Identification and Inference with Many Invalid Instruments*

Michal Kolesár† Raj Chetty‡ John Friedman§ Edward Glaeser¶
Guido W. Imbens‖

July 2011

Abstract

We analyze linear models with a single endogenous regressor in the presence of many instrumental variables. Generalizing the setting typically studied in this literature we allow the number of exogenous regressors to increase with the sample size, and we allow the instruments to have direct effects on the outcome. The setup leads to novel identification strategies and we develop new estimators to exploit those strategies. If the number of exogenous regressors increases with the sample size but all instruments are valid, liml remains consistent. The jive and btsls estimators need to be modified to maintain consistency. If we allow for direct effects of the instruments on the outcome, but assume that the expectation of the product of these effects and the effects of the instruments on the endogenous regressor average is zero, the modified jive and btsls estimators remain consistent but liml is no longer consistent. We argue in the context of two specific examples with a group structure that the key condition that the expectation of product of the direct effects of the instrument on the outcome and the effects of the instrument on the endogenous regressor is zero has substantive content. Furthermore, in one of the examples we consider weaker conditions where the average (over the instruments) of the direct effect of the instruments on the outcome is zero, or where the correlation of the direct effects on outcome and endogenous variable is zero.

JEL Classification:
Keywords:

*Financial support for this research was generously provided through NSF grants 0820361 and 0961707. We are grateful for comments by participants in the econometrics lunch seminar at Harvard University, the Harvard-MIT Econometrics seminar, and in particular for discussions with Gary Chamberlain.

†Dept of Economics, Harvard University. Electronic correspondence: mkolesar@fas.harvard.edu.

‡Department of Economics, Harvard University, M-24 Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: chetty@fas.harvard.edu, http://www.economics.harvard.edu/faculty/chetty.

§, Harvard University. Electronic correspondence: john_friedman@harvard.edu, http://www.hks.harvard.edu/fs/jfriedm/.

¶Department of Economics, Harvard University, M-24 Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: eglaeser@harvard.edu, http://www.economics.harvard.edu/faculty/glaeser.

‖Department of Economics, Harvard University, M-24 Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: imbens@harvard.edu, http://www.economics.harvard.edu/faculty/imbens/imbens.html.

1 Introduction

We analyze linear models with a single endogenous regressor in the presence of instrumental variables. Following Kunitomo (1980), Morimune (1983), Bekker (1994), Hahn (2002), Chamberlain and Imbens (2004), Chao and Swanson (2005), Hansen, Hausman and Newey (2008), and Anderson, Kunitomo, and Matsushita (2010), we focus on the case with many instrumental variables. Each instrument is weak in the Staiger-Stock (1997) sense, but collectively the instruments have substantial predictive power. This case has generated a considerable amount of attention, especially following the empirical study by Angrist and Krueger (1991, AK from hereon). Morimune (1983) and Bekker (1994) showed that the two-stage-least-squares (tsls) estimator is inconsistent under a sequence of parametrizations where the number of instruments increases proportional to the sample size, whereas the limited-information-maximum-likelihood (liml) estimator remains consistent and efficient in that setting, as shown by Chioda and Jansson (2009) and Anderson, Kunitomo and Matsuhita (2010). Additional estimators that have been shown to be consistent under such asymptotic sequences include a bias-corrected-two-stage-least-squares (btsls) estimator (Newey and Donald, 2001), and the jackknive-instrumental-variables-estimator (jive) (Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999).

In the current paper we build on this literature. We consider generalizations of the settings typically studied in the many-instrument literature. In these new settings the estimators that are consistent in the standard many-instrument setting now differ in their probability limits and in particular it is no longer true that the liml estimator is generally the preferred choice. We derive novel identification strategies and develop new estimators that exploit these identification strategies.

Our first result generalizes the current large sample results by allowing not only the number of instruments, but also the number of exogenous regressors to increase with the sample size. The motivation for this extension is that in many applications the number of exogenous variables is of the same order as the number of instruments. For example, in the AK study most of the instruments are generated by interacting a small number of basic instruments with a large number of exogenous regressors so that the number of instruments and exogenous covariates is of the same order of magnitude. In one of the analyses in the Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011, CFHSSY from hereon) study the instruments are classroom indicators and exogenous variables include school indicators so that again the number of instruments and exogenous regressors is of the same order of magnitude. We show that the liml estimator remains consistent in the presence of many exogenous regressors, although its standard error needs to be modified. The btsls and jive estimators require some modification in order to to maintain consistency.

Next, in our main contribution, we extend the many-instruments literature by considering forms of misspecification where each instrument potentially has a direct effect on the outcome. In that case liml loses consistency. However, as long as the correlation between the direct effect of the instrument and the strength of that instrument is zero, then our proposed modified btsls and jive estimators remain consistent. The intuition is that the liml estimator attempts to impose proportionality of all the reduced form coefficients, whereas the other, tsls-like, estimators can be

interpreted as instrumental variables estimators with a single, constructed, instrument.

We then study in detail two leading cases that motivate the set up and illustrate the range and applicability of the identifying assumption that the direct effects of the instrument on the outcome and the effect of the instrument on the endogenous regressor are uncorrelated. Both cases have a clustering structure where the population can be partitioned into subpopulations. As is common in clustering settings, asymptotic approximations are based on the number of subpopulations growing with the sample size while the number of sampled units from each subpopulation remains fixed.

The first of the two special cases arises when the instruments are indicators for membership in a subpopulation. This case is motivated by the CFHSSY study that uses an indicator for (randomly assigned) classroom as an instrument for test scores in an analysis of the effect of test scores on long term outcomes using individuals who participated in the Project Star Experiment. In this setting the assumption that the direct effect of the instrument on the outcome is uncorrelated with the strength of the instrument is sufficient for identification. This assumption can be motivated by taking a clustering perspective where the clusters correspond to the classrooms. Under that assumption liml is not consistent, but the jive and btsls estimators are. In this case jive takes on a particularly intuitive form where the outcome for unit i is regressed on the average value of the endogenous regressor for all units in that cluster other than unit i.

In the second case there is a small number of basic instruments. These basic instruments are interacted with indicators for subpopulation membership to generate a large number of instruments. Here the number of exogenous regressors (indicators for subpopulation membership) increases proportional to the number of instruments. This case is motivated by the AK study where the basic instruments, four quarter of birth indicators, are interacted with year and state of birth indicators to generate additional instruments. Our discussion suggests new identification strategies for this setting. In the first of these identification strategies the average effect of the instruments on the outcome is zero, even though every single instrument is invalid on its own. In the second identification strategy the basic instrument is allowed to have an unrestricted direct effect on the outcome, and only the interactions between the basic instrument and the group indicators satisfy exclusion restrictions.

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 whose large sample properties we study in this paper. In the next section, Section 4 we present the results for many exogenous variables. In Section 5 we discuss the general results allowing for misspecification. In Section 6 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 7. Section 8 concludes.

2 Setup

The equation of interest relates a scalar outcome Y_i, for $i = 1, \ldots, N$, and $N = 1, 2, \ldots$ to a potentially endogenous scalar regressor X_i and a vector of exogenous covariates W_i of dimension L_N:

$$Y_i = X_i\beta + W_i'\delta + \epsilon_i, \tag{2.1}$$

The second equation relates the endogenous regressor X_i to the exogenous regressors W_i and the vector of instruments Z_i with dimension K_N:

$$X_i = Z_i'\pi_{12} + W_i'\pi_{22} + \nu_i. \tag{2.2}$$

We maintain the assumption that the pairs of structural errors (ϵ_i, ν_i) are independent for $i = 1, \dots, N$, and that

$$\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix} \perp Z_i, W_i, \qquad \mathbb{E}\left[\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathbb{E}\left[\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix}\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix}'\right] = \Sigma.$$

This is a fairly standard instrumental variables set up, with the only modification that we index the number of exogenous covariates, L_N, by the sample size. To be precise we should index the random variables and parameter by the sample size N: 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. For ease of notation we drop this index.

We also consider a more fundamental generalization of the conventional outcome equation by allowing the instruments to have a direct effect on the outcome:

$$Y_i = X_i\beta + W_i'\delta + Z_i'\gamma + \epsilon_i. \tag{2.3}$$

Allowing for such direct effects calls into question the identification of β. We will consider assumptions restricting the values of γ that ensure identification.

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, 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 three $N \times N$ matrices, the projection matrix $\mathbf{P_S}$, the matrix $\mathbf{M_S}$ that projects on the orthogonal complement of $\mathbf{S}$, and the diagonal matrix $\mathbf{D_S}$ with diagonal elements equal to those of the projection matrix:

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

We use the subscript $\perp$ as shorthand for taking residuals after regression on the exogenous regressors $\mathbf{W}$, so $\mathbf{Z}_\perp = \mathbf{M_W Z}$, $\mathbf{X}_\perp = \mathbf{M_W X}$, and $\mathbf{Y}_\perp = \mathbf{M_W Y}$.

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

$$\Lambda_N = \begin{pmatrix} \Lambda_{N,11} & \Lambda_{N,12} \\ \Lambda_{N,12} & \Lambda_{N,22} \end{pmatrix} = \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}' \mathbf{Z}_\perp'\mathbf{Z}_\perp \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}. \tag{2.4}$$

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 $\Lambda_{N,22}$. In our more general set up with outcome equation (2.3), all elements of Λ_N can differ from zero. The (2,2) component of the augmented concentration parameter, $\Lambda_{N,22}$ is related to the concentration parameter (Mariano, 1973; Rothenberg, 1984), conventionally 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

Here we introduce the five estimators whose properties we shall study. Three have been introduced previously, and the other two are minor modifications of previously proposed estimators. The first three estimators fit into the k-class (Nagar, 1959; Theil, 1961, 1971; Davidson and MacKinnon, 1993). Given a scalar k, the k-class estimator for (β, α) is

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

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 (Bekker, 1995), and we will not investigate its properties under the various generalizations of the many-instrument set up here. Instead we consider a bias corrected version of the tsls estimator. Nagar (1959) suggested $\hat{k}_{\text{nagar}} = 1 + (K_N - 2)/N$, but the first of the five estimators we focus on is a slightly different version suggested by Donald and Newey (2001), with

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

Under the conventional asymptotics with the number of instruments K fixed the difference between the Nagar and Donald-Newey estimators is small, but under the many-instruments asymptotics with $K_N/N \to \alpha_K > 0$ the difference matters and the Donald-Newey version is more attractive. The second estimator we consider is a further modification of the Donald-Newey bias corrected estimator, with

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

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

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

The fourth estimator we study in the current paper is the jackknife-instrumental-variables estimator (jive, Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999),

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

Alternative versions of jackknife estimators are introduced in Angrist, Imbens and Krueger (1999) and Ackerberg and Devereux (2009). We also study a new version of the jacknife estimator, closely

related to the Ackerberg-Devereux estimator, which we refer to as the modified jive estimator, or mjive:

$$\hat{\beta}_{\text{mjive}} = \left(\mathbf{X}_\perp' \left(\mathbf{I} - \left(1 - \frac{L_N}{N}\right) \mathbf{M}_{\mathbf{Z}_\perp} \left(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\right) \mathbf{X}_\perp\right)^{-1} \times \left(\mathbf{X}_\perp' \left(\mathbf{I} - \left(1 - \frac{L_N}{N}\right) \mathbf{M}_{\mathbf{Z}_\perp} \left(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\right) \mathbf{Y}_\perp\right) \tag{3.3}$$

We analyze 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 Exogenous Regressors

in this section we look at the properties of the five estimators under the assumption that the instruments are all valid $\gamma = 0$), but allowing for many exogenous covariates ($L_N/N \to \alpha_L > 0$). If we fix $\alpha_L = 0$ we are in the many instrument case studied in the literature (e.g., Bekker, 1994; Chamberlain and Imbens, 2004; Chao and Swanson, 2005). If we restrict both $\alpha_L = \alpha_K = 0$ we are back in the case with conventional instrumental variables asymptotics.

Assumption 1. (VALIDITY OF INSTRUMENTS)
All elements of γ are zero.

Assumption 2. (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)*** $(\mathbf{Z}, \mathbf{W})$ *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)*** $\max_{i \le N} |(\mathbf{Z}_\perp)_i' \pi_{12}| / \sqrt{N} \to 0$ *and;*
- ***(v)*** $\sup_N \sup_{i \ge 1} \sum_j |(\mathbf{P}_{\mathbf{Z}_\perp})_{ij}| < C$ *and* $\sup_N \sup_{i \ge 1} \sum_j |(\mathbf{P}_{\mathbf{W}})_{ij}| < C$ *for some* $C < \infty$ *with probability one*

Assumption 3. (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.*

Assumption 4. (NUMBER OF INSTRUMENTS AND EXOGENOUS REGRESSORS)
For some $0 \le \alpha_K < 1$ *and* $0 \le \alpha_L < 1$,

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

Assumption 5. (CONCENTRATION PARAMETER)
For some positive semidefinite 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 result establishes consistency in the case with many exogenous covariates and many instruments.

Theorem 1. (CONSISTENCY WITH MANY VALID INSTRUMENTS AND MANY EXOGENOUS REGRESSORS)
Suppose Assumptions 1, 2 (i)–(iii), 3 (i), 4 and 5 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:*

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

(ii) *(liml)*

$$\beta_{\text{liml}} = \beta, \qquad k_{\text{liml}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L},$$

(iii) *(btsls)*

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

(iv) *(mbtsls)*

$$\beta_{\text{mbtsls}} = \beta, \qquad k_{\text{mbtsls}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L},$$

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

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

(vi) *(mjive)*

$$\beta_{\text{mjive}} = \beta,$$

The key insight of this theorem is that the bias corrected tsls estimator and jive are not consistent when the number of exogenous covariates increases proportional to the sample size. These two estimators are consistent if only the number of instruments increases with the sample size. There exist simple modifications, however, to achieve consistency. Especially for the btsls estimator the modification is very modest relative to the adjustment for the many instruments, changing the value for k from $1/(1-\alpha_K)$ to $(1-\alpha_L)/(1-\alpha_K-\alpha_L)$. The liml estimator is not affected by either the presence of multiple instruments or multiple exogenous covariates.

The second result establishes asymptotic normality for the estimators that are consistent in this setting with many exogenous variables and many covariates. The presence of many exogenous covariates does affect the large sample variances.

Theorem 2. (Asymptotic Normality with Many Valid Instruments and Many Exogenous Regressors)
Suppose Assumptions 1–5 hold. Then:
(i) *(liml)*

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

(ii) *(mbtsls)*

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

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

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

The conventional results for these estimators under many instruments asymptotics correspond to the special case with $\alpha_L=0$. Note that the presence of many covariates increases the variance for the liml and mbtsls estimators because $(1-\alpha_L)(1-\alpha_K)/(1-\alpha_K-\alpha_L)>1$ if $0<\alpha_L<1$.

5 Many Invalid Instruments

In this section we extend the previous results to allow for direct effects of the instruments on the outcome. First we establish the probability limits of the previously considered estimators under the same assumptions as before other than that we allow γ to be different from zero.

Theorem 3. (Consistency with Many Invalid Instruments and Many Exogenous Regressors)
Suppose Assumptions 2 (i)–(iii), 3 (i), 4 and 5 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:*

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

(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)*

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

The key insight from this result 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. On the other hand, in order for liml to be consistent for all values of Σ, it has to be the case that Λ_{11} is equal to zero (and that immediately implies that $\Lambda_{12} = 0$. The intuition appears to be that 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, which holds under correct specificaction, is never exploited. On the other hand, the liml estimator tries to impose proportionality of the reduced forms. This leads to efficiency if proportionality holds, but if it does not hold, the estimator need not be consistent.

Next we consider asymptotic normality. We consider two additional assumptions. The first restricts the direct effects of the instrument on the outcome. Combined with the earlier assumptions it implies consistency of the mbtsls and mjive estimators.

Assumption 6. (ORTHOGONAL EFFECTS)
The augmented concentration parameter is diagional, or $\Lambda_{12} = 0$.

The second additional assumption puts a random effects structure on the direct effects of the instrument on the outcome and the endogenous regressor. This assumption is less common in the instrumental variables literature. The issue is that the estimators (mbtsls and mjive) depend 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 mbtsls and mjive, and we need to put sufficient structure on it to be able to determine this distribution. A random effects structure is a natural way to do so, although not necessarily the only one. First we redefine the parameters by orthogonalizing them with respect to $\mathbf{Z}_\perp$ as

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

Then we consider the following assumption

Assumption 7. (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 7 is that

$$\Lambda = \text{plim}\left(\frac{\Lambda_N}{N}\right) = \text{plim}\left(\frac{1}{N} \left(\begin{array}{cc} \gamma & \pi_{12} \end{array} \right) (\mathbf{Z}_\perp' \mathbf{Z}_\perp) \left(\begin{array}{c} \gamma' \\ \pi_{12}' \end{array} \right)\right) = \left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right) \left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right)' + \Xi.$$

This representation of Λ will be useful in interpreting the assumption concerning its off-diagonal value. We defer further discussion of this assumption to the next section where we consider two special cases.

Theorem 4. (ASYMPTOTIC NORMALITY WITH MANY INVALID INSTRUMENTS AND MANY EXOGENOUS REGRESSORS)
Suppose that Assumptions 2–7 hold. Suppose in addition that $\mu_\gamma = 0$*. Then:*
(i) *(mbtsls)*

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

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

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

6 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, with the subpopulation indicator $G_i \in= \{1, 2, \ldots, G_N\}$. 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. Let the number of units in group g be N_g, with $N = \sum_{g=1}^{G_N} N_g$. For convenience, let us initially assume that the number of unit sampled from each subpopulation is the same for all subpopulations, $N_g = N/G_N$ for all g.

6.1 Special Case I: Subpopulations with Clustering

To focus on the key issues, let us assume there are no exogenous regressors beyond the intercept, $L_N = 1$. In the first special case the instruments are indicators for the groups, $Z_i =$

$(Z_{i1}, \ldots, Z_{iK_{Z,N}-1})$, where $Z_{ik} = \mathbf{1}_{G_i=k}$, for $k = 1, \ldots, G_N$, and the number of instruments K_N equals the number of groups G_N. The general model in (2.2) and (2.3) can now be written as

$$Y_i = \delta + \beta X_i + \sum_{k=1}^{K_N-1} \gamma_k Z_{ik} + \epsilon_i, \tag{6.1}$$

$$X_i = \pi_{11} + \sum_{k=1}^{K_N-1} \pi_{12,k} Z_{ik} + \nu_i. \tag{6.2}$$

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

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

where

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

Therefore

$$\Lambda = \text{plim}(\Lambda/N) = \Xi.$$

Now let us consider Assumption 7 and interpret in this context with a group structure. Suppose we have a large population of groups. Let $\mu_{Y,g}$ and $\mu_{X,g}$ be the population means of $Y_i - \beta X_i$ and X_i in group g, and let μ_Y and μ_X be the overall population means of $Y_i - \beta X_i$ and X_i, the average of the group means. The natural place to put a random effects structure on the parameters would be to assume that the pairs $(\mu_{Y,k}, \mu_{X,k})$ are independent and

$$\begin{pmatrix} \mu_{Y,k} \\ \mu_{X,k} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Phi\right). \tag{6.3}$$

In terms of the specification in (6.1) and (6.2), we have $\mu_{Y,k} = \delta + \gamma_k$, $\mu_{Y,K} = \delta$, $\mu_{X,k} = \pi_{11} + \pi_{12,k}$ $\mu_{X,K} = \pi_{11}$. We can invert this to

$$\delta = \mu_{Y,K}, \qquad \gamma_k = \mu_{Y,k} - \mu_{Y,K}, \qquad \pi_{11} = \mu_{X,K}, \qquad \text{and} \quad \pi_{12,k} = \mu_{X,k} - \mu_{X,K}.$$

Now consider the transformation to $(\tilde{\gamma}\ \tilde{\pi}_{12})$. Because in this case with equal group sizes and only the intercept as an exogenous regressor, we have $\mathbf{Z}_\perp'\mathbf{Z}_\perp = \mathbf{I}_{K_N-1} - \iota_{K_N-1}\iota_{K_N-1}'/K_N$, we have

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

$$= \alpha_K^{1/2}\left(\mathbf{I}_{K_N-1} - \frac{1}{K_N-1}\left(1 + \frac{1}{\sqrt{K_N}}\right)\iota_{K_N-1}\iota_{K_N-1}'\right)\begin{pmatrix} \mu_{Y,1} - \mu_{Y,K} & \mu_{X,1} - \mu_{X,K} \\ \vdots & \vdots \\ \mu_{Y,K-1} - \mu_{Y,K} & \mu_{X,K-1} - \mu_{X,K} \end{pmatrix}$$

$$= \alpha_K^{1/2} B \begin{pmatrix} \mu_{Y,1} & \mu_{X,1} \\ \vdots & \vdots \\ \mu_{Y,K} & \mu_{X,K} \end{pmatrix},$$

where the $(K-1)\times K$ matrix B equals

$$B = \left(\mathbf{I}_{K_N-1} - \frac{1}{K_N-1}\left(1+\frac{1}{\sqrt{K_N}}\right)\iota_{K_N-1}\iota'_{K_N-1} \quad : \quad \iota_{K_N-1}\cdot\frac{1}{\sqrt{K_N}} \right),$$

satisfying

$$B\iota_{K_N} = 0, \qquad \text{and} \quad BB' = c\cdot \mathbf{I}_{K_N-1}.$$

Thus, a random effects specification on $(\mu_{Y,k},\mu_{X,k})$ as in (6.3) implies a random effects specification on $(\tilde\gamma\ \tilde\pi_{12})$, namely

$$\left.\begin{pmatrix} \tilde\gamma_k \\ \tilde\pi_{12,k} \end{pmatrix}\right| \mathbf{Z},\mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Xi\right),$$

with $\Xi = c\cdot\Phi$.

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.

There is an important alternative interpretation to the set up in (6.1)–(6.2). In this alternative interpretation the γ_k are viewed as random effects reflecting clustering. Let us write the outcome equation (6.1) as

$$Y_i = \delta + \beta X_i + \eta_i,$$

where the composite residual η_i has the form

$$\eta_i = \xi_{G_i} + \epsilon_i,$$

where the cluster-specific component ξ_k is equal to the direct effect of the instrument:

$$\xi_k = \gamma_k.$$

Then assuming the γ_k have expectation zero and denoting their variance by $\Xi_{11} > 0$, we can think of the residuals η_i having a clustering structure

$$\mathbb{E}\left[\eta_i \mid \mathbf{Z}\right] = 0 \qquad \text{and} \quad \mathbb{E}\left[\eta_i\eta_j' \mid \mathbf{Z}\right] = \begin{cases} \sigma_\varepsilon^2 + \Xi_{11} & \text{if } i=j, \\ \Xi_{11} & \text{if } G_i = G_j, i\neq j \\ 0 & \text{otherwise.} \end{cases}$$

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

$$X_i = \pi_{11} + \zeta_i, \qquad \text{where} \quad \zeta_i = \varphi_{G_i} + \nu_i, \qquad \text{and} \quad \varphi_{G_k} = \pi_{12,k}.$$

Now we have

$$\mathbb{E}\left[\zeta_i \mid \mathbf{Z}\right] = 0 \qquad \text{and} \quad \mathbb{E}\left[\zeta_i\zeta_j' \mid \mathbf{Z}\right] = \begin{cases} \sigma_\nu^2 + \Xi_{22} & \text{if } i=j, \\ \Xi_{22} & \text{if } G_i = G_j, i\neq j \\ 0 & \text{otherwise.} \end{cases}$$

The critical assumption that Λ_{12} is equal to zero (equivalent to $\Xi_{12} = 0$) now amounts to assuming that the cluster component in the outcome equation, ξ_k, is uncorrelated with the cluster component in the first stage, φ_k. This assumption is not innocuous, but assumptions about zero correlations for cluster components are often made in clustering settings.

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

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

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

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

With a finite number of units per cluster omitting unit i can make a 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 3-4, specializing it to the case with $L_N = 1$ so that $\alpha_L = 0$.

Corollary 1. *Suppose Assumptions 2-5 and 7 hold. Then*
(i) (k-class)

$$\hat{\beta}_k \to \beta + \frac{\{1 - (1 - \alpha_Z)\, k\}\, \Sigma_{21} + \Lambda_{12}}{\{1 - (1 - \alpha_Z)\, k\}\, \Sigma_{22} + \Lambda_{22}},$$

so that a sufficient condition for consistency of k-class estimators is that $\hat{k} \to 1/(1 - \alpha_Z)$ and $\Lambda_{12} = 0$.
(ii), (liml) liml is not consistent if $\rho_{\pi\gamma} = 0$, unless $\sigma_\gamma^2 = 0$:

$$\hat{k}_{\text{liml}} \to \frac{1}{1 - \alpha_Z} \cdot \left(1 + \min\text{eig}\left(\Sigma^{-1}\Lambda\right)\right),$$

and

$$\hat{\beta}_{\text{liml}} = \beta + \frac{\Lambda_{12} - \min\text{eig}\left(\Sigma^{-1}\Lambda\right)\Sigma_{12}}{\Lambda_{22} - \min\text{eig}\left(\Sigma^{-1}\Lambda\right)\Sigma_{22}},$$

(iii) (btsls and bmtsls) the btsls and mbtsls estimators are consistent if $\Lambda_{12} = 0$,

$$\hat{\beta}_{\text{btsls}} \to \beta + \frac{\Lambda_{12}}{\Lambda_{22}}, \qquad \hat{\beta}_{\text{mbtsls}} \xrightarrow{p} \beta + \frac{\Lambda_{12}}{\Lambda_{22}},$$

(iv) (jive and mjive) $\hat{\beta}_{\text{jive}}$ and $\hat{\beta}_{\text{mjive}}$ are consistent if $\Lambda_{12} = 0$,

$$\hat{\beta}_{\text{jive}} \longrightarrow \beta + \frac{\Lambda_{12}}{\Lambda_{22}}, \qquad \hat{\beta}_{\text{mjive}} \longrightarrow \beta + \frac{\Lambda_{12}}{\Lambda_{22}}.$$

6.2 Special Case II: Groups with Interactions

In the second case we maintain the subpopulation structure: each unit i belongs to a subpopulation $G_i \in \{1, 2, \ldots, G_N\}$. 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 group indicator with this binary instrument. We include the group indicators as exogenous covariates, $W_{i,k} = \mathbf{1}_{G_i=k}$, so that $K_N = L_N = G_N$. Again for ease of exposition let us assume that the groups are all equal size, $N_g = N/G_N$ for all g, and that the fraction of $Q_i = 1$ units in each subsample 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

$$Y_i = \beta X_i + \sum_{k=1}^{K_N} \delta_k W_{ik} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i, \tag{6.4}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Z_{ik} + \sum_{k=1}^{K_N} \pi_{22,k} W_{ik} + \nu_i. \tag{6.5}$$

In this case

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

We can directly apply the results from Section 5, 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 a plausible assumption. It would require that $\mu_\gamma \cdot \mu_\pi + \Xi_{12} = 0$. We can in fact relax the sufficient conditions for identification in this special setting. We consider two specific alternatives. First, suppose 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.

Under the first assumption, $\mu_\gamma = 0$, we can simply use the just-identified iv estimator with Q_i as the single instrument. Adding the interactions of the type $Q_i \cdot \mathbf{1}_{G_i=k}$ as additional instruments would lead to inconsistency if we use liml, btsls, mbtsls, jive or mjive.

Theorem 5. (Zero Mean)
Suppose the model in (6.4)-(6.5) holds. Suppose also that Assumptions 2–4 and 7 hold. Suppose that in addition $\mu_\gamma = 0$ and $\mu_\pi \neq 0$. Then the just-identified iv estimator with exogenous covariate $W_i = 1$, endogenous regressor X_i and instrument $Z_i = Q_i$ is consistent for β and satisifies

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

In the second case with $\mu_\gamma \neq 0$ and $\Xi_{12} = 0$, again using all interactions as instruments does not lead to consistency if we use liml, btsls, mbtsls, jive or mjive. However, in this case we can base a consistent estimator on a strategy where we can 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:

$$Y_i = X_i\beta + Q_i\delta_0 + \sum_{i=1}^{K_N} W_{ik}\delta_k + \epsilon_i \tag{6.6a}$$

$$X_i = Q_i\pi_{22,0} + \sum_{k=1}^{K_N} W_{ik}\pi_{22,k} + \sum_{k=1}^{K_N-1} \pi_{12,k}Q_iW_{ik} + \nu_i \tag{6.6b}$$

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

Theorem 6. (Interactions)
Suppose that the model (6.4)–(6.5) *holds. Suppose also that Assumptions 2–4 and 7 hold and that* $\Xi_{12} = 0$. *Then the mbtsls and mjive estimators based on the model* (6.6) *that treats* Q_i *as an exogenous regressor are consistent for* β. *Moreoever, under those assumptions:*

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

and

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

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

7 An Application

We apply some of the methods to a subset of the Angrist-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.

We look at six estimators, the five studied in this paper and tsls. We consider three sets of instruments. First, a single binary instrument, an indicator for being born in the fourth quarter. No exogenous covariates beyond the intercept. Second, we interact the qob dummy with state of year times year of birth dummies, for a total of 500 instruments, and 500 exogenous regressors. Finally, 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.

Table 1 presents some estimates.

8 Conclusion

Appendices

We first define some additional notation. Write the reduced-form based on Equations (2.2) and (2.3) 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 $V_i = (\epsilon_i + \nu_i\beta, \nu_i)'$. Denote the upper $K_N \times 2$ submatrix of the matrix of reduced-form coefficients by $\Pi_1 = (\pi_{11}, \pi_{12})$. Let:

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

Let $\Omega = \operatorname{cov}(V_i)$ 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 V denote the $N \times 2$ matrix of reduced-form errors with ith row equal to V_i'. Let $\overline{\mathbf{Y}} = (\mathbf{Y}, \mathbf{X})$ denote the $N \times 2$ matrix of endogenous regressors, and let $\overline{\mathbf{Y}}_\perp = (\mathbf{Y}_\perp, \mathbf{X}_\perp)$ denote the matrix of endogenous variables after regressing them on $\mathbf{W}$.

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 $\mathbf{S}$.

Let $\iota_N \in \mathbb{R}^N$ be an N-vector of ones.

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 $U = (u_1, \dots, u_N)'$, 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 $d_C = \operatorname{diag}(C)$. Then:*

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

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

If the distribution of u_i is Normal, the second and third lines of the variance expression equals zero.

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

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

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:*

$$\sqrt{N}\,(Qa/N - \mathbb{E}Qa/N) \Rightarrow \mathcal{N}\,(0, V)$$

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 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 $b \in \mathbb{R}^S$:

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

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

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

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:

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

Pulling together the results yields:

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

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 resut. □

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 $a \in \mathbb{R}^S$*

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

Proof. By definition of a non-central Wishart distribution, we can decompose $X_N = (U+M)'(U+M)$, where $U = (u_1, \dots, u_{J_N})'$, $u_j \sim N(0, \Omega)$ 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:

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

which yields the result. □

Lemma A.3. *Let $\overline{\mathbf{Y}}_\perp = (\mathbf{Y}_\perp, \mathbf{X}_\perp)$. Then under Assumptions 2, 3 (i), 4 and 5:*

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

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

$$\overline{\mathbf{Y}}'_\perp \mathbf{H}_{\overline{\mathbf{Z}}} \overline{\mathbf{Y}}_\perp/N \xrightarrow{p} \Omega \qquad \mathbf{H}_{\overline{\mathbf{Z}}} = (\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}\mathbf{M}_{\overline{\mathbf{Z}}} \tag{A.5c}$$

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 $V'\mathbf{P}_{\mathbf{Z}_\perp}V/N$. By Lemma A.1 (i):

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

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

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

Combining Equations (A.7) and (A.8) with Assumption 4 yields :

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

By similar arguments:

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

Next, by Assumption 3 (i), $\mathbb{E}[\Pi_1'\mathbf{Z}_\perp'V/n \mid \mathbf{Z}_\perp] = 0$, so that:

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

where the last equality follows by Assumption 5. Consequently:

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

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

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

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

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

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

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

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 2. Moreover, $\sum_i (\mathbf{H}_{\overline{\mathbf{Z}}})_{ii}^2 = \sum_i 1^2 = N$. Hence, for any $a \in \mathbb{R}^{G+1}$:

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

Therefore, by Chebyshev's inequality:

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

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 $\hat{k} \xrightarrow{p} k$ st $k < \frac{1-\alpha_L}{1-\alpha_L-\alpha_K} + \frac{\Lambda_{22}/\Sigma_{22}}{1-\alpha_L-\alpha_K}$. Then under Assumptions 2, 3 (i), 4 and 5:*

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

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:

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

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

$$\left((1-\hat{k})\mathbf{X}'_{\perp}\mathbf{X}_{\perp}/N + \hat{k}\mathbf{X}'_{\perp}\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{X}_{\perp}/N\right)^{-1} = (\Lambda_{22} + (1-\alpha_L-(1-\alpha_K-\alpha_L)k)\Sigma_{22})^{-1} + o_p(1) \tag{A.14}$$

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

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

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

$$\begin{aligned}\hat{\beta}_{\hat{k}} &= ((1-\hat{k})\mathbf{X}'_{\perp}\mathbf{X}_{\perp}/n + \hat{k}\mathbf{X}'_{\perp}\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{X}_{\perp}/n)^{-1}((1-\hat{k})\mathbf{X}'_{\perp}\mathbf{Y}_{\perp}/n + \hat{k}x'_{\perp}\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{Y}_{\perp}) \\ &= \beta + \frac{\Lambda_{12} + ((1-k)(1-\alpha_L)+\alpha_K k)\Sigma_{12}}{\Lambda_{22} + ((1-k)(1-\alpha_L)+\alpha_K k)\Sigma_{22}} + o_p(1) \quad \square\end{aligned}$$

B Proofs of Propositions

Proof of Theorem 1. The results are a special case of Theorem 3 with $\Lambda_{12} = 0$ and $\Lambda_{22} = 0$. □

Proof of Theorem 3. 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. $\hat{k}_{\text{liml}}$ is given by the minimum of the liml objective function:

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

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 3 (i) guaranees that the denominator is non-zero for any value of ϕ. The minimum of $Q(\phi)$ is achieved at:

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

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. Since 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. Since $\overline{\mathbf{Y}}'_{\perp}M_{\mathbf{Z}_{\perp}}\overline{\mathbf{Y}}_{\perp}/n\xrightarrow{p}T_{\perp}>0$, $\phi'\overline{\mathbf{Y}}'_{\perp}M_{\mathbf{Z}_{\perp}}\overline{\mathbf{Y}}_{\perp}\phi/N>0$ 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. Define

$$\mathbf{H}_{\overline{\mathbf{Z}}} = (\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}\mathbf{M}_{\overline{\mathbf{Z}}} = \mathbf{M}_{\mathbf{W}}\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}\mathbf{M}_{\mathbf{W}}$$

Then jive can be written as:

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

mjive can be written as:

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

Applying Lemma A.3, we get:

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

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

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

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

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

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-oder condition is given by:

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

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:

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

Since the limit of $\hat{g}'_N(\hat{\beta})$ doesn't 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 $\Lambda_{11} = \Lambda_{12} = 0$ by Assumption 1. 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:

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

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 a (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:

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

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 $C = (\mathbf{M_W} - \hat{k}_{\text{mbtsls}}\mathbf{M}_{\overline{\mathbf{Z}}})$. 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:

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

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

Proof of Part (iii) Write the estimator as:

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

where

$$\mathbf{H}_{\overline{\mathbf{Z}}} = \mathbf{M_W}\mathbf{M}_{\mathbf{Z}_\perp}(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1} = \mathbf{M}_{\overline{\mathbf{Z}}}(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}$$

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 a (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 $C = \mathbf{M_W} - (1 - L_N/N)\mathbf{H}_{\overline{\mathbf{Z}}}$. Since $\text{tr}(\mathbf{H}_{\overline{\mathbf{Z}}}(\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1}) = \text{tr}((\mathbf{I} - \mathbf{D}_{\overline{\mathbf{Z}}})^{-1})$, 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 2:

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

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

Proof of Theorem 5. Since $\tilde{\gamma}_k = \sqrt{q(1-q)}\gamma_k$, we can write the estimator as:

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

By law of large numbers, we have:

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

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

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

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

Proof of Theorem 6. Denote the matrices of instruments and exogenous regressors in the model (6.6) 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 $\bar{\nu}_k = \frac{K_N}{N}\sum_{i\colon k(i)=k}\nu_i$ denote group averages, let $\bar{\nu}_{1,k} = \frac{K_N}{qN}\sum_{i\colon Q_i=1,k(i)=k}\nu_i$ denote group averages for individuals with $Q_i = 1$, and let $\bar{\nu}_{0,k} = \frac{K_N}{(1-q)N}\sum_{i\colon Q_i=0,k(i)=k}\nu_i$ denote group averages for individuals with $Q_i = 0$. Define:

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

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

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:

$$\frac{1}{K_N}\sum_k \hat{\Sigma}_{22,k} \xrightarrow{p} (1-\alpha_K)\Sigma_{22} \qquad \frac{1}{K_N}\sum_k \nu_{10,k}^2 \xrightarrow{p} \alpha_K\Sigma_{22} \tag{A.11a}$$

$$\frac{1}{K_N}\sum_k (s_k^{\pi_{12},\nu})^2 \xrightarrow{p} \Xi_{22} + \alpha_K\Sigma_{22} \qquad \frac{1}{K_N}\sum_k s_k^{\pi_{12},\nu} \xrightarrow{p} 0 \tag{A.11b}$$

Hence:

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

The nominator can be written as:

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

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 $\mathcal{F}_k=\sigma(\gamma_k,\pi_{12,k},\{\epsilon_i\colon k(i)=k\},\{\nu_i\colon k(i)=k\})$.

The next step is to show that:

$$\frac{\sqrt{N}}{K_N}\sum_k D_{k,\hat{k}_{\text{mbtsls}}}\Rightarrow\mathcal{N}\left(0,\Xi_{11}\Xi_{22}/\alpha_K+\Xi_{11}\Sigma_{22}+\Xi_{22}\Sigma_{11}+\frac{(1-\alpha_K)\alpha_K}{(1-2\alpha_K)}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right)\right) \tag{A.13}$$

by applying the martingale central limit theorem. The claim of the theorem for mbtsls will then follow by comining (A.13) with (A.12). To show (A.13), we first need to check that:

$$\frac{1}{K_N}\sum_{k=1}^{K_N}\mathbb{E}[D_{k,\hat{k}_{\text{mbtsls}}}^2\mid\mathcal{F}_{N,k-1}]\xrightarrow{p}\Xi_{11}\Xi_{22}+\alpha_K(\Xi_{11}\Sigma_{22}+\Xi_{22}\Sigma_{11})+\frac{(1-\alpha_K)\alpha_K^2}{(1-2\alpha_K)}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right) \tag{A.14}$$

Expanding the left-hand side yields:

$$\begin{aligned}\mathbb{E}[D_{k,\hat{k}_{\text{mbtsls}}}^2\mid\mathcal{F}_{N,k-1}]&=\Xi_{11}\Xi_{22}+\frac{K_N}{N}\Xi_{11}\Sigma_{22}+\frac{K_N}{N}\Xi_{22}\Sigma_{11}+2\frac{K_N}{N}\Sigma_{12}\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k}s_l^{\pi_{12},\nu}s_m^{\gamma,\epsilon}\\&+(\Xi_{11}+\frac{K_N}{N}\Sigma_{11})\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k}s_l^{\pi_{12},\nu}s_m^{\pi_{12},\nu}+(\Xi_{22}+\frac{K_N}{N}\Sigma_{22})\frac{1}{K_N^2}\sum_{l<k}\sum_{m<k}s_l^{\gamma,\epsilon}s_m^{\gamma,\epsilon}\\&+2\hat{k}_{\text{mbtsls}}(1-\hat{k}_{\text{mbtsls}})\mathbb{E}\epsilon_{10,k}\nu_{10,k}\hat{\Sigma}_{12,k}+\hat{k}_{\text{mbtsls}}^2(K_N/N)^2(\Sigma_{11}\Sigma_{22}+2\Sigma_{12}^2)+(1-\hat{k}_{\text{mbtsls}})^2\mathbb{E}\hat{\Sigma}_{12,k}^2 xpan\end{aligned} \tag{A.15}$$

where the expecations in the last line evaluate as:

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

We can therefore write:

$$\frac{1}{K_N}\sum_{k=1}^{K_N}\mathbb{E}[D^2_{k,\hat{k}_{\text{mbtsls}}}\mid \mathcal{F}_{N,k-1}] = \Xi_{11}\Xi_{22}+\frac{K_N}{N}\Xi_{11}\Sigma_{22}$$

$$+\frac{K_N}{N}\Xi_{22}\Sigma_{11}+\frac{(1-K_N/N)(K_N/N)^2}{(1-2(K_N/N))^2}\left(\Sigma_{11}\Sigma_{22}+\Sigma_{12}^2\right)+2\frac{K_N}{N}\Sigma_{12}\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k}s_l^{\pi_{12},\nu}s_m^{\gamma,\epsilon}$$

$$+(\Xi_{11}+\frac{K_N}{N}\Sigma_{11})\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k}s_l^{\pi_{12},\nu}s_m^{\pi_{12},\nu}+(\Xi_{22}+\frac{K_N}{N}\Sigma_{22})\frac{1}{K_N^3}\sum_k\sum_{l<k}\sum_{m<k}s_l^{\gamma,\epsilon}s_m^{\gamma,\epsilon}$$

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

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

Threfore, the last three terms are $o_p(1)$, which proves (A.14). One can also show that $K_N^{-2}\sum_k\mathbb{E}D^4_{k,\hat{k}_{\text{mbtsls}}}\to 0$, which implies (A.13). Next consider the mjive estimator. Let

$$\hat{\Sigma}^1_{12,k}=\frac{K_N}{N}\sum_{i:\,k(i)=k}Q_i v_i\epsilon_i\qquad \Sigma^0_{12,k}=\frac{K_N}{N}\sum_{i:\,k(i)=k}(1-Q_i)v_i\epsilon_i$$

$$t_k^{12}=\frac{q}{q-K_N/N}\left(\hat{\Sigma}^1_{12,k}-q\bar{\nu}_{1,k}\bar{\epsilon}_{1,k}\right)+\frac{1-q}{1-q-K_N/N}\left(\hat{\Sigma}^0_{12,k}-(1-q)\bar{\nu}_{0,k}\bar{\epsilon}_{0,k}\right)$$

Then we can write the mjive estimator as:

$$\hat{\beta}_{\text{mjive}}=\frac{\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{Y}-\left(1-\frac{K_N}{N}\right)\mathbf{X}'\mathbf{M}_{\overline{\mathbf{Z}}}\left(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\mathbf{Y}}{\mathbf{X}'\mathbf{M}_{\tilde{\mathbf{W}}}\mathbf{X}-\left(1-\frac{K_N}{N}\right)\mathbf{X}'\mathbf{M}_{\overline{\mathbf{Z}}}\left(\mathbf{I}-\mathbf{D}_{\overline{\mathbf{Z}}}\right)^{-1}\mathbf{X}}$$

$$=\beta+\frac{\frac{1}{K_N}\sum_k\left(s_k^{\gamma,\epsilon}s_k^{\pi_{12},\nu}+\hat{\Sigma}_{12,k}-\epsilon_{10,k}\nu_{10,k}-(1-\frac{K_N}{N})t_k^{12}\right)-K_N^{-2}\sum_k\sum_l s_k^{\gamma,\epsilon}s_l^{\pi,\nu}}{\frac{1}{K_N}\sum_k\left((s_k^{\pi_{12},\nu})^2+\hat{\Sigma}_{22,k}-\nu^2_{10,k}-(1-\frac{K_N}{N})t_k^{22}\right)-K_N^{-2}\sum_k\sum_l s_k^{\pi,\nu}s_l^{\pi,\nu}}$$

Using (A.11) and the fact that by the weak law of large numbers $t_k^{22}\xrightarrow{p}\Sigma_{22}$, we get:

$$\left(\frac{1}{K_N}\sum_k\left((s_k^{\pi_{12},\nu})^2+\hat{\Sigma}_{22,k}-\nu^2_{10,k}-(1-\frac{K_N}{N})t_k^{22}\right)-K_N^{-2}\sum_k\sum_l s_k^{\pi,\nu}s_l^{\pi,\nu}\right)^{-1}=\Xi_{22}+o_p(1) \tag{A.16}$$

We can rewrite the nominator as:

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

where:

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

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 $\mathcal{F}_k = \sigma(\gamma_k, \pi_{12,k}, \{\epsilon_i \colon k(i) = k\}, \{\nu_i \colon k(i) = k\})$. To prove the claim of the theorem for mjive, it therefore remains to check that:

$$\frac{1}{K_N}\sum_{k=1}^{K_N} \mathbb{E}[\tilde{D}_k^2 \mid \mathcal{F}_{N,k-1}] \xrightarrow{p} \\ \Xi_{11}\Xi_{22} + \alpha_K(\Xi_{11}\Sigma_{22} + \Xi_{22}\Sigma_{11}) + \alpha_K(1-\alpha_K)((1-\alpha_K)\tau - 1)\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right) \quad \text{(A.17)}$$

and that:

$$K_N^{-2}\sum_k \mathbb{E}\tilde{D}_k^4 \to 0 \quad \text{(A.18)}$$

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.17). Using the expansion in (A.15), we get that:

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

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.17). It can also be shown that (A.18) holds, which proves the result. □

References

Ackerberg, D. and P. Devereux (2009): "Improved JIVE estimators for overidentified linear models with and without heteroskedasticity?" *Review of Economics and Statistics*, 91 (2), 351–362.

Anatolyev, S. and N. Gospodinov (2011): "Specification testing in models with many instruments" *Econometric Theory*, forthcoming.

Anderson, T., (1976), "Estimation of Linear Functional Relationships: Approximate Distributions and Connections with Simultaneous Equations in Econometrics," *The Journal of the Royal Statistical Sociery*, Series B, 38, 1-36.

Anderson, T., and H. Rubin (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570-582.

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

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

Andrews, D., and J. Stock (2007), "Testing with many weak instruments," *Journal of Econometrics*, 138, 24-46.

Angrist, J., G. Imbens, and A. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14 (1), 57–67.

Angrist, J. and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106 (4), 979–1014.

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. (1957): "A Generalized Classical Method of Linear Estimation in a Structural Equation," *Econometrica*, 25(1), 77–83.

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

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

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

Caner, M. (2007): "Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics," Working paper, North Carolina State University.

Cattaneo, M., M. Jansson, and W. Newey, (2010): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," Working paper, MIT

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

Chamberlain, G., and M. Moreira (2009): "Decision Theory Applied to a Linear Panel Data Model" *Econometrica*, 77(1), 107-133.

Chao, J., and N. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73 (5), 1673–1692.

CHAO, J., N. SWANSON, J. HAUSMAN, W. NEWEY, AND T. WOUTERSEN (2009): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," Working paper, MIT

CHIODA, L., AND M. JANSSON (2009): "Optimal Invariant Inference When the Number of Instruments Is Large," *Econometric Theory*, 25(3): 793-805.

CHETTY, R., FRIEDMAN, J., N. HILGER, E. SAEZ, D. SCHANZENBACH, AND D. YAGAN, (2011): "How does your Kindergarten Classroom affect your earnings? Evidence from Project Star," Unpublished Manuscript.

DAVIDSON, R., AND J. MACKINNON. (1993): *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

DONALD, S., AND W. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69 (5), 1161–1191

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

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

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

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.

HANSEN, C., J. HAUSMAN, AND W. NEWEY (2008): "Estimation With Many Instrumental Variables," *Jornal of Business and Economic Statistics*, 26 (4), 398–422.

HAUSMAN, J., W. NEWEY, T. WOUTERSEN, CHAO, J., AND N. SWANSON (2009): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," Working paper, MIT

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.

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

MOREIRA, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models. Econometrica" *Econometrica*, 71(4), 1027-1048.

MOREIRA, M. (2009): "A maximum likelihood method for the incidental parameter problem" *The Annals of Statistics* , 37 (6A), 36603696.

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. (1959): "The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations," *Econometrica*, 27 (4), 575–595.

NEVO, A., AND ROSEN, A. (2010): "Identification with Imperfect Instruments," *Review of Economics and Statistics*, forthcoming.

PHILLIPS, G., AND C. HALE, (1977) "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems,", *International Economic Review*, 18, 219-228.

Rothenberg, T. (1984), “Approximating the Distributions of Econometrics Estimators and Test Statistics,” in: *Handbook of Econometrics*, Vol 2, Griliches and Intrilligator (eds.) North Holland, Amsterdam.

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

Theil, H. (1961): *Economic Forecasts and Policy*, North Holland Publishing Company, Amsterdam.

Theil, H. (1971): *Principles of Econometrics*, John Wiley and Sons, New York.

Table 1: ESTIMATES FOR ANGRIST-KRUEGER DATA ($N = 162, 487$)

		standard errors			
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)			
btsls	0.097	(0.017)	0.039	(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 as 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)