t_0,t,q,x) \in \mathcal{T}^3 \times \mathcal{Q} \times \mathcal{X}$, $\mathbb{E}[Y_i(t_1) - Y_i(t_0) \mid Q_i = q, T_i = t, X_i = x] = (t_1 - t_0)\beta$.

Assumption CTE restricts the treatment effects in two ways. First, although it allows the individual treatment effects to vary, it requires that the source of heterogeneity in the individual treatment effects be unrelated to observables. In other words, it imposes that $\mathbb{E}[Y_i(t_1) - Y_i(t_0) \mid Q_i = q, T_i = t, X_i = x] = \mathbb{E}[Y_i(t_1) - Y_i(t_0)]$. In particular, it does not allow individuals' treatment status to be correlated with gains from treatment, ruling out what Heckman, Urzúa and Vytlacil (2006) call essential heterogeneity, or sorting on gains from treatment. This makes it implausible in many empirical applications—I will relax it in the next section when I introduce the LATE framework of Imbens and Angrist (1994).

Second, it restricts the treatment effect to be linear, so that $\mathbb{E}[Y_i(t_1) - Y_i(t_0)] = (t_1 - t_0)\beta$. This linearity condition is only restrictive when the treatment is multi-valued (see Appendix A for discussion of this case). With binary treatment, the condition is moot. The parameter β now corresponds to the population average treatment effect.

Some textbook discussions of the classic linear IV model (Wooldridge, 2002; Angrist and Pischke, 2009) use a stronger version of this assumption by imposing $Y_i(t_1) - Y_i(t_0) = (t_1 - t_0)\beta$

for all i, ruling out *any* heterogeneity in the treatment effect, but such restrictive assumption is not needed.

Assumption CTE implies that

$$\begin{aligned} 0 &= \mathbb{E}[Y_i(t) - Y_i(0) - t\beta \mid Q_i = q, T_i = t, X_i = x] \\ &= \mathbb{E}[Y_i(T_i) - Y_i(0) - T_i\beta \mid Q_i = q, T_i = t, X_i = x] \\ &= \mathbb{E}[Y_i - Y_i(0) - T_i\beta \mid Q_i = q, X_i = x], \end{aligned} \tag{8}$$

where the last line follows the Law of iterated expectations. To turn Equation (8) into the moment condition (7), we need that

$$\mathbb{E}^*[Y_i(0) \mid Z_i, W_i] = \mathbb{E}^*[Y_i(0) \mid W_i]. \tag{9}$$

If there are no covariates beyond the intercept, so that $W_i = 1$, then this equality holds automatically. However, since Assumption IV allows for cases in which the assignment of instrument is only random conditional on covariates, it only implies that $\mathbb{E}[Y_i(0) \mid Z_i, X_i] = \mathbb{E}[Y_i(0) \mid X_i]$. If the conditional expectation $\mathbb{E}[Y_i(0) \mid X_i]$ is not linear in W_i, then controlling for W_i in a linear way does not fully control for the effect of the covariates on $Y_i(0)$. Consequently, $\tilde{Z}_i = Z_i - \mathbb{E}^*[Z_i \mid W_i]$ (part of Z_i orthogonal to W_i) may be correlated with $Y_i(0) - \mathbb{E}^*[Y_i(0) \mid W_i]$, and the coefficient on Z_i on the left-hand side of (9) may be non-zero. Therefore, some textbook discussions (Wooldridge, 2002; Angrist and Pischke, 2009) make the assumption that $\mathbb{E}[Y_i(0) \mid X_i] = W_i'\delta$, so that controlling for W_i in a linear way controls fully for the effect of the covariates on $Y_i(0)$. Unfortunately, this assumption has the undesirable implication that, in principle, sufficient variation in the covariates alone is enough to identify β since non-linear functions of W_i, such as squares of W_i, are valid instruments. Moreover, since it involves potential, rather than observed outcomes, it is not directly testable.

Here I focus on the other way we can ensure that $\tilde{Z}_i$ is not correlated with $Y_i(0) - \mathbb{E}^*[Y_i(0) \mid W_i]$—by restricting the expectation of Z_i conditional on X_i to be linear in W_i:[2]

Assumption L (Linearity). $\mathbb{E}[Z_i \mid X_i] = \mathbb{E}^*[Z_i \mid W_i]$.

Assumption L ensures that controlling for the effect of covariates on the instruments by a linear projection on W_i is as good as conditioning on X_i. There are three important special cases

[2] By the residual regression formula (9) holds iff $\mathbb{E}[Y_i(0)\tilde{Z}_i] = 0$. Assumption L implies that $\mathbb{E}[\tilde{Z}_i \mid X_i] = 0$. Therefore, by the law of iterated expectations, we have $\mathbb{E}[Y_i(0)\tilde{Z}_i \mid X_i] = \mathbb{E}[\mathbb{E}[Y_i(0) \mid X_i, Z_i]\mathbb{E}[\tilde{Z}_i \mid X_i, Z_i] \mid X_i] = \mathbb{E}[Y_i(0) \mid X_i]\mathbb{E}[\tilde{Z}_i \mid X_i] = 0$ where the second equality follows from Assumption IV.

in which Assumption L holds automatically. First, if there are no covariates. Second, if X_i is discrete and W_i is saturated, consisting of dummies for different values of X_i. Third, if Z_i is a function of Q_i only, and Q_i is independent of X_i, in which case $\mathbb{E}[Z_i \mid X_i] = \mathbb{E}[Z_i]$. This happens, for example, when Q_i is some randomly assigned encouragement to take the treatment, and the covariates are added after the randomization to increase precision of inference.

Abadie (2003) shows that another consequence of Assumption L is that the parameter δ in Equation (7) can now be interpreted as providing the best linear approximation to $\mathbb{E}[Y_i(0) \mid X_i]$ in the sense of minimizing the mean-square error $\mathbb{E}[(\mathbb{E}[Y_i(0) \mid X_i] - W_i'\delta)^2]$.

3.2 Two-step IV estimators

An implication of the moment condition (7) is that β can be identified using a single instrument $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i] = \tilde{Z}_i'\pi_2$, the linear approximation to the propensity score (4) with the covariates partialled out. $\tilde{P}_i^L$ can be thought of as an approximation to $\mathbb{E}[T_i \mid Q_i, X_i] - \mathbb{E}[T_i \mid X_i] = P_i - \mathbb{E}[P_i \mid X_i]$, which measures how strong the instrument assigned to individual i is (in terms of how likely it is to induce an individual into taking the treatment), relative to other instruments they could have been assigned, holding the covariates fixed. Since $\tilde{P}_i^L$ is linear in Z_i and W_i, the moment condition implies that

$$0 = \mathbb{E}^*[Y_i - W_i'\delta - T_i\beta \mid \tilde{P}_i^L] = \mathbb{E}^*[Y_i - T_i\beta \mid \tilde{P}_i^L],$$

where the second equality follows from $\mathbb{E}[W_i\tilde{P}_i^L] = 0$. Rearranging this expression, we obtain

$$\beta = \frac{\mathbb{E}[\tilde{P}_i^L Y_i]}{\mathbb{E}[\tilde{P}_i^L T_i]},$$

so that the IV estimator that uses $\tilde{P}_i^L$ as a single instrument, $\hat{\beta}_{\text{IV}} = \sum_i \tilde{P}_i^L Y_i / \sum_i \tilde{P}_i^L T_i$, is consistent for β. Moreover, if the error $\epsilon_i = Y_i - W_i'\delta - T_i\beta$ is homoscedastic, so that $\operatorname{var}(\epsilon_i^2 \mid X_i, Q_i) = \sigma^2$, then this estimator is asymptotically efficient.

Since $\tilde{P}_i^L$ is not directly observed, such an estimator is not feasible. Two step IV estimators implement a feasible version of $\hat{\beta}_{\text{IV}}$. In the first-step, they construct an estimate $\hat{P}_i$ of $\tilde{P}_i^L$. In the second step, an IV estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect:

$$\hat{\beta}_{\hat{\mathbf{P}}} = \frac{\hat{\mathbf{P}}'\mathbf{Y}}{\hat{\mathbf{P}}'\mathbf{T}}. \tag{10}$$

The class of two-step IV estimators is given by all estimators that admit this representation, where $\hat{\mathbf{P}}$ is a function of $\mathbf{T}$, $\mathbf{W}$ and $\mathbf{Z}$, including:

- The two-stage least squares (TSLS) estimator, which replaces π_2 and ψ_2 in (4) by their least-squares estimates, leading to $\hat{\mathbf{P}} = \mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T}$;

- The bias-corrected two-stage least squares estimator (Nagar, 1959; Donald and Newey, 2001), which adjusts the TSLS propensity score estimator to $\hat{\mathbf{P}} = ((1-k)\mathbf{M}_{\mathbf{W}} + k\mathbf{H}_{\mathbf{Z}_\perp})\mathbf{T}$ to improve its finite-sample properties, where $k = 1/(1-(K-2)/n)$;

- The jackknife instrumental variables estimator (Phillips and Hale, 1977; Angrist *et al.*, 1999), with $\hat{\mathbf{P}} = \mathbf{M}_{\mathbf{W}}\big(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}\mathbf{M}_{(\mathbf{Z},\mathbf{W})}\big)\mathbf{T}$.

Under regularity conditions, the estimation error in the first step does not matter, and all of these estimators are consistent for β, and asymptotically efficient under homoscedasticity.

3.3 Reverse two-step IV estimators

The classic linear IV model is symmetric in Y and T; instead of instrumenting for T in Equation (7) like two-step estimators do, we can multiply the moment condition by $1/\beta$ (provided $\beta \neq 0$), instrument for Y, and take the reciprocal of the resulting estimator.

All (forward) two-step IV estimators have reverse counterparts. In particular, we can swap the role of the outcome and the treatment in the first step to obtain an estimate $\hat{\mathbf{R}}$ of $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] - \mathbb{E}^*[Y_i \mid W_i] = \tilde{Z}_i'\pi_1$, is the linear approximation to (3) with the covariates partialled out. In the second step, we use $\hat{\mathbf{R}}$ to instrument for the outcome, and take the reciprocal, obtaining

$$\hat{\beta}_{\hat{\mathbf{R}},\text{reverse}} = \left(\frac{\hat{\mathbf{R}}'\mathbf{T}}{\hat{\mathbf{R}}'\mathbf{Y}}\right)^{-1}.$$

For example, the reverse two-stage least squares estimator (RTSLS) uses $\mathbf{R}_{\text{RTSLS}} = \mathbf{H}'_{\mathbf{Z}_\perp}\mathbf{Y}$ in the first step. Under regularity conditions, this class of estimators is also asymptotically efficient for β under homoscedasticity.

Although it is rarely used in practice, this class will prove useful in understanding the behavior of the minimum distance class of estimators, which I introduce next, under treatment effect heterogeneity.

3.4 Minimum distance estimators

Another implication of the conditional moment restriction (7) is that if we project Y_i and T_i onto Z_i and W_i, the coefficients on Z_i will be proportional to each other. To see this, by linearity of linear projections, we obtain:

$$\mathbb{E}^*[Y_i \mid Z_i, W_i] = W_i'\delta + \mathbb{E}^*[T_i \mid Z_i, W_i]\beta. \tag{11}$$

Therefore, the coefficients in the linear projections (3)–(4) are related to the coefficients (β, δ) by $\delta = \psi_1 - \psi_2\beta$, and

$$\pi_1 = \pi_2\beta. \tag{12}$$

This proportionality restriction can be imposed directly in estimation of β by using a minimum distance objective function

$$(\operatorname{vec}(\hat{\Pi}) - a \otimes \pi_2)'\hat{\Phi}(\operatorname{vec}(\hat{\Pi}) - a \otimes \pi_2), \qquad a = \begin{pmatrix} \beta \\ 1 \end{pmatrix}, \tag{13}$$

where $\hat{\Pi} = (\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1}\mathbf{Z}_\perp'(\mathbf{Y}, \mathbf{T})$ is an unrestricted least-squares estimator of $\Pi = (\pi_1, \pi_2)$, and $\hat{\Phi}$ is some weight matrix. Malinvaud (1966, Chapter 20) and Goldberger and Olkin (1971) show that the limited information maximum likelihood (LIML) estimator minimizes this objective function if the weight matrix is given by

$$\hat{\Phi} = \hat{\Omega}^{-1} \otimes \mathbf{Z}_\perp'\mathbf{Z}_\perp/n, \qquad \hat{\Omega} = \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix}' \mathbf{M}_{(\mathbf{Z},\mathbf{W})} \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix} / (n - K - L).$$

Here $\hat{\Omega}$ an estimator of the covariance matrix of the reduced-form errors $V_{1i} = Y_i - \mathbb{E}^*[Y_i \mid Z_i, W_i]$ and $V_{2i} = T_i - \mathbb{E}^*[T_i \mid Z_i, W_i]$ based on the unrestricted least-squares residuals. To understand the sensitivity of minimum distance estimators to departures from the assumption of constant treatment effects (Assumption CTE), it is helpful to work with a slightly different minimum distance objective function. Define a 2-by-2 matrix

$$\Xi = \Pi'\mathbb{E}[\tilde{Z}_i\tilde{Z}_i']\Pi. \tag{14}$$

In Section 5, I will show that this matrix plays a key role in understanding the behavior of classic linear IV estimators under the LATE framework. The proportionality restriction (12) implies a rank restriction on Ξ, namely that $\Xi = \Lambda aa'$, where $\Lambda = \Xi_{22} = \pi_2'\mathbb{E}[\tilde{Z}_i\tilde{Z}_i']\pi_2$. This restriction is essentially a restriction on the second moments of $\hat{\Pi}$ if $\mathbb{E}[\tilde{Z}_i\tilde{Z}_i']$ is proportional to the identity matrix. If the weight matrix $\hat{\Phi}$ has a Kronecker structure, $\hat{\Phi} = \hat{S}^{-1} \otimes \mathbf{Z}_\perp'\mathbf{Z}_\perp/n$ for some positive definite matrix $\hat{S} \in \mathbb{R}^{2\times 2}$, minimizing the objective function (13) yields the same estimator of β as a minimum distance estimator based on the rank-restriction on Ξ given by[3]

$$\hat{\mathcal{D}}(\beta, \Lambda) = \operatorname{vec}(\hat{\Xi} - \Lambda aa')'(\hat{S}^{-1} \otimes \hat{S}^{-1}) \operatorname{vec}(\hat{\Xi} - \Lambda aa'), \tag{15}$$

where $\hat{\Xi} = (\mathbf{Y}, \mathbf{T})'\mathbf{H}_{\mathbf{Z}_\perp}(\mathbf{Y}, \mathbf{T})/n = \hat{\Pi}'(\mathbf{Z}_\perp'\mathbf{Z}_\perp/n)\hat{\Pi}$ is an unrestricted estimator of Ξ. The class of minimum distance estimators is given by all estimators that minimize the objective function (15) for some unrestricted estimator $\hat{\Xi}$ of Ξ and some weight matrix $\hat{S}^{-1} \otimes \hat{S}^{-1}$. These estimators can be written as

$$\hat{\beta}_{\hat{\Xi},\hat{S}} = \frac{\hat{\Xi}_{12} - \hat{S}_{12} \min\operatorname{eig}(\hat{S}^{-1}\hat{\Xi})}{\hat{\Xi}_{22} - \hat{S}_{22} \min\operatorname{eig}(\hat{S}^{-1}\hat{\Xi})}. \tag{16}$$

Apart from LIML, the class of minimum distance estimators includes:

- Ω-class estimators of Keller (1975), which, like LIML, set $\hat{\Xi} = (\mathbf{Y}, \mathbf{T})'\mathbf{H}_{\mathbf{Z}_\perp}(\mathbf{Y}, \mathbf{T})/n$, but S is free to be any positive definite matrix. The choice $\hat{S} = \mathbf{I}_2$ leads to the symmetrically normalized two-stage least squares estimator studied in Keller (1975), Hillier (1990) and Alonso-Borrego and Arellano (1999).

- The symmetric jackknife estimator of (Bekker and Crudu, 2012), which sets

$$\hat{\Xi} = \frac{1}{n}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'\left(\mathbf{H}_{\mathbf{Z}_\perp} - \frac{1}{2}\left(\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{C} + \mathbf{C}'\mathbf{H}_{\mathbf{Z}_\perp}\right) - \frac{1}{4}\mathbf{C}'\mathbf{H}_{\mathbf{W}}\mathbf{C}\right)\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix},$$
$$\hat{S} = \frac{1}{n-K-L}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'\mathbf{M}_{(\mathbf{Z},\mathbf{W})}\mathbf{D}_{(\mathbf{Z},\mathbf{W})}(\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}\mathbf{M}_{(\mathbf{Z},\mathbf{W})}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix},$$

  where $\mathbf{C} = \mathbf{D}_{(\mathbf{Z},\mathbf{W})}(\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}\mathbf{M}_{(\mathbf{Z},\mathbf{W})}$.

- If there are no covariates W_i, then the HLIM estimator of Hausman *et al.* (2012) also admits this minimum distance representation, with

$$\hat{\Xi} = \frac{1}{n}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'(\mathbf{H}_\mathbf{Z} - \mathbf{D}_\mathbf{Z})\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}, \quad \hat{S} = \frac{1}{n-K}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'(\mathbf{M}_\mathbf{Z} + \mathbf{D}_\mathbf{Z})\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}.$$

Under homoscedasticity, any weight matrix S produces an asymptotically efficient estimator (Alonso-Borrego and Arellano, 1999). Consequently all of these estimators are asymptotically

[3] See Kolesár (2013) for derivation.

efficient under these conditions, and first-order asymptotically equivalent to the optimal forward and reverse two-step IV estimators.

4 Local average treatment effects approach

Instead of restricting treatment effect heterogeneity, the local average treatment effects framework of Imbens and Angrist (1994) replaces Assumption CTE by a monotonicity assumption that restricts how a treatment response to changing the value of the instrument may vary *across* people:

Assumption M (Monotonicity). For all $q, q' \in \mathcal{Q}$ either $\mathbb{P}(T_i(q) \geq T_i(q') \mid X_i) = 1$ or $\mathbb{P}(T_i(q) \leq T_i(q') \mid X_i) = 1$ a.s.

This assumption maintains that changing the instruments from q to q' affects all individuals with the same value of X_i in the same direction—it rules out situations in which, in response to a change in Q_i, some people drop out of treatment and others select into it. If Q_i is an encouragement to take the treatment, for example, then monotonicity requires that encouraging people to take the treatment makes everyone more likely to take it. Vytlacil (2002) shows that Assumption M is equivalent to assuming a latent index model as first proposed by Heckman (1976), in which selection into the treatment is modeled by a latent index crossing a threshold.[4]

For each value $x \in \mathcal{X}$ and for each pair (q, q'), define a local average treatment effect (LATE):

$$\tau(q, q'; x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i(q) \neq T_i(q'), X_i = x]. \tag{17}$$

This is the treatment effect averaged over individuals with $X_i = x$ who change their treatment status if we change their instrument from q to q'. Angrist, Imbens and Rubin (1996) refer to this set of individuals as compliers. Imbens and Angrist (1994) show that under Assumptions IV and M, so long as $\mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) > 0$, these local average treatment effects can be identified from the reduced form regressions:

$$\tau(q, q'; x) = \frac{r(q, x) - r(q', x)}{p(q, x) - p(q', x)}. \tag{18}$$

If $\mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) = 0$, then the set of compliers that the local average treatment effect (17) conditions on is empty, $p(q, x) = p(q', x)$, and $\tau(q, q'; x)$ is not identified. Since

[4] In the Heckman (1976) model, the index is given by $T_i^* = p(Q_i, X_i) - U_i$, where U_i is an unobserved random variable, distributed independently of (Q_i, X_i). T_i^* is interpreted as the expected net utility of selecting into treatment, so that $T_i = 1$ if $T_i^* \geq 0$.

Assumption IV (iii) implies that the distribution of P_i conditional on X_i is non-degenerate with positive probability, it ensures that at least some local average treatment effects are identified. On the other hand, the population average treatment effect $\mathbb{E}[\tau_i]$ is no longer identified once Assumption CTE is dropped unless the instrument Q_i is sufficiently strong to change everyone's treatment status (known as "identification at infinity"). The reason is that without restricting treatment effect heterogeneity, we have no way of computing the treatment effect for individuals who don't change their treatment status in response to a change in Q_i.

To facilitate expressing estimands or estimators based on the linear IV model in terms of local average treatment effects, it will be useful to write $\tau(q, q'; x)$ in terms of functions of the propensity score. Because of the equivalence between monotonicity and single index models, the instruments Q_i enter the model only through the propensity score (Heckman and Vytlacil, 1999; Heckman *et al.*, 2006). Therefore, $r(q, x) = \mathbb{E}[Y_i \mid P_i = p(q, x), X_i = x]$. Let $\mathcal{P}_x$ be the support of P_i conditional on $X_i = x$. Suppose that Q_i is discrete, so that $\mathcal{P}_x$ has finitely many support points. Let J_x be the number of support points, with $\mathcal{P}_x = \{p_{1,x} < \ldots < p_{J_x,x}\}$. Define a *marginal* local average treatment effect:

$$\alpha(p_{j,x}; x) = \frac{\mathbb{E}[Y_i \mid P_i = p_{j+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{j,x}, X = x]}{p_{j+1,x} - p_{j,x}}, \qquad j = 1, \ldots, J_x - 1. \tag{19}$$

$\alpha(p_{j,x}; x)$ is the is the local average treatment effect for individuals who get treated when the instrument they receive corresponds to propensity score with rank higher than j but not otherwise. We can express every local average treatment effect (17) for which the set of compliers is non-empty in terms of these marginal LATES. In particular, let $p(q, x) = p_{j,x}$ and that $p(q', x) = p_{j',x}$, and suppose that $j > j'$. Then we obtain:

$$\begin{aligned}\tau(q, q'; x) &= \frac{\sum_{m=j'}^{j-1} \left(\mathbb{E}[Y_i \mid P_i = p_{m+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{m,x}, X_i = x]\right)}{p_{j,x} - p_{j',x}} \\ &= \sum_{m=j'}^{j-1} \frac{p_{m+1,x} - p_{m,x}}{p_{j,x} - p_{j',x}} \alpha(p_m; x).\end{aligned} \tag{20}$$

If $j' = j - 1$, then $\tau(q, q'; x) = \alpha(p_{j'}; x)$.

If the support of $\mathcal{P}_x$ is continuous, with $\mathcal{P}_x = [\underline{p}_x, \bar{p}_x]$, a similar result obtains if we replace the marginal LATES $\alpha(p_{j,x}; x)$ by their limit as $q \to q'$, the marginal treatment effect (Heckman, 1997):

$$\tau(q, q'; x) = \frac{1}{p(q,x) - p(q',x)} \int_{p(q',x)}^{p(q,x)} \mathrm{mte}(p; x)\,\mathrm{d}p, \qquad \mathrm{mte}(p; x) = \frac{\partial}{\partial p} \mathbb{E}[Y_i \mid P_i = p, X_i = x],$$

where the equality follows from Equation (18) and the fundamental theorem of calculus. To keep the exposition simple, I will focus on the case with discrete instruments and finite support $\mathcal{P}_x$. The results in this paper generalize easily to the continuous case by replacing $\alpha(p_m; x)$ with the marginal treatment effect, $(p_{m+1,x} - p_{m,x})$ with $\mathrm{d}p$, and replacing sums with integrals.

5 Estimands under the LATE framework

This section presents the first main result of the paper: the estimands of two-step IV estimators and minimum distance estimators when we do not restrict treatment effect heterogeneity.

I derive this result in two steps. First, in Lemma 1 below, express their probability limits in terms of the reduced-form parameter Ξ, defined in Equation (14)—this result does not require any modeling assumptions. Second, I assume the local average treatment effects framework, and I express these reduced-form limits in terms of local average treatment effects.

5.1 Reduced-form limits

Lemma 1. *Suppose that the data $\{Y_i, T_i, Q_i, Z_i, X_i, W_i\}_{i=1}^n$ are i.i.d with finite second moments.*

(i) *Consider a two-step IV estimator $\hat{\beta}_{\hat{\mathbf{P}}}$ that satisfies $\hat{\mathbf{P}}'\mathbf{Y}/n \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L Y_i]$ and $\hat{\mathbf{P}}'\mathbf{T}/n \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L T_i] \neq 0$, where $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i]$. Then:*

$$\hat{\beta}_{\hat{\mathbf{P}}} \xrightarrow{p} \frac{\mathbb{E}[\tilde{P}_i^L Y_i]}{\mathbb{E}[\tilde{P}_i^L T_i]} = \frac{\Xi_{12}}{\Xi_{22}}.$$

(ii) *Consider a reverse two-step IV squares estimator $\hat{\beta}_{\hat{\mathbf{R}},reverse}$ that satisfies $\hat{\mathbf{R}}'\mathbf{Y}/n \xrightarrow{p} \mathbb{E}[\tilde{R}_i^L Y_i]$ and $\hat{\mathbf{R}}'\mathbf{T}/n \xrightarrow{p} \mathbb{E}[\tilde{R}_i^L T_i] \neq 0$. Then:*

$$\hat{\beta}_{\hat{\mathbf{R}},reverse} \xrightarrow{p} \frac{\mathbb{E}[\tilde{R}_i^L Y_i]}{\mathbb{E}[\tilde{R}_i^L T_i]} = \frac{\Xi_{11}}{\Xi_{12}},$$

where $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] - \mathbb{E}^*[Y_i \mid W_i]$.

(iii) *Consider a minimum distance estimator $\hat{\beta}_{\hat{\Xi},\hat{S}}$ that satisfies $\hat{S} \xrightarrow{p} S$ for some positive definite matrix S, and $\hat{\Xi} \xrightarrow{p} \Xi$. Suppose that $\Xi_{22} \neq \min\operatorname{eig}(S^{-1}\Xi)S_{22}$. Then $\hat{\beta}_{\hat{\Xi},\hat{S}} \xrightarrow{p} \beta_S$, where β_S minimizes* *the objective function*

$$\mathcal{D}_S(\beta, \Lambda) = \operatorname{vec}(\Xi - aa'\Lambda)'(S^{-1} \otimes S^{-1}) \operatorname{vec}(\Xi - aa'\Lambda), \tag{21}$$

and it is given by

$$\beta_S = \frac{\Xi_{12} - S_{12} \min \operatorname{eig}(S^{-1}\Xi)}{\Xi_{22} - S_{22} \min \operatorname{eig}(S^{-1}\Xi)}.$$

Lemma 1 shows that understanding how the reduced-form parameter Ξ relates to local average treatment effects is the key to understanding the properties of estimators based on the classic linear IV model.

In particular, Part (i) shows that the probability limit of TSLS and other two-step IV estimators is simply given by Ξ_{12}/Ξ_{22}, the estimand of an IV estimator that uses the linear predictor of the treatment (with the effect of the covariates partialled out), $\tilde{P}_i^L$, as a single instrument. It makes a high-level assumption that the first-step estimator $\hat{P}_i$ converges to its population target, $\tilde{P}_i^L$. The primitive conditions for this depend on the estimator, but for TSLS, a sufficient condition is that $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank.

Part (ii) shows that the probability limit of the reverse two-step IV estimators is given by Ξ_{11}/Ξ_{12}, the estimand of an IV estimator that uses the linear predictor of the outcome (again with the covariates partialled out), $\tilde{R}_i^L$, as a single instrument.

Part (iii) shows that one that one way of thinking about what a minimum distance estimand tries to do is to think of it as trying to be close to both Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12}, using the weight matrix S as a distance metric. The regularity condition $\Xi_{22} \neq \min \operatorname{eig}(S^{-1}\Xi)S_{22}$ ensures that the limiting objective function has a well-defined minimum. Again, the primitive conditions for $\hat{\Xi} \xrightarrow{p} \Xi$ depend on the estimator, but for LIML, a sufficient condition is that $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank.

The rationale for trying to equate the two-step IV estimand Ξ_{12}/Ξ_{22} with the RTSLS estimand Ξ_{11}/Ξ_{12} is that the classic linear IV model is symmetric in Y and T. Both TSLS and RTSLS converge to the same probability limit, equal to the population average treatment effect, so that $\Xi_{11}/\Xi_{12} = \Xi_{12}/\Xi_{22} = \beta$. As a result, Ξ is reduced rank, and there are no trade-offs in how close we can be to Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12}; the weight matrix S does not matter, $\min \operatorname{eig}(S^{-1}\Xi) = 0$ for any positive-definite weight matrix S and all minimum distance estimators converge to the population average treatment effect β. By pooling the information about β contained in TSLS with the information contained in RTSLS, minimum distance estimators have more attractive finite sample properties in classic IV model than two-step IV estimators, which don't use information about β contained in RTSLS (Phillips, 1983; Hillier, 1990). They are also more efficient under many instrument asymptotics (Hausman *et al.*, 2012).

5.2 Interpreting the reduced-form limits under the LATE framework

The key question is how the interpretation of two-step IV, RTSLS, and minimum distance estimands changes under the LATE framework when Assumption CTE in the classic IV model is replaced by Assumption M. I first answer this question for two-step IV and RTSLS estimands in Theorem 1 and Corollary 1 below by expressing the two ratios Ξ_{11}/Ξ_{12} and Ξ_{12}/Ξ_{22} in terms of the marginal local average treatment effects $\alpha(\cdot)$ defined in Equation (19).

Theorem 1. *Suppose that Assumptions IV, L and M hold. Let F^X denote the distribution of X_i. Then*

$$\frac{\Xi_{12}}{\Xi_{22}} = \int \sum_{j=1}^{J_x-1} \frac{\theta_j(x)}{\int \sum_{j=1}^{J_x-1} \theta_j(x)\,\mathrm{d}F^X(x)} \alpha(p_{j,x}; x)\,\mathrm{d}F^X(x),$$

and, if $\Xi_{12} \neq 0$

$$\frac{\Xi_{11}}{\Xi_{12}} = \int \sum_{j=1}^{J_x-1} \frac{\zeta_j(x)}{\int \sum_{j=1}^{J_x-1} \zeta_j(x)\,\mathrm{d}F^X(x)} \alpha(p_{j,x}; x)\,\mathrm{d}F^X(x),$$

where

$$\begin{aligned}\theta_j(x) &= (p_{j+1,x} - p_{j,x})\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}],\\ \zeta_j(x) &= (p_{j+1,x} - p_{j,x})\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{j,x}].\end{aligned}$$

Theorem 1 shows that both Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12} can be expressed as an affine combination of (marginal) local average treatment effects (the weights integrate to one, but are not necessarily positive).

For the two-step IV weights $\theta_j(x)/\int \sum_{j=1}^{J_x-1} \theta_j(x)\,\mathrm{d}F^X(x)$ to be positive, we need that the single instrument $\tilde{P}_i^L$ is monotone in the propensity score P_i. This ensures that the last term $\mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}]$ is always positive. In other words, we need the linear approximation P_i^L to the true propensity score P_i to be good enough in the sense that changing the value of $\tilde{P}_i^L$ does not induce two-way flows in and out of treatment (see Heckman and Vytlacil (2005) and Heckman *et al.* (2006) for discussion of this issue). If the linear approximation to the propensity score is exact, so that $P_i = P_i^L$, then the weights are guaranteed to be positive. A leading case in which this condition holds automatically is when Q_i and X_i are both finite, and we estimate a saturated model in which the instruments Z_i are generated by interacting indicators for different values of Q_i with indicators for different values of X_i. In this case, Angrist and Imbens (1995) obtain a similar expression for the weights $\theta_j(x) / \int \sum_{j=1}^{J_x-1} \theta_j(x) \, \mathrm{d}F^X(x)$.

Similarly, the RTSLS weights $\zeta_j(x) / \int \sum_{j=1}^{J_x-1} \zeta_j(x) \, \mathrm{d}F^X(x)$ are positive if the single instrument $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid W_i, Z_i] - \mathbb{E}^*[Y_i \mid W_i]$ used by RTSLS is monotone in the propensity score P_i. The next corollary gives a necessary and sufficient condition for this condition to hold if the linear approximations (3)–(4) are exact.

Corollary 1. *Suppose that the linear approximations (3)–(4) are exact, so that $\mathbb{E}[Y_i \mid Q_i, X_i] = \mathbb{E}^*[Y_i \mid Z_i, W_i]$ and $\mathbb{E}[T_i \mid Q_i, X_i] = \mathbb{E}^*[T_i \mid Z_i, W_i]$, and that Assumptions IV, L and M hold. Then the weights $\theta_j(x) / \int \sum_{j=1}^{J_x-1} \theta_j(x) \, \mathrm{d}F^X(x)$ are positive, and the weights $\zeta_j(x) / \int \sum_{j=1}^{J_x-1} \zeta_j(x) \, \mathrm{d}F^X(x)$ are positive if all marginal LATEs $\{\alpha(p_j(x); x)\}$ have the same sign. In the special case that $J_x = 2$ for all x,*

$$\theta_1(x) = \operatorname{var}(P_i \mid X_i = x), \qquad \zeta_1(x) = \operatorname{var}(P_i \mid X_i = x)\alpha(p_{1,x}; x).$$

The proof relies on the fact that if the linear approximations (3)–(4) are exact, then the conditional expectation of $R_i = R_i^L$ can be decomposed as

$$\mathbb{E}[R_i^L \mid P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

so that R_i^L is only monotone in the propensity score if the marginal LATEs $\alpha(\cdot)$ all have the same sign. The other implication of this decomposition is that it demonstrates that the conditional expectation of the instrument R_i^L, and hence the term $\mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{j,x}]$ depend on the size of the marginal treatment effects $\alpha(\cdot)$. In the special case that Q_i is binary, so that $J_x = 2$, and the instruments Z_i are generated by interacting Q_i with the covariates, this results in the weights ζ to be exactly equal to the product of the marginal treatment effect with the two-step IV weights θ. Therefore, larger local average treatment effects receive more weight, and negative local average treatment effects receive a negative weight in this case.

Taken together, Lemma 1, Theorem 1 and Corollary 1 show that under Assumptions IV, L and M, two-step IV estimators estimate a convex combination of local average treatment effects, so long as the linear approximation P_i^L to the true propensity score P_i is monotone in P_i. In the special case with a binary Q_i, Corollary 1 shows that these weights are given by the variance of the propensity score, so that better identified LATES receive more weight. If in fact all LATES are equal, then this weighting scheme ensures that under homoscedasticity, asymptotic variance of two-step IV estimators is minimized. On the other hand, the weighting used by the RTSLS estimand is different, depends on the size of the local average treatment effects, and may result in an estimand outside of the convex hull of LATES if some LATES are positive and some are negative.

Because it gives more weight to larger LATES, the RTSLS estimand will always be larger than the two-step iv estimand. This result holds in general by the Cauchy-Schwarz inequality since Ξ is a covariance matrix of $(\tilde{R}_i^L, \tilde{P}_i^L)$,

$$\Xi = \mathbb{E}[(\tilde{R}_i^L, \tilde{P}_i^L)'(\tilde{R}_i^L, \tilde{P}_i^L)],$$

so that $\Xi_{11}\Xi_{22} \geq \Xi_{12}^2$. Hence, $|\Xi_{11}/\Xi_{12}| \geq |\Xi_{12}/\Xi_{22}|$, with equality only if $\tilde{P}_i^L$ is perfectly correlated with $\tilde{R}_i^L$, in which case the RTSLS weights are proportional to the two-step IV weights. There are only two ways how this can happen: either all local average treatment effects are equal, or else the dimension of Z_i is one, so that the IV model (7) is exactly identified. In general with more than one instrument, it will be the case that $\Xi_{11}/\Xi_{12} > \Xi_{12}/\Xi_{22}$.

The result that the RTSLS estimand is in general different from the two-step IV estimand has important implications for minimum distance estimators. On the one hand, combining RTSLS with TSLS leads to more attractive properties of minimum distance estimators in the classic IV model under which $\Xi_{11}/\Xi_{12} = \Xi_{12}/\Xi_{22}$. On the other hand, trying to equate RTSLS and TSLS when their estimands are in fact different makes minimum distance estimators unattractive under treatment effect heterogeneity; as I discuss next, it may cause the minimum distance estimands to no longer correspond to a causal effect.

If the local average treatment effects are not all equal, then $\Xi_{11}/\Xi_{12} \neq \Xi_{12}/\Xi_{22}$, and the probability limit of a minimum distance estimator depends on the weight matrix S. If the weight matrix is diagonal, then the minimum distance estimand lies between TSLS and RTSLS estimands—this was first shown in Zellner (1970) in an errors-in-variables context. Therefore, the symmetrically normalized two-stage least squares estimator (see page 14 for definition), for example, which uses the identity matrix as a weight matrix will always lie between two-step IV and RTSLS estimands. The relative weight given to the RTSLS and TSLS estimands depends on the ratio S_{11}/S_{22}. In particular if the ratio S_{11}/S_{22} is small, then the penalty from being far away from Ξ_{11}/Ξ_{12} is large, so the minimum distance estimand will be close to the RTSLS estimand. On the other hand, if S_{11}/S_{22} is large, then the minimum distance estimand will be close to the two-step IV estimand (see Zellner (1970) and Keller (1975) for a detailed discussion). Heuristically, if we concentrate Λ out of the objective function $\mathcal{D}_S(\beta, \Lambda)$ given in (21), we obtain that

$$\beta_S = \operatorname*{argmin}_{\beta} \frac{\Xi_{22}\beta^2 - 2\Xi_{12}\beta + \Xi_{11}}{S_{11} + S_{22}\beta^2}.$$

Now, if we set $S_{22} = 0$, then $\beta_S = \Xi_{12}/\Xi_{22}$, and if we set $S_{11} = 0$, then we obtain $\beta_S = \Xi_{11}/\Xi_{12}$.

If S is non-diagonal, however, then the minimum distance estimand may lie outside the interval formed by the two-step IV and RTSLS estimands. This is typically the case for LIML, for which S equals the covariance matrix of the reduced-form errors, which is typically non-diagonal. To see how this may happen, consider a simple model in which we observe draws of a vector A_i, distributed according to the bivariate Normal distribution with mean $(\mu_1, \mu_2)'$ and covariance matrix Ω. If $\mu_1 = \mu_2$, and Ω is known, then the optimal estimator is given by:

$$\hat{\mu} = \frac{\iota'\Omega^{-1}\overline{A}}{\iota'\Omega^{-1}\iota'}, \qquad \iota = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where $\overline{A} = n^{-1}\sum_{i=1}^{n} A_i$. The probability limit of this estimator is given by

$$\mu_\Omega = \frac{(\Omega_{22} - \Omega_{12})\mu_1 + (\Omega_{11} - \Omega_{12})\mu_2}{\Omega_{22} + \Omega_{11} - 2\Omega_{12}}.$$

If $\Omega_{12} = 0$, then μ_Ω lies between μ_1 and μ_2. If, however, Ω is non-diagonal, then this may no longer be the case—if, for example, $\mu_2 = 0$ and μ_1 is positive, then μ_Ω will be negative if $\Omega_{22} < \Omega_{12}$.

There are two ways, therefore, in which a minimum distance estimand may end up being outside of the convex hull of LATEs. First, if some LATEs are positive and some are negative, and the RTSLS estimand is outside of the convex hull, then so long as the weight matrix S gives sufficient weight to RTSLS, the minimum distance estimand will also be outside of the convex hull. Second, even if the RTSLS estimand lies inside the convex hull, if the weight matrix S is non-diagonal, the minimum distance estimand may end up being outside of the convex hull. These possibilities make LIML and other minimum distance estimators an unattractive estimator choice in settings with possible treatment effect heterogeneity.

6 Estimation with many instruments

In this section, I derive the second main result of the paper that a version of the jackknife instrumental variables estimator, the unbiased jackknife instrumental variables estimator (UJIVE), is consistent for a convex combination of LATEs under a many instrument asymptotic sequence in which both the number of instruments and the number of covariates is allowed to increase in proportion with the sample size. In settings with many instruments and treatment effect heterogeneity, UJIVE is therefore a more attractive estimator than TSLS, which is inconsistent under many instrument asymptotics. It is also more attractive than LIML, the standard alternative to TSLS when many instruments are used, since, as shown in Section 5, LIML may converge to a quantity outside of the convex hull of local average treatment effects even under standard asymptotics.

To illustrate the issues that arise with a large number of instruments, as well as to motivate UJIVE, I first discuss a simple example in which the instruments Z_i are indicators for group membership. I then give the general consistency theorem.

6.1 A simple example with groups as instruments

Consider the special case in which the instruments are indicators for group membership, $Z_{ik} = \mathbb{1}_{Q_i=k}$, where $Q_i \in \{1, \dots, K+1\}$ indexes groups (Z_i omits the indicator for the last group so that we can include the intercept).

For instance, the instruments could be judge indicators as in Aizer and Doyle, Jr. (2011) and Nagin and Snodgrass (2011) to instrument for length of sentence or incarceration. The identification strategy in these papers relies on the fact that cases are randomly assigned to judges who vary in their sentencing severity.[5] . In this context, the monotonicity assumption requires that the judges can be ordered in terms of how strict they are. The local average treatment effects are defined for each pair of judges and correspond to the average treatment effect for individuals who would get incarcerated if assigned to the stricter judge of the two, but would not get incarcerated if assigned to the more lenient judge. If the effect of incarceration for more serious offenders (who get incarcerated unless assigned to the most lenient judges) is different from the effect for individuals who committed less serious crimes (who only get incarcerated if assigned to the strictest judges), then these LATEs will differ.

In the absence of covariates (beyond the intercept), the propensity score for individual i

[5] A similar strategy is also used in Dobbie and Song (2012), who study the effect of being granted bankruptcy protection on subsequent earnings, using judge indicators as instruments.

is given by $P_i = \mathbb{E}[T_i \mid Q_i]$. Because the first stage is saturated, the linear approximation (4) is exact, and $\tilde{P}^L_i = \tilde{P}_i = P_i - \mathbb{E}[P_i]$. In the judges example, $\tilde{P}_i$ measures how strict the judge assigned to individual i is compared to other judges.

Let J_k denote the number observations in group k. The two-stage least squares estimator of $\hat{P}_{i,\textsc{tsls}} = (\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T})_i$ of $\tilde{P}_i$ can be written as

$$\hat{P}_{i,\textsc{tsls}} = \hat{T}_{i,\textsc{tsls}} - n^{-1}\sum_{j=1}^{n} \hat{T}_{j,\textsc{tsls}} = \hat{T}_{i,\textsc{tsls}} - n^{-1}\sum_{j=1}^{n} T_j,$$

where $\hat{T}_{i,\textsc{tsls}} = J_{Q_i}^{-1}\sum_{j\colon Q_j=Q_i} T_j$ is the predictor of T_i based on least-squares estimation of the first-stage (4), and it is the simplest estimator of P_i. The resulting TSLS estimator is given by

$$\hat{\beta}_{\textsc{tsls}} = \frac{n^{-1}\sum_i \hat{P}_{i,\textsc{tsls}} Y_i}{n^{-1}\sum_i \hat{P}_{i,\textsc{tsls}} T_i}. \tag{22}$$

There are two basic ways of doing asymptotics in this setting. The first option is to let the number of observations per group grow to infinity while keeping the number groups fixed. This corresponds to the standard asymptotics. As J_{Q_i} increases, $\hat{T}_{i,\textsc{tsls}} \xrightarrow{p} P_i$, and the numerator and the denominator in (22) converge to $\mathbb{E}[\tilde{P}_i Y_i]$ and $\mathbb{E}[\tilde{P}_i T_i]$, respectively. By Lemma 1 and Theorem 1, $\hat{\beta}_{\textsc{tsls}}$ therefore converges to a weighted average of local average treatment effects. However, with a large number of groups and a small number of observations per group, these asymptotics do not capture the finite-sample properties of the estimator very well.

The other possibility is to keep J_k fixed, and let the number of groups $K \to \infty$. This corresponds to the many instrument asymptotics (Kunitomo, 1980; Morimune, 1983; Bekker, 1994) that let the dimension of Z_i increase in proportion with the sample size. Under these asymptotics, P_i can no longer be consistently estimated, and so the exact way in which it is estimated will matter. The problem with the TSLS estimator $\hat{T}_{i,\textsc{tsls}}$ is that since it includes own observation T_i, its estimation error is correlated with Y_i and T_i. As a result, the numerator and the denominator in (22) no longer converge to $\mathbb{E}[\tilde{P}_i Y_i]$ and $\mathbb{E}[\tilde{P}_i T_i]$. To see this, let $V_{1,i} = Y_i - R_i$ and $V_{2,i} = T_i - P_i$ denote errors in the reduced form (1)–(2), and let $K/n \to \kappa > 0$. Then we can write $\hat{T}_{i,\text{TSLS}} = P_i + J_{Q_i}^{-1} \sum_{j\colon Q_j = Q_i} V_{2,j}$. We have

$$\begin{aligned}
\frac{1}{n}\sum_i \hat{P}_{i,\text{TSLS}} Y_i &= \frac{1}{n}\sum_i \hat{T}_{i,\text{TSLS}} Y_i - \frac{1}{n^2}\sum_i\sum_j T_j Y_i \\
&= \frac{K}{n}\left(\frac{1}{K}\sum_k \frac{1}{J_k}\sum_{j\colon Q_j=k} V_{2,j} \sum_{i\colon Q_i=k} Y_i\right) + \left(\frac{1}{n}\sum_i P_i Y_i\right) - \left(\frac{1}{n}\sum_j T_j\right)\left(\frac{1}{n}\sum_i Y_i\right) \qquad (23)\\
&\xrightarrow{p} \kappa \operatorname{cov}(V_{2,i}, V_{1,i}) + \mathbb{E}[Y_i \tilde{P}_i],
\end{aligned}$$

where the last line follows from the law of large numbers applied to all four expressions in parentheses, and the fact that $\mathbb{E}[\frac{1}{J_k}\sum_{j\colon Q_j=k} V_{2,j} \sum_{i\colon Q_i=k} Y_i] = \mathbb{E}[V_{2,i} Y_i] = \mathbb{E}[V_{2,i} V_{1,i}]$. Similarly, for the denominator, $n^{-1}\sum_i \hat{P}_{i,\text{TSLS}} T_i \xrightarrow{p} \kappa \operatorname{var}(V_{2,i}) + \mathbb{E}[T_i \tilde{P}_i]$. Therefore, TSLS is inconsistent for its target, $\mathbb{E}[\tilde{P}_i Y_i]/\mathbb{E}[\tilde{P}_i T_i]$.

There are two basic ways of adjusting the TSLS estimator to make it work under many instruments. First is to estimate the unconditional covariance matrix of $V_i = (V_{1,i}, V_{2,i})$ and subtract an estimate of the bias. This is exactly the idea behind the bias-corrected two-stage least squares estimator of Nagar (1959) and Donald and Newey (2001). Unfortunately, the estimator of the bias is only consistent under homoscedasticity (Bekker and van der Ploeg, 2005; Ackerberg and Devereux, 2009), and it is unclear how to estimate $\operatorname{var}(V_i)$ consistently when $\operatorname{var}(V_i \mid Q_i, X_i)$ is heteroscedastic.

The second approach is to change the estimator of P_i so that it does not include own observation T_i. This is the idea behind the (leave-one-out) jackknife instrumental variables estimator (JIVE, Phillips and Hale, 1977; Angrist *et al.*, 1999). It replaces $\hat{T}_{i,\text{TSLS}}$ with $\hat{T}_{i,\text{JIVE}} = (J_{Q_i} - 1)^{-1}\sum_{j\colon Q_j = Q_i, j\neq i} T_j$.The JIVE estimator of $\tilde{P}_i$ is given by

$$\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1}\sum_{j=1}^{n} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1}\sum_{j=1}^{n} T_j. \qquad (24)$$

The estimation error $P_i - \hat{T}_{i,\text{JIVE}} = \sum_{j\colon Q_j=Q_i, j\neq i} V_j$ is no longer correlated with T_i or Y_i, and the JIVE estimator is consistent for a convex combination of LATEs under both types of asymptotics.

So far, the discussion has abstracted from the presence of covariates. In the judges example, however, the judges are only randomly assigned at the county level. Therefore, with data from several counties, we need to include county indicators (sometimes called "fixed effects") as covariates. With L counties, $\dim(W_i) = L$, and $W_{i\ell} = \mathbb{1}_{X_i = \ell}$, where X_i indexes counties. The propensity score P_i still corresponds to the incarceration propensity of judge Q_i. However, we now have $\tilde{P}_i^L = \tilde{P}_i = P_i - \mathbb{E}[P_i \mid X_i]$, so that $\tilde{P}_i$ measures how strict judge Q_i is compared to other judges that individual i could have been assigned in the county.

The JIVE estimator of P_i now becomes

$$\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - m_{X_i}^{-1} \sum_{j\colon X_j = X_i} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - m_{X_i}^{-1} \sum_{j\colon X_j = X_i} T_j,$$

where m_ℓ is the number of cases in county ℓ. With a large number of counties, a natural way of thinking about the sampling is to let the number of counties $L \to \infty$, while keeping the number of judges per county and the number of cases per judge fixed. This is similar to the many instrument asymptotics in that the number of judges increases in proportion to the sample size, $K/n \to \kappa > 0$, except that instead of keeping the number of counties fixed, we also let them to grow in proportion with sample size, so that $L/n \to \lambda$. This modification of the many instrument asymptotic sequence was proposed by Anatolyev (2011) and Kolesár *et al.* (2011), and it is also used in Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011).

Under these asymptotics, the JIVE estimator is biased. The problem is not its estimate of the propensity score—we still have that $n^{-1}\sum_i \hat{T}_{i,\text{JIVE}} Y_i \xrightarrow{p} \mathbb{E}[P_i Y_i]$, and $n^{-1}\sum_i \hat{T}_{i,\text{JIVE}} T_i \xrightarrow{p} \mathbb{E}[P_i T_i]$. Instead, the source of bias comes from its estimate of the average strictness of judges in county X_i, $m_{X_i}^{-1}\sum_{j\colon X_j = X_i} T_j$. By the same logic as in the case of TSLS with many instruments, the problem is that this estimate includes own observation T_i, so that the estimation error $\mathbb{E}[P_i \mid X_i] - m_{X_i}^{-1}\sum_{j\colon X_j = X_i} T_j$ is correlated with Y_i and T_i. By arguments similar to those used to derive Equation (23), we have

$$\hat{\beta}_{\text{JIVE}} = \frac{n^{-1}\sum_i \hat{P}_{i,\text{JIVE}} Y_i}{n^{-1}\sum_i \hat{P}_{i,\text{JIVE}} T_i} \xrightarrow{p} \frac{\mathbb{E}[\tilde{P}_i Y_i] - \lambda \operatorname{cov}(V_{1,i} V_{2,i})}{\mathbb{E}[\tilde{P}_i T_i] - \lambda \operatorname{var}(V_{2,i})}. \tag{25}$$

This probability limit may differ substantially from the target $\mathbb{E}[\tilde{P}_i Y_i]/\mathbb{E}[\tilde{P}_i T_i]$, especially in settings in which the concentration parameter $\mathbb{E}[\tilde{P}_i T_i]/\operatorname{var}(V_{2,i})$ is small. As a result, JIVE can be severely biased in finite samples.

The unbiased jackknife instrumental variables estimator (UJIVE) that I propose solves the bias problem of JIVE by also leaving out own observation when estimating $\mathbb{E}[P_i \mid X_i]$:

$$\hat{P}_{i,\text{UJIVE}} = \hat{T}_{i,\text{JIVE}} - \frac{1}{m_{X_i} - 1} \sum_{j\colon X_j = X_i, j \neq i} T_j.$$

Intuitively, $\hat{P}_i$ is a sample measure of how strict judge Q_i is relative to other judges in country X_i in a sample that excludes individual i. This estimator of $\hat{P}_i$ was first used in Chetty *et al.* (2011) in a setting with the same formal structure as the current example. In particular, Chetty *et al.* (2011) used classroom indicators as instruments for test score, conditioning on schools. The next subsection gives a general formula for UJIVE, and proves that it is consistent under many instrument asymptotics that also allow for many covariates.

6.2 Consistency of UJIVE under many instruments

Consider now the general case. Let ϕ denote the coefficient on W_i in the linear projection $\mathbb{E}^*[T_i \mid W_i]$. To define UJIVE, decompose $\tilde{P}_i^L$, the linear approximation to the propensity score with the effect of covariates partialled out, as

$$\begin{aligned} \tilde{P}_i^L &= \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i] \\ &= Z_i'\pi_2 + W_i'\psi_2 - W_i'\phi. \end{aligned}$$

Let $\hat{\pi}_{2\setminus i}$ and $\hat{\psi}_{2\setminus i}$ be the least-squares estimates of π_2 and ψ_2 based on a sample with observation i removed. Similarly, let $\hat{\phi}_{\setminus i}$ be the least-squares estimate of ϕ based on a sample with observation i removed. The UJIVE estimator is a two-step IV estimator with the first-step estimator of $\tilde{P}_i^L$ given by

$$\hat{P}_{i,\text{UJIVE}} = Z_i'\hat{\pi}_{2\setminus i} + W_i'\hat{\psi}_{2\setminus i} - W_i'\hat{\phi}_{\setminus i}.$$

In matrix notation

$$\hat{\mathbf{P}}_{\text{UJIVE}} = \hat{\mathbf{T}}_{\text{UJIVE}} - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}})^{-1}(\mathbf{H}_{\mathbf{W}} - \mathbf{D}_{\mathbf{W}})\mathbf{T},$$

where $\hat{\mathbf{T}}_{\text{UJIVE}} = (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}(\mathbf{H}_{(\mathbf{Z},\mathbf{W})} - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})\mathbf{T}$. Using $\hat{\mathbf{P}}_{\text{UJIVE}}$ as a single instrument in an IV estimator then yields

$$\hat{\beta}_{\text{UJIVE}} = \frac{\hat{\mathbf{P}}_{\text{UJIVE}}'\mathbf{Y}}{\hat{\mathbf{P}}_{\text{UJIVE}}'\mathbf{T}}.$$

In contrast, while the JIVE estimator of $\mathbb{E}^*[Y_i \mid Z_i, W_i]$ is identical to $\hat{\mathbf{T}}_{\text{UJIVE}}$, its estimator of $\mathbb{E}^*[Y_i \mid W_i]$ is given by a sample projection of $\hat{\mathbf{T}}_{\text{UJIVE}}$ onto $\mathbf{W}$, so that $\hat{\mathbf{P}}_{\text{JIVE}} = \hat{\mathbf{T}}_{\text{UJIVE}} - \mathbf{H}_{\mathbf{W}}\hat{\mathbf{T}}_{\text{UJIVE}}$ (see the JIVE formula on page 12).

To formally define the many instrument asymptotic framework, I need to allow the distribution of random variables to change with the sample size. To reflect this, let the random variables be indexed by n, so that, for instance, $\mathbf{Y}_n = (Y_{n,1}, \ldots, Y_{n,n})'$ denotes the vector of observed outcomes when the sample size is n. In addition, let $P_{n,i}^X = \mathbb{E}[T_{n,i} \mid X_{n,i}]$ and $R_{n,i}^X = \mathbb{E}[Y_{n,i} \mid X_{n,i}]$ denote the expectations of $T_{n,i}$ and $Y_{n,i}$ conditional on $X_{n,i}$ only, so that $P_{n,i}^X = \mathbb{E}[P_{n,i} \mid X_{n,i}]$ and

$R^X_{n,i} = \mathbb{E}[R_{n,i} \mid X_{n,i}]$. The many instrument asymptotic framework I consider is summarized by the following assumptions:

Assumption R (Regularity conditions).

(i) $\{(Y_{n,i}, T_{n,i}, X_{n,i}, Q_{n,i}) \colon i = 1, \dots, n\}_{n \geq 1}$ is a triangular array of i.i.d. random variables, the nth row having distribution $F_n^{Y,T,X,Q}$. $F_n^{Y,T,X,Q}$ converges in distribution to $F^{Y,T,X,Q}$;

(ii) There is a positive constant C_1, such that $\sup_n \sup_{i \leq n} \operatorname{var}(Y_{n,i} \mid Q_{n,i}, X_{n,i}) \leq C_1$, and $\sup_n \sup_{i \leq n} \operatorname{var}(Y_{n,i} \mid X_{n,i}) \leq C_1$ a.s. Also, as $n \to \infty$,

$$\mathbb{E}[(R^2_{n,i}, P^2_{n,i}, |R_{n,i}P_{n,i}|)] \to \mathbb{E}[(R^2, P^2, |RP|] < \infty,$$
$$\mathbb{E}[((R^X_{n,i})^2, (P^X_{n,i})^2, |R^X_{n,i}P^X_{n,i}|)] \to \mathbb{E}[((R^X)^2, (P^X)^2, |R^X P^X|] < \infty,$$

where (R, P, R^X, P^X) is distributed according to the limiting distribution F^{R,P,R^X,P^X}; and

(iii) $\operatorname{rank}(\mathbf{Z}_n, \mathbf{W}_n) = K + L$ and $(\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)})_{ii} < C_2$ for some $C_2 < 1$ a.s., where $Z_{n,i} = z(Q_{n,i}, X_{n,i})$ and $W_{n,i} = w(X_{n,i})$, with $\dim(Z_{n,i}) = K$ and $\dim(W_{n,i}) = L$. The functions z and w may depend on n.

Assumption MI (Many instruments). As $n \to \infty$:

(i) $K/n \to \kappa$ and $L/n \to \lambda$ for some $\kappa, \lambda \geq 0$;

(ii) $\sum_i (\mathbb{E}[T_{n,i} \mid X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid W_{n,i}])^2/n \to 0$ a.s.; and

(iii) $\sum_i (\mathbb{E}[T_{n,i} \mid Q_{n,i}, X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid Z_{n,i}, W_{n,i}])^2/n \to 0$ a.s.

Assumption R (i) allows the distribution of the data to change with the sample size, converging to some limiting distribution $F^{Y,T,X,Q}$. Part (ii) requires that the second moments of conditional expectations of $Y_{n,i}$ and $T_{n,i}$ exist and are well-behaved in the limit. It is necessary for sample averages such as $n^{-1}\sum_{i=1}^n R^2_{n,i}$ to have a well-specified probability limit. The restriction $\operatorname{rank}(\mathbf{Z}, \mathbf{W}) = K + L$ in Part (iii) is a normalization. The assumption that $(\mathbf{H}_{(\mathbf{Z},\mathbf{W})})_{ii} < C_2$ requires that no single observation has too much leverage. If the instruments are group indicators, then we need at least two observations per group. It implies that $(K + L)/n < C_2$ since $n^{-1}\sum_i (\mathbf{H}_{(\mathbf{Z},\mathbf{W})})_{ii} = (K + L)/n$.

Assumption MI (i) generalizes the many instrument asymptotic sequence by also allowing the number of covariates to increase with the sample size. In terms of the incarceration example, the original Bekker (1994) many instruments sequence keeps the number of counties as well as the number of cases per judge fixed, and lets the number of judges per county increase to infinity. Under Assumption MI, we can think of generating the data by sampling L counties form some large population of counties. In Angrist and Krueger (1991), where Z_i is generated by interacting quarter of birth with L state of birth and year of birth indicators, Assumption MI (i) lets the number of states and years $L \to \infty$, while keeping the number of individuals observed in each state and year fixed. Finally, Assumption MI also accommodates models in which z and w are some approximating functions, such as splines or polynomials in the basic instruments and covariates Q_i and X_i. This corresponds to fixing the distribution of the data, so that $F_n^{Y,T,X,Q} = F^{Y,T,X,Q}$, and letting the number of terms in the approximating functions w and z increase with the sample size. Parts (ii)–(iii) then require that these approximating functions get to their population targets in the limit, and allow me to relax the requirement imposed by Assumption L that expectation of Z_i conditional on X_i is exactly linear in W_i in the sample. These conditions are similar to the assumptions in Bekker (1994) and Hansen, Hausman and Newey (2008).

Note that I do not make any assumptions about the coefficients on $Z_{n,i}$ and $W_{n,i}$ in the projections $\mathbb{E}^*[T_{n,i} \mid W_{n,i}, Z_{n,i}]$ and $\mathbb{E}^*[T_{n,i} \mid W_{n,i}]$. Under additional assumptions, such as sparsity (only few coefficients in these linear projections matter), approximations to $\tilde{P}^L_{n,i}$ other than $\hat{P}_{n,i,\textsc{ujive}}$ will work (see, for example, Belloni, Chen, Chernozhukov and Hansen, 2012).

Theorem 2. *Suppose that Assumptions R and MI hold, and that the limiting distribution $F^{Y,T,X,Q}$ of the data satisfies Assumptions IV and M. Then:*

$$\begin{aligned}\hat{\beta}_{\textsc{ujive}} &\xrightarrow{p} \frac{\mathbb{E}[Y(P - \mathbb{E}[P \mid X])]}{\mathbb{E}[T(P - \mathbb{E}[P \mid X])]} \\ &= \int \sum_{j=1}^{J_x - 1} \frac{\theta_j(x)}{\int \sum_{j=1}^{J_x-1} \theta_j(x)\, \mathrm{d}F^X(x)} \alpha(p_{j,x}; x)\, \mathrm{d}F^X(x),\end{aligned}$$

where (Y, T, P, X) are distributed according to the limiting distribution $F^{Y,T,P,X}$, and

$$\theta_j(x) = (p_{j+1,x} - p_{j,x})\mathbb{P}(P > p_{j,x} \mid X = x)\left(\mathbb{E}[P \mid X = x, P > p_{j,x}] - \mathbb{E}[P \mid X = x]\right).$$

Thus, UJIVE estimates a convex combination of local average treatment effects. This conclusion is robust to many instruments, many covariates, and heteroscedasticity.

7 A small simulation study

To illustrate the main implications of the theoretical results, I conducted a small Monte Carlo experiment.

I consider the case in which the covariates are group indicators, $W_{i\ell} = \mathbb{1}_{X_i=\ell}$, and the basic instrument Q_i is binary. The constructed instrument Z_i is given by $Q_i W_i$. For example, Q_i could be quarter of birth indicators and W_i state of birth indicators, as in Angrist and Krueger (1991). Alternatively, Q_i could be an indicator for being assigned the first judge in the judges example when there are only two judges per country X_i. In the simulations, half of individuals within each group are assigned $Q_i = 1$; the other half are assigned $Q_i = 0$.

The data generating process is given by

$$\begin{aligned} Y_i(t) &= t\beta_{X_i} + W_i'\delta + \epsilon_i, \\ T_i(q) &= qW_i'\pi + W_i'\psi_2 + V_{2i}, \end{aligned}$$

with $(\epsilon_i, V_{2i}) \sim \mathcal{N}_2(0, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right))$, $\delta = \psi_2 = 0$, and $\pi = 1$. The design is constructed so that it corresponds to a classic linear IV model with the only exception that β, the marginal treatment effect, may now vary between covariate groups.

I consider two designs for the instruments. In the first design (few instruments), there are only $L = 2$ and $K = 2$ instruments. To create an unbalanced design, I let half of the groups have size $m_1 = 500$, and the other half $m_2 = 100$, so that the sample size is given by $n = 600$. The first-stage F statistic equals approximately 76 on average. In this case, the standard asymptotics should perform well. In the second design (many instruments), I let $L = 20$, $m_1 = 50$, and $m_2 = 10$, so that the sample size is still $n = 600$, with F now only equal to about 8.5 on average.

I consider six estimators: LIML, the reverse TSLS estimator (RTSLS), and four two-step estimators: TSLS, JIVE, UJIVE, and the bias-corrected TSLS estimator, BTSLS (See Section 3 for definitions of these estimators).

As a baseline, Table 1 reports the results for the case when β is constant across groups and equal to 0, so that there is no heterogeneity in the treatment effects. Panel I reports the results for the first instrument design. In this case, all estimators perform well (except RTSLS, which does not converge since $\Xi_{11} = \Xi_{12} = 0$). Panel II reports the result for the second instrument design. The median for the estimators considered is very close to their probability limit under the many-instrument asymptotics. In particular, TSLS is (median) biased to due to the presence of many instruments, and JIVE is biased due to the presence of many covariates. In addition, JIVE is very dispersed. LIML, BTSLS, and UJIVE all perform well with LIML being the least dispersed.

I then set $\beta_\ell = 2$ in the small groups, and $\beta_\ell = 0$ in the large groups. In this case, the two-step IV estimand equals $1/3$, the weighted average of β_ℓ weighted by group size. The

Table 1: *Simulation: No heterogeneity in treatment effects, $\beta_\ell = 0$*

Estimator	Median	Estimand	plim	9DR	IQR
Panel I: few instruments					
LIML	-0.00	0	0	0.27	0.11
TSLS	0.01	0	0	0.27	0.11
BTSLS	0.01	0	0	0.27	0.11
JIVE	-0.02	0	0	0.29	0.12
UJIVE	-0.01	0	0	0.28	0.11
RTSLS	0.04	–	–	1.34	0.30
Panel II: many instruments					
LIML	0.00	0	0	0.28	0.11
TSLS	0.10	0	0.09	0.22	0.09
BTSLS	0.01	0	0.01	0.30	0.12
JIVE	-0.15	0	-0.12	0.47	0.18
UJIVE	-0.01	0	0	0.32	0.13
RTSLS	0.97	–	1.25	9.25	0.97

Median, nine-decile range (9DR), and inter-quantile range (IQR) for different estimators. Estimand refers to the estimand as given by Lemma 1, and plim to refers to the probability limit under standard (Panel I), or many-instrument asymptotics (Panel II).
50,000 simulation draws.

reverse two-step IV estimand puts more weight on the larger treatment effects (see discussion following Corollary 1); in this case put puts all the weight on the non-zero effects, so that the estimand equals 2. Even though all the treatment effects are non-negative, the LIML estimand is negative and equal -0.02 since it uses a non-diagonal weight matrix to equate the forward and reverse two-step IV estimands. Table 2, Panel I reports the results for the design with few instruments: all estimators are median-unbiased relative to their estimands. Panel II reports the results for the design with many instruments. Again, the many-instrument asymptotic limit of the estimators is very close to their finite-sample median. The four two-step IV estimands all behave differently. TSLS and JIVE are median-biased due to the presence of many instruments and many covariates. The Donald and Newey (2001) bias-corrected TSLS estimator is biased due to the presence of heteroscedasticity. Even though the structural errors are homoscedastic, the reduced-form errors aren't since they depend on β_ℓ which varies between groups. This heteroscedasticity in the reduced-form errors biases the estimator upward. As predicted by

Table 2: *Simulation: Heterogeneous treatment effects, $\beta_\ell = 0$ for large groups, $\beta_\ell = 2$ for small groups.*

Estimator	Median	Estimand	plim	9DR	IQR
Panel I: few instruments					
LIML	−0.06	−0.02	−0.02	0.53	0.19
TSLS	0.34	0.33	0.33	0.49	0.20
BTSLS	0.34	0.33	0.33	0.49	0.20
JIVE	0.30	0.33	0.33	0.51	0.21
UJIVE	0.32	0.33	0.33	0.50	0.20
RTSLS	2.04	2.00	2.00	2.12	0.68
Panel II: many instruments					
LIML	−0.25	−0.01	−0.20	1.01	0.34
TSLS	0.51	0.33	0.51	0.44	0.18
BTSLS	0.43	0.33	0.44	0.52	0.21
JIVE	0.08	0.33	0.11	0.80	0.32
UJIVE	0.34	0.33	0.33	0.60	0.24
RTSLS	2.24	2.00	2.23	1.18	0.45

Median, nine-decile range (9DR), and inter-quantile range (IQR) for different estimators. Estimand refers to the estimand as given by Lemma 1, and plim to refers to the probability limit under standard (Panel I), or many-instrument asymptotics (Panel II).

50,000 simulation draws.

Theorem 2, UJIVE remains unbiased. It is also considerably less dispersed than JIVE. In addition to estimating a quantity that is hard to interpret, LIML is now more, rather than dispersed than any of the two-step estimators.

8 Conclusion

In this paper, I derived estimands of estimators based on a classic linear IV model under treatment effect heterogeneity. I assumed that the instruments satisfy the monotonicity condition of Angrist and Imbens (1995), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). If the LATEs for all possible instrument pairs are all equal to each other, then all classic estimators estimate this common local average treatment effect. If the LATEs vary, then, under mild assumptions, estimators in the class of two-step IV estimators estimate the same convex combination of them. This class includes the two-stage least squares estimator (TSLS). The estimand of LIML, however, is different, depends on the reduced-form covariance matrix, and may be outside of the convex hull of the local average treatment effects. This possibility makes LIML unattractive in settings with treatment effect heterogeneity.

Unfortunately, the TSLS estimator is inconsistent under many instrument asymptotics, making it a poor choice of estimator in settings with a large number of instruments. I showed that a different two-step IV estimator, the unbiased jackknife IV estimator (UJIVE), on the other hand, remains consistent for a convex combination of LATEs under a many instrument asymptotic sequence that allows for heteroscedasticity, and lets the number of instruments and covariates increase in proportion with the sample size. I therefore recommend that in settings with many instruments, empirical researchers use UJIVE instead of LIML or TSLS.

Appendix A Multi-valued Treatments

Suppose that instead of being binary, the set of possible treatments is given by an ordered set $\mathcal{T} = \{0, 1, \ldots, t_{\max}\}$. Angrist and Imbens (1995) show that under Assumptions IV and M, the Wald estimand (18) identifies a weighted average of per-unit treatment effects

$$\tau(q, q'; x) = \sum_{t=1}^{t_{\max}} \omega_t(x) \mathbb{E}[Y_i(t) - Y_i(t-1) \mid T(q) \geq t > T(q'), X_i = x],$$

where the weights $\omega_t(x)$ are given by

$$\omega_t(x) = \frac{\mathbb{P}(T_i(q) \geq t > T_i(q') \mid X_i = x)}{\sum_{t'=1}^{t_{\max}} \mathbb{P}(T_i(q) \geq t' > T_i(q') \mid X_i = x)}.$$

Angrist and Imbens (1995) refer to the parameter $\tau(q, q'; x)$ as an average causal response (ACR). If $t_{\max} = 1$, then the expression reduces to (17). Define a marginal ACR by $\alpha(p_m; x) = \tau(q_1, q_2; x)$, where $p(q_1, x) = p_{m+1,x}$, and $p(q_2, x) = p_{m,x}$ (if there is another pair (q_1', q_2') that satisfies this condition, then under Assumptions IV and M, it must be that $\tau(q_1, q_2; x) = \tau(q_1', q_2'; x)$, so that $\alpha(p_m; x)$ is well-defined). We can write all ACRs in terms of these marginal ACRs as in (20).

By Theorem 1, forward and reverse two-step IV estimators estimate a weighted average of these marginal ACRs, with weights given by $\theta_j(x)$ and $\zeta_j(x)$. Compared to the binary treatment case, the only difference is that now the marginal ACR $\alpha(p_m; x)$ is itself a weighted average of per-unit treatment effects.

Consequently, the TSIV and reverse TSIV estimands may differ even if there is no heterogeneity in the treatment effects if the causal response function is non-linear. To see this, suppose that $\mathbb{E}[Y_i(t_1) \mid T_i = t, Q_i = q, X_i = x] = g(t_1)$, for some non-linear function $g(\cdot)$, as in Newey and Powell (2003) or Darolles, Fan, Florens and Renault (2011). In this case Assumption CTE fails unless $g(\cdot)$ is linear. Consequently, different instruments will in general estimate different averages of the per-unit treatment effects $g(t) - g(t-1)$.

Appendix B Auxiliary Lemmata

First I define some notation and collect some basic results that I use throughout Appendices B and C. Let $\mathcal{Z}_n = \{Q_i, X_i\}_{i=1}^n$ denote the collection of covariates and instruments. Also, let

$$\mathbf{G}_n = (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n, \mathbf{W}_n)})^{-1}(\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)} - \mathbf{D}_{(\mathbf{Z}_n, \mathbf{W}_n)}) - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n}). \qquad (26)$$

Then $\hat{\mathbf{P}}_{\text{UJIVE}} = \mathbf{G}_n \mathbf{T}_n$.

Lemma 2. *Let $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i]$, and let $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] - \mathbb{E}^*[Y_i \mid W_i]$. Then:* ***(i)*** $\Xi_{12} = \mathbb{E}[Y_i \tilde{P}_i^L]$ ***(ii)*** $\Xi_{12} = \mathbb{E}[T_i \tilde{R}_i^L]$ ***(iii)*** $\Xi_{11} = \mathbb{E}[Y_i \tilde{R}_i^L]$ ***(iv)*** $\Xi_{22} = \mathbb{E}[T_i \tilde{P}_i^L]$

Proof. Consider Part (i). Observe that

$$\begin{aligned}
\mathbb{E}[Y_i \tilde{P}_i^L] &= \mathbb{E}[(Z_i'\pi_1 + W_i'\psi_1)'\tilde{P}_i^L] \\
&= \mathbb{E}[(Z_i'\pi_1 + W_i'\psi_1)'\tilde{Z}_i'\pi_2] \\
&= \mathbb{E}[\pi_1 Z_i \tilde{Z}_i \pi_2] = \mathbb{E}[\pi_1 \tilde{Z}_i \tilde{Z}_i \pi_2] \\
&= \Xi_{12},
\end{aligned}$$

where the first line follows from Equation (3) and the fact that $\tilde{P}_i^L$ is linear in Z_i and W_i, the second line follows from $\tilde{P}_i^L = \tilde{Z}_i'\pi_1$, the third line follows from $\mathbb{E}[W_i \tilde{Z}_i] = 0$, and the last line follows by definition of Ξ_{12}. Parts (ii)–(iv) follow by similar arguments, using the substitutions $\tilde{R}_i^L = \tilde{Z}_i'\pi_1$ and $\tilde{P}_i^L = \tilde{Z}_i'\pi_2$. □

Lemma 3. *Let $A_i = a(Q_i, X_i)$ be some function of the instruments and covariates such that $\mathbb{E}[A_i \mid X_i] = 0$. Then, under Assumptions IV and M*

$$\mathbb{E}[Y_i A_i] = \int \sum_{j=1}^{J_x - 1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x})\mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\,\mathrm{d}F^X(x),$$

$$\mathbb{E}[T_i A_i] = \int \sum_{j=1}^{J_x - 1} (p_{j,x} - p_{j-1,x})\mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\,\mathrm{d}F^X(x).$$

Proof. First consider $\mathbb{E}[Y_i A_i]$. By the Law of iterated expectations,

$$\begin{aligned}
\mathbb{E}[Y_i A_i] &= \int \sum_j \sum_a a\,\mathbb{E}[Y_i \mid X_i = x, A_i = a, P_i = p_{j,x}]\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\,\mathrm{d}F^X(x) \\
&= \int \sum_j \sum_a a\,\mathbb{E}[Y_i \mid X_i = x, P_i = p_{j,x}]\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\,\mathrm{d}F^X(x)
\end{aligned} \tag{27}$$

where the second line follows from the fact that under Assumptions IV and M, the conditional expectation of Y_i depends only on P_i and X_i. Using the substitution

$$\mathbb{E}[Y_i \mid X_i = x, P_i = p_{j,x}] = \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

we can expand the expression (27) as

$$\begin{aligned}\mathbb{E}[Y_iA_i] &= \int \sum_j \sum_a a \sum_{j'=1}^{j-1} \alpha(p_{j',x};x)(p_{j'+1,x}-p_{j',x})\mathbb{P}(P_i=p_{j,x}, A_i=a \mid X_i=x)\,\mathrm{d}F^X(x)+ \\ &\quad + \int \mathbb{E}[A_i \mid X_i=x]\mathbb{E}[Y_i \mid X_i=x, P_i=p_{1,x}]\,\mathrm{d}F^X(x) \\ &= \int \sum_j \sum_a a \sum_{j'=1}^{j-1} \alpha(p_{j',x};x)(p_{j'+1,x}-p_{j',x})\mathbb{P}(P_i=p_{j,x}, A_i=a \mid X_i=x)\,\mathrm{d}F^X(x) \\ &= \int \sum_{j'=1}^{J_x-1} \alpha(p_{j',x};x)(p_{j'+1,x}-p_{j',x}) \sum_{j>j'} \sum_a a\mathbb{P}(P_i=p_{j,x}, A_i=a \mid X_i=x)\,\mathrm{d}F^X(x),\end{aligned}$$

where the second line follows since $\mathbb{E}[A_i \mid X_i]=0$ by assumption, and the last line follows from changing the order of summation. Therefore, by definition of conditional expectation:

$$\begin{aligned}\mathbb{E}[Y_iA_i] &= \int \sum_{j=1}^{J_x-1} \alpha(p_{j,x};x)(p_{j,x}-p_{j-1,x}) \sum_{j>j'} \mathbb{E}[A_i \mid P_i=p_{j,x}, X_{i=x}]\mathbb{P}(P_i=p_{j,x} \mid X_i=x)\,\mathrm{d}F^X(x) \\ &= \int \sum_{j=1}^{J_x-1} \alpha(p_{j,x};x)(p_{j,x}-p_{j-1,x})\mathbb{E}[A_i \mid X_i=x, P_i>p_{j,x}]\mathbb{P}(P_i>p_{j,x} \mid X_i=x)\,\mathrm{d}F^X(x).\end{aligned} \tag{28}$$

The expression for $\mathbb{E}[T_iA_i]$ can be derived using the same arguments, except that we substitute

$$\mathbb{E}[T_i \mid X_i=x, P_i=p_{j,x}] = p_{1,x} + \sum_{j'=1}^{j-1}(p_{j'+1,x}-p_{j',x}). \qquad \square$$

I use the following results from Chao, Swanson, Hausman, Newey and Woutersen (2012) and Politis, Romano and Wolf (1999) to prove Lemma 6 and Lemma 7 below:

Lemma 4 (Chao *et al.*, 2012, Lemma A.1). *Suppose that, conditional on some set of random variables $\mathcal{F}$, $\{(A_i, B_i)\}_{i=1}^n$ is independent a.s., where A_i and B_i are some scalars random variables. Let $\mathbf{H}$ be a symmetric idempotent matrix with rank K. Let $\mathbb{E}[A_i \mid \mathcal{F}] = \bar{a}_i$, $\mathbb{E}[B_i \mid \mathcal{F}] = \bar{b}_i$, and $\sigma_A^2 = \max_{i\leq n}\operatorname{var}(A_i \mid \mathcal{F})$, $\sigma_B^2 = \max_{i\leq n}\operatorname{var}(B_i \mid \mathcal{F})$. Then there exists a positive constant C such that*

$$\mathbb{E}\left[\left(\sum_i\sum_{j\neq i}(A_iH_{ij}B_j - \bar{a}_iH_{ij}\bar{b}_j)\right)^2 \;\middle|\; \mathcal{F}\right] \leq C(K\sigma_A^2\sigma_B^2 + \sigma_A^2\bar{b}'\bar{b} + \sigma_B^2\bar{a}'\bar{a}).$$

Lemma 5 (Lemma 1.3.2., Politis *et al.*, 1999). *Suppose that $(A_{n,1},\ldots,A_{n,n})$ is a triangular array of i.i.d. random variables, the nth row having distribution F_n^A. Assume F_n^A converges in distribution to F^A, and $\mathbb{E}[|A_{n,1}|] \to \mathbb{E}[|A|] < \infty$, as $n\to\infty$, where A is distributed according to F^A. Then $n^{-1}\sum_{i=1}^n A_{n,i} \to E[A]$ as $n\to\infty$.*

Lemma 6. *Suppose that Assumptions R and MI hold. Then*

(i) *$\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n/n = \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n + o_p(1)$; and*

(ii) $\mathbf{Y}_n'\mathbf{G}_n\mathbf{T}_n/n = \mathbf{R}_n'\mathbf{G}_n\mathbf{P}_n/n + o_p(1)$,

where $\mathbf{G}_n$ *is defined in Equation* (26).

Proof. I will prove Part (i), Part (ii) follows by similar arguments. Let $A_{n,i}^W = (1-(\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}T_{n,i}$, and let $A_{n,i}^{(Z,W)} = (1-(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii})^{-1}T_{n,i}$, and denote by $\bar{a}_{n,i}^W = (1-(\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}P_{n,i}$ and $\bar{a}_{n,i}^{(Z,W)} = (1-(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii})^{-1}P_{n,i}$ their expectations conditional on $\mathcal{Z}_n$. Note that since $0 \leq P_{n,i} \leq 1$, it follows that $\operatorname{var}(P_{n,i} \mid \mathcal{Z}_n) \leq 1$. Then we can write:

$$\begin{aligned}\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n = &\sum_i\sum_{j\neq i}\left(A_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii}T_j - \bar{a}_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii}P_j\right)\\ &+\sum_i\sum_{j\neq i}\left(A_{n,i}^{(\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{W}_n)})_{ii}T_j - \bar{a}_{n,i}^{(\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{W}_n)})_{ii}P_j\right).\end{aligned}$$

Therefore, obtain that

$$\begin{aligned}\mathbb{E}[(\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n)^2/n^2 \mid \mathcal{Z}_n] \leq\ &\mathbb{E}\left[\left(\sum_i\sum_{j\neq i}\left(A_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii}T_j - \bar{a}_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii}P_j\right)/n\right)^2 \,\middle|\, \mathcal{Z}_n\right]\\ &+\mathbb{E}\left[\left(\sum_i\sum_{j\neq i}\left(A_{n,i}^{(\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{W}_n)})_{ii}T_j - \bar{a}_{n,i}^{(\mathbf{W}_n)}(\mathbf{H}_{(\mathbf{W}_n)})_{ii}P_j\right)/n\right)^2 \,\middle|\, \mathcal{Z}_n\right]\\ \leq\ &\frac{C}{n^2}\left(\frac{K+L}{(1-C_2)^2}+\frac{2}{(1-C_2)^2}\mathbf{P}_n'\mathbf{P}_n\right)+\frac{C}{n^2}\left(\frac{L}{(1-C_2)^2}+\frac{2}{(1-C_2)^2}\mathbf{P}_n'\mathbf{P}_n\right),\end{aligned}$$

where the first line follows from triangle inequality, and the second line follows from applying Lemma 4 with $\mathcal{F} = \mathcal{Z}_n$, and the implication of Assumption R that

$$\sup_n\sup_{i\leq n}\operatorname{var}\left(A_{n,i}^{(Z,W)} \mid \mathcal{Z}_n\right) \leq \frac{1}{(1-C_2)^2}, \qquad \sup_n\sup_{i\leq n}\operatorname{var}\left(A_{n,i}^{(W)} \mid \mathcal{Z}_n\right) \leq \frac{1}{(1-C_2)^2},$$

$$\sum_{i=1}^n A_{n,i}^W A_{n,i}^W \leq \frac{1}{1-C_2}\sum_{i=1}^n P_{n,i}^2, \qquad \sum_{i=1}^n A_{n,i}^{(Z,W)} A_{n,i}^{(Z,W)} \leq \frac{1}{1-C_2}\sum_{i=1}^n P_{n,i}^2.$$

Next, by Assumption R, we can apply the law of large numbers given in Lemma 5 to $n^{-1}\sum_{i=1}^n P_{n,i}^2$ to get $n^{-1}\sum_{i=1}^n P_{n,i}^2 = \mathbb{E}[P^2] + o_p(1) = O_p(1)$. Also, $(K+L)/n^2 = o(1)$ by Assumption MI, so that

$$\mathbb{E}[(\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n)^2/n^2 \mid \mathcal{Z}_n] \leq o_p(1).$$

Therefore, by Markov inequality and the dominated convergence theorem,

$$\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n/n = \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n + o_p(1),$$

which proves assertion (i). □

Lemma 7. *Suppose Assumption* R *and Assumption* MI *hold. Let* (Y, T, P, R) *be distributed according to the limiting distribution* $F^{Y,T,R,P}$. *Then*

(i) $\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbb{E}[T(P-\mathbb{E}[P \mid X])] + o_p(1)$*; and*

(ii) $\mathbf{R}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbb{E}[Y(P-\mathbb{E}[P \mid X])] + o_p(1)$, *where $\mathbf{G}_n$ is defined in Equation (26).*

Proof. Again, I will only prove Part (i), Part (ii) follows by similar arguments. To this end, write $\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n$ as

$$\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n - \mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n.$$

I will prove the assertion in two steps. First, I will prove that

$$\mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n = \mathbb{E}[TP] + o_p(1). \tag{29}$$

Second, I will prove that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = \mathbb{E}[T\mathbb{E}[P \mid X]] + o_p(1). \tag{30}$$

Combining (29) with (30) then yields the result.

To prove (29), note that

$$\begin{aligned}\|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/\sqrt{n}\|^2 &= \|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)/\sqrt{n}\|^2 \\ &= \mathrm{tr}((\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) - \mathrm{tr}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) \\ &\leq \mathrm{tr}((\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) = \sum_{i=1}^{n}(P_{n,i}^L - P_{n,i})^2/n \to 0 \qquad \text{a.s.,}\end{aligned} \tag{31}$$

where the first equality follows from $\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n^L = 0$, the second equality follows from the definition of Euclidean norm, and the last line follows from the fact that $\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}$ is positive semi-definite so that $\mathrm{tr}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) \geq 0$ and Assumption MI. Therefore, we obtain

$$\begin{aligned}|\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/n| &\leq \left(n^{-1}\sum_i \frac{P_{n,i}^2}{(1 - (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii})^2}\right)^{1/2} \|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/\sqrt{n}\| \\ &\leq \frac{1}{1 - C_2}\left(n^{-1}\sum_i P_{n,i}^2\right)^{1/2} o_p(1) \\ &= o_p(1),\end{aligned} \tag{32}$$

where the first line follows by the Cauchy-Schwarz inequality, the second line follows from the result (31) and Assumption R, and the last line follows from applying the law of large numbers given in Lemma 5 to $n^{-1}\sum_{i=1}^n P_{n,i}^2$. Therefore, we obtain

$$\mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n = \mathbf{P}_n'\mathbf{P}_n/n + o_p(1).$$

Since by Assumption R and Lemma 5, $n^{-1}\sum_{i=1}^n P_{n,i}^2 \to \mathbb{E}[P^2] = \mathbb{E}[TP]$, assertion (29) follows.

Now I prove assertion (30). Let $A_{n,i} = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}P_{n,i}$, and denote by $\bar{a}_{n,i}^W = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}P_{n,i}^X$ its expectation conditional on $\{X_{n,i}\}_{i=1}^n$, where $P_{n,i}^X = \mathbb{E}[P_{n,i} \mid X_{n,i}]$. Note that since $0 \leq P_{n,i} \leq 1$, it follows that

$\text{var}(P_{n,i} \mid X_{n,i}) \leq 1$. Therefore, applying Lemma 4 with $\mathcal{F} = \{X_{n,i}\}_{i=1}^{n}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_i\sum_{j\neq i}\left(A_{n,i}(\mathbf{H}_{\mathbf{W}_n})_{ij}P_j - a_{n,i}(\mathbf{H}_{\mathbf{W}_n})_{ij}\mathbb{E}[P_j \mid W_j]\right)\right)^2 \,\middle|\, \mathcal{F}\right] &\leq \frac{C}{n^2}\left(\frac{L}{(1-C_2)^2} + \frac{2}{(1-C_2)^2}\sum_i \mathbb{E}[P_{n,i} \mid X_{n,i}]^2\right) \\ &\leq \frac{C}{n^2}\left(\frac{L}{(1-C_2)^2} + \frac{2n}{(1-C_2)^2}\right) \\ &= o_p(1),\end{aligned}$$

where the first line follows from the implication of Assumption R, $(1-(\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1} \leq 1/(1-C_2)$, the second line follows from $|P_{n,i}| \leq 1$, and the last line follows from $L \leq n$. It therefore follows by Markov inequality and the dominated convergence theorem that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = (\mathbf{P}_n^X)'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n^X/n + o_p(1)$$

where $\mathbf{P}_n^X$ is an n-vector with ith element given by $P_{n,i}^X$. Let $P_{n,i}^{X,L} = \mathbb{E}^*[P_{n,i} \mid W_{n,i}]$. Now, by arguments as in Equations (31) and (32) with $\mathbf{P}_n$ replaced by $\mathbf{P}_n^X$ and $\mathbf{P}_n^L$ replaced by $\mathbf{P}_n^{X,L}$, we have that:

$$|(\mathbf{P}_n^X)'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}\mathbf{M}_{\mathbf{W}_n}\mathbf{P}_n^X/n| = o_p(1).$$

Since $(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n}) = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}\mathbf{M}_{\mathbf{W}_n}$, it follows that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = n^{-1}\sum_i \mathbb{E}[P_{n,i} \mid X_{n,i}]\mathbb{E}[P_{n,i} \mid X_{n,i}] + o_p(1).$$

By Assumption R, we can apply Lemma 5 to obtain

$$\begin{aligned}\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n &= \mathbb{E}\left[\mathbb{E}[P \mid X]^2\right] + o_p(1) \\ &= \mathbb{E}\left[T\mathbb{E}[P \mid X]\right] + o_p(1),\end{aligned}$$

which prove assertion (30). □

Appendix C Proofs

Proof of Lemma 1. First consider part (i). Since $\mathbb{E}[\tilde{P}_i^L T_i] \neq 0$, it follows by the continuous mapping theorem that $\hat{\beta}_{\hat{\mathbf{P}}} \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L Y_i]/\mathbb{E}[\tilde{P}_i^L T_i]$. Part (i) then follows by Lemma 2. Part (ii) follows by similar arguments.

Finally, to prove Part (iii), it suffices to show that

$$\min \text{eig}(\hat{S}^{-1}\hat{\Xi}) \xrightarrow{p} \min \text{eig}(S^{-1}\Xi), \tag{33}$$

since then $\hat{\beta}_{\hat{S},\hat{\Xi}} \xrightarrow{p} \beta_S$ by the continuous mapping theorem. To show (33), note that $\min \text{eig}(\hat{S}^{-1}\hat{\Xi})$ is the minimum of the function

$$\hat{\mathcal{D}}_{\mathcal{S}}(\omega) = \frac{\omega'\hat{\Xi}\omega}{\omega'\hat{S}\omega}, \qquad \omega \in \mathcal{S}^1,$$

where $\mathcal{S}^1$ denotes the unit circle in $\mathbb{R}^2$, a compact space. Therefore, if $\hat{\mathcal{D}}_{\mathcal{S}}(\omega)$ converges uniformly to the limiting function $\mathcal{D}_{\mathcal{S}}(\omega) = \omega'\Xi\omega/(\omega' S\omega)$, then $\min_\omega \hat{\mathcal{D}}_{\mathcal{S}}(\omega) \xrightarrow{p} \min_\omega \mathcal{D}_{\mathcal{S}}(\omega)$ by standard arguments (see, for example Newey and McFadden, 1994). To prove uniform convergence, I will use the arguments in Chao and Swanson (2005). Fix some $\omega \in \mathcal{S}^1$, and note that:

$$\begin{aligned}
|\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| &= \left| \frac{\omega'\hat{\Xi}\omega}{\omega'\hat{S}\omega} - \mathcal{D}_{\mathcal{S}}(\omega)\frac{\omega'\hat{S}\omega}{\omega'\hat{S}\omega} \right| = \frac{1}{|\omega'\hat{S}\omega|} \left|\omega'\hat{\Xi}\omega - \mathcal{D}_{\mathcal{S}}(\omega)\omega'\hat{S}\omega\right| \\
&= \frac{1}{|\omega'\hat{S}\omega|} \left|\omega'(\hat{\Xi} - \Xi)\omega - \mathcal{D}_{\mathcal{S}}(\omega)\omega'(\hat{S} - S)\omega\right| \\
&\leq \frac{1}{|\omega'\hat{S}\omega|} \left(\left|\omega'(\hat{\Xi} - \Xi)\omega\right| + \mathcal{D}_{\mathcal{S}}(\omega) \left|\omega'(\hat{S} - S)\omega\right| \right),
\end{aligned}$$

where the first line follows from the definition of $\hat{\mathcal{D}}_{\hat{\mathcal{S}}}$, the second line follows from the definition of $\mathcal{D}_{\hat{\mathcal{S}}}$, and the third line follows by triangle inequality. I now bound all three terms in the last the expression uniformly in ω. Since the trace operator is an inner product under Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(AA')}$, by Cauchy-Schwarz inequality:

$$\begin{aligned}
|\omega'(\hat{\Xi} - \Xi)\omega| &= \left|\mathrm{tr}\left(\omega\omega'(\hat{\Xi} - \Xi)\right)\right| \leq \sqrt{\mathrm{tr}(\omega\omega'\omega\omega')}\|\hat{\Xi} - \Xi\|_F \\
&= \|\hat{\Xi} - \Xi\|_F = o_p(1),
\end{aligned}$$

where the second line follows from $\omega'\omega = 1$ since $\omega \in \mathcal{S}^1$ and $\hat{\Xi} \xrightarrow{p} \Xi$ so that $\|\hat{\Xi} - \Xi\|_F = o_p(1)$. By an identical argument, we also have $|\omega'(\hat{S} - S)\omega| = o_p(1)$. Finally, to bound $1/|\omega'\hat{S}\omega|$, note that since $\hat{S} \xrightarrow{p} S > 0$, $\omega'\hat{S}\omega > 0$ with probability approaching 1, so that $1/|\omega'\hat{S}\omega| < C$ for some $C < \infty$ with probability approaching 1. Hence:

$$|\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| \leq o_p(1) + o_p(1)\mathcal{D}_{\mathcal{S}}(\omega),$$

since $\mathcal{D}_{\mathcal{S}}(\omega)$ is bounded by $\max \mathrm{eig}(S^{-1}\Xi)$, it follows that $\sup_\omega |\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| = o_p(1)$ as required. □

Proof of Theorem 1. By Lemma 1, $\Xi_{12} = \mathbb{E}[Y_i\tilde{P}_i^L]$ and $\Xi_{22} = \mathbb{E}[T_i\tilde{P}_i^L]$. Since by Assumption L, $\mathbb{E}[\tilde{P}_i^L \mid X_i] = 0$, we can apply Lemma 3 with $A_i = \tilde{P}_i^L$ to get

$$\frac{\Xi_{12}}{\Xi_{22}} = \frac{\int \sum_{j=1}^{J_x - 1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x})\mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\,\mathrm{d}F^X(x)}{\int \sum_{j=1}^{J_x - 1}(p_{j,x} - p_{j-1,x})\mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\,\mathrm{d}F^X(x)},$$

which yields the result for Ξ_{12}/Ξ_{22}.

Second, by Lemma 1, $\Xi_{12} = \mathbb{E}[T_i\tilde{R}_i^L]$ and $\Xi_{11} = \mathbb{E}[Y_i\tilde{R}_i^L]$. Since by Assumption L, $\mathbb{E}[\tilde{R}_i^L \mid X_i] = 0$, applying Lemma 3 with $A_i = \tilde{R}_i^L$ yields the result for Ξ_{11}/Ξ_{12}. □

Proof of Corollary 1. Let $\mathbb{P}(P_i = p_{j,x} \mid X_i = x) = s_{j,x}$. If the linear approximations (3)–(4) are exact, then $P_i = P_i^L$ and $R_i = R_i^L$. We can therefore write:

$$\mathbb{E}[R_i^L \mid P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

so that

$$\mathbb{E}[R_i^L \mid X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m=j'+1}^{J_x} s_{m,x},$$

and

$$
\begin{aligned}
\mathbb{E}[\tilde{R}_i^L \mid P_i > p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) + \\
+ \frac{1}{\sum_{m=j+1}^{J_x} s_{m,x}} \sum_{j'=j}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m=j'+1}^{J_x} s_{m,x}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}[\tilde{R}_i^L \mid P_i > p_{j,x}, X_i = x]\mathbb{P}(P_i > p_{j,x} \mid X_i = x) = \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=j+1}^{J_x} s_{m',x} \sum_{m=1}^{j'} s_{m,x} \\
+ \sum_{j'=j}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=1}^{j} s_{m',x} \sum_{m=j'+1}^{J_x} s_{m,x}.
\end{aligned}
$$

By Theorem 1, we therefore have:

$$
\begin{aligned}
\zeta_j(x) = (p_{j+1,x} - p_{j,x}) \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=j+1}^{J_x} s_{m',x} \sum_{m=1}^{j'} s_{m,x} \\
+ (p_{j+1,x} - p_{j,x}) \sum_{j'=j}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=1}^{j} s_{m',x} \sum_{m=j'+1}^{J_x} s_{m,x}.
\end{aligned}
$$

If $\alpha(p_{j,x}; x) \geq 0$ for all j and x, then all the terms in this expression are non-negative, so that $\zeta_j(x)$ is non-negative.

To obtain the expressions for $\zeta_1(x)$ and $\theta_1(x)$ in the special case that $J_x = 2$, note that since $\mathbb{E}[P_i \mid X_i = x] = s_{1,x}p_{1,x} + s_{2,x}p_{2,x}$, we have

$$
\begin{aligned}
\theta_1(x) &= (p_{2,x} - p_{1,x})[p_{2,x} - \mathbb{E}[P_i \mid X_i = x]]s_{2,x} \\
&= (p_{2,x} - p_{1,x})[p_{2,x}(1 - s_{2,x}) - p_{1,x}s_{1,x}]s_{2,x} = (p_{2,x} - p_{1,x})^2 s_{1,x}s_{2,x}.
\end{aligned}
$$

On the other hand,

$$
\operatorname{var}(P_i \mid X_i = x) = (p_{2,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{2,x} + (p_{1,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{1,x} = (p_{2,x} - p_{1,x})^2 s_{1,x}s_{2,x}
$$

Secondly, since $R_i^L = R_i$,

$$
\begin{aligned}
\mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] &= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] \\
\mathbb{E}[Y_i \mid X_i = x, P_i = p_{2,x}] &= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}),
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{1,x}] &= \mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i = p_{2,x}] \\
&= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{0,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}) - \mathbb{E}[Y_i \mid X_i = x] \\
&= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x})(1 - s_{2,x}) \\
&= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x})s_{1,x}.
\end{aligned}
$$

Therefore, it follows that:

$$\zeta_1(x) = (p_{2,x} - p_{1,x})^2 s_{1,x} s_{2,x} \alpha(p_{1,x}; x),$$

which completes the proof. □

Proof of Theorem 2. Using the matrix notation from Equation (26),

$$\hat{\beta}_{\text{UJIVE}} = \frac{\mathbf{Y}_n' \mathbf{G}_n' \mathbf{T}_n / n}{\mathbf{T}_n \mathbf{G}_n' \mathbf{T}_n / n}$$

By Lemma 6 and Lemma 7,

$$\mathbf{Y}_n' \mathbf{G}_n' \mathbf{T}_n / n = \mathbb{E}[Y(P - \mathbb{E}[P \mid X])] + o_p(1), \qquad \mathbf{T}_n \mathbf{G}_n' \mathbf{T}_n / n = \mathbb{E}[T(P - \mathbb{E}[P \mid X])] + o_p(1).$$

Next, since the limiting distribution of the data satisfies Assumption IV (iii), $\mathbb{E}[T(P - \mathbb{E}[P \mid X])] > 0$, so that by the continuous mapping theorem,

$$\frac{\mathbf{T}_n' \mathbf{G}_n \mathbf{Y}_n}{\mathbf{T}_n' \mathbf{G}_n \mathbf{T}_n} = \frac{\mathbb{E}[Y(P - \mathbb{E}[P \mid X])]}{\mathbb{E}[T(P - \mathbb{E}[P \mid X])]} + o_p(1).$$

The assertion of the Theorem then follows by applying Lemma 3 with $A_i = P - \mathbb{E}[P \mid X]$. □

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, **113** (2), 231–263. 5, 11

Ackerberg, D. A. and Devereux, P. J. (2009). Improved Jive estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics*, **91** (2), 351–362. 4, 25

Aizer, A. and Doyle, Jr., J. J. (2011). Effects of Juvenile Incarceration: Evidence from Randomly-Assigned Judges. 23

Alonso-Borrego, C. and Arellano, M. (1999). Symmetrically normalized instrumental-variable estimation using panel data. *Journal of Business & Economic Statistics*, **17** (1), 36–49. 14

Anatolyev, S. (2011). Instrumental variables estimation and inference in the presence of many exogenous regressors. 4, 26

Anderson, T. W., Kunitomo, N. and Matsushita, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics*, **157** (2), 191–204. 4

— and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63. 2

Angrist, J. D. and Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. *Journal of the American Statistical Association*, **90** (430), 431–442. 20, 33, 34

—, — and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67. 4, 12, 25

—, — and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91** (434), 444–455. 15

— and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014. 7, 29, 30

— and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. 9, 10

Balke, A. and Pearl, J. (1997). Bounds on Treatment Effects From Studies With Imperfect Compliance. *Journal of the American Statistical Association*, **92** (439), 1171–1176. 5

Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, **62** (3), 657–681. 4, 24, 28, 29

— and Crudu, F. (2012). Symmetric Jackknife Instrumental Variable Estimation. 4, 14

— and van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, **59** (3), 239–267. 25

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. B. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, (forthcoming). 29

Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90** (430), 443 – 450. 3

Chamberlain, G. (2007). Decision theory applied to an instrumental variables model. *Econometrica*, **75** (3), 609–652. 3

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, **73** (5), 1673–1692. 40

—, —, Hausman, J. A., Newey, W. K. and Woutersen, T. (2012). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. *Econometric Theory*, **12** (1), 42–86. 36

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? *Quarterly Journal of Economics*, **126** (4), 1593–1660. 26

Chioda, L. and Jansson, M. (2009). Optimal Invariant Inference When the Number of Instruments Is Large. *Econometric Theory*, **25** (3), 793–805. 4

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011). Nonparametric Instrumental Regression. *Econometrica*, **79** (5), 1541–1565. 34

Dobbie, W. and Fryer, R. G. (2011). Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, **3** (3), 158–187. 7

— and Song, J. (2012). Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection. 23

Donald, S. G. and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica*, **69** (5), 1161–1191. 12, 25, 31

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, **139** (1), 35–75. 5

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, **45** (4), 939–953. 4

Goldberger, A. S. and Olkin, I. (1971). A minimum-distance interpretation of limited-information estimation. *Econometrica*, **39** (3), 635–639. 3, 13

Hansen, C. B., Hausman, J. A. and Newey, W. K. (2008). Estimation With Many Instrumental Variables. *Journal of Business and Economic Statistics*, **26** (4), 398–422. 29

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, **3** (2), 211–255. 4, 14, 19

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **4** (5), 475–492. 15

— (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, **32** (3), 441–452. 2, 16

—, Urzúa, S. and Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, **88** (3), 389–432. 9, 16, 19

— and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America*, **96** (8), 4730–4734. 16

— and — (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, **73** (3), 669–738. 19

Hillier, G. H. (1990). On the normalization of structural equations: Properties of direction estimators. *Econometrica*, **58** (5), 1181–1194. 14, 19

Hirano, K., Imbens, G. W., Rubin, D. B. and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1** (1), 69–88. 5

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** (2), 467–475. 1, 2, 5, 8, 9, 15

Keller, W. J. (1975). A new class of limited-information estimators for simultaneous equations systems. *Journal of Econometrics*, **3** (1), 71–92. 14, 22

Kitagawa, T. (2009). Identification region of the potential outcome distributions under instrument independence. 5

Kolesár, M. (2013). Integrated Likelihood Approach to Inference With Many Instruments. 14

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. and Imbens, G. W. (2011). Identification and Inference with Many Invalid Instruments. 4, 26

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association*, **75** (371), 693–700. 4, 24

Machado, C., Shaikh, A. M. and Vytlacil, E. J. (2013). Instrumental Variables and the Sign of the Average Treatment Effect. 5

Malinvaud, E. (1966). *Statistical Methods of Econometrics*. Amsterdam: North-Holland. 3, 13

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica*, **51** (3), 821–841. 4, 24

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, **27** (4), 575–595. 8, 12, 25

Nagin, D. and Snodgrass, M. G. (2011). The Effect of Incarceration on Offending: Evidence from a Natural Experiment in Pennsylvania. 23

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, *Chapter 36*, Elsevier, pp. 2111–2245. 40

— and Powell, J. L. (2003). Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71** (5), 1565–1578. 34

Phillips, G. D. A. and Hale, C. (1977). The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems. *International Economic Review*, **18** (1), 219–228. 4, 12, 25

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Elsevier, vol. 1, pp. 449–516. 19

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics, New York: Springer-Verlag. 36

Theil, H. (1961). *Economic Forecasts and Policy*. Amsterdam: Horth-Holland, 2nd edn. 8

— (1971). *Principles of Econometrics*. New York: John Wiley & Sons. 8

Vytlacil, E. J. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, **70** (1), 331–341. 15

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press. 8, 9, 10

Yau, L. H. and Little, R. J. A. (2001). the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, **96** (456), 1232–1244. 5

Zellner, A. (1970). Estimation of regression relationships containing unobservable independent variables. *International Economic Review*, **11** (3), 441–454. 3, 21, 22