NBER WORKING PAPER SERIES

IDENTIFICATION AND INFERENCE WITH MANY INVALID INSTRUMENTS

Michal Kolesár
Raj Chetty
John N. Friedman
Edward L. Glaeser
Guido W. Imbens

Working Paper 17519
http://www.nber.org/papers/w17519

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2011

We thank the National Science Foundation for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Identification and Inference with Many Invalid Instruments
Michal Kolesár, Raj Chetty, John N. Friedman, Edward L. Glaeser, and Guido W. Imbens
NBER Working Paper No. 17519
October 2011
JEL No. C01,C2,C26,C36

ABSTRACT

We analyze linear models with a single endogenous regressor in the presence of many instrumental variables. We weaken a key assumption typically made in this literature by allowing all the instruments to have direct effects on the outcome. We consider restrictions on these direct effects that allow for point identification of the effect of interest. The setup leads to new insights concerning the properties of conventional estimators, novel identification strategies, and new estimators to exploit those strategies. A key assumption underlying the main identification strategy is that the product of the direct effects of the instruments on the outcome and the effects of the instruments on the endogenous regressor has expectation zero. We argue in the context of two specific examples with a group structure that this assumption has substantive content.

Michal Kolesár
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
kolesarmi@gmail.com

Raj Chetty
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
and NBER
chetty@fas.harvard.edu

John N. Friedman
Harvard Kennedy School
Taubman 356
79 JFK St.
Cambridge, MA 02138
and NBER
john_friedman@harvard.edu

Edward L. Glaeser
Department of Economics
315A Littauer Center
Harvard University
Cambridge, MA 02138
and NBER
eglaeser@harvard.edu

Guido W. Imbens
Department of Economics
Littauer Center
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
imbens@fas.harvard.edu

1 Introduction

A key condition underlying identification of the causal effect in instrumental variable models is the assumption that the instruments only affect the outcome of interest through their effect on the endogenous variable. However, in many empirical applications, there is a concern that instruments may also affect the outcome directly. To address this concern, this paper establishes conditions under which the effects of interest are identified in settings with direct effects of instruments on the outcome. Following Kunitomo (1980), Morimune (1983), Bekker (1994), Hahn (2002), Chamberlain and Imbens (2004), Chao and Swanson (2005), Hansen, Hausman and Newey (2008), Chioda and Jansson (2009), Anderson, Kunitomo and Matsushita (2010), and others, we focus on the case with many instruments where each individual instrument is weak in the Staiger and Stock (1997) sense but collectively the instruments have substantial predictive power.

In the absence of direct effects of the instruments the limited-information-maximum-likelihood (liml) estimator is consistent (Bekker, 1994) and efficient (Chioda and Jansson, 2009; Anderson *et al.*, 2010) under the Bekker many-instrument asymptotic sequence given homoscedasticity. The two-stage-least-squares (tsls) estimator is inconsistent (Kunitomo, 1980; Bekker, 1994), but a bias-corrected version, known as the bias-corrected-two-stage-least-squares (btsls) (Donald and Newey, 2001), estimator remains consistent. Another consistent estimator in this setting is the jackknife-instrumental-variables-estimator (jive) (Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999). Motivated by our leading examples, and as in Anatolyev (2011), we also allow the number of exogenous covariates to increase in proportion with the sample size. This requires some minor modification of the btsls and jive estimators (denoted by mbtsls and mjive), but does not affect the consistency of liml.

We examine the robustness of these five estimators (liml, btsls, jive, mbtsls, and mjive) to the presence of direct effects in this many-instrument setting. We show that liml loses consistency if direct effects are present. The intuition is that the liml estimator attempts to impose proportionality of *all* the reduced form coefficients. This explains the efficiency of liml in the absence of direct effects, but because the reduced form coefficients are no longer proportional when direct effects are present, it makes liml sensitive to their presence. On the other hand, under the assumption that the product of the direct effects of the instruments on the outcome and the direct effects on the endogenous regressor has expectation zero, the btsls and jive estimators (in the case with a fixed number of exogenous variables) or their many-exogenous-variables modifications mbtsls and mjive (in general) remain consis-

tent. We argue through some examples and a link with the clustering literature that this identifying assumption, although not innocuous, substantively weakens existing identification conditions. The intuition for the robustness compared to liml is that the btsls, jive, mbtsls, and mjive estimators, like the tsls estimator, can be thought of as two-stage estimators. In the first stage a single instrument is constructed as a function of only instruments and endogenous regressors, not involving the outcome variable. This constructed instrument is then used in the second stage to estimate the parameter of interest using methods for just-identified settings. Identification only requires validity of the constructed instrument, not of all the individual instruments.

We then study in detail two leading cases that motivate the set up and illustrate the range and applicability of our new identifying assumptions. Both cases have a clustering structure where the instruments are related to the cluster indicators. Such settings are often the reason for the presence of many instruments. The many-instrument asymptotic approximation implies that, as is common in clustering settings, large sample approximations are based on the number of clusters growing with the sample size while the number of sampled units from each cluster remains fixed.

The first of the two special cases arises when the instruments are cluster indicators. For example, Fryer (2011) and Levitt, List, Neckermann and Sadoff (2011) conducted a series of experiments where students in randomly selected schools were given varying financial incentives to improve achievement on test scores. Suppose we are interested in the effect of test score achievement on outcomes later in life as in Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011). One could use the school indicators study as instruments to capture the fact that the incentives varied between schools. However, one might be concerned that schools also affect outcomes directly, not just through test scores. Our results suggest that a sufficient, and, because of the randomization, substantially weaker condition for identification is that the direct effects of the school on the outcomes are uncorrelated with the effects of the school on test scores.

In another example within this class, Aizer and Doyle, Jr. (2011) and Nagin and Snodgrass (2011) study the effect of incarceration on subsequent outcomes. Defendants are randomly assigned to one of a relatively large number of judges. Judges vary in their propensity to sentence individuals to jail terms. The judge assignment is used as an instrument. One might be concerned that judges have direct effects on outcomes beyond those mediated through the effect on incarceration. Our critical identification assumption is that these direct effects of the judges are uncorrelated with the judges' propensity to incarcerate. This is

a substantive assumption that may or may not hold in practice, but shifts the discussion of the validity of inference away from the substantially stronger assumption that judges have no direct effect on outcomes whatsoever.

In the second case we have a small number of basic instruments. These basic instruments are interacted with cluster indicators to generate a large number of instruments. Here the number of exogenous regressors (which includes the cluster indicators) increases proportional to the number of instruments. This case is motivated by the Angrist and Krueger (1991, AK from hereon) study where the basic instruments, four quarter of birth indicators, are interacted with year and state of birth indicators to generate additional instruments. In the context of this set up our approach suggests new identification strategies that allow for direct effects of the instruments on the outcome. In the first of these identification strategies, the average direct effect of the instruments on the outcome is zero. In the second identification strategy, the average direct effect (equal to the direct effect of the basic instrument) is unrestricted, but the direct effects are uncorrelated with the effect of the instruments on the endogenous regressor. Again these are not innocuous assumptions, but they substantively weaken the assumption that all instruments are valid.

The results in this paper contribute to two strands of literature. First, we contribute to the recent many-instrument literature that has extended the earlier work by Kunitomo (1980), Morimune (1983), Bekker (1994), and Chao and Swanson (2005). In recent work Anatolyev (2011) also relaxes the assumption of fixed number of exogenous regressors. Hausman, Newey, Woutersen, Chao and Swanson (2009); Chao, Swanson, Hausman, Newey and Woutersen (2010) and Ackerberg and Devereux (2009) relax the assumption of homoscedasticity. Hansen *et al.* (2008), Belloni, Chen, Chernozhukov and Hansen (2011) and Gautier and Tsybakov (2011) allow the first stage to be estimated non-parametrically. This paper takes a complementary approach: we relax the assumption of no direct effects, but keep the rest of the model simple to maintain tractability.

Second, we contribute to the literature studying properties of instrumental variables methods allowing for direct effects in settings with a fixed number of instruments. The focus of this literature has been on correcting size distortions of tests, biases of estimators, sensitivity analyses, and bounds in the presence of direct effects. Fisher (1961, 1966, 1967), Caner (2007); Berkowitz, Caner and Fang (2008) and Guggenberger (2010) analyze the implications of local (small) violations of exogeneity assumption. Hahn and Hausman (2005) compare biases for different estimators in the presence of direct effects. Conley, Hansen and Rossi (2007); Ashley (2009) and Kraay (2008) propose sensitivity analyses in the presence of

possibly invalid instruments. Nevo and Rosen (2010) consider assumptions about the sign of the direct effects of the instruments on the outcome to derive bounds on the parameters of interest. Reinhold and Woutersen (2011) and Flores and Flores-Lagunes (2010) also derive bounds allowing for direct effects of the instruments on the outcome. The current paper is the first to derive (point) identification results in the presence of non-local departures from the no-direct-effects assumption or exclusion restriction.

The rest of the paper is organized as follows. In Section 2 we introduce the set up and the notation. In Section 3 we introduce the estimators. In Section 4 we present the main formal results allowing for direct effects of the instruments. In Section 5 we discuss in detail two leading cases with a clustering structure. We apply the methods developed in this paper to the data analyzed by AK in Section 6. Section 7 concludes.

2 Set Up

We consider the following instrumental variables model:

$$\begin{aligned} Y_i &= X_i\beta + W_i'\delta + Z_i'\gamma + \epsilon_i. \\ X_i &= Z_i'\pi_{12} + W_i'\pi_{22} + \nu_i. \end{aligned} \tag{2.1}$$

The first equation relates a scalar outcome Y_i, $i = 1, \ldots, N$, to a potentially endogenous scalar regressor X_i. W_i is a vector of exogenous regressors with dimension L_N, and Z_i is a vector of instruments with dimension K_N. The second equation relates the endogenous regressor X_i to the exogenous regressors W_i the instruments Z_i. The object of interest is the coefficient β on the endogenous regressor in the outcome equation.

The model (2.1) modifies the conventional many-instruments model (e.g., Bekker (1994)) in two ways. First, we allow γ to be non-zero, thus allowing for direct effects of the instrument on the outcome. If we restrict $\gamma = 0$, then the exclusion restriction holds, and the instruments are valid. If we leave γ unrestricted, then β, the coefficient of interest, is not identified. In this paper, we will be concerned with determining assumptions on γ that are weaker than $\gamma = 0$, but that still allow us to identify β. Second, like Anatolyev (2011), we allow the number of exogenous regressors, L_N, to change with the sample size. The main motivation for this extension is that often the presence of a large number of instruments is the result of interacting a few basic instruments with many exogenous covariates. We discuss such an example in detail in Section 5.2.

Because the number of instruments and the number of exogenous variables changes with

the sample size, the distribution of some of the random variable also changes with the sample size. To be precise, we should therefore index the random variables and parameters by the sample size N. For ease of notation we drop this index.

We assume that the pairs of structural errors (ϵ_i, ν_i) are mutually independent, and conditionally homoscedastic:

$$\mathbb{E}\left[\begin{pmatrix}\epsilon_i\\ \nu_i\end{pmatrix}\begin{pmatrix}\epsilon_i\\ \nu_i\end{pmatrix}' \,\middle|\, Z_1,\ldots,Z_N,W_1,\ldots,W_N\right] = \Sigma$$

Recent papers by Chao *et al.* (2010) and Hausman *et al.* (2009) investigate the implications of heteroscedasticity in the setting with many valid instruments, and show that liml loses some of its attractive properties in that case. Our results complement theirs in the sense that our results highlight a different potential concern with liml. To simplify the derivation of distributional results, we will assume in addition that the structural errors Normally distributed. We do not require Normality for consistency arguments.

In the remainder of this section we introduce some additional notation. Let $\mathbf{Y}$ be the N-component vector with ith element Y_i, $\mathbf{X}$ the N-component vector with ith element X_i, ϵ the N-component vector with ith element ϵ_i, ν the N-component vector with ith element ν_i, $\mathbf{W}$ the $N \times L_N$ matrix with ith row equal to W_i', and $\mathbf{Z}$ the $N \times K_N$ matrix with ith row equal to Z_i'. Let $\overline{\mathbf{X}} = (\mathbf{X}, \mathbf{W})$ be the full matrix of endogenous and exogenous regressors, let $\overline{\mathbf{Y}} = (\mathbf{Y}, \mathbf{X})$ be the full matrix of endogenous variables, and let $\overline{\mathbf{Z}} = (\mathbf{Z}, \mathbf{W})$ be the full matrix of exogenous variables. Define for an arbitrary $N \times J$ matrix $\mathbf{S}$ the following four $N \times N$ matrices, the projection matrix $\mathbf{P_S}$, the matrix $\mathbf{M_S}$ that projects on the orthogonal complement of $\mathbf{S}$, the diagonal matrix $\mathbf{D_S}$ with diagonal elements equal to those of the projection matrix, and the product of $\mathbf{M_S}$ and $(1 - \mathbf{D_S})^{-1}$:

$$\mathbf{P_S} = (\mathbf{S}\,(\mathbf{S}'\mathbf{S})^{-1}\,\mathbf{S}', \quad \mathbf{M_S} = \mathbf{I} - (\mathbf{S}\,(\mathbf{S}'\mathbf{S})^{-1}\,\mathbf{S}', \quad \mathbf{D_S} = \mathrm{Diag}(\mathbf{P_S}), \quad \text{and } \mathbf{H_S} = \mathbf{M_S}(1 - \mathbf{D_S})^{-1}.$$

We use the subscript $\perp$ as shorthand for taking residuals after regression on the exogenous regressors $\mathbf{W}$, so $\mathbf{Z}_\perp = \mathbf{M_W}\mathbf{Z}$, $\mathbf{X}_\perp = \mathbf{M_W}\mathbf{X}$, $\mathbf{Y}_\perp = \mathbf{M_W}\mathbf{Y}$, and $\overline{\mathbf{Y}}_\perp = \mathbf{M_W}\overline{\mathbf{Y}}$. We also denote by ι_N and N-dimensional vector of ones.

Define the augmented concentration parameter, the two by two matrix Λ_N:

$$\Lambda_N = \begin{pmatrix}\Lambda_{N,11} & \Lambda_{N,12}\\ \Lambda_{N,12} & \Lambda_{N,22}\end{pmatrix} = \begin{pmatrix}\gamma & \pi_{12}\end{pmatrix}' \mathbf{Z}_\perp'\mathbf{Z}_\perp \begin{pmatrix}\gamma & \pi_{12}\end{pmatrix}. \tag{2.2}$$

The $(2,2)$ element of Λ_N, denoted by $\Lambda_{N,22}$ is a key measure of the strength of the instru-

ments. The (1,1) element, $\Lambda_{N,11}$, measures the degree of misspecification. In the case with valid instruments, $\gamma = 0$, $\Lambda_{N,11} = \Lambda_{N,12} = 0$ and the only non-zero element of Λ_N is $\Lambda_{N,22}$. The $(2,2)$ element $\Lambda_{N,22}$ is closely related to the conventional concentration parameter (Mariano, 1973; Rothenberg, 1984), defined as $\Lambda_{N,22}/\Sigma_{22}$. Here, following Andrews, Moreira and Stock (2006), we use the version without dividing by the structural variance Σ_{22} because that will simplify the discussion later.

3 Estimators

In this section we introduce the five estimators for β whose properties we shall study. All five have asymptotically equivalent in the setting with a fixed number of valid instruments and a fixed number of exogenous regressors. Four of these estimators have been introduced previously, and the fifth is a minor modification of a previously proposed estimator. The first three estimators fit into the k-class (Nagar, 1959; Theil, 1961, 1971; Davidson and MacKinnon, 1993). Given a scalar k, a k-class estimator for (β, δ) is given by:

$$\begin{pmatrix} \hat{\beta}_k \\ \hat{\delta}_k \end{pmatrix} = \left(\overline{\mathbf{X}}'(\mathbf{I} - k\mathbf{M}_{\overline{\mathbf{Z}}})\overline{\mathbf{X}} \right)^{-1} \left(\overline{\mathbf{X}}'(\mathbf{I} - k\mathbf{M}_{\overline{\mathbf{Z}}})\mathbf{Y} \right).$$

We are primarily interested in the estimator for β, which can be written using the $\perp$ notation as

$$\hat{\beta}_k = \left(\mathbf{X}_\perp'(\mathbf{I} - k\mathbf{M}_{\mathbf{Z}_\perp})\mathbf{X}_\perp \right)^{-1} \left(\mathbf{X}_\perp'(\mathbf{I} - k\mathbf{M}_{\mathbf{Z}_\perp})\mathbf{Y}_\perp \right). \tag{3.1}$$

A prominent member of the k-class is the two-stage-least-squares (tsls Basmann, 1957; Theil, 1961) estimator, with $\hat{k}_{\text{tsls}} = 1$. This estimator has been shown to be inconsistent under many-instrument asymptotics, see Kunitomo (1980) and Bekker (1994). We therefore do not further investigate its properties under the various generalizations of the many-instrument setup here. Instead we consider a bias-corrected version of the tsls estimator. Nagar (1959) suggested the correction $\hat{k}_{\text{nagar}} = 1 + (K_N - 2)/N$, but the first of the five estimators we focus on is a slightly different version suggested by Donald and Newey (2001), with

$$\hat{k}_{\text{btsls}} = \frac{1}{1 - (K_N - 2)/N}.$$

Although in samples with a moderate number of instruments the difference between the Nagar and Donald-Newey estimators is small, this difference does not go away under many-

instruments asymptotics with $K_N/N \to \alpha_K > 0$, and only the Donald-Newey version is consistent. As we will show in the next section, once we allow L_N to increase with sample size, btsls also loses consistency. To address this issue, the second estimator we consider is a further modification of the Donald-Newey bias-corrected estimator that achieves consistency even when $L_N/N \to \alpha_L > 0$:

$$\hat{k}_{\text{mbtsls}} = \frac{1 - L_N/N}{1 - K_N/N - L_N/N}.$$

This estimator is also considered in Anatolyev (2011).

The third estimator we consider is the limited-information-maximum-likelihood estimator (liml, Anderson and Rubin, 1949), with

$$\hat{k}_{\text{liml}} = \min_{\beta} \frac{(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{M}_{\mathbf{W}} (\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{M}_{\overline{\mathbf{Z}}} (\mathbf{Y} - \mathbf{X}\beta)}.$$

This estimator has been shown to be asymptotically efficient under many-instrument asymptotics (Chioda and Jansson, 2009; Anderson *et al.*, 2010).

The fourth estimator we study in the current paper is the jackknife-instrumental-variables estimator (jive Phillips and Hale, 1977; Angrist *et al.*, 1999):

$$\hat{\beta}_{\text{jive}} = \left(\mathbf{X}_\perp' \left(\mathbf{M}_{\mathbf{W}} - \mathbf{H}_{\overline{\mathbf{Z}}}\right) \mathbf{X}_\perp\right)^{-1} \left(\mathbf{X}_\perp' \left(\mathbf{M}_{\mathbf{W}} - \mathbf{H}_{\overline{\mathbf{Z}}}\right) \mathbf{Y}_\perp\right). \tag{3.2}$$

Ackerberg and Devereux (2009) present simulation evidence that this estimator is biased when the number of exogenous regressors is large, and suggest a bias-corrected version. We study a new version of the jackknife estimator, closely related to the Ackerberg-Devereux estimator, which we refer to as the modified jive estimator, or mjive:

$$\hat{\beta}_{\text{mjive}} = \left(\mathbf{X}_\perp' \left(\mathbf{M}_{\mathbf{W}} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}}\right) \mathbf{X}_\perp\right)^{-1} \left(\mathbf{X}_\perp' \left(\mathbf{M}_{\mathbf{W}} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}}\right) \mathbf{Y}_\perp\right). \tag{3.3}$$

We will show that unlike the original jive estimator, this estimator remains consistent even if $L_N/N \to \alpha_L > 0$.

The focus of the current paper is on the properties of these five estimators, that is, $\hat{\beta}_{\text{btsls}}$, $\hat{\beta}_{\text{mbtsls}}$, $\hat{\beta}_{\text{liml}}$, $\hat{\beta}_{\text{jive}}$, and $\hat{\beta}_{\text{mjive}}$, under various assumptions about the rates at which the number of instruments and exogenous regressors increase with the sample size, K_N, L_N, and the assumptions about the parameters governing the misspecification, γ.

4 Many Invalid Instruments

In this section we look at the properties of the five estimators allowing for many exogenous covariates ($L_N/N \to \alpha_L > 0$), and allowing for direct effects of the instruments ($\gamma \neq 0$). If we fix $\alpha_L = 0$ and $\gamma = 0$, we are in the many instrument case studied in the literature (e.g. Bekker, 1994; Morimune, 1983; Hahn, 2002; Chao and Swanson, 2005). If we also restrict $\alpha_K = 0$, we are back in the case with conventional instrumental variables asymptotics discussed in most textbooks (e.g. Wooldridge, 2002; Angrist and Pischke, 2009).

We make the following assumptions.

Assumption 1.(Instruments and exogenous variables)

(i) *$Z_i \in \mathbb{R}^{K_N}$, $W_i \in \mathbb{R}^{L_N}$, $\epsilon_i \in \mathbb{R}, \nu_i \in \mathbb{R}$, for $i = 1, \ldots, N$, $N = 1, \ldots$ are triangular arrays of random variables with $(Z_i, W_i, \epsilon_i, \nu_i)$, $i = 1, \ldots, N$ exchangeable.*

(ii) *$\overline{\mathbf{Z}}$ is full column rank with probability one.*

(iii) *$(\mathbf{P}_{\overline{\mathbf{Z}}})_{ii} < c$ for some $c < 1$ for all $i = 1, \ldots, N$ with probability one.*

(iv) *$\max_{i \leq N} |(\mathbf{Z}_\perp)_i' \pi_{12}| / \sqrt{N} \to 0$ and;*

(v) *$\sup_N \sup_{i \geq 1} \sum_j |(\mathbf{P}_{\mathbf{Z}_\perp})_{ij}| < C$ and $\sup_N \sup_{i \geq 1} \sum_j |(\mathbf{P}_{\mathbf{W}})_{ij}| < C$ for some $C < \infty$ with probability one*

The first two parts of this assumption are standard, with a minor adaption to allow for many exogenous variables. The remaining three parts are technical assumptions we use to deal with the jive and mjive estimators.

Assumption 2.(Model)

(i) *$(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ are iid with mean zero, positive definite covariance matrix Σ, and finite fourth moments;*

(ii) *The distribution of $(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ is Normal.*

For consistency we only use the first part of this assumption. For the distributional results we use Normality to highlight the specific modifications to the asymptotic distributions coming from the direct effects of the instruments.

Assumption 3.(Number of instruments and exogenous regressors)
For some $0 \leq \alpha_K < 1$ and $0 \leq \alpha_L < 1$,

$$K_N/N = \alpha_K + o(N^{-1/2}), \qquad \text{and} \qquad L_N/N = \alpha_L + o(N^{-1/2}).$$

The first part of this assumption is standard in the many-instrument literature. The second part is identical to the corresponding assumption in Anatolyev (2011).

Assumption 4.(Concentration parameter)
For some positive semi-definite Λ *with* $\Lambda_{22} > 0$,

$$\Lambda_N/N \xrightarrow{p} \Lambda, \qquad \text{and} \quad \mathbb{E}\left[\Lambda_N/N\right] \to \Lambda.$$

The first part of assumption 4 is a natural extension of the assumption underlying the Bekker many-instrument asymptotics. The second part of the assumption strengthens this slightly by also requiring the expectation of the concentration parameter to converge to its probability limit.

The first main result establishes the probability limit of the estimators.

Theorem 1.(Consistency with Many Invalid Instruments)
Suppose Assumptions 1(i)–(iii), 2(i), 3 and 4 hold. Then:

(i) *(k-class) if* $\hat{k} \xrightarrow{p} k$ *with* $k < \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\Lambda_{22}}{\Sigma_{22}(1-\alpha_K-\alpha_L)}$, *then:*

$$\hat{\beta}_{\hat{k}} \xrightarrow{p} \beta_k = \beta + \frac{\Lambda_{12} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{12}}{\Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22}},$$

(ii) *(liml) Suppose* $\min \operatorname{eig}(\Sigma^{-1}\Lambda) < \Lambda_{22}/\Sigma_{22}$. *Then:*

$$\beta_{\text{liml}} = \beta + \frac{\Lambda_{12} - \min \operatorname{eig}(\Sigma^{-1}\Lambda)\Sigma_{12}}{\Lambda_{22} - \min \operatorname{eig}(\Sigma^{-1}\Lambda)\Sigma_{22}}, \quad k_{\text{liml}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\min \operatorname{eig}(\Sigma^{-1}\Lambda)}{1-\alpha_K-\alpha_L},$$

(iii) *(btsls)*

$$\beta_{\text{btsls}} = \beta + \frac{\Lambda_{12} + \{\alpha_K\alpha_L/(1-\alpha_K)\}\Sigma_{12}}{\Lambda_{22} + \{\alpha_K\alpha_L/(1-\alpha_K)\}\Sigma_{22}}, \quad k_{\text{btsls}} = \frac{1}{1-\alpha_K},$$

(iv) *(mbtsls)*

$$\beta_{\text{mbtsls}} = \beta + \frac{\Lambda_{12}}{\Lambda_{22}}, \quad k_{\text{mbtsls}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L},$$

(v) *(jive) Suppose* $\alpha_L < \Lambda_{22}/\Sigma_{22}$. *Then:*

$$\beta_{\text{jive}} = \beta + \frac{\Lambda_{12} - \alpha_L\Sigma_{21}}{\Lambda_{22} - \alpha_L\Sigma_{22}},$$

(vi) *(mjive)*

$$\beta_{\text{mjive}} = \beta + \frac{\Lambda_{12}}{\Lambda_{22}},$$

If we impose $\Lambda_{11} = 0$ (implying $\Lambda_{12} = 0$) and $\alpha_L = 0$, the condition for consistency of $\hat{\beta}_{\hat{k}}$ is the same as in Chao and Swanson (2005), namely that $\hat{k} \to 1/(1 - \alpha_K)$. Having many exogenous regressors changes the condition on $\hat{k}$ to $\hat{k} \to (1 - \alpha_L)/(1 - \alpha_K - \alpha_L)$.

A key finding is the robustness of the mbtsls and mjive estimators relative to the liml estimator. Specifically, if Λ_{12} is equal to zero, then mbtsls and mjive are consistent even if Λ_{11} differs from zero. If the number of exogenous variables is fixed, then btsls and jive are also consistent if $\Lambda_{12} = 0$. In order for liml to be consistent for all values of Σ, then it has to be the case that Λ_{11} is equal to zero (and that immediately implies that $\Lambda_{12} = 0$). To provide some intuition, consider the reduced-form based on the model (2.1):

$$\begin{aligned} Y_i &= Z_i'(\pi_{12}\beta + \gamma) + W_i'(\delta + \pi_{22}\beta) + (\nu_i\beta + \epsilon_i), \\ X_i &= Z_i'\pi_{12} + W_i'\pi_{22} + \nu_i. \end{aligned}$$

If the instruments are valid, so that $\gamma = 0$, then the vector of reduced-form coefficients on Z_i in the first equation is proportional to π_{12}, the vector of reduced-form coefficients in the second equation. The liml estimator tries to impose this proportionality. This leads to efficiency if proportionality holds (Chioda and Jansson, 2009; Anderson *et al.*, 2010). However, if $\gamma \neq 0$, then the proportionality does not hold in the population, and liml loses consistency. On the other hand, mbtsls and mjive, like tsls, can be thought of as two stage estimators. In the first stage composite instruments are constructed, one for each regressor (endogenous or exogenous) based on the data on the endogenous regressor, the exogenous variables, and the instruments alone. These instruments are then used to estimate the parameters of interest using a method for just-identified settings, possibly with some adjustment. In this procedure proportionality of the reduced forms is never exploited. This explains why $\Lambda_{12} = 0$ is a sufficient condition for consistency, although it results in efficiency loss relative to liml when proportionality does hold.

Next we consider large sample approximations to the distribution of the estimators. We make use of Assumption 2(ii), which puts Normality on the error terms. If instead of Normality we only assumed finite fourth moments (Assumption 2 (i)), then the asymptotic variance terms would depend on the third and fourth moments of the error terms (Hansen *et al.*, 2008; van Hasselt, 2010). Assuming Normality leads to simpler asymptotic formulae

that will allow us to better focus on the effect of relaxing the standard assumptions that $\gamma = 0$ and $\alpha_L = 0$ and highlight the substantive differences. To put the main results for the case with direct effects in perspective we first present the distributional results for the case with $\gamma = 0$, but α_L possibly positive. See Anatolyev (2011) for asymptotic variances for liml and mbtsls without normality.

Theorem 2.(Asymptotic Normality with Many Exogenous Regressors)
Suppose Assumptions 1–4 hold. Suppose in addition that $\gamma = 0$. Then:

(i) *(liml)*

$$\sqrt{N}\left(\hat{\beta}_{\text{liml}} - \beta\right) \mid \overline{\mathbf{Z}} \xrightarrow{d} \mathcal{N}\left(0, \Lambda_{22}^{-2} \cdot \left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right)\right)\right)$$

(ii) *(mbtsls)*

$$\sqrt{N}\left(\hat{\beta}_{\text{mbtsls}} - \beta\right) \mid \overline{\mathbf{Z}} \xrightarrow{d} \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right)$$

(iii) *(mjive) Suppose in addition that $N^{-1}\sum_i \frac{1}{1-(\mathbf{D}_{\overline{\mathbf{Z}}})_{ii}} \to \tau$*

$$\sqrt{N}\left(\hat{\beta}_{\text{mjive}} - \beta\right) \mid \overline{\mathbf{Z}} \xrightarrow{d} \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + (1-\alpha_L)\left((1-\alpha_L)\tau - 1\right)\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right). \quad (4.1)$$

The presence of many exogenous variables increases the asymptotic variance of liml and mbtsls since $(1-\alpha_K)\alpha_K/(1-\alpha_L-\alpha_K) > \alpha_K/(1-\alpha_K)$ if $\alpha_L > 0$, but the conclusion that liml is more efficient than mbtsls does not change. Also, by Jensen's inequality $\tau \geq \frac{1}{1-\alpha_K-\alpha_L}$, so that mbtsls has smaller asymptotic variance than mjive.

If we want to determine the asymptotic distribution when γ is allowed to differ from zero, it no longer suffices to simply condition on $\overline{\mathbf{Z}}$ and treat the sequence of parameters γ as constant. The reason is because the stochastic behaviour of the estimators now depends on $\Lambda_{N,12}$. Even if the limit $\Lambda_{12} = 0$, if γ differs from zero (and thus $\Lambda_{11} > 0$) it will generally be the case that $\Lambda_{N,12}$ differs from zero for finite N. The stochastic behavior of $\Lambda_{N,12}$ affects the large sample distribution of the estimators, and we need to put sufficient structure on it to be able to determine this distribution.

The assumption below puts a random effects structure on the direct effects of the instrument on the outcome and the endogenous regressor similar to that in Chamberlain and Imbens (2004). This provides a natural way of determining the stochastic behaviour of $\Lambda_{N,12}$, although it is not necessarily the only way of doing so.

First we redefine the parameters by orthogonalizing them with respect to $\mathbf{Z}_\perp$ as

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_{12} \end{pmatrix} = \left(\alpha_K \mathbf{Z}_\perp'\mathbf{Z}_\perp\right)^{1/2}\begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}.$$

Then we consider the following assumption

Assumption 5.(Incidental parameters)

The pairs $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$*, for* $k = 1, 2, \ldots, K_N$*, are iid with distribution*

$$\left(\begin{array}{c} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{array} \right) \Bigg| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right), \Xi \right).$$

A key implication of Assumption 5 is that

$$\begin{aligned} \Lambda &= \text{plim}\left(\frac{\Lambda_N}{N}\right) = \text{plim}\left(\frac{1}{N}\begin{pmatrix}\gamma' \\ \pi_{12}'\end{pmatrix}\mathbf{Z}_\perp'\mathbf{Z}_\perp\begin{pmatrix}\gamma & \pi_{12}\end{pmatrix}\right) \\ &= \text{plim}\left(\frac{1}{K_N}\sum_{k=1}^{K_N}\begin{pmatrix}\tilde{\gamma}_k \\ \tilde{\pi}_{12,k}\end{pmatrix}\begin{pmatrix}\tilde{\gamma}_k & \tilde{\pi}_{12,k}\end{pmatrix}\right) = \begin{pmatrix}\mu_\gamma \\ \mu_\pi\end{pmatrix}\begin{pmatrix}\mu_\gamma \\ \mu_\pi\end{pmatrix}' + \Xi. \end{aligned}$$

Hence, if we rule out the knife-edge case $\Xi_{12} = -\mu_\gamma\mu_\pi$, then under Assumption 5, the identification condition $\Lambda_{12} = 0$ is equivalent to $\mu_\gamma = 0$ and $\Xi_{12} = 0$. This equivalence will be useful in determining more primitive conditions that imply the condition $\Lambda_{12} = 0$. We defer further discussion of this assumption, and in particular the motivation for making the independent random effects assumption in terms of $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$ (instead of in terms of $(\gamma_k, \pi_{12,k})$) to the next section where we consider two special cases. Next we present the large sample distribution theory for the case with $\gamma \neq 0$.

Theorem 3.(Asymptotic Normality with Many Invalid Instruments)

Suppose that Assumptions 1(i)–(iii), 2–5 hold. Suppose in addition that $\mu_\gamma = \Xi_{12} = 0$*. Then:*

(i) *(mbtsls)*

$$\begin{aligned} &\sqrt{N}\left(\hat{\beta}_{\text{mbtsls}} - \beta\right) \xrightarrow{d} \\ &\quad \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right) + \Lambda_{11}\left(\Sigma_{22} + \frac{\Lambda_{22}}{\alpha_K}\right)\right)\right), \end{aligned}$$

(ii) *(mjive) Suppose that in addition* $N^{-1}\sum_i \frac{1}{1-(\mathbf{D}_{\overline{\mathbf{Z}}})_{ii}} \to \tau$*. Then:*

$$\begin{aligned} &\sqrt{N}\left(\hat{\beta}_{\text{mjive}} - \beta\right) \xrightarrow{d} \\ &\mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + (1-\alpha_L)((1-\alpha_L)\tau - 1)\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right) + \Lambda_{11}\left(\Sigma_{22} + \frac{\Lambda_{22}}{\alpha_K}\right)\right)\right), \end{aligned}$$

Compared to Theorem 2 (ii)–(iii), allowing for direct effects leads to an additional term in the asymptotic variance which is proportional to Λ_{11}, which measures the extent of mis-

specification. If $\Lambda_{11} = 0$, then the asymptotic variance of mbtsls and mjive reduces to that in Theorem 2(ii)–(iii). Note that the extra term decreases in the number of instruments. The intuition is that as the number of instruments increases, we are better able to deal with the presence of direct effects, as the product of the direct effects and the effects of the instruments on the endogenous variable gets averaged out to identify β.

5 Two Special Cases

In this section we consider two special cases with additional structure on the data generating process. In both cases each unit i belongs to a subpopulation or cluster, with cluster indicator $G_i \in \{1, 2, \ldots, G_N\}$. These clusters are closely related to the instruments. We are interested in large sample approximations where the number of units sample from each subpopulation is finite, and the number of subpopulations increases proportional to the sample size, leading to the many-instruments setting. Let the number of units in group g be N_g, with $N = \sum_{g=1}^{G_N} N_g$. For convenience, let us assume that the number of unit sampled from each subpopulation is the same for all subpopulations, $N_g = N/G_N$ for all g.

5.1 Special Case I: Clustering

To focus on the conceptual issues, let us assume there are no exogenous regressors beyond the intercept, $L_N = 1$. In the first special case the instruments are the cluster indicators, $Z_{ik} = \mathbf{1}_{G_i=k}$, for $k = 1, \ldots, G_N - 1$, so that the number of instruments is the number of clusters minus one, $K_N = G_N - 1$. The general model in (2.1) can now be written as

$$Y_i = \delta + \beta X_i + \sum_{k=1}^{G_N-1} \gamma_k \mathbf{1}_{G_i=k} + \epsilon_i, \tag{5.1}$$

$$X_i = \pi_{22} + \sum_{k=1}^{G_N-1} \pi_{12,k} \mathbf{1}_{G_i=k} + \nu_i. \tag{5.2}$$

Exploiting the special structure here, in combination with the equal cluster size, the augmented concentration parameter can be written as the sample covariance matrix of $(\gamma_k, \pi_{12,k})$:

$$\Lambda_N = \frac{N}{G_N} \sum_{k=1}^{G_N-1} \begin{pmatrix} (\gamma_k - \overline{\gamma})^2 & (\gamma_k - \overline{\gamma})(\pi_{12,k} - \overline{\pi}_{12}) \\ (\gamma_k - \overline{\gamma})(\pi_{12,k} - \overline{\pi}_{12}) & (\pi_{12,k} - \overline{\pi}_{12})^2 \end{pmatrix},$$

where

$$\overline{\gamma} = \frac{1}{G_N}\sum_{k=1}^{G_N-1}\gamma_k, \qquad \text{and} \qquad \overline{\pi}_{12} = \frac{1}{G_N}\sum_{k=1}^{G_N-1}\pi_{12,k}.$$

Now let us consider Assumption 5 and interpret it in this context. Suppose we have a large population of clusters. Let $\mu_{Y,g}$ and $\mu_{X,g}$ be the population means of $Y_i - \beta X_i$ and X_i in cluster g, and let μ_Y and μ_X be the overall population means. In terms of the original parametrization, we have: $\pi_{22} = \mu_{X,G_N}$, $\pi_{12,k} = \mu_{X,k} - \mu_{X,G_N}$, $\delta = \mu_{Y,G_N}$ and $\gamma_k = \mu_{Y,k} - \mu_{Y,G_N}$.

The natural way to impose a random effects structure on the parameters would be to assume that the cluster means $(\mu_{Y,k}, \mu_{X,k})$ are independent and

$$\begin{pmatrix} \mu_{Y,k} \\ \mu_{X,k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Phi\right). \tag{5.3}$$

This implies

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_{12} \end{pmatrix} = \sqrt{\frac{G_N-1}{G_N}} B \begin{pmatrix} \mu_{Y,1} & \mu_{X,1} \\ \vdots & \vdots \\ \mu_{Y,1} & \mu_{X,1} \end{pmatrix}, \qquad B = \left(\begin{array}{c|c} I_{G_N-1} - \frac{1-1/\sqrt{G_N}}{G_N-1}\iota_{G_N-1}\iota'_{G_N-1} & -\frac{1}{\sqrt{G_N}}\iota_{G_N-1} \end{array} \right)$$

where the $(G_N - 1) \times G_N$ matrix B satisfies $B\iota_{G_N} = 0$, and $BB' = \mathbf{I}_{G_N-1}$. Thus, a random effects specification on $(\mu_{Y,k}, \mu_{X,k})$ as in (5.3) implies a random effects specification on $(\tilde{\gamma}, \tilde{\pi}_{12})$, namely

$$\left.\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{pmatrix}\right| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Xi\right), \qquad \text{with } \Xi = \frac{G_N - 1}{G_N} \cdot \Phi.$$

On the other hand, because $(\gamma_k, \pi_{12,k})$ measure the effect relative to the last group, G_N, assuming independence of $(\gamma_k, \pi_{12,k})$ of $(\gamma_l, \pi_{12,l})$ is not attractive. The random effects assumption on $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$ is therefore more reasonable than a random effects assumption on $(\gamma_k, \pi_{12,k})$ would be. Moreover, the augmented concentration parameter can be expressed as a sample covariance matrix of $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$:

$$\Lambda_N = \frac{N}{G_N}\sum_{k=1}^{G_N}\begin{pmatrix} \left(\tilde{\gamma}_k - \overline{\tilde{\gamma}}\right)^2 & \frac{1}{G_N}\sum_{k=1}^{G_N}\left(\tilde{\gamma}_k - \overline{\tilde{\gamma}}\right)\left(\tilde{\pi}_{12,k} - \overline{\tilde{\pi}}_{12}\right) \\ \left(\tilde{\gamma}_k - \overline{\tilde{\gamma}}\right)\left(\tilde{\pi}_{12,k} - \overline{\tilde{\pi}}_{12}\right) & \left(\tilde{\pi}_{12,k} - \overline{\tilde{\pi}}_{12}\right)^2 \end{pmatrix},$$

where $\overline{\tilde{\gamma}} = \frac{1}{G_N}\sum_{g=1}^{G_N}\tilde{\gamma}_k$ and $\overline{\tilde{\pi}}_{12} = \frac{1}{G_N}\sum_{g=1}^{G_N}\tilde{\pi}_{12,k}$.

There is an alternative representation to the set up in (5.1)–(5.2) that ties it in more closely to the clustering literature. In this alternative representation the demeaned direct effects $\mu_{Y,g} - \mu_Y$ are viewed as random effects reflecting clustering. Let us write the outcome equation (5.1) as

$$Y_i = \mu_Y + \beta X_i + \eta_i, \qquad \text{where } \eta_i = \Big(\mu_{Y,G_i} - \mu_Y\Big) + \epsilon_i,$$

is the composite residual. The cluster-specific component is equal to the direct effect of the instrument. Hence, we can think of the residuals η_i having a clustering structure associated with the instruments

$$\mathbb{E}\left[\eta_i \middle| \mathbf{Z}\right] = 0 \qquad \mathbb{E}\left[\eta_i\eta_j \middle| \mathbf{Z}\right] = \begin{cases} \Sigma_{11} + \Phi_{11} & \text{if } i = j, \\ \Phi_{11} & \text{if } G_i = G_j, i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Analogously we can write the second equation with a clustering structure:

$$X_i = \mu_X + \zeta_i \qquad \text{where } \zeta_i = \Big(\mu_{X,G_i} - \mu_X\Big) + \nu_i,$$

and

$$\mathbb{E}\left[\zeta_i \middle| \mathbf{Z}\right] = 0 \qquad \mathbb{E}\left[\zeta_i\zeta_j \middle| \mathbf{Z}\right] = \begin{cases} \Sigma_{22} + \Phi_{22} & \text{if } i = j, \\ \Phi_{22} & \text{if } G_i = G_j, i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

In addition, let $\Phi_{12} = \mathbb{E}\left[\zeta_i\eta_j | G_i = G_j\right]$. The critical assumption that Λ_{12} is equal to zero (equivalent to $\Phi_{12} = 0$) in this representation amounts to assuming that the cluster component in the outcome equation is uncorrelated with the cluster component in the first stage. This assumption is not innocuous, but assumptions about zero correlations for cluster components are often made in clustering settings. It is obviously substantively weaker than assuming the absence of clustering effects in the outcome equation, or $\Phi_{11} = 0$.

In this case with the instruments equal to the group dummies the original jive estimator has an interesting form. The predicted value for X_i underlying the tsls estimator is the average value of X_j for all units in the cluster,

$$\widehat{X}_i^{\text{tsls}} = \frac{1}{N_{G_i}} \sum_{j:G_j=G_i} X_j$$

The jive estimator modifies that to the average over all units in the cluster, excluding unit i itself:

$$\widehat{X}_i^{\text{jive}} = \frac{1}{N_{G_i} - 1} \sum_{j:G_j=G_i, j\neq i} X_j.$$

With a finite number of units per cluster omitting unit i can make a substantial difference, and this is reflected in the inconsistency of tsls in this setting.

The properties of the previously discussed estimators liml, btsls, mbtsls, jive, and mjive follow as a special case of Theorems 1-3, specializing it to the case with $L_N = 1$ so that $\alpha_L = 0$. In this case there is no difference asymptotically between jive and mjive and between btsls and mbtsls because the number of exogenous variables is fixed.

5.2 Special Case II: Clusters with Interactions

In the second case we maintain the cluster structure with cluster indicator $G_i \in \{1, 2, \ldots, G_N\}$. For each unit there is a binary indicator Q_i that serves as the basic instrument. More generally we could have a number of basic instruments, and allow these to be discrete or continuous. This special case is motivated by the Angrist-Krueger analysis where the basic instruments are quarter of birth indicators. We generate additional instruments by interacting the cluster indicator with this binary instrument. We include the cluster indicators as exogenous covariates, $W_{i,k} = \mathbf{1}_{G_i=k}$, so that again $K_N = L_N = G_N$. Again for ease of exposition let us assume that the clusters are all equal size, $N_g = N/G_N$ for all g, and that the fraction of $Q_i = 1$ units in each cluster is equal to $q = \sum_i Q_i \cdot \mathbf{1}_{G_i=g}/N_g$ for all g. The model can now be written as

$$Y_i = \beta X_i + \sum_{k=1}^{K_N} \delta_k W_{ik} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i, \tag{5.4}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Z_{ik} + \sum_{k=1}^{K_N} \pi_{22,k} W_{ik} + \nu_i. \tag{5.5}$$

In this case the limit of the augmented concentration parameter is

$$\Lambda = \begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix} \begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}' + \Xi. \tag{5.6}$$

We can directly apply the results from Section 4, which imply that mjive and mbtsls are consistent and asymptotically normally distributed if Λ_{12} is equal to zero. In this case $\Lambda_{12} = 0$ is not necessarily an attractive assumption. It would require that $\mu_\gamma \cdot \mu_\pi + \Xi_{12} = 0$, which essentially requires that both μ_γ and Ξ_{12} are zero. We can in fact relax the sufficient conditions for identification in this special setting. We consider two specific alternatives. First, we assume that $\mu_\gamma = 0$, allowing Ξ_{12} to be different from zero. Second, we consider the assumption that $\Xi_{12} = 0$, allowing μ_γ to be different from zero. In both cases $\Lambda_{12} \neq 0$, yet the parameter of interest is identified.

Under the first assumption, $\mu_\gamma = 0$, we can simply use the Wald estimator with Q_i as the single instrument and $W_i = 1$ as a single exogenous regressor:

$$\hat{\beta}_{\text{wald}} = \frac{\frac{1}{Nq}\sum_{i:\, Q_i=1} Y_i - \frac{1}{N(1-q)}\sum_{i:\, Q_i=0} Y_i}{\frac{1}{Nq}\sum_{i:\, Q_i=1} X_i - \frac{1}{N(1-q)}\sum_{i:\, Q_i=0} X_i}$$

Adding the interactions of the type $Q_i \cdot \mathbf{1}_{G_i=k}$ as additional instruments would lead to inconsistency if we use the liml, btsls, mbtsls, jive or mjive estimators.

Theorem 4.(Zero Mean)
Suppose the model in (5.4)-(5.5) holds. Suppose also that Assumptions 1–3 and 5 hold. Suppose that in addition $\mu_\gamma = 0$ and $\mu_\pi \neq 0$. Then $\hat{\beta}_{\text{wald}}$ is consistent for β and satisfies

$$\sqrt{N}\left(\hat{\beta}_{\text{wald}} - \beta\right) \xrightarrow{d} \mathcal{N}(0, (\Xi_{11}/\alpha_K + \Sigma_{11})/\mu_\pi^2)$$

In the second case with $\mu_\gamma \neq 0$ and $\Xi_{12} = 0$, again using all interactions as instruments does not lead to consistency whether we use liml, btsls, mbtsls, jive or mjive. However, in this case we can base a consistent estimator on a strategy where we treat Q_i as an exogenous regressor instead of an instrument, and only use the remaining $K_N - 1$ interactions of the type $Q_i \cdot \mathbf{1}_{G_i=k}$ as instruments with the mbtsls or mjive estimators:

$$Y_i = X_i\beta + Q_i\delta_0 + \sum_{i=1}^{K_N} W_{ik}\delta_k + \epsilon_i \tag{5.7a}$$

$$X_i = Q_i\pi_{22,0} + \sum_{k=1}^{K_N} W_{ik}\pi_{22,k} + \sum_{k=1}^{K_N-1} \pi_{12,k} Q_i W_{ik} + \nu_i \tag{5.7b}$$

This allows for a direct (common) effect of the original basic instrument, but rules out interaction effects.

Theorem 5.(INTERACTIONS)
Suppose that the model (5.4)–(5.5) *holds. Suppose also that Assumptions 1–5 hold and that* $\Xi_{12} = 0$. *Then the mbtsls and mjive estimators based on the model* (5.7) *are consistent for* β. *Moreover, under those assumptions:*

$$\sqrt{N}(\hat{\beta}_{\text{mbtsls}} - \beta) \xrightarrow{d}$$
$$\mathcal{N}\left(0, \Xi_{22}^{-2}\left(\Xi_{11}\Xi_{22}/\alpha_K + \Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11} + \frac{(1-\alpha_K)\alpha_K}{(1-2\alpha_K)}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right)$$

and

$$\sqrt{N}(\hat{\beta}_{\text{mjive}} - \beta) \xrightarrow{d}$$
$$\mathcal{N}\left(0, \Xi_{22}^{-2}\left(\Xi_{11}\Xi_{22}/\alpha_K + \Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11} + (1-\alpha_K)((1-\alpha_K)\tau - 1)\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right)$$

where $\tau = \frac{q^2}{q-\alpha_K} + \frac{(1-q)^2}{1-q-\alpha_K}$ *is the probability limit of* $\text{tr}((\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}/N)$.

6 An Application

We apply some of the methods to a subset of the Angrist and Krueger (1991) data. We use individuals born in the first and fourth quarter (so we have a single binary basic instrument, although this is not essential), dropping observations from Alaska because there are some years birth quarters with no observations, leaving us with observations on 162,487 individuals.

Let W_{ik}, for $k = 1, \ldots, G_N$ be the cluster indicators, corresponding to year of birth times state of birth interactions, so that $G_N = 500$, and let Q_i be the binary quarter of birth indicator. The general model we consider is

$$Y_i = \beta X_i + \sum_{k=1}^{K_N} \delta_k W_{ik} + \sum_{k=1}^{K_N} \gamma_k Q_i W_{ik} + \epsilon_i, \tag{6.1}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Q_i W_{ik} + \sum_{k=1}^{K_N} \pi_{22,k} W_{ik} + \nu_i. \tag{6.2}$$

$$\left(\begin{array}{c} \gamma_k \\ \pi_{12,k} \end{array} \right) \Bigg| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right), \Xi \right).$$

We look at six estimators, the five studied in this paper and the two-stage-least-squares (tsls) estimator. We consider three sets of instruments and exogenous variables.

In the first setting, we use a single binary instrument, an indicator for being born in the fourth quarter, $Z_i = Q_i$. There are no exogenous covariates beyond the intercept. The properties of this estimator are captured by Theorem 4. In particular, in this just-identified case the iv estimator is valid here if the average direct effect of the instruments is zero, $\mu_\gamma = 0$.

In the second case we interact the qob dummy with state of year times year of birth dummies, for a total of 500 instruments, and 500 exogenous regressors. Here Theorems 1 and 3 contain the relevant results. In this case liml is not consistent unless Ξ_{11}, Ξ_{12} and μ_γ are zero. The mjive and mbtsls estimators are consistent under the weaker condition that the linear combination $\Lambda_{12} = \Xi_{12} + \mu_\gamma \cdot \mu_\pi$ is equal to zero. Within the context of the model this setting requires the strongest conditions.

In the third case we only use the interactions as instruments and treat the basic quarter of birth dummy as an exogenous variable rather than as an excluded instrument. We also include the year of birth times quarter of birth dummies as exogenous covariates. For this case Theorem 5 has the appropriate results. Here liml is not consistent unless both Ξ_{11} and Ξ_{12} are equal to zero. The mbtsls and mjive estimators are consistent under the weaker condition that $\Xi_{12} = 0$.

Table 1 presents the estimates and standard errors under various assumptions. Liml, mbtsls, and mjive yield similar point estimates, irrespective of the set of instruments. Jive yields smaller point estimates under the designs which include many exogenous regressors, which is consistent with Theorem 1. On the other hand, the bias of btsls under these designs appears small.

The standard errors are quite different though for the different estimators when we use a large number of instruments. Taking into account the large number of exogenous variables does not appear to matter very much. Neither does taking into account non-zero values for μ_γ, Ξ_{12} or Ξ_{11}. In this specific case this appears to be due to the fact that point estimates for Λ_{11} conditional on $\Lambda_{12} = 0$ are close to zero: for this data set there is little evidence for direct effects of the instruments, consistent with the validity of the instruments.

7 Conclusion

In this paper we analyze linear models with a single endogenous and many instruments. Departing from the current literature we allow for direct effects of the instruments on the outcome. Such direct effects have very different impacts on standard estimators. The liml estimator, efficient in the many-valid-instrument case, is inconsistent in the presence of such effects. The btsls and jive estimators are consistent if the direct effects are uncorrelated with the effects of the instruments on the endogenous regressor. This condition is not innocuous. In many cases direct effects of the instruments on the outcome may well be correlated with effects on the endogenous regressor. However, it does shift the discussion of identification issues in instrumental variables away from the focus on the requirement that none of the instruments have any direct effects whatsoever, which in cases with many instruments may be unrealistic, and as this paper shows, unnecessarily restrictive. The results in the paper also suggest a re-assesment of the merits of liml versus other estimators in the many-instrument setting.

Appendices

We first define some additional notation. Write the reduced-form based on Equations (2.1) as:

$$\begin{pmatrix} Y_i & X_i \end{pmatrix} = \begin{pmatrix} Z_i & W_i \end{pmatrix} \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} + V_i'$$

where $\pi_{11} = \gamma + \pi_{12}\beta$ and $\pi_{21} = \delta + \pi_{12}\beta$, and $V_i = (\epsilon_i + \nu_i\beta, \nu_i)'$, and let $\mathbf{V}$ be the N by 2 matrix with ith row equal to V_i'. Denote the upper $K_N \times 2$ submatrix of the matrix of reduced-form coefficients by $\Pi_1 = (\pi_{11}, \pi_{12})$. Let:

$$\Gamma = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix}$$

Let $\Omega = \mathbb{E}[V_i V_i']$ denote the reduced-form covariance matrix. Then:

$$\Omega = \Gamma^{-1\prime}\Sigma\Gamma^{-1} = \begin{pmatrix} \Sigma_{11} + 2\Sigma_{12}\beta + \Sigma_{22}\beta^2 & \Sigma_{12} + \Sigma_{22}\beta \\ \Sigma_{21} + \Sigma_{22}\beta & \Sigma_{22} \end{pmatrix}$$

Let $\mathcal{W}_d(f, V, V^{-1}M)$ denote a d-dimensional non-central Wishart distribution with f degrees of freedom, scale parameter V, and non-centrality parameter M. Let $\mathbf{S}^{1/2}$ denote the symmetric square root of a symmetric positive semi-definite matrix $\mathbf{S}$.

Appendix A Auxilliary Lemmata

Lemma A.1.
Consider the quadratic form $Q = (M + U)'C(M + U)$, where $M \in \mathbb{R}^{N\times S}, C \in \mathbb{R}^{N\times N}$ are non-stochastic, C is symmetric, and $U = (u_1, \ldots, u_N)'$, with $u_i \sim [0, \Omega]$ iid. Let $a \in \mathbb{R}^S$ be a non-stochastic vector. Assume u_i has finite fourth moments. Denote $d_C = \mathrm{diag}(C)$. Then:

(i) *(Lemma 1, Bekker and van der Ploeg, 2005)*

$$\begin{aligned} \mathbb{E}[Q \mid C] &= M'CM + \mathrm{tr}(C)\Omega \\ \mathrm{var}(Qa \mid C) &= a'\Omega a M'C^2M + a'M'C^2Ma\Omega + \Omega aa'M'C^2M + MC^2Maa'\Omega \\ &\quad + \mathrm{tr}(C^2)(a'\Omega a\Omega + \Omega aa'\Omega) \\ &\quad + d_C'd_C[\mathbb{E}(a'u)^2uu' - a'\Omega aa'\Omega - a'\Omega a\Omega] + 2d_C'CMa\mathbb{E}[(a'u)uu'] \\ &\quad + M'Cd_C\mathbb{E}[(a'u)^2u'] + \mathbb{E}[(a'u)^2u]d_C'CM \end{aligned}$$

If the distribution of u_i is Normal, the last two lines of the variance expression equals zero.

(ii) *Suppose that the distribution of u_i is Normal, and that, as $N \to \infty$:*

$$M'C^2M/N \to Q_{CM} \qquad \text{tr}(C^2)/N \to \tau_{C^2}$$

where the elements c_{is} of C may depend on N. Suppose also that $\max_{i\leq N}\|m_{is}\|/\sqrt{N} \to 0$ and $\sup_N \max_{i\leq N}\sum_{j=1}^N |c_{ij}| = D_C < \infty$. Then:

$$\sqrt{N}\left(Qa/N - \mathbb{E}Qa/N\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

where

$$V = a'\Omega a Q_{CM} + a'Q_{CM}a\Omega + \Omega aa'Q_{CM} + Q_{CM}aa'\Omega + \tau_{C^2}(a'\Omega a\Omega + \Omega aa'\Omega).$$

Proof. We only prove Part (ii). We follow the arguments in van Hasselt (2010), who proves asymptotic Normality of Qa/N when u_i are non-normal, but imposes slightly stronger regularity conditions. By the Cramér-Wold device, it suffices to prove that for any vector $b \in \mathbb{R}^S$:

$$N^{-1/2}\left(b'Qa - \mathbb{E}\left[b'Qa\right]\right) \xrightarrow{d} \mathcal{N}\left(0, b'Vb\right). \tag{A.1}$$

Let $m^b = Mb$ be an N-vector with the ith element equal to $\sum_{s=1}^S m_{is}b_s$, and similarly for m^a, u^b and u^a. Let also $\Omega_{p,r} = p'\Omega r$, for $p, r \in \{a, b\}$. Then the left-hand side of (A.1) can be written as:

$$N^{-1/2}\left(b'Qa - \mathbb{E}\left[b'Qa\right]\right) = \sum_i \sum_j c_{ij}(u_i^a m_i^b + m_i^a u_i^b + u_i^a u_i^b) - \sum_i c_{ii}\Omega_{a,b} = N^{-1/2}\sum_i D_i^{a,b},$$

where, using the fact that $c_{ij} = c_{ji}$:

$$D_{N,i}^{a,b} = c_{ii}(u_i^a u_i^b - \Omega_{a,b}) + u_i^b \sum_{j<i} c_{ij}u_j^a + u_i^a \sum_{j<i} c_{ij}u_j^b + u_i^b \sum_j c_{ij}m_j^a + u_i^a \sum_j c_{ij}m_j^b. \tag{A.2}$$

$\{N^{-1/2}D_{N,i}^{a,b}, 1 \leq i \leq N\}$ is a martingale-difference sequence with respect to the filtration $\mathcal{F}_{N,i} = \sigma(u_1, \ldots, u_i)$. To apply a martingale central limit theorem, we need to verify that:

$$N^{-1}\sum_{i=1}^N \mathbb{E}\left[(D_{N,i}^{a,b})^2 \mid \mathcal{F}_{N,i-1}\right] \xrightarrow{p} b'Vb \tag{A.3}$$

Expanding the expression yields:

$$
\begin{aligned}
N^{-1}\sum_i \mathbb{E}\left[(D_{N,i}^{a,b})^2 \mid \mathcal{F}_{n,i-1}\right] &= N^{-1}\sum_i c_{ii}^2(\Omega_{a,a}\Omega_{b,b}+\Omega_{a,b}^2) + \Omega_{b,b}N^{-1}\sum_i\sum_{j<i}\sum_{k<i} c_{ij}c_{ik}u_j^a u_k^a \\
&+ \Omega_{a,a}N^{-1}\sum_i\sum_{j<i}\sum_{k<i} c_{ij}c_{ik}u_j^b u_k^b + 2\Omega_{a,b}N^{-1}\sum_i\sum_{j<i}\sum_{k<i} c_{ij}c_{ik}u_j^a u_k^b \\
&+ \Omega_{b,b}a'M'C^2Ma/N + \Omega_{a,a}b'MC^2Mb/N + 2\Omega_{a,b}b'MC^2Ma/N \\
&+ 2\Omega_{b,b}N^{-1}\sum_i\sum_{j<i}\sum_{k} c_{ij}c_{ik}m_k^a u_j^a + 2\Omega_{a,b}N^{-1}\sum_i\sum_{j<i}\sum_{k} c_{ij}c_{ik}m_k^b u_j^a \\
&+ 2\Omega_{a,b}N^{-1}\sum_i\sum_{j<i}\sum_{k} c_{ij}c_{ik}m_k^a u_j^b + 2\Omega_{a,a}N^{-1}\sum_i\sum_{j<i}\sum_{k} c_{ij}c_{ik}m_k^b u_j^b
\end{aligned}
\tag{A.4}
$$

The last four terms are $o_p(1)$ since their variance converges to zero. This follows from writing them as:

$$
N^{-1}\sum_i\sum_{j<i}\sum_k c_{ij}c_{ik}m_k^p u_j^r = \sum_i\left(N^{-1}\sum_{j>i}\sum_k c_{ij}c_{jk}m_k^p\right)u_i^r \qquad p,r\in\{a,b\}
$$

and noting that

$$
\sum_i\left(N^{-1}\sum_{j>i}\sum_k c_{ij}c_{jk}m_k^p\right)^2 \le (\max_{i\le N} m_i^p/\sqrt{N})^2 N^{-1}\sum_i\left(\sum_j c_{ij}\sum_k c_{ik}\right)^2 \le (\max_{i\le N} m_i^p/\sqrt{N})^2 C_M^4 \to 0
$$

Now consider the terms of the form:

$$
\begin{aligned}
N^{-1}\sum_i\sum_{j<i}\sum_{k<i} c_{ij}c_{ik}u_j^p u_k^r &= N^{-1}\sum_i\sum_{j<i} c_{ij}^2 u_j^p u_j^r + N^{-1}\sum_i\sum_{j<i}\sum_{k<j} c_{ij}c_{ik}(u_j^p u_k^r + c_{ij}c_{ik}u_j^r u_k^p) \\
&= \frac{1}{N}\sum_j\left(\sum_{i>j} c_{ij}^2 + \frac{1}{2}c_{ii}^2\right)u_j^p u_j^r + N^{-1}\sum_i\sum_{j<i}\sum_{k<j} c_{ij}c_{ik}(u_j^p u_k^r + u_j^r u_k^p) - \frac{1}{2N}\sum_i c_{ii}^2 u_i^p u_i^r \\
&= \frac{1}{2}\tau_{C^2}p'\Omega r - \frac{1}{2N}\sum_i c_{ii}^2 u_i^p u_i^r + o_p(1)
\end{aligned}
$$

The last line follows from applying Chebyshev inequality to the first two terms, and noting that:

$$\begin{aligned}\operatorname{var}\left(\tfrac{1}{N}\textstyle\sum_j\left(\sum_{i>j}c_{ij}^2+\tfrac{1}{2}c_{ii}^2\right)u_j^pu_j^r\right) &= \operatorname{var}(u_j^pu_j^r)\cdot N^{-2}\sum_j\left(\sum_{i>j}c_{ij}^2+\frac{1}{2}c_{ii}^2\right)^2\\ &\leq \operatorname{var}(u_j^pu_j^r)N^{-2}t_{C^2}D_C^2\to 0\\ \operatorname{var}\left(\tfrac{1}{N}\textstyle\sum_i\sum_{j<i}\sum_{k<j}c_{ij}c_{ik}u_j^pu_k^r\right) &= N^{-2}p'\Omega pr'\Omega p\sum_j\sum_{k<j}\left(\sum_{i>j}c_{ij}c_{ik}\right)^2\leq O(N^{-2}D_C^4)\to 0\end{aligned}$$

Pulling together the results yields:

$$\begin{aligned}N^{-1}\sum_i\mathbb{E}\left[(D_{N,i}^{a,b})^2\mid\mathcal{F}_{n,i-1}\right] &= b'Vb+\\ &N^{-1}\sum_i c_{ii}^2(\Omega_{a,a}b'\Omega b+(\Omega_{a,b})^2-\Omega_{a,a}u_i^bu_i^b/2-b'\Omega bu_i^au_i^a/2-\Omega_{a,b}u_i^au_i^b)\end{aligned}$$

This establishes (A.3), since the second term is $o_p(1)$ as $\max_i c_{ii}^2/N\to 0$.

Secondly, it is possible to show that $N^{-2}\sum_i\mathbb{E}(D_{N,i}^{a,b})^4\to 0$, so that the Lindeberg condition holds. Hence, a martingale central limit theorem applies, which yields the result. □

Lemma A.2.
Consider a sequence of random matrices $\{X_N\}_{N=1}^\infty$ such that $X_N\sim\mathcal{W}_S(J_N,\Omega,\Omega^{-1}\Xi_N)$. Suppose that $\Xi_N/N\to\Xi$, and that $J_N/N=\alpha+o(N^{-1/2})$, $\alpha>0$. Then, for any vector $a\in\mathbb{R}^S$

$$\begin{aligned}&N^{-1/2}\left(X_Na/N-(\Xi_N/N+\alpha\Omega)a\right)\\ &\qquad\xrightarrow{d}\mathcal{N}\left(0,(a'\Omega a\Xi+a'\Xi a\Omega+\Omega aa'\Xi+\Xi aa'\Omega)+\alpha(a'\Omega a\Omega+\Omega aa'\Omega)\right)\end{aligned}$$

Proof. By definition of a non-central Wishart distribution, we can decompose $X_N=(U+M)'(U+M)$, where $U=(u_1,\ldots,u_{J_N})'$, $u_j\sim N(0,\Omega)$ iid, $M'M=\Xi_N$, and $\Xi_N/J_N\to\Xi/\alpha$. Hence, we can apply Lemma A.1 (ii) with $C=\mathbf{I}_{J_N}$ to get:

$$\begin{aligned}&J_N^{-1/2}(X_Na-(\Xi_N+J_N\Omega)a)\\ &\qquad\xrightarrow{d}\mathcal{N}\left(0,\alpha^{-1}(a'\Omega a\Xi+a'\Xi a\Omega+\Omega aa'\Xi+\Xi aa'\Omega)+a'\Omega a\Omega+\Omega aa'\Omega\right)\end{aligned}$$

which yields the result. □

Lemma A.3.
Suppose Assumptions 1, 2(i), 3 and 4 hold. Then:

$$\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N\xrightarrow{p}\Psi+(1-\alpha_L)\Omega \tag{A.5a}$$

$$\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N\xrightarrow{p}\Psi+\alpha_K\Omega \tag{A.5b}$$

$$\overline{\mathbf{Y}}_\perp'\mathbf{H}_{\overline{\mathbf{Z}}}\overline{\mathbf{Y}}_\perp/N\xrightarrow{p}\Omega \tag{A.5c}$$

where

$$\Psi = \begin{pmatrix} \Lambda_{11} + 2\Lambda_{12}\beta + \Lambda_{22}\beta^2 & \Lambda_{12} + \Lambda_{22}\beta \\ \Lambda_{12} + \Lambda_{22}\beta & \Lambda_{22} \end{pmatrix} \tag{A.6}$$

These probability limits also hold conditional on $\overline{\mathbf{Z}}$.

Proof. First we establish the probability limit of $\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N$. By Lemma A.1 (i):

$$\mathbb{E}[\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \mid \mathbf{Z}_\perp] = (K_N/N)\Omega \tag{A.7}$$

Fix $a \in \mathbb{R}^2$. Since $\mathbf{P}_{\mathbf{Z}_\perp}$ is a projection matrix, $0 \leq (\mathbf{P}_{\mathbf{Z}_\perp})_{ii} \leq 1$. Hence, $\sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii}^2 \leq \sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii} \leq K_N$. Therefore:

$$\begin{aligned} \operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N) &= \mathbb{E}\operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N \mid \mathbf{P}_{\mathbf{Z}_\perp}) \\ &= \mathbb{E}\left[\operatorname{tr}(\mathbf{P}_{\mathbf{Z}_\perp}/N^2)\right](a'\Omega a\Omega + \Omega a a'\Omega) \\ &\quad + \mathbb{E}\left[N^{-2}\textstyle\sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii}^2\right][\mathbb{E}(a'V_i)^2 V_i V_i' - a'\Omega a a'\Omega - a'\Omega a\Omega] \\ &\leq \frac{K_N}{N^2}(a'\Omega a\Omega + \Omega a a'\Omega) + \frac{K_N}{N^2}[\mathbb{E}(a'v_i)^2 v_i v_i' - a'\Omega a a'\Omega - a'\Omega a\Omega] \\ &= O(K_N/N^2) \end{aligned} \tag{A.8}$$

Combining Equations (A.7) and (A.8) with Assumption 3 yields :

$$\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \xrightarrow{p} \alpha_K \Omega \tag{A.9}$$

By similar arguments:

$$\mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \xrightarrow{p} (1-\alpha_L)\Omega \tag{A.10}$$

Next, by Assumption 2 (i), $\mathbb{E}[\Pi_1'\mathbf{Z}_\perp'\mathbf{V}/N \mid \mathbf{Z}_\perp] = 0$, so that:

$$\begin{aligned} \operatorname{var}\left(\Pi_1'\mathbf{Z}_\perp'\mathbf{V}a/N\right) &= \mathbb{E}\left[\operatorname{var}\left(\Pi_1'\mathbf{Z}_\perp'\mathbf{V}a/N \mid \mathbf{Z}_\perp\right)\right] = (a'\Omega a)\mathbb{E}\left[\Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N^2\right] \\ &= (a'\Omega a)\Gamma^{-1\prime}\mathbb{E}\left[\Lambda_N/N^2\right]\Gamma^{-1} = O(1/N) \end{aligned}$$

where the last equality follows by Assumption 4. Consequently:

$$\Pi_1\mathbf{Z}_\perp'\mathbf{V}/N \xrightarrow{p} 0 \tag{A.11}$$

Combining the representation $\mathbf{Y}_\perp = \mathbf{Z}_\perp\Pi_1 + \mathbf{V}_\perp$ with the limits in Equations (A.10) and (A.11), and Assumption 4 establishes (A.5a):

$$\begin{aligned} \overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N &= \Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N + \Pi_1'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi_1/N + \mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \\ &= \Gamma^{-1}\Lambda_N\Gamma^{-1}/N + (1-\alpha_L)\Omega + o_p(1) \\ &= \Psi + (1-\alpha_L)\Omega \end{aligned}$$

Claim (A.5b) follows by similar arguments from Equations (A.9) and (A.11):

$$\begin{aligned}\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= \Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N + \Pi_1'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi_1/N + \mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \\ &\xrightarrow{p} \Psi + \alpha_K\Omega\end{aligned}$$

Next we prove (A.5c). As an intermediate step, we need to find the probability limit of $\mathbf{V}'\mathbf{H}_{\overline{\mathbf{Z}}}\mathbf{V}$. Because $\mathbf{H}_{\overline{\mathbf{Z}}}$ is symmetric, we can apply Lemma A.1 (i), so that:

$$\mathbb{E}[\mathbf{V}'\mathbf{H}_{\overline{\mathbf{Z}}}\mathbf{V}/N] = \mathbb{E}\,\mathrm{tr}(\mathbf{H}_{\overline{\mathbf{Z}}}/N)\Omega = \Omega$$

since $\mathrm{tr}(\mathbf{H}_{\overline{\mathbf{Z}}}) = N$. Denoting $t = \mathrm{tr}(\mathbf{H}_{\overline{\mathbf{Z}}}^2)$, we have $t = \mathrm{tr}(\mathbf{M}_{\overline{\mathbf{Z}}}(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}})^{-2}) = \mathrm{tr}((\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}})^{-1}) \leq \frac{N}{1-c}$ by Assumption 1. Moreover, $\sum_i (\mathbf{H}_{\overline{\mathbf{Z}}})_{ii}^2 = \sum_i 1^2 = N$. Hence, for any $a \in \mathbb{R}^{G+1}$:

$$\begin{aligned}\mathrm{var}(\mathbf{V}'\mathbf{H}_{\overline{\mathbf{Z}}}\mathbf{V}a/N) &= \mathbb{E}\,\mathrm{var}(\mathbf{V}'\mathbf{H}_{\overline{\mathbf{Z}}}\mathbf{V}a/N \mid \overline{\mathbf{Z}}) \\ &= \mathbb{E}[t]\cdot(a'\Omega a\Omega + \Omega aa'\Omega)/N^2 + \mathbb{E}\left[\sum_i (\mathbf{H}_{\overline{\mathbf{Z}}})_{ii}^2\right]\cdot[\mathbb{E}(a'v_i)^2 v_i v_i' - a'\Omega aa'\Omega - a'\Omega a\Omega]/N^2 \\ &\leq \frac{1}{1-c}(a'\Omega a\Omega + \Omega aa'\Omega)/N + [\mathbb{E}(a'v_i)^2 v_i v_i' - a'\Omega aa'\Omega - a'\Omega a\Omega]/N \\ &= O(N^{-1})\end{aligned}$$

Therefore, by Chebyshev's inequality:

$$\overline{\mathbf{Y}}'\mathbf{H}_{\overline{\mathbf{Z}}}\overline{\mathbf{Y}}/N = \mathbf{V}'\mathbf{H}_{\overline{\mathbf{Z}}}\mathbf{V}/N \xrightarrow{p} \Omega \tag{A.12}$$

Finally, the same calculations go through even if we condition on $\overline{\mathbf{Z}}$, so that the probability limits hold also conditional on $\overline{\mathbf{Z}}$. □

Lemma A.4.
Consider a k-class estimator with $\hat{k} \xrightarrow{p} k$ subject to $k < \frac{1-\alpha_L}{1-\alpha_L-\alpha_K} + \frac{\Lambda_{22}/\Sigma_{22}}{1-\alpha_L-\alpha_K}$. Then under Assumptions 1, 2 (i), 3 and 4:

$$\hat{\beta}_{\hat{k}} \xrightarrow{p} \beta + \frac{\Lambda_{12} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{12}}{\Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22}}$$

Proof. Combining Lemma A.3 with the condition $\hat{k} = k + o_p(1)$ yields:

$$\begin{aligned}(1-\hat{k})\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N + \hat{k}\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= (1-k)(\Psi + (1-\alpha_L)\Omega) + k(\Psi + \alpha_K\Omega) + o_p(1) \\ &= \Psi + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Omega + o_p(1)\end{aligned} \tag{A.13}$$

The (2,2) element of (A.13) is given by:

$$(1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N = \Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22} + o_p(1)$$

becausee $\Sigma_{22} = \Omega_{22}$. By the condition on k, $\Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22} > 0$, so that:

$$\left((1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N\right)^{-1} = (\Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22})^{-1} + o_p(1)$$

(A.14)

The (1,2) element in Equation (A.13) is given by:

$$(1-\hat{k})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp/N = \Lambda_{12} + \Lambda_{22}\beta + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Omega_{12} + o_p(1)$$
$$= \Lambda_{12} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{12} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22}\beta + \Lambda_{22}\beta + o_p(1) \tag{A.15}$$

Applying Equations (A.14) and (A.15) to $\hat{\beta}_{\hat{k}}$:

$$\hat{\beta}_{\hat{k}} = \frac{(1-\hat{k})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{k}x_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp}{(1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N} = \beta + \frac{\Lambda_{12} + ((1-k)(1-\alpha_L)+\alpha_K k)\Sigma_{12}}{\Lambda_{22} + ((1-k)(1-\alpha_L)+\alpha_K k)\Sigma_{22}} + o_p(1). \ \square$$

Appendix B Proofs of Theorems

Proof of Theorem 1. The results for a general k-class estimator, btsls and mbtsls follows directly from Lemma A.4. We therefore just need to derive the results for liml, jive and mjive.

First, we establish the result for liml. Define

$$\hat{Q}_N(\phi) = \frac{\phi'\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N\phi}{\phi'\overline{\mathbf{Y}}_\perp'M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N\phi}.$$

Then

$$\hat{k}_{\text{liml}} = \min_{\tilde{\beta}} \frac{(1,-\tilde{\beta})\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N(1,-\tilde{\beta})'}{(1,-\tilde{\beta})\overline{\mathbf{Y}}_\perp'M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N(1,-\tilde{\beta})'} = \min_{\phi\in S^1}\hat{Q}_N(\phi)$$

where S^1 denotes the unit circle in $\mathbb{R}^2$. Applying Lemma A.3 yields:

$$\hat{Q}_N(\phi) \xrightarrow{p} \frac{\phi'(\Psi+(1-\alpha_L)\Omega)\phi}{(1-\alpha_L-\alpha_K)\phi'\Omega\phi} \equiv \frac{\phi'T\phi}{\phi'T_\perp\phi} \equiv Q(\phi)$$

where we define $T = \Psi + (1-\alpha_L)\Omega$ and $T_\perp = (1-\alpha_L-\alpha_K)\Omega$. Assumption 2 (i) guarantees that the denominator is non-zero for any value of ϕ. The minimum of $Q(\phi)$ is achieved at:

$$\min_{\phi\in S^1} Q(\phi) = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{1}{1-\alpha_L-\alpha_K}\min_{\phi\in S^1}\frac{\phi'\Psi\phi}{\phi'\Omega\phi}$$
$$= \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\min\text{eig}(\Sigma^{-1}\Lambda)}{1-\alpha_K-\alpha_L} = k_{\text{liml}}$$

where the last line follows since the eigenvalues of $\Omega^{-1}\Psi$ correspond to the eigenvalues of $\Sigma^{-1}\Lambda$. The minimand ϕ_{liml} is given by the eigenvector corresponding to the smallest eigenvalue of the matrix:

$$\frac{1}{1-\alpha_K-\alpha_L}\Omega^{-1}\left(\Psi+(1-\alpha_L)\Omega\right)$$

We now need to show that:

$$\hat{k}_{\text{liml}} - k_{\text{liml}} = \min_{\phi \in S^1} \hat{Q}_N(\phi) - Q(\phi_{\text{liml}}) \xrightarrow{p} 0 \tag{A.1}$$

To this end, we first show that the convergence of the objective function is uniform:

$$\sup_{\phi \in S^1} |\hat{Q}_N(\phi) - Q(\phi)| \xrightarrow{p} 0 \tag{A.2}$$

Fix $\phi \in S^1$. By triangle inequality:

$$\begin{aligned}
|\hat{Q}_N(\phi) - Q(\phi)| &\leq \frac{1}{|\phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left| \phi' \overline{\mathbf{Y}}'_\perp \overline{\mathbf{Y}}_\perp \phi / N - Q(\phi) \phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N \right| \\
&= \frac{1}{|\phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left| \phi' (\overline{\mathbf{Y}}'_\perp \overline{\mathbf{Y}}_\perp / N - T) \phi - Q(\phi) \phi' \left(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \right) \phi \right| \\
&\leq \frac{1}{|\phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left(\left| \phi' (\overline{\mathbf{Y}}'_\perp \overline{\mathbf{Y}}_\perp / N - T) \phi \right| + Q(\phi) \left| \phi' \left(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \right) \phi \right| \right)
\end{aligned} \tag{A.3}$$

We now need to bound all three terms in the expression uniformly in ϕ. Because the trace operator is the inner product under Frobenius norm, by Cauchy-Schwarz inequality:

$$\begin{aligned}
|\phi' (\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \phi| &= \left| \text{tr} \left((\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \phi \phi' \right) \right| \\
&\leq \sqrt{\text{tr}((\phi \phi')^2)} \| (\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \|_F \\
&= \| (\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \|_F \\
&= o_p(1)
\end{aligned}$$

where the third line follows since $\|\phi\|_2 = 1$, and the last line follows since $\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N \xrightarrow{p} T_\perp$ by Lemma A.3. By similar argument

$$|\phi' (\overline{\mathbf{Y}}'_\perp \overline{\mathbf{Y}}_\perp / N - T) \phi| = o_p(1)$$

Finally, we bound the denominator. Because $\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N \xrightarrow{p} T_\perp > 0$, $\phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N > 0$ wpa1, so that wpa1 $|\phi' \overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N| < C$ for some $C < \infty$. Applying these bounds and the fact that $Q(\phi)$ is bounded implies that the right-hand side in (A.3) is $o_p(1)$, which implies (A.2).

Next, denote the argmin of $\hat{Q}_N(\phi)$ by $\hat{\phi}$. Note that $\hat{k}_{\text{liml}}$ and hence $\hat{\phi}$ exists wpa1. We can now establish (A.1), using the uniform convergence result (A.2):

$$\begin{aligned}
Q(\phi_{\text{liml}}) \leq Q(\hat{\phi}) &= \hat{Q}_N(\hat{\phi}) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \leq \hat{Q}(\phi_{\text{liml}}) + (Q(\hat{\phi}) - \hat{Q}_n(\hat{\phi})) \\
&= Q(\phi_{\text{liml}}) + (\hat{Q}_N(\phi_{\text{liml}}) - Q(\phi_{\text{liml}})) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \\
&= Q(\phi_{\text{liml}}) + o_p(1)
\end{aligned}$$

The probability limit for liml then follows by Lemma A.4.

It remains to establish the results for jive and mjive. Applying Lemma A.3, we get:

$$\overline{\mathbf{Y}}'(\mathbf{M_W} - \mathbf{H}_{\overline{\mathbf{Z}}})\overline{\mathbf{Y}}/N \xrightarrow{p} \Psi - \alpha_L\Omega \tag{A.4}$$

$$\overline{\mathbf{Y}}'(\mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\overline{\mathbf{Y}}/N \xrightarrow{p} \Psi \tag{A.5}$$

Because $\Lambda_{22} > \alpha_L\Sigma_{22}$, it follows from the (2,2) element of (A.4) that:

$$(\mathbf{X}'(\mathbf{M_W} - \mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X})^{-1} = (\Lambda_{22} - \alpha_L\Sigma_{22}) + o_p(1)$$

$$(\mathbf{X}'(\mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X})^{-1} = \Lambda_{22} + o_p(1)$$

Combining these with an expansion of the (2,1) element in (A.4) and (A.5) yields the results for jive and mjive. □

Proof of Theorem 2. All probability statements are conditional on $\overline{\mathbf{Z}}$. We omit the conditioning for ease of notation.

Proof of part (i) The liml estimator is given by the minimand of the objective function:

$$\hat{Q}_N(\tilde{\beta}) = \frac{(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})'(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})}{(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})'\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})}$$

The associated first-order condition is proportional to $\hat{g}_N(\hat{\beta}_{\text{liml}}) = 0$, where

$$\hat{g}_N(\tilde{\beta}) = -\frac{1}{N}\mathbf{X}'_\perp(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta}) + \frac{\hat{Q}_N(\tilde{\beta})}{N}\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})$$

The derivative of the first-order condition is given by:

$$\hat{g}'_N(\tilde{\beta}) = \frac{\mathbf{X}'_\perp\mathbf{X}_\perp}{N} - \hat{Q}_N(\tilde{\beta})\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\mathbf{X}_\perp + \frac{2\hat{g}_N(\tilde{\beta})}{(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})'\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})}\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{Y}_\perp - \mathbf{X}_\perp\tilde{\beta})$$

We will show that for any estimator $\hat{\beta}$ with $\hat{\beta} \xrightarrow{p} \beta$:

$$\hat{g}'_N(\hat{\beta}) \xrightarrow{p} \Lambda_{22} \tag{A.6}$$

Secondly, we will show that at the true value:

$$\sqrt{N}\hat{g}_N(\beta) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{12}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right)\right) \tag{A.7}$$

Because the limit of $\hat{g}'_N(\hat{\beta})$ does not depend on β and it is positive, and since $\hat{\beta}_{\text{liml}} \xrightarrow{p} \beta$ is consistent by Theorem 1, assertion ((i)) the theorem will follow (see Newey and McFadden, 1994)

We first prove (A.6). Let $\phi = (1, -\beta)$. By Lemma A.3 and consistency of $\hat{\beta}$:

$$\hat{Q}_N(\hat{\beta}) \xrightarrow{p} \frac{\phi'(\Psi + (1-\alpha_L)\Omega)\phi}{(1-\alpha_L-\alpha_K)\phi'\Omega\phi} = \frac{1-\alpha_L}{1-\alpha_L-\alpha_K} \equiv k_{\text{liml}}$$

$$\hat{g}_N(\hat{\beta}) \xrightarrow{p} -(1-\alpha_L)\Sigma_{12} + \frac{1-\alpha_L}{1-\alpha_L-\alpha_K}(1-\alpha_K-\alpha_L)\Sigma_{12} = 0$$

where we use the fact that $\Lambda_{11} = \Lambda_{12} = 0$ since $\gamma = 0$. Hence:

$$\hat{g}_N'(\hat{\beta}) \xrightarrow{p} \Lambda_{22} + (1 - \alpha_L)\Sigma_{22} - k\Sigma_{22} + 0 = \Lambda_{22}$$

which proves (A.6). It remains to show that $\hat{g}_N(\beta_{\text{liml}})$ satisfies a central limit theorem. Let $\tilde{\nu} = \nu - \varrho\epsilon$, where $\varrho = \Sigma_{12}/\Sigma_{11}$ be a projection of ν onto space orthogonal to ϵ. We have:

$$\begin{aligned}
\sqrt{N}\hat{g}_N(\beta) &= N^{-1/2}\left(\nu'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon\frac{\epsilon'\mathbf{M}_{\mathbf{W}}\epsilon}{\epsilon'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon} - \mathbf{X}'\mathbf{M}_{\mathbf{W}}\epsilon\right) \\
&= N^{-1/2}\left(\tilde{\nu}'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon\frac{\epsilon'\mathbf{M}_{\mathbf{W}}\epsilon}{\epsilon'\mathbf{M}_{\mathbf{Z}}\epsilon} - (\mathbf{Z}_\perp\pi_{12} + \tilde{\nu})'\mathbf{M}_{\mathbf{W}}\epsilon\right) \\
&= N^{-1/2}\left(\tilde{\nu}'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon \cdot k_{\text{liml}} - (\mathbf{Z}_\perp\pi_{12} + \tilde{\nu})'\mathbf{M}_{\mathbf{W}}\epsilon\right) + o_p(1)
\end{aligned}$$

where the third line follows since $\frac{\epsilon'\mathbf{M}_{\mathbf{W}}\epsilon}{\epsilon'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon} = k_{\text{liml}} + o_p(1)$ by arguments in Lemma A.3, and $N^{-1/2}\tilde{\nu}'\mathbf{M}_{\overline{\mathbf{Z}}}\epsilon$ is $O_p(1)$. Therefore, we can write:

$$\sqrt{N}\hat{g}_N(\beta) = N^{-1/2}(\mathbf{Z}_\perp\pi_{12} + \tilde{\nu})'\left(k_{\text{liml}}\mathbf{M}_{\overline{\mathbf{Z}}} - \mathbf{M}_{\mathbf{W}}\right)\epsilon$$

This expression is the (2,1) element of the quadratic form:

$$N^{-1/2}\begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \tilde{\nu}\end{pmatrix}' C \begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \tilde{\nu}\end{pmatrix}$$

where $C = k_{\text{liml}}\mathbf{M}_{\overline{\mathbf{Z}}} - \mathbf{M}_{\mathbf{W}}$. To establish (A.7), we need to check the assumptions of Lemma A.1(ii). We have:

$$\text{tr}(C) = o(N^{-1/2}) \qquad \tau_{C^2} = \frac{\alpha_K(1 - \alpha_L)}{1 - \alpha_L - \alpha_K} \tag{A.8a}$$

$$Q_{CM} = \begin{pmatrix}0 & 0 \\ 0 & \Lambda_{22}\end{pmatrix} \qquad \text{cov}\begin{pmatrix}\epsilon_i \\ \tilde{\nu}_i\end{pmatrix} = \begin{pmatrix}\Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{12}^2/\Sigma_{11}\end{pmatrix} \tag{A.8b}$$

Applying Lemma A.1 (ii) then yields (A.7).

Proof of part (ii) We can write:

$$\sqrt{N}\left(\hat{\beta}_{\text{mbtsls}} - \beta\right) = \left(\mathbf{X}'(\mathbf{M}_{\mathbf{W}} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})\mathbf{X}/N\right)^{-1} N^{-1/2}\left(\mathbf{X}'(\mathbf{M}_{\mathbf{W}} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})\epsilon\right)$$

By Lemma A.3, we have:

$$\left(\mathbf{X}'(\mathbf{M}_{\mathbf{W}} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})\mathbf{X}/N\right)^{-1} = \Lambda_{22} + o_p(1) \tag{A.9}$$

The second term is a (2,1) element of the quadratic form:

$$N^{-1/2}\begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \nu\end{pmatrix}' C \begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \nu\end{pmatrix}$$

where $C = (\mathbf{M}_{\mathbf{W}} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})$. Applying Lemma A.1 (ii) with $\text{tr}(C), \tau_{C^2}$ and Q_{CM} given by

Equation (A.8), and $\text{cov}(\epsilon_i, \nu_i) = \Sigma$ then yields:

$$N^{-1/2}\left(\mathbf{X}'(\mathbf{M_W} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})\epsilon\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_L-\alpha_K}(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2)\right)$$

Combining this result with (A.9) yields part ((ii)) in the Theorem.

Proof of Part (iii) Write the estimator as:

$$\sqrt{N}\left(\hat{\beta}_{\text{mjive}} - \beta\right) = (\mathbf{X}'(\mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X}/N)^{-1}N^{-1/2}\mathbf{X}'(\mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\epsilon.$$

By Lemma A.3, the first term satisfies:

$$(\mathbf{X}'(\mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X}/N)^{-1} = \Lambda_{22}^{-1} + o_p(1) \tag{A.10}$$

The second term is the (2,1) element of:

$$N^{-1/2}\begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \nu\end{pmatrix}' C \begin{pmatrix}\epsilon & \mathbf{Z}_\perp\pi_{12} + \nu\end{pmatrix}$$

where $C = \mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}}$. Because $\text{tr}(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}) = \text{tr}((\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1})$, we have:

$$\begin{aligned}\text{tr}(C) &= N - L_N - (1 - L_N/N)N = 0\\ \text{tr}(C^2/N) &= (L_N/N - 1) + (1 - L_N/N)^2\,\text{tr}((\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}/N)\\ &\xrightarrow{p} (\alpha_L - 1) + (1 - \alpha_L)^2\tau\end{aligned}$$

Q_{CM} is given by Equation (A.8), and $\text{cov}(\epsilon_i, \nu_i) = \Sigma$. Moreover, by Assumption 1:

$$\begin{aligned}\sup_N \max_{i\le N}\sum_{j=1}^N |c_{ij}| &\le 1 + \sup_N \max_{i\le N}\sum_{j=1}^N |(\mathbf{P_W})_{ij}|\\ &\quad + (1 - L_N/N)\sup_N \max_{i\le N}\sum_{j=1}^N |(\mathbf{I}_{ij} - (\mathbf{P_W})_{ij} - (\mathbf{P}_{\mathbf{Z}_\perp})_{ij})||((\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1})_{jj}|\\ &\le 1 + C_P + \frac{C_P}{1 - C_D} < \infty\end{aligned}$$

Applying Lemma A.1 (ii) and combining it with (A.10) then yields the result. □

Proof of Theorem 3. Under Assumption 2, we have:

$$\begin{aligned}\sqrt{\alpha_K}\begin{pmatrix}(\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}_\perp'\mathbf{Y}\\ (\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}_\perp'\mathbf{X}\end{pmatrix} \Big|\, \overline{\mathbf{Z}} &\sim \mathcal{N}\left(\begin{pmatrix}\tilde{\pi}_{12}\beta + \tilde{\gamma}\\ \tilde{\pi}_{12}\end{pmatrix}, \alpha_K\Omega \otimes \mathbf{I}_{K_N}\right)\\ \overline{\mathbf{Y}}_\perp'\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp \mid \overline{\mathbf{Z}} &\sim \mathcal{W}_2(N - K_N - L_N, \Omega)\end{aligned}$$

Moreover, these two statistics are independent. Let $b = (1, -\beta)'$ and $a = (\beta, 1)$. Assumption 5

then implies that unconditionally:

$$\overline{\mathbf{Y}}'_\perp \mathbf{P}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \sim \mathcal{W}_2(K_N, \Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega, \left(\Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega\right)^{-1} K_N a a' \mu_\pi^2/\alpha_K)$$
$$\overline{\mathbf{Y}}'_\perp \mathbf{M}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \sim \mathcal{W}_2(N - K_N - L_N, \Omega)$$

with the independence property preserved. Applying Lemma A.2 then after some algebra yields:

$$N^{1/2}\left(\mathbf{X}'_\perp \mathbf{M}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp b/N - (1-\alpha_K-\alpha_L)\Sigma_{12}\right) \xrightarrow{d} \mathcal{N}\left(0, (1-\alpha_K-\alpha_L)V_\Sigma\right) \tag{A.11a}$$
$$N^{1/2}\left(\mathbf{X}'_\perp \mathbf{P}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp /Nb - (\alpha_K \Sigma_{12})\right) \xrightarrow{d} \mathcal{N}\left(0, \alpha_K V_\Sigma + V_\Xi\right) \tag{A.11b}$$

where

$$V_\Sigma = \Sigma_{22}\Sigma_{11} + \Sigma_{12}^2$$
$$V_\Xi = \Lambda_{22}\Sigma_{11} + \Lambda_{11}\Sigma_{22} + \alpha_K^{-1}\Lambda_{22}\Lambda_{11}$$

Equations (A.11) imply:

$$N^{1/2}\left(\mathbf{X}'_\perp \mathbf{P}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp/N + (1-k_{\text{mbtsls}})\mathbf{X}'_\perp \mathbf{M}_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp/N\right) b \xrightarrow{d} \mathcal{N}\left(0, V_\Xi + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L} V_\Sigma\right)$$

Because, by Lemma A.3, $(\mathbf{X}'_\perp \mathbf{P}_{\mathbf{Z}_\perp} \mathbf{X}_\perp N + (1-k_{\text{mbtsls}})\mathbf{X}'_\perp \mathbf{M}_{\mathbf{Z}_\perp} \mathbf{X}_\perp/N)^{-1} \xrightarrow{p} \Lambda_{22}^{-1} + o_p(1)$, this yields the claim in the theorem.

Now consider mjive. Write the estimator as:

$$\sqrt{N}\left(\hat{\beta}_{\text{mjive}} - \beta\right) = (\mathbf{X}'(\mathbf{M}_\mathbf{W} - (1-L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X}/N)^{-1} N^{-1/2} \mathbf{X}'(\mathbf{M}_\mathbf{W} - (1-L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})(\mathbf{Z}_\perp\gamma + \epsilon).$$

By Lemma A.3, the first term satisfies:

$$(\mathbf{X}'(\mathbf{M}_\mathbf{W} - (1-L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\mathbf{X}/N)^{-1} = \Lambda_{22}^{-1} + o_p(1) \tag{A.12}$$

Let where $\tilde{\epsilon} = (\mathbf{Z}'_\perp \mathbf{Z}_\perp)^{-1/2}\mathbf{Z}'_\perp \epsilon$ and $\tilde{\nu} = (\mathbf{Z}'_\perp \mathbf{Z}_\perp)^{-1/2}\mathbf{Z}'_\perp \nu$. The second term can be rewritten as:

$$\begin{aligned}
&N^{-1/2}\mathbf{X}'(\mathbf{M}_\mathbf{W} - (1-L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})(\mathbf{Z}_\perp\gamma + \epsilon) \\
&\quad = N^{-1/2}\left(\pi'_{12}\mathbf{Z}'_\perp\mathbf{Z}_\perp\gamma + \pi'_{12}\mathbf{Z}'_\perp\epsilon + \gamma'\mathbf{Z}'_\perp\nu + \nu'(\mathbf{M}_\mathbf{W} - (1-L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}})\epsilon\right) \\
&\quad = N^{-1/2}\left((\alpha_K^{-1/2}\tilde{\pi}_{12} + \tilde{\nu})'(\alpha_K^{-1/2}\tilde{\pi}_{12} + \tilde{\epsilon}) + \nu'((\mathbf{I} - (1-L_N/N)(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1})\mathbf{M}_\mathbf{W}\mathbf{M}_{\mathbf{Z}_\perp})\epsilon\right)
\end{aligned}$$

Because $(\tilde{\epsilon}, \tilde{\nu})$ is independent of $\mathbf{M}_{\mathbf{Z}_\perp}(\epsilon, \nu)$, the two terms are independent. The distribution of the first term is given by the (2,1) element of a random variable with distribution

$$\mathcal{W}_2\left(K_N, \Omega + \alpha_K^{-1}\Xi, (\Omega + \alpha_K^{-1}\Xi)^{-1}\begin{pmatrix} 0 & 0 \\ 0 & \frac{K_N}{\alpha_K}\mu_\pi^2 \end{pmatrix}\right)$$

so that by Lemma A.2:

$$N^{-1/2}\left((\alpha_K^{-1/2}\tilde{\pi}_{12}+\tilde{\nu})'(\alpha_K^{-1/2}\tilde{\pi}_{12}+\tilde{\epsilon})-K_N\Omega_{12}\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,\Sigma_{11}\Lambda_{22}+\alpha_K(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2)+\Lambda_{11}(\Sigma_{11}+\Lambda_{22}/\alpha_K)\right)$$

Applying Lemma A.1(ii) to the second term yields:

$$N^{-1/2}\left(\nu'((\mathbf{I}-(1-L_N/N)(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}})^{-1})\mathbf{M}_{\mathbf{W}}\mathbf{M}_{\mathbf{Z}_\perp})\epsilon+K_N\Omega_{12}\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,(\alpha_L-1-\alpha_K+(1-\alpha_L)^2\tau)(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2)\right)$$

Adding the variances of these limit distributions yields the result. □

Proof of Theorem 4. Because $\tilde{\gamma}_k=\sqrt{q(1-q)}\gamma_k$ and $\tilde{\pi}_{12,k}=\sqrt{q(1-q)}\pi_{12,k}$, we can write the estimator as:

$$\hat{\beta}_{\text{wald}}=\beta+\frac{\frac{1}{K_N}\sum_k\tilde{\gamma}_k+\sqrt{q(1-q)}\left(\frac{1}{Nq}\sum_{i\colon Q_i=1}\epsilon_i-\frac{1}{N(1-q)}\sum_{i\colon Q_i=0}\epsilon_i\right)}{\frac{1}{K_N}\sum_k\tilde{\pi}_{12,k}+\sqrt{q(1-q)}\left(\frac{1}{Nq}\sum_{i\colon Q_i=1}\nu_i-\frac{1}{N(1-q)}\sum_{i\colon Q_i=0}\nu_i\right)}$$

By law of large numbers, we have:

$$\frac{1}{K_N}\sum_k\tilde{\pi}_{12,k}+\sqrt{q(1-q)}\left(\frac{1}{Nq}\sum_{i\colon Q_i=1}\nu_i-\frac{1}{N(1-q)}\sum_{i\colon Q_i=0}\nu_i\right)\xrightarrow{p}\mu_\pi$$

Therefore, because $\mu_\pi\neq 0$:

$$\sqrt{N}\left(\hat{\beta}_{\text{wald}}-\beta\right)=\frac{1}{\mu_\pi}\left(\frac{\sqrt{N/K_N}}{\sqrt{K_N}}\sum_k\tilde{\gamma}_k+\frac{\sqrt{(1-q)}}{\sqrt{Nq}}\sum_{i\colon Q_i=1}\epsilon_i-\frac{\sqrt{q}}{\sqrt{N(1-q)}}\sum_{i\colon Q_i=0}\epsilon_i\right)+o_p(1)$$

All three terms are Normally distributed and mutually independent. Adding up the variances yields the result. □

Proof of Theorem 5. Denote the matrices of instruments and exogenous regressors in the model (5.7) by $\sim$, so that $\tilde{\mathbf{W}}=[\mathbf{Q},\mathbf{W}]$, where $\mathbf{Q}$ is an N-vector of basic instruments, $\tilde{\mathbf{Z}}$ is the matrix of first K_N-1 columns of $\mathbf{Z}$, and $\tilde{\mathbf{Z}}_\perp=\mathbf{M}_{\tilde{\mathbf{W}}}\tilde{\mathbf{Z}}$. Then $\mathbf{P}_{\tilde{\mathbf{W}}}=\mathbf{P}_{\mathbf{W}}+\mathbf{P}_{\mathbf{Q}_\perp}$, where $(\mathbf{P}_{\mathbf{Q}_\perp})_{ij}=\frac{(Q_i-q)(Q_j-q)}{Nq(1-q)}$. Note that $\overline{\mathbf{Z}}$ remains the same.

Let $\bar{\nu}_k=\frac{K_N}{N}\sum_{i\colon G_i=k}\nu_i$ denote group averages, let $\bar{\nu}_{1,k}=\frac{K_N}{qN}\sum_{i\colon Q_i=1,G_i=k}\nu_i$ denote group averages for individuals with $Q_i=1$, and let $\bar{\nu}_{0,k}=\frac{K_N}{(1-q)N}\sum_{i\colon Q_i=0,G_i=k}\nu_i$ denote group averages

for individuals with $Q_i = 0$. Define:

$$\hat{\Sigma}_{12,k} = \frac{K_N}{N} \sum_{i:\, G_i=k} \nu_i \epsilon_i - \bar{\nu}_k \bar{\epsilon}_k \qquad \hat{\Sigma}_{22,k} = \frac{K_N}{N} \sum_{i:\, G_i=k} \nu_i^2 - \bar{\nu}_k^2$$

$$\epsilon_{10,k} = \frac{K_N}{N} \sum_{i:\, G_i=k} \frac{Q_i - q}{\sqrt{(1-q)q}} \epsilon_i = \sqrt{q(1-q)}(\bar{\epsilon}_{1,k} - \bar{\epsilon}_{0,k}) \qquad s_k^{\gamma,\epsilon} = \tilde{\gamma}_k - \mu_\gamma + \epsilon_{10,k}$$

Some tedious algebra shows that the mbtsls estimator is given by:

$$\begin{aligned}\hat{\beta}_{\text{mbtsls}} &= \frac{(1-\hat{k}_{\text{mbtsls}})\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{Y} + \hat{k}_{\text{mbtsls}}\mathbf{X}'\mathbf{P}_{\tilde{\mathbf{Z}}_\perp}\mathbf{Y}}{(1-\hat{k}_{\text{mbtsls}})\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{X} + \hat{k}_{\text{mbtsls}}\mathbf{X}'\mathbf{P}_{\tilde{\mathbf{Z}}_\perp}\mathbf{X}} \\ &= \beta + \frac{\frac{1}{K_N}\sum_k \left(s_k^{\gamma,\epsilon} s_k^{\pi_{12},\nu} + (1-\hat{k}_{\text{mbtsls}})(\hat{\Sigma}_{12,k} - \epsilon_{10,k}\nu_{10,k})\right) - \frac{1}{K_N^2}\sum_k \sum_l s_l^{\pi_{12},\nu} s_k^{\gamma,\epsilon}}{\frac{1}{K_N}\sum_k \left((s_k^{\pi_{12},\nu})^2 + (1-\hat{k}_{\text{mbtsls}})(\hat{\Sigma}_{22,k} - \nu_{10,k}^2)\right) - \frac{1}{K_N^2}\sum_k \sum_l s_l^{\pi_{12},\nu} s_k^{\pi_{12},\nu}}\end{aligned}$$

By the weak law of large numbers, we have:

$$\frac{1}{K_N}\sum_k \hat{\Sigma}_{22,k} \xrightarrow{p} (1-\alpha_K)\Sigma_{22} \qquad \frac{1}{K_N}\sum_k \nu_{10,k}^2 \xrightarrow{p} \alpha_K \Sigma_{22} \tag{A.13a}$$

$$\frac{1}{K_N}\sum_k (s_k^{\pi_{12},\nu})^2 \xrightarrow{p} \Xi_{22} + \alpha_K \Sigma_{22} \qquad \frac{1}{K_N}\sum_k s_k^{\pi_{12},\nu} \xrightarrow{p} 0 \tag{A.13b}$$

Hence:

$$\left(\frac{1}{K_N}\sum_k \left((s_k^{\pi_{12},\nu})^2 + (1-\hat{k}_{\text{mbtsls}})(\hat{\Sigma}_{22,k} - \nu_{10,k}^2)\right) - \frac{1}{K_N^2}\sum_k \sum_l s_l^{\pi_{12},\nu} s_k^{\pi_{12},\nu}\right)^{-1} = \Xi_{22} + o_p(1) \tag{A.14}$$

The nominator can be written as:

$$\begin{aligned}&\frac{1}{K_N}\sum_k \left(s_k^{\gamma,\epsilon} s_k^{\pi_{12},\nu} + (1-\hat{k}_{\text{mbtsls}})(\hat{\Sigma}_{12,k} - \epsilon_{10,k}\nu_{10,k})\right) - \frac{1}{K_N^2}\sum_k \sum_l s_l^{\pi_{12},\nu} s_k^{\gamma,\epsilon} = \\ &\qquad \frac{1}{K_N}\sum_k D_{k,\hat{k}_{\text{mbtsls}}} - \frac{1}{K_N^2}\sum_k s_k^{\gamma,\epsilon} s_k^{\pi_{12},\nu} = \frac{1}{K_N}\sum_k D_{k,\hat{k}_{\text{mbtsls}}} + O_p(1/K_N)\end{aligned}$$

where:

$$D_{k,\hat{k}_{\text{mbtsls}}} = s_k^{\gamma,\epsilon} s_k^{\pi_{12},\nu} + (1-\hat{k}_{\text{mbtsls}})(\hat{\Sigma}_{12,k} - \epsilon_{10,k}\nu_{10,k}) - \frac{1}{K_N}\sum_{l<k}\left(s_l^{\pi_{12},\nu} s_k^{\gamma,\epsilon} + s_k^{\pi_{12},\nu} s_l^{\gamma,\epsilon}\right)$$

Note that under the Assumption that $\Xi_{12} = 0$, $\{K_N^{-1/2} D_{k,\hat{k}_{\text{mbtsls}}}\}_{k\geq 1}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_k = \sigma(\gamma_k, \pi_{12,k}, \{\epsilon_i \colon G_i = k\}, \{\nu_i \colon G_i = k\})$.

The next step is to show that:

$$\frac{\sqrt{N}}{K_N}\sum_k D_{k,\hat{k}_{\text{mbtsls}}} \xrightarrow{d} \mathcal{N}\left(0, \Xi_{11}\Xi_{22}/\alpha_K + \Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11} + \frac{(1-\alpha_K)\alpha_K}{(1-2\alpha_K)}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right)\right) \tag{A.15}$$

by applying the martingale central limit theorem. The claim of the theorem for mbtsls will then follow by combining (A.15) with (A.14). To show (A.15), we first need to check that:

$$\frac{1}{K_N}\sum_{k=1}^{K_N}\mathbb{E}[D^2_{k,\hat{k}_{\text{mbtsls}}} \mid \mathcal{F}_{N,k-1}] \xrightarrow{p} \Xi_{11}\Xi_{22} + \alpha_K(\Xi_{11}\Sigma_{22}+\Xi_{22}\Sigma_{11}) + \frac{(1-\alpha_K)\alpha_K^2}{(1-2\alpha_K)}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right) \tag{A.16}$$

Expanding the left-hand side yields:

$$\begin{aligned}\mathbb{E}[D^2_{k,\hat{k}_{\text{mbtsls}}} \mid \mathcal{F}_{N,k-1}] &= \Xi_{11}\Xi_{22} + \frac{K_N}{N}\Xi_{11}\Sigma_{22} + \frac{K_N}{N}\Xi_{22}\Sigma_{11} + 2\frac{K_N}{N}\Sigma_{12}\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k} s_l^{\pi_{12},\nu}s_m^{\gamma,\epsilon} \\ &+ (\Xi_{11}+\frac{K_N}{N}\Sigma_{11})\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k} s_l^{\pi_{12},\nu}s_m^{\pi_{12},\nu} + (\Xi_{22}+\frac{K_N}{N}\Sigma_{22})\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k} s_l^{\gamma,\epsilon}s_m^{\gamma,\epsilon} \\ &+ 2\hat{k}_{\text{mbtsls}}(1-\hat{k}_{\text{mbtsls}})\mathbb{E}\epsilon_{10,k}\nu_{10,k}\hat{\Sigma}_{12,k} + \hat{k}^2_{\text{mbtsls}}(K_N/N)^2(\Sigma_{11}\Sigma_{22}+2\Sigma_{12}^2) + (1-\hat{k}_{\text{mbtsls}})^2\mathbb{E}\hat{\Sigma}^2_{12,k}\end{aligned} \tag{A.17}$$

where the expectations in the last line equal

$$\begin{aligned}\mathbb{E}\hat{\Sigma}^2_{12,k} &= \frac{K_N}{N}(1-\frac{K_N}{N})\Sigma_{11}\Sigma_{22} + (1-\frac{K_N}{N})\Sigma_{12}^2 \\ \mathbb{E}\hat{\epsilon}_{10,k}\nu_{10,k}\Sigma_{12,k} &= \frac{{K_N}^2}{N}\Sigma_{11}\Sigma_{22} + \alpha\Sigma_{12}^2\end{aligned}$$

We can therefore write:

$$\begin{aligned}\frac{1}{K_N}\sum_{k=1}^{K_N}\mathbb{E}[D^2_{k,\hat{k}_{\text{mbtsls}}} \mid \mathcal{F}_{N,k-1}] &= \Xi_{11}\Xi_{22} + \frac{K_N}{N}\Xi_{11}\Sigma_{22} \\ &+ \frac{K_N}{N}\Xi_{22}\Sigma_{11} + \frac{(1-K_N/N)(K_N/N)^2}{(1-2(K_N/N))^2}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right) + 2\frac{K_N}{N}\Sigma_{12}\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k} s_l^{\pi_{12},\nu}s_m^{\gamma,\epsilon} \\ &+ (\Xi_{11}+\frac{K_N}{N}\Sigma_{11})\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k} s_l^{\pi_{12},\nu}s_m^{\pi_{12},\nu} + (\Xi_{22}+\frac{K_N}{N}\Sigma_{22})\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k} s_l^{\gamma,\epsilon}s_m^{\gamma,\epsilon}\end{aligned}$$

Now, for $a, b \in \{(\gamma,\epsilon),(\pi_{12},\nu)\}$, note that:

$$\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k} s_l^b s_m^a = \frac{1}{K_N^3}\sum_l (K_N-l)s_l^b s_l^a + \frac{1}{K_N^3}\sum_l (K_N-l)\sum_{m<l}(s_l^b s_m^a + s_m^b s_l^a) = o_p(1)$$

Therefore, the last three terms are $o_p(1)$, which proves (A.16). One can also show that $K_N^{-2}\sum_k \mathbb{E}D^4_{k,\hat{k}_{\text{mbtsls}}} \to 0$, which implies (A.15). Next consider the mjive estimator. Let

$$\hat{\Sigma}^1_{12,k} = \frac{K_N}{N}\sum_{i:\, G_i=k} Q_i v_i \epsilon_i \qquad \Sigma^0_{12,k} = \frac{K_N}{N}\sum_{i:\, G_i=k}(1-Q_i)v_i\epsilon_i$$
$$t^{12}_k = \frac{q}{q-K_N/N}\left(\hat{\Sigma}^1_{12,k} - q\bar{\nu}_{1,k}\bar{\epsilon}_{1,k}\right) + \frac{1-q}{1-q-K_N/N}\left(\hat{\Sigma}^0_{12,k} - (1-q)\bar{\nu}_{0,k}\bar{\epsilon}_{0,k}\right)$$

Then we can write the mjive estimator as:

$$\begin{aligned}\hat{\beta}_{\text{mjive}} &= \frac{\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{Y} - \left(1-\frac{K_N}{N}\right)\mathbf{X}'\mathbf{M}_{\overline{\mathbf{Z}}}\left(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\mathbf{Y}}{\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{X} - \left(1-\frac{K_N}{N}\right)\mathbf{X}'\mathbf{M}_{\overline{\mathbf{Z}}}\left(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\mathbf{X}} \\ &= \beta + \frac{\frac{1}{K_N}\sum_k\left(s_k^{\gamma,\epsilon}s_k^{\pi_{12},\nu} + \hat{\Sigma}_{12,k} - \epsilon_{10,k}\nu_{10,k} - (1-\frac{K_N}{N})t^{12}_k\right) - K_N^{-2}\sum_k\sum_l s_k^{\gamma,\epsilon}s_l^{\pi,\nu}}{\frac{1}{K_N}\sum_k\left((s_k^{\pi_{12},\nu})^2 + \hat{\Sigma}_{22,k} - \nu^2_{10,k} - (1-\frac{K_N}{N})t^{22}_k\right) - K_N^{-2}\sum_k\sum_l s_k^{\pi,\nu}s_l^{\pi,\nu}}\end{aligned}$$

Using (A.13) and the fact that by the weak law of large numbers $t^{22}_k \xrightarrow{p} \Sigma_{22}$, we get:

$$\left(\frac{1}{K_N}\sum_k\left((s_k^{\pi_{12},\nu})^2 + \hat{\Sigma}_{22,k} - \nu^2_{10,k} - (1-\frac{K_N}{N})t^{22}_k\right) - K_N^{-2}\sum_k\sum_l s_k^{\pi,\nu}s_l^{\pi,\nu}\right)^{-1} = \Xi_{22} + o_p(1) \tag{A.18}$$

We can rewrite the nominator as:

$$\frac{1}{K_N}\sum_k\left(s_k^{\gamma,\epsilon}s_k^{\pi_{12},\nu} + \hat{\Sigma}_{12,k} - \epsilon_{10,k}\nu_{10,k} - (1-\frac{K_N}{N})t^{12}_k\right) - K_N^{-2}\sum_k\sum_l s_k^{\gamma,\epsilon}s_l^{\pi,\nu} = \frac{1}{K_N}\sum_k\tilde{D}_k + o_p(K_N^{-1})$$

where:

$$\tilde{D}_k = D_{k,0} - (1-\frac{K_N}{N})t^{12}_k$$

Like in the case of mbtsls, under the Assumption that $\Xi_{12}=0$, $\{K_N^{-1/2}D_{k,\hat{k}_{\text{mbtsls}}}\}_{k\geq 1}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_k = \sigma(\gamma_k, \pi_{12,k}, \{\epsilon_i\colon G_i = k\}, \{\nu_i\colon G_i = k\})$. To prove the claim of the theorem for mjive, it therefore remains to check that:

$$\begin{aligned}&\frac{1}{K_N}\sum_{k=1}^{K_N}\mathbb{E}[\tilde{D}_k^2 \mid \mathcal{F}_{N,k-1}] \xrightarrow{p} \\ &\qquad \Xi_{11}\Xi_{22} + \alpha_K(\Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11}) + \alpha_K(1-\alpha_K)((1-\alpha_K)\tau - 1)\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\end{aligned} \tag{A.19}$$

and that:

$$K_N^{-2} \sum_k \mathbb{E}\tilde{D}_k^4 \to 0 \tag{A.20}$$

These two conditions will allow us to apply the martingale central limit theorem to $K_N^{-1} \sum_k \tilde{D}_k$. We first establish (A.19). Using the expansion in (A.17), we get that:

$$\begin{aligned}\frac{1}{K_N} \sum_{k=1}^{K_N} \mathbb{E}[\tilde{D}_k^2 \mid \mathcal{F}_{N,k-1}] = \Xi_{11}\Xi_{22} + \alpha_K(\Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11}) + \alpha_K(1-\alpha_K)\Sigma_{11}\Sigma_{22} + (1-\alpha_K)\Sigma_{12}^2 \\ + (1 - K_N/N)^2 \mathbb{E}[t_k^{12} t_k^{12}] - 2(1 - K_N/N)\mathbb{E}[D_{k,0} t_k^{12}] + o_p(1)\end{aligned}$$

The remaining expectations are given by:

$$\begin{aligned}\mathbb{E}[D_{k,0} t_k^{12} \mid \mathcal{F}_{N,k-1}] &= \frac{K_N}{N}\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2 \\ \mathbb{E}[t_k^{12} t_k^{12} \mid \mathcal{F}_{N,k-1}] &= \left(\frac{q^2}{1 - K_N/N} + \frac{(1-q)^2}{1 - q - K_N/N} \right) \frac{K_N}{N} (\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2) + \Sigma_{12}^2\end{aligned}$$

Substituting them in the expansion above yields (A.19). It can also be shown that (A.20) holds, which proves the result. □

References

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.

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

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

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.

— 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.

Andrews, D. W. K., Moreira, M. J. and Stock, J. H. (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, **74** (3), 715–752.

Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67.

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

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

Ashley, R. (2009). Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics*, **24** (2), 325–337.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, **25** (1), 77–83.

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

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

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. B. (2011). Sparse models and methods for optimal instruments with an application to eminent domain, unpublished manuscript.

Berkowitz, D., Caner, M. and Fang, Y. (2008). Are "Nearly Exogenous Instruments" reliable? *Economics Letters*, **101** (1), 20–23.

Caner, M. (2007). Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics, unpublished manuscript.

Chamberlain, G. and Imbens, G. W. (2004). Random Effects Estimators with Many Instrumental Variables. *Econometrica*, **72** (1), 295–306.

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

—, —, Hausman, J. A., Newey, W. K. and Woutersen, T. (2010). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. *Econometric Theory*, (forthcoming).

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*, (forthcoming).

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

Conley, T. G., Hansen, C. B. and Rossi, P. E. (2007). Plausibly Exogenous, unpublished manuscript.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

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

Fisher, F. M. (1961). On the cost of approximate specification in simultaneous equation estimation. *Econometrica*, **29** (2), 139–170.

— (1966). The relative sensitivity to specification error of different k-class estimators. *Journal of the American Statistical Association*, **61** (314), 345–356.

— (1967). Approximate Specification and the Choice of a k-Class Estimator. *Journal of the American Statistical Association*, **62** (320), 1265–1276.

Flores, C. A. and Flores-Lagunes, A. (2010). Partial Identification of Local Average Treatment Effects with an Invalid Instrument, unpublished manuscript.

Fryer, R. G. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, (forthcoming).

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets, unpublished manuscript.

Guggenberger, P. (2010). On the Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption, unpublished manuscript.

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory*, **18** (1), 140–168.

— and Hausman, J. A. (2005). IV Estimation with Valid and Invalid Instruments. *Annales d'Économie et de Statistique*, (79/80), 25–57.

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.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2009). Instrumental Variable Estimation with Heteroskedasticity and Many Instruments, unpublished manuscript.

Kraay, A. (2008). Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions, unpublished manuscript.

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.

Levitt, S. D., List, J. A., Neckermann, S. and Sadoff, S. (2011). The Impact of Short-term Incentives on Student Performance, unpublished manuscript.

Mariano, R. S. (1973). Approximations to the Distribution Functions of Theil's k-Class Estimators. *Econometrica*, **41** (4), 715–721.

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.

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.

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

Nevo, A. and Rosen, A. M. (2010). Identification with Imperfect Instruments. *Review of Economics and Statistics*, (forthcoming).

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.

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.

Reinhold, S. and Woutersen, T. (2011). Endogeneity and Imperfect Instruments in Applied Work : Deriving Bounds in a Semiparametric Model, unpublished manuscript.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of econometrics*, vol. 2, *Chapter 15*, Elsevier, pp. 881–935.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65** (3), 557–586.

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

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

van Hasselt, M. (2010). Many Instruments Asymptotic Approximations Under Nonnormal Error Distributions. *Econometric Theory*, **26** (02), 633–645.

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

Table 1: Estimates for Angrist-Krueger Data ($N = 162,487$)

		Standard Error			
Estimator	$\hat{\beta}$	classic	bekker	many exo	$\Lambda_{11} > 0$
single qob dummy					
tsls	0.089	(0.021)			
liml	0.089	(0.021)	(0.021)	(0.021)	
btsls	0.089	(0.021)	(0.021)		
mbtsls	0.089	(0.021)	(0.021)	(0.021)	(0.021)
jive	0.090	(0.021)	(0.021)		
mjive	0.089	(0.021)	(0.021)	(0.021)	(0.021)
qob interacted with year and state of birth					
tsls	0.073	(0.017)			
liml	0.095	(0.017)	(0.042)	(0.042)	
btsls	0.097	(0.017)	(0.039)		
mbtsls	0.098	(0.017)	(0.040)	(0.040)	(0.039)
jive	0.056	(0.017)	(0.053)		
mjive	0.096	(0.017)	(0.054)	(0.040)	(0.040)
qob interacted with year and state of birth, qob exogenous variable					
tsls	0.069	(0.033)			
liml	0.093	(0.034)	(0.128)	(0.128)	
btsls	0.099	(0.034)	(0.131)		
mbtsls	0.102	(0.034)	(0.132)	(0.132)	(0.132)
jive	0.064	(0.033)	(0.180)		
mjive	0.096	(0.034)	(0.184)	(0.133)	(0.133)