Instrumental Variable Estimation with Heteroskedasticity and Many Instruments*

Jerry A. Hausman
Department of Economics
M.I.T.

Whitney K. Newey
Department of Economics
M.I.T.

Tiemen Woutersen
Department of Economics
Johns Hopkins University

John Chao
Department of Economics
University of Maryland

Norman Swanson
Department of Economics
Rutgers University

August 2006
Revised September 2009

Abstract

This paper gives a relatively simple, well behaved solution to the problem of many instruments in heteroskedastic data. Such settings are common in microeconometric applications where many instruments are used to improve efficiency and allowance for heteroskedasticity is generally important. The solution is a Fuller (1977) like estimator and standard errors that are robust to heteroskedasticity and many instruments. We show that the estimator has finite moments and high asymptotic efficiency in a range of cases. The standard errors are easy to compute, being like White's (1982), with additional terms that account for many instruments. They are consistent under standard, many instrument, and many weak instrument asymptotics. Based on a series of Monte Carlo experiments, we find that the estimators perform as well as LIML or Fuller (1977) under homoskedasticity, and have much lower bias and dispersion under heteroskedasticity, in nearly all cases considered.

JEL Classification: C12, C13, C23

Keywords: Instrumental Variables, Heteroskedasticity, Many Instruments, Jackknife

*The NSF provided financial support for this paper under Grant No. 0136869. Helpful comments were provided by A. Chesher and participants in seminars at CalTech, CEMMAP, Harvard, MIT, Pittsburgh, UC Berkeley, UCL, and USC. Capable research assistance was provided by H. Arriizumi, S. Chang, A. Kowalski, R. Lewis, and K. Menzel. K. Menzel derived the vectorized form of the variance estimator.

1 Introduction

This paper gives a relatively simple, well behaved solution to the problem of many instruments in heteroskedastic data. Such settings are common in microeconometric applications where many instruments are used to improve efficiency and allowance for heteroskedasticity is generally important. The solution is a Fuller (1977) like estimator and standard errors that are robust to heteroskedasticity and many instruments. We show that the estimator has finite moments and high asymptotic efficiency in a range of cases. The standard errors are easy to compute, being like White's (1982), with additional terms that account for many instruments.They are consistent under standard, many instrument, and many weak instrument asymptotics. They extend Becker's (1994) standard errors to the heteroskedastic case.

The estimator, that we refer to as HFUL, is based on a jackknife version of LIML (HLIM), that will be described below. Because HFUL has finite moments it does not have the large dispersion that can occur with HLIM for weak identification, an advantage analogous to that of the Fuller (1977) estimator over LIML with homoskedasticity. Hahn, Hausman, and Kuersteiner (2004) pointed out this problem for LIML and we follow them in referring to it as the "moments problem." Because of its jackknife form, HFUL is robust to heteroskedasticity and many instruments, as are jackknife instrumental variable (JIV) estimators, see Phillips and Hale (1977), Blomquist and Dahlberg (1999), Angrist, Imbens, and Krueger (1999), Ackerberg and Deveraux (2003) and Chao and Swanson (2004). However, HFUL is as efficient as LIML under many weak instruments and homoskedasticity, and so overcomes the efficiency problems for JIV noted in Davidson and MacKinnon (2006). Thus, HFUL provides a relatively efficient estimator for many instruments with heteroskedasticity that does not suffer from the moments problem.

Bekker and van der Ploeg (2005) proposed interesting consistent estimators with many dummy instrumental variables and group heteroskedasticity, but these results are restrictive. For high efficiency it is often important to use instruments that are not dummy variables. For example, linear instrumental variables can be good first approximations to optimal nonlinear instruments.. Also, disturbance variances that are constant within groups is too restrictive for most econometric applications. HFUL allows for general instrumental variables and unrestricted heteroskedasticity, as does the asymptotics given here.

Newey and Windmeijer (2009) showed that the continuously updated GMM (CUE) and other generalized empirical likelihood estimators are robust to heteroskedasticity and many weak instruments, and asymptotically efficient under that asymptotics relative to JIV. However this efficiency depends on using a heteroskedasticity consistent weighting matrix that can degrade the finite sample performance of CUE with many instruments, as shown in Monte Carlo experiments here. HFUL continues to have good properties under many instrument asymptotics, rather than just many weak instruments. The properties of CUE are likely to be poor under many instruments asymptotics due to the heteroskedasticity consistent weighting matrix. Also CUE is quite difficult to compute and tends to have large dispersion under weak identification, which HFUL does not. Thus, relative to CUE, HFUL provides a computationally simpler solution with better finite sample properties.

The need for HFUL is motivated by the inconsistency of LIML and the Fuller (1977) estimator under heteroskedasticity and many instruments. The inconsistency of LIML was pointed out by Bekker and van der Ploeg (2005) and Chao and Swanson (2004) in special cases. We give a characterization of the inconsistency here, showing the precise restriction on the heteroskedasticity that would be needed for LIML to be consistent.

The asymptotic theory we consider allows for many instruments as in Kunitomo (1980) and Bekker (1994) or many weak instruments as in Chao and Swanson (2004, 2005), Stock and Yogo (2005), Han and Phillips (2006), and Andrews and Stock (2007). The asymptotic variance estimator will be consistent for any of standard, many instrument, or many weak instrument asymptotics. Asymptotic normality is obtained via a central limit theorem that imposes weak conditions on instrumental variables, given by Chao, Swanson, Hausman, Newey, and Woutersen (2009). Although the inference methods will not be valid under the weak instrument asymptotics of Staiger and Stock (1997), we do not consider this to be very important. Hansen, Hausman, and Newey's (2008) survey of the applied literature suggests that the weak instrument approximation is not needed very often in microeconomic data, where we focus our attention.

In Section 2, the model is outlined and a practitioner's guide to the estimator is given. We give there simple formulae for HFUL and its variance estimator. Section 3 motivates HLIM and HFUL as jackknife forms of LIML and Fuller (1977) estimators and discusses some of their properties. Section 5 shows HFUL has finite moments. Monte Carlo findings are presented in Section 6. The asymptotic theory proofs are given in the Appendix and the proof of existence of moments can be found at http://econweb.umd.edu/˜chao/.

2 The Model and HFUL

The model we consider is given by

$$\begin{aligned} \underset{n\times 1}{y} &= \underset{n\times G}{X}\,\underset{G\times 1}{\delta_0} + \underset{n\times 1}{\varepsilon}, \\ X &= \Upsilon + U, \end{aligned}$$

where n is the number of observations, G is the number of right-hand side variables, Υ is a matrix of observations on the reduced form, and U is the matrix of reduced form disturbances. For our asymptotic approximations, the elements of Υ will be implicitly allowed to depend on n, although we suppress dependence of Υ on n for notational convenience. Estimation of δ_0 will be based on an $n \times K$ matrix, Z, of instrumental variable observations with $rank(Z) = K$. We will assume that Z is nonrandom and that observations (ε_i, U_i) are independent across i and have mean zero. Alternatively, we could allow Z to be random, but condition on it, as in Chao et al. (2009).

In this model some columns of X may be exogenous, with the corresponding columns of U being zero. Also, this model allows for Υ to be a linear combination of Z, i.e. $\Upsilon = Z\pi$ for some $K \times G$ matrix π. The model also permits Z to approximate the reduced form. For example, let X_i', Υ_i', and Z_i' denote the i^{th} row (observation) of X, Υ, and Z respectively. We could let $\Upsilon_i = f_0(w_i)$ be a vector of unknown functions of a vector w_i of underlying instruments, and $Z_i = (p_{1K}(w_i), ..., p_{KK}(w_i))'$ be approximating functions $p_{kK}(w)$, such as power series or splines. In this case, linear combinations of Z_i may approximate the unknown reduced form (e.g. as in Newey, 1990).

To describe HFUL, let

$$P = Z(Z'Z)^{-1}Z',$$

P_{ij} denote the ij^{th} element of P, and $\bar{X} = [y, X]$. Let

$$\tilde{\alpha} \text{ be the smallest eigenvalue of } (\bar{X}'\bar{X})^{-1}(\bar{X}'P\bar{X} - \sum_{i=1}^{n} P_{ii}\bar{X}_i\bar{X}_i').$$

Although this matrix is not symmetric it has real eigenvalues because it is a product of symmetric, positive semi-definite matrices. For a constant C let

$$\hat{\alpha} = [\tilde{\alpha} - (1 - \tilde{\alpha})C/T]/[1 - (1 - \tilde{\alpha})C/T].$$

In the Monte Carlo results given below we try different values of C and recommend $C = 1$. HFUL is given by

$$\hat{\delta} = \left(X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \hat{\alpha}X'X \right)^{-1} \left(X'Py - \sum_{i=1}^{n} P_{ii}X_iy_i - \hat{\alpha}X'y \right). \quad (2.1)$$

Thus, HFUL can be computed by finding the smallest eigenvalue of a matrix and then using this explicit formulae.

To describe the asymptotic variance estimator, let $\hat{\varepsilon}_i = y_i - X_i'\hat{\delta}$, $\hat{\gamma} = X'\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon}$, $\hat{X} = X - \hat{\varepsilon}\hat{\gamma}'$, $\dot{X} = P\hat{X}$, and $\tilde{Z} = Z(Z'Z)^{-1}$. Also let

$$\hat{H} = X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \hat{\alpha}X'X,$$

$$\hat{\Sigma} = \sum_{i=1}^{n}(\dot{X}_i\dot{X}_i' - \hat{X}_iP_{ii}\dot{X}_i' - \dot{X}_iP_{ii}\hat{X}_i')\hat{\varepsilon}_i^2 + \sum_{k=1}^{K}\sum_{\ell=1}^{K}\left(\sum_{i=1}^{n}\tilde{Z}_{ik}\tilde{Z}_{i\ell}\hat{X}_i\hat{\varepsilon}_i\right)\left(\sum_{j=1}^{n}Z_{jk}Z_{j\ell}\hat{X}_j\hat{\varepsilon}_j\right)',$$

The formula for $\hat{\Sigma}$ is vectorized in such a way that it can easily be computed even when the sample size is n is very large. The asymptotic variance estimator is

$$\hat{V} = \hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1}.$$

This asymptotic variance estimator will be consistent under standard, many instrument, and many weak instrument asymptotics.

This asymptotic variance estimator can be used to do large sample inference in the usual way under the conditions of Section 4. This is done by treating $\hat{\delta}$ as if it were normally distributed with mean δ_0 and variance $\hat{V}$. Asymptotic t-ratios $\hat{\delta}_j/\sqrt{\hat{V}_{jj}}$ will be asymptotically normal. Also, defining q_α as the $1-\alpha/2$ quantile of a $N(0,1)$ distribution, an asymptotic $1-\alpha$ confidence interval for δ_{0k} is given by $\hat{\delta}_k \pm q_\alpha \sqrt{\hat{V}_{kk}}$. More generally, a confidence interval for a linear combination $c'\delta$ can be formed as $c'\hat{\delta} \pm q_\alpha \sqrt{c'\hat{V}c}$. We find in the Monte Carlo results that these asymptotic confidence intervals are very accurate in a range of finite sample settings.

3 Consistency with Many Instruments and Heteroskedasticity

In this Section we explain the HFUL estimator, why it has moments, is robust to heteroskedasticity and many instruments, and why it has high efficiency under homoskedasticity. We also compare it with other estimators and briefly discuss some of their properties. To do so it is helpful to consider each estimator as a minimizer of an objective function. As usual, the limit of the minimizer will be the minimizer of the limit under appropriate regularity conditions, so estimator consistency can be analyzed using the limit of the objective function. This amounts to modern version of method of moments interpetations of consistency, that has now become common in econometrics; Amemiya (1973, 1984), Newey and McFadden (1994).

To motivate HFUL it is helpful to begin with two-stage least squares (2SLS). The 2SLS estimator minimizes

$$\hat{Q}_{2SLS}(\delta) = (y - X\delta)'P(y - X\delta)/n.$$

The limit of this function will equal the limit of its expectation under general conditions.

With independent observations

$$E[\hat{Q}_{2SLS}(\delta)] = (\delta-\delta_0)'A_n(\delta-\delta_0)+\sum_{i=1}^{n}P_{ii}E[(y_i-X_i'\delta)^2]/n,$$

$$A_n = \Upsilon'P\Upsilon/n-\sum_{i=1}^{n}P_{ii}\Upsilon_i\Upsilon_i'/n.$$

The matrix A_n will be positive definite under conditions given below, so that the first term $(\delta-\delta_0)'A_n(\delta-\delta_0)$ will be minimized at δ_0. The second term $\sum_{i=1}^{n}P_{ii}E[(y_i-X_i'\delta)^2]/n$ is an expected squared residual that will not be minimized at δ_0 due to endogeneity. With many (weak) instruments P_{ii} does not shrink to zero, so that the second term does not vanish asymptotically (relative to the first). Hence, with many (weak) instruments, 2SLS is not consistent, even under homoskedasticity, as pointed out by Bekker (1994). This objective function calculation for 2SLS is also given in Han and Phillips (2006), though the following analysis is not.

A way to modify the objective function so it gives a consistent estimator is to remove the term whose expectation is not minized at δ_0. This leads to an objective function of the form

$$\hat{Q}_{JIV}(\delta)=\sum_{i\neq j}(y-X\delta)'P_{ij}(y-X\delta)/n.$$

The expected value of this objective function is

$$E[\hat{Q}_{JIV}(\delta)]=(\delta-\delta_0)'A_n(\delta-\delta_0),$$

which is minimized at $\delta=\delta_0$. Thus, the estimator minimizing $\hat{Q}_{JIV}(\delta)$ should be consistent. Solving the first order conditions gives

$$\hat{\delta}_{JIV}=\left(\sum_{i\neq j}X_i'P_{ij}X_j\right)^{-1}\sum_{i\neq j}X_i'P_{ij}y_j.$$

This is the JIVE2 estimator of Angrist, Imbens, and Krueger (1999). Since the objective function for $\hat{\delta}_{JIV}$ has expectation minimized at δ_0 we expect that $\hat{\delta}_{JIV}$ is consistent, as has already been shown by Ackerberg and Deveraux (2003) and Chao and Swanson (2004). Other JIV estimators have also been shown to be consistent in these papers.

So far we have only used the objective function framework to describe previously known consistency results. We now use it to motivate the form of HFUL (and HLIM).

A problem with JIV estimators, pointed out by Davidson and MacKinnon (2006), is that they can have low efficiency relative to LIML under homoskedasticity. This problem can be avoided by using a jackknife version of LIML. The LIML objective function is

$$\hat{Q}_{LIML}(\delta) = \frac{(y - X'\delta)'P(y - X'\delta)}{(y - X\delta)'(y - X\delta)}.$$

The numerator of $\hat{Q}_{LIML}(\delta)$ is $n\hat{Q}_{2SLS}(\delta)$. If we replace this numerator with $n\hat{Q}_{JIV}(\delta)$ we obtain

$$\hat{Q}_{HLIM}(\delta) = \frac{\sum_{i \neq j}(y - X\delta)'P_{ij}(y - X\delta)}{(y - X\delta)'(y - X\delta)}.$$

The minimizer of this objective function is the HLIM estimator that we denote by $\tilde{\delta}$. This estimator is consistent with many instruments and heteroskedasticity. It is also as efficient asymptotically and performs as in our Monte Carlo resuts as LIML under homoskedasticity, thus overcoming the Davidson and MacKinnon (2006) objection to JIV.

The use of the JIV objective function in the numerator makes this estimator consistent with heteroskedasticity and many instruments. In large samples the HLIM objective function will be close to

$$\frac{E[n\hat{Q}_{JIV}(\delta)]}{E[(y - X\delta)'(y - X\delta)]} = \frac{(\delta - \delta_0)'A_n(\delta - \delta_0)}{E[(y - X\delta)'(y - X\delta)]}.$$

This function is minimized at $\delta = \delta_0$ even with heteroskedasticity and many instruments, leading to consistency of HLIM.

Computation of HLIM is straightforward. For $\bar{X} = [y, X]$, the minimized objective function $\tilde{\alpha} = \hat{Q}_{HLIM}(\tilde{\delta})$ is the smallest eigenvalue of $(\bar{X}'\bar{X})^{-1}(\bar{X}'P\bar{X} - \sum_{i=1}^{n} P_{ii}\bar{X}_i\bar{X}_i')$. Solving the first order conditions gives

$$\tilde{\delta} = \left(X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \tilde{\alpha}X'X\right)^{-1}\left(X'Py - \sum_{i=1}^{n} P_{ii}X_iy_i - \tilde{\alpha}X'y\right).$$

The formula for HLIM is exactly analogous to that of LIML where the own observation terms have been removed from the double sums involving P. Also, HLIM is invariant to normalization, similarly to LIML, although HFUL is not. The vector $\tilde{d} = (1, -\tilde{\delta}')'$ solves

$$\min_{d:d_1=1} \frac{d'\left(\bar{X}'P\bar{X} - \sum_{i=1}^n P_{ii}\bar{X}_i\bar{X}_i'\right)d}{d'\bar{X}'\bar{X}d}.$$

Because of the ratio form of the objective function, another normalization, such as imposing that another d is equal to 1, would produce the same estimator, up to the normalization.

Like LIML, the HLIM estimator suffers from the moments problem, having large dispersion with weak instruments, as shown in the Monte Carlo results below. Hahn, Hausman, Kuersteiner (2005) suggested the Fuller (1977) estimator as a solution to this problem for LIML. We suggest the HFUL as a solution to this potential problem with HLIM. HFUL is obtained exactly analogously to Fuller (1977) by replacing the eigenvalue $\tilde{\alpha}$ in the HLIM estimator with $\hat{\alpha} = [\tilde{\alpha} - (1-\tilde{\alpha})C/T]/[1-(1-\tilde{\alpha})C/T]$, giving the HFUL estimator of equation (2.1). We show that this estimator does have moments and low dispersion with weak instruments, thus providing a solution to the moments problem.

HFUL, HLIM, and JIV are members of a class of estimators of the form

$$\bar{\delta} = \left(X'PX - \sum_{i=1}^n P_{ii}X_iX_i' - \bar{\alpha}X'X\right)^{-1}\left(X'Py - \sum_{i=1}^n P_{ii}X_iy_i - \bar{\alpha}X'y\right).$$

This might be thought of a type of k-class estimator that is robust to heteroskedasticity and manyh instruments. HFUL takes this form as in equation (2.1), HLIM does with $\bar{\alpha} = \tilde{\alpha}$, and JIV with $\bar{\alpha} = 0$.

HLIM can also be interpreted as a jackknife version of the continuously updated GMM estmator and as an optimal linear combination of forward and reverse JIV estimators, analogously to Hahn and Hausman's (2002) interpretation of LIML as an optimal linear combination of forward and reverse bias corrected estimators. For brevity we do not give these interpretations here.

HFUL is motivated by the inconsistency of LIML and Fuller (1977) with many instruments and heteroskedasticity. To give precise conditions for LIML inconsistency, note that in large samples the LIML objective function will be close to

$$\frac{E[\hat{Q}_{2SLS}(\delta)]}{E[(y-X\delta)'(y-X\delta)]} = \frac{(\delta-\delta_0)'A_n(\delta-\delta_0)}{E[(y-X\delta)'(y-X\delta)]} + \frac{\sum_{i=1}^n P_{ii}E[(y_i - X_i'\delta)^2]}{E[(y-X\delta)'(y-X\delta)]}.$$

The first term following the equality will be minimized at δ_0. The second term may not have a critical value at δ_0, and so the objective function will not be minimized at δ_0. To see this let $\sigma_i^2 = E[\varepsilon_i^2]$, $\gamma_i = E[X_i\varepsilon_i]/\sigma_i^2$, and $\bar{\gamma} = \sum_{i=1}^n E[X_i\varepsilon_i]/\sum_{i=1}^n \sigma_i^2 = \sum_i \gamma_i\sigma_i^2/\sum_i \sigma_i^2$. Then

$$\begin{aligned} \frac{\partial}{\partial\delta}\frac{\sum_{i=1}^n P_{ii}E[(y_i - X_i\delta)^2]}{\sum_{i=1}^n E[(y_i - X_i\delta)^2]}\bigg|_{\delta=\delta_0} &= \frac{-2}{\sum_{i=1}^n \sigma_i^2}\left[\sum_{i=1}^n P_{ii}E[X_i\varepsilon_i] - \sum_{i=1}^n P_{ii}\sigma_i^2\bar{\gamma}\right] \\ &= \frac{-2\sum_{i=1}^n P_{ii}(\gamma_i - \bar{\gamma})\sigma_i^2}{\sum_{i=1}^n \sigma_i^2} = -2\widehat{Cov_{\sigma^2}(P_{ii}, \gamma_i)}, \end{aligned}$$

where $\widehat{Cov_{\sigma^2}(P_{ii}, \gamma_i)}$ is the covariance between P_{ii} and γ_i, for the distribution with probability weight $\sigma_i^2/\sum_{i=1}^n \sigma_i^2$ for the i^{th} observation. When

$$\lim_{n\longrightarrow\infty} \widehat{Cov_{\sigma^2}(P_{ii}, \gamma_i)} \neq 0,$$

the LIML objective function will not have zero derivative at δ_0 asymptotically so that it is not minimized at δ_0. Bekker and van der Ploeg (2005) and Chao and Swanson (2004) pointed out that LIML can be inconsistent with heteroskedasticity; the contribution here is to give the exact condition $\widehat{Cov_{\sigma^2}(P_{ii}, \gamma_i)} = 0$ for consistency of LIML.

Note that $\widehat{Cov_{\sigma^2}(P_{ii}, \gamma_i)} = 0$ when either γ_i or P_{ii} does not depend on i. Thus, it is variation in $\gamma_i = E[X_i\varepsilon_i]/\sigma_i^2$, the coefficients from the projection of X_i on ε_i, that leads to inconsistency of LIML, and not just any heteroskedasticity. Also, the case where P_{ii} is constant occurs with dummy instruments and equal group sizes. It was pointed out by Bekker and van der Ploeg (2005) that LIML is consistent in this case, under heteroskedasticity. Indeed, when P_{ii} is constant,

$$\hat{Q}_{LIML}(\delta) = \hat{Q}_{HLIM}(\delta) + \frac{\sum_i P_{ii}(y_i - X_i'\delta)^2}{(y - X\delta)'(y - X\delta)} = \hat{Q}_{HLIM}(\delta) + P_{11},$$

so that the LIML objective function equals the HLIM objective function plus a constant, and hence HLIM equals LIML.

Bekker and van der Ploeg (2005, BP) proposed estimators that are consistent with dummy instruments and group heteroskedasticity. To explain why these estimators do not apply with general instruments and heteroskedasticity we briefly describe their MM estimator. BP assume that the instrumental variables are dummy varianbes with exactly one instrumental variable being equal to one for each observation. The "groups" then correspond to instrumental variables with $\mathcal{I}_j = \{i : Z_{ij} = 1\}$ indexing the j^{th} "group" and $n_j = \#\mathcal{I}_j$ be the number of observations in a group, $j = 1, ..., K$. Let $\varepsilon_i(\delta) = y_i - X_i'\delta$ and $\bar{\varepsilon}_j(\delta) = \sum_{i\in\mathcal{I}_j} \varepsilon_i(\delta)/n_j$. The objective function for the MM estimator is

$$\frac{(y - X\delta)'P(y - X\delta)}{\sum_{j=1}^{K} \frac{1}{n_j - 1} \sum_{i\in\mathcal{I}_j} [\varepsilon_i(\delta) - \bar{\varepsilon}_j(\delta)]^2}.$$

The denominator of this function is the sum of estimated group variances and so the MM estimator clearly depends on the instrumental variables being dummies. Also the denominator corresponds to constant within group variances as is imposed in the BP asymptotics. It may be interesting to consider the properties of this estimator with general heteroskedasticity (and dummy instruments), but this is beyond the scope of this paper, and in any case HFUL has good properties for general instruments.

4 Asymptotic Theory

Theoretical justification for the estimators is provided by asymptotic theory where the number of instruments grows with the sample size. Some regularity conditions are important for this theory. Let $Z_i', \varepsilon_i, U_i'$, and Υ_i' denote the i^{th} row of Z, ε, U, and Υ respectively. Here, we will consider the case where Z is constant, which can be viewed as conditioning on Z (see e.g. Chao et al. 2009).

Assumption 1: Z includes among its columns a vector of ones, $rank(Z) = K$, and there is a constant C such that $P_{ii} \leq C < 1$, $(i = 1, ..., n)$, $K \longrightarrow \infty$.

The restriction that $rank(Z) = K$ is a normalization that requires excluding redundant columns from Z. It can be verified in particular cases. For instance, when w_i is a continuously distributed scalar, $Z_i = p^K(w_i)$, and $p_{kK}(w) = w^{k-1}$, it can be shown that $Z'Z$ is nonsingular with probability one for $K < n$.[1] The condition $P_{ii} \leq C < 1$ implies

[1]The observations $w_1, ..., w_n$ are distinct with probability one and therefore, by $K < n$, cannot all be roots of a K^{th} degree polynomial. It follows that for any nonzero a there must be some i with $a'Z_i = a'p^K(w_i) \neq 0$, implying that $a'Z'Za > 0$.

that $K/n \leq C$, because $K/n = \sum_{i=1}^{n} P_{ii}/n \leq C$.

Assumption 2: $\Upsilon_i = S_n z_i/\sqrt{n}$ where $S_n = \tilde{S}\,\mathrm{diag}\,(\mu_{1n}, ..., \mu_{Gn})$ and $\tilde{S}$ is nonsingular. Also, for each j either $\mu_{jn} = \sqrt{n}$ or $\mu_{jn}/\sqrt{n} \longrightarrow 0$, $\mu_n = \min_{1 \leq j \leq G} \mu_{jn} \longrightarrow \infty$, and $\sqrt{K_n}/\mu_n^2 \longrightarrow 0$. Also, there is $C > 0$ such that $\|\sum_{i=1}^{n} z_i z_i'/n\| \leq C$ and $\lambda_{\min}\left(\sum_{i=1}^{n} z_i z_i'/n\right) \geq C$, for n sufficiently large.

The S_n matrix in Assumption 2 determines the convergence rate of the estimators. We will show that $S_n'(\hat{\delta} - \delta_0)$ and $S_n'(\tilde{\delta} - \delta_0)$ are asymptotically normal under conditions given here. The S_n matrix has a complicated form that seems necessary to cover important cases, as discussed below. However, one need not even know the form of S_n to perform inference. Under the conditions given here the standard errors we have provided can be used to do large sample inference in the usual way without knowing the form of S_n, as shown in Theorem 5 below.

Assumption 2 and the S_n matrix are designed to accomodate a linear model where included instruments (e.g. a constant) have fixed reduced form coefficients and excluded instruments have coefficients that can shrink as the sample size grows. Such a model has a linear structural equation of the form

$$y_i = Z_i^{1\prime}\delta_0^1 + X_i^{2\prime}\delta_0^2 + \varepsilon_i$$

where Z_i^1 is a $G_1 \times 1$ vector of included instruments (e.g. a constant) and X_i^2 is a $G_2 \times 1$ vector of endogenous variables with $G_1 + G_2 = G$. Let the reduced form be partitioned conformably with δ, $\Upsilon_i = (\Upsilon_i^{1\prime}, \Upsilon_i^{2\prime})'$ and $U_i = (U_i^{1\prime}, U_i^{2\prime})'$. The corresponding reduced form for the included instruments is $Z_i^1 = \Upsilon_i^1$ with $U_i^1 = 0$. Suppose that the reduced form for X_i^2 is

$$X_{i2} = \Upsilon_i^2 + U_i^2, \Upsilon_i^2 = \pi^1 Z_i^1 + \left(\mu_n/\sqrt{n}\right) z_i^2,$$

where z_i^2 are instruments that are excluded from the structural equation and $\mu_n \leq \sqrt{n}$. Here any reduced form coefficients in z_i^2 are subsumed in z_i^2. Let $z_i = (Z_i^{1\prime}, z_i^{2\prime})'$ and impose Assumption 2, so that the second moment matrix of z_i is bounded and bounded away from zero. This is a normalization that makes the strength of identification of δ^2

be determined by μ_n. Indeed, $1/\mu_n$ will be the convergence rate for estimators of δ^2. Assumption 2 also allows for a diagnonal matrix in place of $(\mu_n/\sqrt{n})\, I$, which would correspond to different convergence rates for estimators of different components of δ^2. In this example we maintain the scalar matrix form of the cofficients of z_i^2 for simplicity.

For this model $\Upsilon_i = S_n z_i$ with

$$S_n = \begin{bmatrix} I & 0 \\ \pi^1 & \mu_n/\sqrt{n} \end{bmatrix} \sqrt{n} = \begin{bmatrix} I & 0 \\ \pi^1 & I \end{bmatrix} diag(\sqrt{n}, ..., \sqrt{n}, \mu_n, ..., \mu_n).$$

This S_n has the form given in Assumption 2 with

$$\tilde{S} = \begin{bmatrix} I & 0 \\ \pi^1 & I \end{bmatrix}, \mu_{jn} = \sqrt{n}, 1 \leq j \leq G_1, \mu_{jn} = \mu_n, G_1 + 1 \leq j \leq G.$$

This complicated form of S_n is needed to accomodate fixed reduced form coefficients for included instruments and coefficients for excluded instruments that depend on n. We have been unable to simplify S_n while maintaining the generality needed for these important cases.

In this example $\mu_n \longrightarrow \infty$ must hold for Assumption 2 to be satisfied. This implies that δ^2 is asymptotically identfied. If μ_n were bounded we would be in a weak instrument setting similar Staiger and Stock (1997), where δ^2 is not asymptotically identified and limiting distributions of estimators are different than those given here.

The excluded instruments z_i^2 may be an unknown linear combination of the instrumental variables $Z_i = (Z_i^{1\prime}, Z_i^{2\prime})'$, where z_i^2 implicitly depends on n. For example, we could have $z_i^2 = \sum_{j=1}^{K_n - G_1} \pi_j^2 Z_{ij}^2 / \sqrt{K_n - G_1}$, where Z_{ij}^2 have variances that are bounded uniformly in K_n and $1/\sqrt{K_n - G_1}$ is included to normalize the variance of z_i^2 to be bounded. The many weak instrument example of Chao and Swanson (2005) is then included by taking $\mu_n = \sqrt{K_n - G_1}$, in which case the reduced form for X_i^2 is

$$\Upsilon_i^2 = \pi^1 Z_i^1 + \sum_{j=1}^{K_n - G_1} \pi_j^2 Z_{ij}^2 / \sqrt{n}.$$

The excluded instrument z_i^2 may also be an unknown function that is being approximated by a linear combination of Z_i. For instance, suppose that $z_i^2 = f_0(w_i)$ for an unknown function $f_0(w_i)$ of variables w_i. In this case we could let the instrumental variables include a vector $p_K(w_i) \overset{def}{=} (p_{1K}(w_i), ..., p_{K-G_1,K}(w_i))'$ of approximating functions, such as polynomials or splines. Here the vector of instrumental variables would be $Z_i = (Z_i^{1\prime}, p^K(w_i)')'$. For $\mu_n = \sqrt{n}$ this example is like Newey (1990) where Z_i includes approximating functions for the reduced form but the number of instruments can grow as fast as the sample size. Alternatively, if $\mu_n/\sqrt{n} \longrightarrow 0$, it is a modified version where δ^2 is weakly identified.

In Assumption 2 we can think of μ_n^2 as being proportional to the concentration parameter. For $\mu_n^2 \sim n$, we have asymptotic theory as in Kunitomo (1980), Morimune (1984), and Bekker (1994), where the number of instruments K_n can grow as fast as the sample size. For μ_n^2 growing slower than n we have the many weak instrument asymptotics of Chao and Swanson (2005).

The fundamental rate condition $\sqrt{K_n}/\mu_n^2 \longrightarrow 0$ given in Assumption 2 is needed to ensure that the stochastic part of the objective function for the estimator does not dominate the identifying part.

Assumption 3: There is a constant, $C > 0$ such that $(\varepsilon_1, U_1), ..., (\varepsilon_n, U_n)$ are independent, with $E[\varepsilon_i] = 0$, $E[U_i] = 0$, $E[\varepsilon_i^2] < C$, $E[\|U_i\|^2] \leq C$, $Var((\varepsilon_i, U_i')') = diag(\Omega_i^*, 0)$, and $\sum_{i=1}^n \Omega_i^*/n$ is uniformly nonsingular.

This assumption requires second conditional moments of disturbances to be bounded. It also imposes uniform nonsingularity of the variance of the reduced form disturbances, that is useful in the consistency proof, to help the denominator of the objective function say away from zero.

Assumption 4: There is a π_{Kn} such that $\sum_{i=1}^n \|z_i - \pi_{Kn} Z_i\|^2 / n \longrightarrow 0$.

This condition and $P_{ii} \leq C < 1$ will imply that

$$A_n = \Upsilon' P \Upsilon/n - \sum_{i=1}^n \Upsilon_i \Upsilon_i'/n = \sum_{i=1}^n (1 - P_{ii}) \Upsilon_i \Upsilon_i'/n + o(1) \geq (1 - C) \sum_{i=1}^n \Upsilon_i \Upsilon_i'/n + o(1),$$

so that A_n is positive definite in large enough samples. Also, Assumption 4 is not very restrictive because flexibility is allowed in the specification of Υ_i. If we simply make Υ_i the expectation of Y_i given the instrumental variables then Assumption 4 holds automatically.

These conditions imply estimator consistency:

THEOREM 1: *If Assumptions 1-4 are satisfied and* $\hat{\alpha} = o_p(\mu_n^2/n)$ *or* $\hat{\delta}$ *is HLIM or HFUL then* $\mu_n^{-1} S_n'(\hat{\delta} - \delta_0) \xrightarrow{p} 0$ *and* $\hat{\delta} \xrightarrow{p} \delta_0$.

This result gives convergence rates for linear combinations of $\hat{\delta}$. For instance, in the above example, it implies that $\hat{\delta}_1$ is consistent and that $\pi_{11}'\hat{\delta}^1 + \hat{\delta}^2 = o_p(\mu_n/\sqrt{n})$.

For asymptotic normality it is helpful to strengthen the conditions on moments.

Assumption 5: There is a constant, $C > 0$, such that with probability one, $\sum_{i=1}^n \|z_i\|^4 / n^2 \longrightarrow 0$, $E[\varepsilon_i^4] \le C$ and $E[\|U_i\|^4] \le C$.

To state a limiting distribution result it is helpful to also assume that certain objects converge and to allow for two cases of growth rates of K relative to μ_n^2. Also, the asymptotic variance of the estimator will depend on the growth rate of K relative to μ_n^2. Let $\sigma_i^2 = E[\varepsilon_i^2]$, $\gamma_n = \sum_{i=1}^n E[U_i\varepsilon_i] / \sum_{i=1}^n \sigma_i^2$, $\tilde{U} = U - \varepsilon\gamma_n'$, having i^{th} row $\tilde{U}_i'$; and let $\tilde{\Omega}_i = E[\tilde{U}_i\tilde{U}_i']$.

Assumption 6: Either I) K/μ_n^2 is bounded and $\sqrt{K} S_n^{-1} \longrightarrow S_0$ or; II) $K/\mu_n^2 \longrightarrow \infty$ and $\mu_n S_n^{-1} \longrightarrow \bar{S}_0$. Also $H_P = \lim_{n\longrightarrow\infty} \sum_{i=1}^n (1 - P_{ii}) z_i z_i'/n$, $\Sigma_P = \lim_{n\longrightarrow\infty} \sum_{i=1}^n (1 - P_{ii})^2 z_i z_i' \sigma_i^2/n$ and $\Psi = \lim_{n\longrightarrow\infty} \sum_{i\neq j} P_{ij}^2 \left(\sigma_i^2 E[\tilde{U}_j\tilde{U}_j'] + E[\tilde{U}_i\varepsilon_i]E[\varepsilon_j\tilde{U}_j'] \right) / K$ exist.

This convergence condition can be replaced by an assumption that certain matrices are uniformly positive definite without affecting the limiting distribution result for t-ratios given in Theorem 3 below (see Chao et al. 2009).

We can now state the asymptotic normality results. In Case I we have that

$$S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, \Lambda_I), \tag{4.2}$$

where

$$\Lambda_I = H_P^{-1}\Sigma_P H_P^{-1} + H_P^{-1} S_0 \Psi S_0' H_P^{-1}.$$

In Case II, we have that

$$(\mu_n/\sqrt{K}) S_n'(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, \Lambda_{II}), \tag{4.3}$$

where

$$\Lambda_{II} = H_P^{-1} \bar{S}_0 \Psi \bar{S}_0' H_P^{-1}.$$

The asymptotic variance expressions allow for the many instrument sequence of Kunitomo (1980) and Bekker (1994) and the many weak instrument sequence of Chao and Swanson (2004, 2005). In Case I, the first term in the asymptotic variance, Λ_I, corresponds to the usual asymptotic variance, and the second is an adjustment for the presence of many instruments. In Case II, the asymptotic variance, Λ_{II}, only contains the adjustment for many instruments. This is because K is growing faster than μ_n^2. Also, Λ_{II} will be singular when included exogenous variables are present.

We can now state an asymptotic normality result.

THEOREM 2: *If Assumptions 1-6 are satisfied, $\hat{\alpha} = \tilde{\alpha} + O_p(1/T)$ or $\hat{\delta}$ is HLIM or HFUL, then in Case I, equation (4.2) is satisfied, and in Case II, equation (4.3) is satisfied.*

It is interesting to compare the asymptotic variance of the HFUL estimator with that of LIML when the disturbances are homoskedastic. First, note that the disturbances are not restricted to be Gaussian and that the asymptotic variance does not depend on third or fourth moments of the disturbances. In contrast, the asymptotic variance of LIML does depend on third and fourth moment terms for non Gaussian disturbances; see Bekker and van der Ploeg (2005) and van Hasselt (2000). This makes estimation of the asymptotic variance simpler for HFUL than for LIML. It appears that the the jackknife form of the numerator has this effect on HFUL. Deleting the own observation terms in effect removes moment conditions that are based on squared residuals. Bekker and van der Ploeg (2005) also found that the limiting distribution of their MM estimator for dummy instruments and group heteroskedasticity did not depend on third and fourth moments.

Under homoskedasticity the variance of $Var((\varepsilon_i, U_i'))$ will not depend on i (e.g. so that $\sigma_i^2 = \sigma^2$). Then, $\gamma_n = E[X_i\varepsilon_i]/\sigma^2 = \gamma$ and $E[\tilde{U}_i\varepsilon_i] = E[U_i\varepsilon_i] - \gamma\sigma^2 = 0$, so that

$$\Sigma_P = \sigma^2\tilde{H}_P, \tilde{H}_P = \lim_{n\longrightarrow\infty}\sum_{i=1}^{n}(1-P_{ii})^2 z_i z_i'/n, \Psi = \sigma^2 E[\tilde{U}_j\tilde{U}_j'](1 - \lim_{n\longrightarrow\infty}\sum_{i=1}^{n} P_{ii}^2/K).$$

Focusing on Case I, letting $\Gamma = \sigma^2 S_0 E[\tilde{U}_i\tilde{U}_i']S_0'$, the asymptotic variance of HLIM is then

$$V = \sigma^2 H_P^{-1}\tilde{H}_P H_P^{-1} + \lim_{n\longrightarrow\infty}(1 - \sum_{i=1}^{n} P_{ii}^2/K)H_p^{-1}\Gamma H_P^{-1}.$$

For the variance of LIML, assume that third and fourth moments obey the same restrictions that they do under normality. Then from Hansen, Hausman, and Newey (2008), for $H = \lim_{n\longrightarrow\infty}\sum_{i=1}^{n} z_i z_i'/n$ and $\tau = \lim_{n\longrightarrow\infty} K/n$, the asymptotic variance of LIML is

$$V^* = \sigma^2 H^{-1} + (1-\tau)^{-1}H^{-1}\Gamma H^{-1}.$$

With many weak instruments, where $\tau = 0$ and $\max_{i\leq n} P_{ii} \longrightarrow 0$, we will have $H_P = \tilde{H}_P = H$ and $\lim_{n\longrightarrow\infty}\sum_i P_{ii}^2/K \longrightarrow 0$, so that the asymptotic variances of HLIM and LIML are the same and equal to $\sigma^2 H^{-1} + H^{-1}\Gamma H^{-1}$. This case is most important in practical applications, where K is usually very small relative to n. In such cases we would expect from the asymptotic approximation to find that the variance of LIML and HLIM are very similar. Also, the JIV estimators will be inefficient relative to LIML and HLIM. As shown in Chao and Swanson (2004), under many weak instruments the asymptotic variance of JIV is

$$V_{JIV} = \sigma^2 H^{-1} + H^{-1}S_0(\sigma^2 E[U_i U_i'] + E[U_i\varepsilon_i]E[\varepsilon_i U_i'])S_0' H^{-1},$$

which is larger than the asymptotic variance of HLIM because $E[U_i U_i'] \geq E[\tilde{U}_i\tilde{U}_i']$.

In the many instruments case, where K and μ_n^2 grow as fast as n, it turns out that we cannot rank the asymptotic variances of LIML and HLIM. To show this, consider an example where $p = 1$, z_i alternates between $-\bar{z}$ and $\bar{z}$ for $\bar{z} \neq 0$, $S_n = \sqrt{n}$ (so that $\Upsilon_i = z_i$), and z_i is included among the elements of Z_i. Then, for $\tilde{\Omega} = E[\tilde{U}_i^2]$ and $\kappa = \lim_{n\longrightarrow\infty}\sum_{i=1}^{n} P_{ii}^2/K$ we find that

$$V - V^* = \frac{\sigma^2}{\bar{z}^2(1-\tau)^2}(\tau\kappa - \tau^2)\left(1 - \frac{\tilde{\Omega}}{\bar{z}^2}\right).$$

Since $\tau\kappa - \tau^2$ is the limit of the sample variance of P_{ii}, which we assume to be positive, $V \geq V^*$ if and only if $\bar{z}^2 \geq \tilde{\Omega}$. Here, $\bar{z}^2$ is the limit of the sample variance of z_i. Thus, the asymptotic variance ranking can go either way depending on whether the sample variance of z_i is bigger than the variance of $\tilde{U}_i$. In applications where the sample size is large relative to the number of instruments, these efficiency differences will tend to be quite small, because P_{ii} is small.

For homoskedastic, non-Gaussian disturbances, it is also interesting to note that the asymptotic variance of HLIM does not depend on third and fourth moments of the disturbances, while that of LIML does (see Bekker and van der Ploeg (2005) and van Hasselt (2000)). This makes estimation of the asymptotic variance simpler for HLIM than for LIML.

It remains to establish the consistency of the asymptotic variance estimator, and to show that confidence intervals can be formed for linear combinations of the coefficients in the usual way. The following theorem accomplishes this, under additional conditions on z_i.

THEOREM 3: *If Assumptions 1-6 are satisfied, and $\hat{\alpha} = \tilde{\alpha} + O_p(1/T)$ or $\hat{\delta}$ is HLIM or HFUL, there exists a C with $\|z_i\| \leq C$ for all i, and there exists a π_n, such that $\max_{i \leq n} \|z_i - \pi_n Z_i\| \longrightarrow 0$, then in Case I, $S_n' \hat{V} S_n \xrightarrow{p} \Lambda_I$ and in Case II, $\mu_n^2 S_n' \hat{V} S_n / K \xrightarrow{p} \Lambda_{II}$.. Also, if $c' S_0' \Lambda_I S_0 c \neq 0$ in Case I or $c' \bar{S}_0' \Lambda_{II} \bar{S}_0 c \neq 0$ in Case II, then*

$$\frac{c'(\hat{\delta} - \delta_0)}{\sqrt{c' \hat{V} c}} \xrightarrow{d} N(0,1).$$

This result allows us to form confidence intervals and test statistics for a single linear combination of parameters in the usual way.

5 Existence of Moments of HFUL

In giving some of the results and proofs below, we find it convenient to write the model in terms of its restricted reduced form specification, i.e.

$$\begin{aligned} y &= \Upsilon\delta_0 + v, \\ X &= \Upsilon + U, \end{aligned}$$

or

$$\overline{X} = \Upsilon\Delta + \overline{V}$$

where $\overline{X} = \begin{bmatrix} y & X \end{bmatrix}$, $\Delta = \begin{bmatrix} \delta_0 & I_G \end{bmatrix}$, and $\overline{V} = \begin{bmatrix} v & U \end{bmatrix}$. The following notations are also used in the proofs below. Let $\overline{U} = \begin{bmatrix} \varepsilon & U \end{bmatrix}$ be the matrix of structural form disturbances, then for the restricted reduced form to be compatible with the structural form of the model discussed in section 2, we must have

$$\varepsilon = \overline{V}\delta_{\Delta,0},\ U = \overline{V}F_2, \tag{5.4}$$

with $\delta_{\Delta,0} = \begin{pmatrix} 1 & -\delta_0' \end{pmatrix}'$ and $F_2 = \begin{bmatrix} 0 & I_G \end{bmatrix}'$; or, more succinctly,

$$\overline{U} = \overline{V}D^{-1}, \tag{5.5}$$

where

$$D = \begin{pmatrix} 1 & 0 \\ \delta_0 & I_G \end{pmatrix}.$$

Moreover, let $\overline{V}_i'$ be the i^{th} row of $\overline{V}$, and we take

$$\Xi_i := E\left(\overline{V}_i\overline{V}_i'\right) = D'\Omega_i D, \tag{5.6}$$

where $\Omega_i = Var\left((\varepsilon_i, U_i')'\right)$.

Here let $\mathbb{I}_{\mathcal{A}}$ denote the indicator function of the set $\mathcal{A}$; let $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$ denote, respectively, the minimal and maximal eigenvalue of the matrix B; and let $\|\cdot\|$ denote the Euclidean norm, or the Frobenius norm when applied to matrices so that $\|A\| = \sqrt{tr\{A'A\}}$. Also, the notation $a_n \sim b_n$ means that $\lim_{n\to\infty}(a_n/b_n) = c$ for some constant $c \neq 0$. In addition, M, T and CS denote, respectively, Markov's inequality, the Triangle inequality and the Cauchy-Schwarz inequality.

Assumption 7 $K = O(n^a)$ for some real constant a such that $0 \leq a \leq 1$, if $a = 1$, then $n - K \to \infty$ as $n \to \infty$, and for all n sufficiently large, there exists a positive constant C_P such that $P_{ii} \leq C_P (K/n) < 1 \quad (i = 1, ..., n)$. (b) $\mu_n^2 \sim n^b$ for some real constant b such that $a/2 < b \leq 1$. (c) If K is fixed then $z_i = \pi Z_i$ (d) $\delta_0 \in \mathcal{D} \subset \mathbb{R}^G$, where $\mathcal{D}$ is bounded. (e) $\lambda_{\max}\left(\widetilde{S}'\widetilde{S}\right)$ is bounded.

Next, define

$$\varphi(a,b) = \frac{a \vee (1 - 5\psi(a,b))/2}{\psi(a,b)} \mathbb{I}\left\{\frac{a}{2} < b \leq \frac{1}{2}\right\} + \frac{a \vee (1 - 5(2b-a))/2}{\left\{(2b-a) \wedge \frac{1}{2}\right\}} \mathbb{I}\left\{\frac{1}{2} < b \leq 1\right\}, \tag{5.7}$$

where

$$\psi(a,b) = 2b - a - (b-a)_+$$

with $(b-a)_+ = (b-a) \vee 0$ and where a and b are as specified in Assumptions 7. Here, $x \vee y = \max(x,y)$ and $x \wedge y = \min(x,y)$.

Assumption 8: Let p be a positive integer and η a positive constant and define

$$q = (1+\eta)\left[2G + 1 + \varphi(a,b)\right].$$

$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} Var(\varepsilon_i, V_i')\right)$ is bounded away from zero for n sufficiently large and there is $\widetilde{C} > 0$ such that $E[\left\|\overline{V}_i\right\|^{2pq}] \leq \widetilde{C}$ and $\sum_{i=1}^{n} \|z_i\|^{2pq}/n \leq \widetilde{C}$.

Proving the existence of moments of HFUL requires showing the existence of certain inverse moments of $\det(X_*'MX_*/n)$, where $X_* = \left[\begin{array}{cc} \varepsilon & X \end{array}\right]$. That is, we need to show under more primitive conditions that there exists a constant $\overline{C}$ such that

$$E\left[\det(X_*'MX_*/n)\right]^{-\rho} \leq \overline{C} < \infty \quad \text{for some } \rho > 0.$$

To do this, we need to put conditions on the joint data density in some neighborhood of the set of points where $\det(X_*'MX_*/n) = 0$. This is most conveniently done if we change variables in the following way:

Define $H_Z = Z(Z'Z)^{-1/2} \in V_{K,n}$ and partition

$$H_Z = \begin{pmatrix} Z_{1\cdot}(Z'Z)^{-1/2} \\ Z_{2\cdot}(Z'Z)^{-1/2} \end{pmatrix} = \begin{pmatrix} \underset{K\times K}{H_{Z,1}} \\ \underset{(n-K)\times K}{H_{Z,2}} \end{pmatrix}, \quad (say).$$

Now, define

$$\begin{aligned} H_Z^{\perp} &= \begin{bmatrix} -\left(H'_{Z,1}\right)^{-1} H'_{Z,2} \\ I_{n-K} \end{bmatrix} \left[I_{n-K} + H_{Z,2}\left(H'_{Z,1}H_{Z,1}\right)^{-1} H'_{Z,2}\right]^{-1/2} \\ &= \begin{bmatrix} -\left(Z'_{1\cdot}\right)^{-1} Z'_{2\cdot} \\ I_{n-K} \end{bmatrix} \left[I_{n-K} + Z_{2\cdot}\left(Z'_{1\cdot}Z_{1\cdot}\right)^{-1} Z'_{2\cdot}\right]^{-1/2}. \end{aligned}$$

Note that the implicit assumption that $H_{Z,1}$ is non-singular is really without loss of generality since $rank(Z) = K$ by Assumption 1; and, hence, the invertibility of $Z_{1\cdot}$ (and, thus, $H_{Z,1}$) can always be achieved, if necessary, by a repermutation of the rows of Z. Note also that by construction

$$P = H_Z H'_Z,\ M = H_Z^{\perp} H_Z^{\perp\prime}, \text{ and } \begin{pmatrix} H_Z & H_Z^{\perp} \end{pmatrix} \in O(n),$$

from which we have

$$\frac{X'_* M X_*}{n} = \frac{X'_* H_Z^{\perp} H_Z^{\perp\prime} X_*}{n} = W'W, \tag{5.8}$$

where $W = n^{-1/2} H_Z^{\perp\prime} X_*$[2]. Let $f_n(W)$ denote the joint probability density function of W. Also, let $L = G + 1$, and write $W = (w_1, .., w_L)$ and $f_n(W) = f_n(w_1, .., w_L)$. We give below a transformation of this joint density factorized into a product of conditional

[2]For notational simplicity, we shall suppress the dependence of W on n.

and marginal densities. Consider the joint density factorization

$$\begin{aligned}
& f_n\left(w_1, .., w_L\right) \prod_{\ell=1}^{L}\left(d w_{\ell}\right) \\
= \; & g_n\left(h_1\right)\left(d h_1\right) g_n\left(r_1 \mid s_1\right)\left[r_1\right]^{(n-K-1)} d r_1 \\
& \times g_n\left(q_2^{\perp} \mid s_{2,1}\right)\left(d q_2^{\perp}\right) g_n\left(\widetilde{t}_2 \mid s_{2,2}\right)\left(1-\widetilde{t}_2^2\right)^{(n-K-3)/2} d \widetilde{t}_2 g_n\left(r_2 \mid s_{2,3}\right)\left[r_2\right]^{(n-K-1)} d r_2 \\
& \times \prod_{\ell=3}^{L} g_n\left(q_{\ell}, q_{\ell}^{\perp} \mid s_{\ell,1}\right)\left(d q_{\ell}\right)\left(d q_{\ell}^{\perp}\right) \\
& \times \prod_{\ell=3}^{L} g_n\left(t_{\ell} \mid s_{\ell,2}\right) t_{\ell}^{\ell-2}\left(1-t_{\ell}^2\right)^{(n-K-\ell-1)/2} d t_{\ell} \\
& \times \prod_{\ell=3}^{L} g_n\left(r_{\ell} \mid s_{\ell,3}\right)\left[r_{\ell}\right]^{(n-K-1)} d r_{\ell} \quad a.e. \qquad (5.9)
\end{aligned}$$

where

$$\begin{aligned}
s_1 &= h_1, \\
s_{2,1} &= \left(r_1, s_1'\right)', \; s_{2,2}=\left(q_2^{\perp\prime}, s_{2,1}'\right)', s_{2,3}=\left(\widetilde{t}_2, s_{2,2}'\right)' \\
s_{\ell,1} &= \left(r_{\ell-1}, s_{\ell-1,3}'\right)', \; s_{\ell,2}=\left(q_{\ell}', q_{\ell}^{\perp\prime}, s_{\ell,1}'\right)', s_{\ell,3}=\left(t_{\ell}, s_{\ell,2}'\right)' \quad \text{for } \ell=3, ..., L. (5.10)
\end{aligned}$$

Expression (5.9) has been constructed via a series of recursive polar decompositions performed on the columns of W and defined by the equations:

$$w_{\ell}=h_{\ell} r_{\ell}\;, \;\; r_{\ell}=\left(w_{\ell}' w_{\ell}\right)^{1/2}>0 \; a.s. \;\; \text{for } \ell=1, ..., L;$$

$$\begin{aligned}
h_1 &= \frac{w_1}{\left(w_1' w_1\right)^{1/2}} \in V_{1,n-K}, \; h_2=h_1 \widetilde{t}_2+F\left(h_1\right) q_2^{\perp}\left(1-\widetilde{t}_2^2\right)^{1/2}, \\
h_{\ell} &= H_{[1,\ell-1]} q_{\ell} t_{\ell}+F_{[1,\ell-1]} q_{\ell}^{\perp}\left(1-t_{\ell}^2\right)^{1/2} \;\; \text{for } \ell=3, ..., L,
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{t}_2 &= h_1' h_2, \;\; q_2^{\perp}=\frac{F\left(h_1\right)' h_2}{\left(h_2' F\left(h_1\right) F\left(h_1\right)' h_2\right)^{1/2}} \in V_{1,n-K-1}, \\
t_{\ell} &= \left(h_{\ell}' H_{[1,\ell-1]} H_{[1,\ell-1]}' h_{\ell}\right)^{1/2}, \; q_{\ell}=\frac{H_{[1,\ell-1]}' h_{\ell}}{\left(h_{\ell}' H_{[1,\ell-1]} H_{[1,\ell-1]}' h_{\ell}\right)^{1/2}} \in V_{1,\ell-1}, \\
q_{\ell}^{\perp} &= \frac{F_{[1,\ell-1]}' h_{\ell}}{\left(h_{\ell}' F_{[1,\ell-1]} F_{[1,\ell-1]}' h_{\ell}\right)^{1/2}} \in V_{1,n-K-\ell+1} \;\text{for } \ell=3, ..., L;
\end{aligned}$$

where $F(h_1)$ is an $(n-K)\times(n-K-1)$ matrix chosen so that $\left[\begin{array}{cc} h_1 & F(h_1) \end{array}\right] \in O(n-K)$; and where $H_{[1,\ell-1]} = \underline{H}_{[1,\ell-1]} \left(\underline{H}'_{[1,\ell-1]}\underline{H}_{[1,\ell-1]}\right)^{1/2}$ with $\underline{H}_{[1,\ell-1]} = \left[\begin{array}{cccc} h_1 & h_2 & \cdots & h_{\ell-1} \end{array}\right]$ and $F_{[1,\ell-1]}$ is an $(n-K)\times(n-K-\ell+1)$ matrix chosen so that $\left[\begin{array}{cc} H_{[1,\ell-1]} & F_{[1,\ell-1]} \end{array}\right] \in O(n-K)$ for $\ell = 3, ..., L$. Here, $V_{k,m}$ denotes the Stiefel manifold, so that $V_{k,m} = \{X \ (m \times k) : X'X = I_k\}$, i.e., $V_{k,m}$ is the set (or space) of $m \times k$ matrices such that $X'X = I_k$; and $O(n-K)$ denotes the orthogonal group of $(n-K)\times(n-K)$ orthogonal matrices. In addition, note that under the definition above $-1 < \widetilde{t}_2 < 1$ and $0 < t_\ell < 1$ for $\ell = 3, ..., L$.

A detailed derivation of expression (5.9) is long, and so we have not included it in this paper. It can be found on John Chao's webpage at

http://econweb.umd.edu/˜chao/Research/research.html.

We note that a main reason for transforming the joint density in this way is that, under the new representation, points where $\det(X_*'MX_*/n) = 0$ have now revealed themselves as poles in uni-dimensional integrals, so that it becomes easier to see what additional conditions are needed for the existence of moments and how to specify them. These conditions are stated below.

Assumption 9: For each finite n, let $W = (w_1, .., w_L)$ have density $f_n(w_1, .., w_L)$ with respect to the Lebesgue measure, and let this density be transformed and factorized into the form given by expression (5.9). Suppose that there exist a positive integer N, some real number ϵ with $0 < \epsilon \leq 1$, and a positive constant C_ϵ such that for all n sufficiently large such that $n-K \geq N + L + 4p(1+\eta)/\eta$, the following conditions hold

(i)

$$g_n(r_1|\ s_1)(r_1)^{(n-K-4p[1+\eta]/\eta-1)} \leq C_\epsilon < \infty \quad a.s. \ \ \mathbb{P}_{s_1}$$

for all $r_1 \in [0, \epsilon)$;

(ii)

$$g_n(r_\ell|\ s_{\ell,3})(r_\ell)^{(n-K-4p[1+\eta]/\eta-1)} \leq C_\epsilon < \infty \quad a.s. \ \ \mathbb{P}_{s_{\ell,3}}$$

for all $r_\ell \in [0, \epsilon)$ and $\ell = 2, ..., L$;

(iii)

$$g_n\left(\widetilde{t}_2 \middle|\ s_{2,2}\right)\left(1-\widetilde{t}_2^2\right)^{(n-K-4p[1+\eta]/\eta-3)/2} \leq C_\epsilon < \infty \ \ a.s.\ \ \mathbb{P}_{s_{2,2}}$$

for all $\widetilde{t}_2 \in [-1,-1+\epsilon) \cup (1-\epsilon, 1]$.

(iv)

$$g_n\left(t_\ell \middle|\ s_{\ell,2}\right) t_\ell^{\ell-2}\left(1-t_\ell^2\right)^{(n-K-\ell-4p[1+\eta]/\eta-1)/2} \leq C_\epsilon < \infty \ \ a.s.\ \ \mathbb{P}_{s_{\ell,2}}$$

for all $t_\ell \in (1-\epsilon, 1]$ and $\ell = 3, ..., L$.

Assumption 8 specifies the moment condition on the error process $\left\{\overline{V}_i\right\}$ as dependent on the number of endogenous regressors G, instrument weakness as parameterized by b, and an upper bound on the rate at which the number of instrument grows, as parameterized by a. Although the function $\varphi(a,b)$ which enters into the moment condition seems complicated, it actually depends on a and b in an intuitive way, so that everything else being equal, more stringent moment conditions are needed in cases with weaker instruments and/or faster growing K. More stringent moment conditions are also needed in situations with a larger number of endogenous regressors.

To get more intuition about Assumption 8, consider the following two special cases. First, consider the conventional case where the instruments are strong and the number of instruments is fixed, so that $a = 0$ and $b = 1$. In this case, it is easy to see that $\varphi(a,b) = \varphi(0,1) = 0$, and Assumption 8 requires finite moments up to the order

$$2pq = 2p\,[2G+1]\,(1+\eta)\,.$$

If we further consider the case with one endogenous regressor ($G = 1$) and where η can be taken to be small; then, Assumption 8 requires a bit more than a sixth moment condition (on the errors) for the existence of the first moment of $HFUL$ and a bit more than a twelfth moment condition for the existence of the second moment. Next, consider the many weak instrument case where $a = 1/2$ and $a/2 = 1/4 < b \leq 1/2$. In this case, note that since $b \leq a$, we have

$$\psi(a,b) = \psi(1/2, b) = 2b - 1/2,$$

and

$$\varphi(a,b)=\varphi(1/2,b)=\frac{1}{4b-1}\text{ for }b\in(1/4,1/2],$$

so that $1\leq\varphi(1/2,b)<\infty$ and $\varphi(1/2,b)$ is a decreasing function of b. In particular, note that the strength of the moment condition required grows without bound as b approaches 1/4.

The specification of η involves a trade-off in the stringency of the conditions. If η is taken to be small, then weaker moment conditions are assumed on the error process, but a large sample size n may be needed in order for Assumption 9 to hold and vice versa if η is taken to be large.

THEOREM 4: *If Assumptions 1-4, 7, 8, and 9 are satisfied for some positive p. then there exists a positive constant $\overline{C}$ such that*

$$E\left[\left\|\widehat{\delta}_{HFUL}\right\|^{p}\right]\leq\overline{C}<\infty$$

for n sufficiently large.

A proof of this theorem can be found on John Chao's webpage at
http://econweb.umd.edu/~chao/Research/research.html.

6 Monte Carlo Results

In this Monte Carlo simulation, we provide evidence concerning the finite sample behavior of HLIM and HFUL. The model that we consider is

$$y_i=\delta_{10}+\delta_{20}x_{2i}+\varepsilon_i, x_{2i}=\pi z_{1i}+U_{2i}$$

where $z_{i1}\sim N(0,1)$ and $U_{2i}\sim N(0,1)$. The i^{th} instrument observation is

$$Z_i'=(1,z_{1i},z_{1i}^2,z_{1i}^3,z_{1i}^4,z_{1i}D_{i1},...,z_{1i}D_{i,K-5}),$$

where $D_{ik}\in\{0,1\}$, $\Pr(D_{ik}=1)=1/2$, and $z_{i1}\sim N(0,1)$. Thus, the instruments consist of powers of a standard normal up to the fourth power plus interactions with dummy variables. Only z_1 affects the reduced form, so that adding the other instruments does not improve asymptotic efficiency of HFUL, though the powers of z_{i1} do help with asymptotic efficiency of the CUE.

The structural disturbance, ε, is allowed to be heteroskedastic, being given by

$$\varepsilon = \rho U_2 + \sqrt{\frac{1-\rho^2}{\phi^2 + (0.86)^4}}(\phi v_1 + 0.86 v_2), v_1 \sim N(0, z_1^2), v_2 \sim N(0, (0.86)^2),$$

where v_1 and v_2 are independent of U_2. This is a design that will lead to LIML being inconsistent with many instruments. Here, $E[X_i \varepsilon_i]$ is constant and σ_i^2 is quadratic in z_{i1}, so that $\gamma_i = (C_1 + C_2 z_{i1} + C_3 z_{i1}^2)^{-1} A$, for a constant vector A and constants C_1, C_2, C_3. In this case, P_{ii} will be correlated with $\gamma_i = E[X_i \varepsilon_i]/\sigma_i^2$ so that LIML is not consistent.

We report properties of estimators and t-ratios for δ_2. We set $n = 800$ and $\rho = 0.3$ throughout and let the number of instrumental variables be $K = 2, 30$. For $K = 2$ the instruments are $(1, z_i)$. We choose π so that the concentration parameter is $n\pi^2 = \mu^2 = 8, 32$. We also ran experiments with $K = 10$ and $\mu^2 = 16$. We also choose ϕ so that the R-squared for the regression of ε^2 on the instruments is 0, 0.1, or 0.2.

Below, we report results on median bias, the range between the .05 and .95 quantiles, and nominal .05 rejection frequencies for a Wald test on δ_2, for LIML, HLIM, Fuller (1977), HFUL ($C = 1$), JIVE, and the CUE. Interquartile range results were similar. We find that under homoskedasticity, HFUL is much less dispersed than LIML but slightly more biased. Under heteroskedasticity, HFUL is much less biased and also much less dispersed than LIML. Thus, we find that heteroskedasticity can bias LIML. We also find that the dispersion of LIML is substantially larger than HFUL. Thus we find a lower bias for HFUL under heteroskedasticity and many instruments, as predicted by the theory, as well as substantially lower dispersion, which though not predicted by the theory may be important in practice.

In addition in Tables 3 and 6 we find that the rejection frequencies for HFUL are quite close to their nominal values, being closer than all the rest throughout much of the tables. Thus, the standard errors we have given work very well in accounting for many instruments and heteroskedasticity.

Table One: Median Bias; $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	0.005	0.005	0.042	0.043	−0.034	0.005
8	10	0.024	0.023	0.057	0.057	0.053	0.025
8	30	0.065	0.065	0.086	0.091	0.164	0.071
32	2	0.002	0.002	0.011	0.011	−0.018	0.002
32	10	0.002	0.001	0.011	0.011	−0.019	0.002
32	30	0.003	0.002	0.013	0.013	−0.014	0.006

***Results based on 20,000 simulations.

Table 2: Nine Dec. Range: .05 to .95 $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	1.470	1.466	1.072	1.073	3.114	1.470
8	10	2.852	2.934	1.657	1.644	5.098	3.101
8	30	5.036	5.179	2.421	2.364	6.787	6.336
32	2	0.616	0.616	0.590	0.589	0.679	0.616
32	10	0.715	0.716	0.679	0.680	0.816	0.770
32	30	0.961	0.985	0.901	0.913	1.200	1.156

***Results based on 20,000 simulations.

Table 3: .05 Rejection Frequencies; $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	.025	.026	.021	.034	.051	.012
8	10	.035	.037	.029	.044	.063	.027
8	30	.045	.049	.040	.054	.068	.051
32	2	.041	.042	.037	.044	.038	.030
32	10	.041	.042	.038	.044	.046	.041
32	30	.042	.047	.039	.050	.057	.062

***Results based on 20,000 simulations.

Table 4: Median Bias $\mathcal{R}^2_{\varepsilon^2|z_1^2} = .2$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	−0.001	0.050	0.041	0.078	−0.031	−0.001
8	10	−0.623	0.094	−0.349	0.113	0.039	0.003
8	30	−1.871	0.134	−0.937	0.146	0.148	−0.034
32	2	−0.001	0.011	0.008	0.020	−0.021	−0.001
32	10	−0.220	0.015	−0.192	0.024	−0.021	0.000
32	30	−1.038	0.016	−0.846	0.027	−0.016	−0.017

***Results based on 20,000 simulations.

Table 5: Nine Dec. Range: .05 to .95 $\mathcal{R}^2_{\varepsilon^2|z_1^2} = .2$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	2.219	1.868	1.675	1.494	4.381	2.219
8	10	26.169	5.611	4.776	2.664	7.781	16.218
8	30	60.512	8.191	7.145	3.332	9.975	1.5E+012
32	2	0.941	0.901	0.903	0.868	1.029	0.941
32	10	3.365	1.226	2.429	1.134	1.206	1.011
32	30	18.357	1.815	5.424	1.571	1.678	3.563

***Results based on 20,000 simulations.

Table 6: .05 Rejection Frequecies; $\mathcal{R}^2_{\varepsilon^2|z_1^2} = .2$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$JIVE$	CUE
8	2	.097	.019	.075	.023	.026	.008
8	10	.065	.037	.080	.041	.036	.043
8	30	.059	.051	.118	.055	.046	.094
32	2	.177	.040	.162	.040	.039	.024
32	10	.146	.042	.120	.044	.033	.030
32	30	.128	.049	.107	.051	.039	.073

***Results based on 20,000 simulations.

7 Conclusion

We have considered the situation of many instruments with heteroskedastic data. In this situation both 2SLS and LIML are inconsistent. We have proposed two new estimators, HLIML and HFUL, which are consistent in this situation. We derive the asymptotic normal distributions for both estimators with many instruments and many weak instrument sequences. We find the variances of the asymptotic distributions take a convenient form, which are straightforward to estimate consistently. A problem with the HLIML (and LIML) estimator is the wide dispersion caused by the "moments problem." We demonstrate that HFUL has finite sample moments so that the moments problem does not exist.

In Monte Carlo experiments we find these properties hold. With heteroscedasticity and many instruments we find that both LIML and Fuller have significant median bias (Table 4). We find that HLIM, HFUL, JIVE and CUE do not have this median bias. However, HLIM, JIVE and CUE all suffer from very large dispersion arising from the moments problem (Table 5). Indeed, the nine decile range for CUE exceeds 10^{12}! The dispersion of the HFUL estimate is much less than these alternative consistent estimators. Thus, we recommend that HFUL be used in the many instruments situation when heteroscedasticity is present, which is the common situation in microeconometrics.

8 Appendix: Proofs of Consistency and Asymptotic Normality

Throughout, let C denote a generic positive constant that may be different in different uses and let M, CS, and T denote the conditional Markov inequality, the Cauchy-Schwartz inequality, and the Triangle inequality respectively. The first Lemma is Lemma A0 from Hansen, Hausman, and Newey (2008).

LEMMA A0: *If Assumption 2 is satisfied and* $\left\| S_n'(\hat{\delta}-\delta_0)/\mu_n \right\|^2 / \left(1+\left\|\hat{\delta}\right\|^2\right) \xrightarrow{p} 0$ *then* $\left\| S_n'(\hat{\delta}-\delta_0)/\mu_n \right\| \xrightarrow{p} 0.$

We next give a result from Chao et al. (2007) that is used in the proof of consistency.

LEMMA A1 (LEMMA A1 OF CHAO ET AL., 2009): *If* $(W_i, Y_i), (i = 1, ..., n)$ *are independent,* W_i *and* Y_i *are scalars, and* P *is symmetric, idempotent of rank* K *then for* $\bar{w} = E[(W_1, ..., W_n)']$, $\bar{y} = E[(Y_1, ..., Y_n)']$, $\bar{\sigma}_{Wn} = \max_{i\le n} Var(W_i)^{1/2}$, $\bar{\sigma}_{Yn} = \max_{i\le n} Var(Y_i)^{1/2}$,

$$\sum_{i\neq j} P_{ij} W_i Y_j = \sum_{i\neq j} P_{ij}\bar{w}_i\bar{y}_j + O_p(K^{1/2}\bar{\sigma}_{Wn}\bar{\sigma}_{Yn} + \bar{\sigma}_{Wn}\sqrt{\bar{y}'\bar{y}} + \bar{\sigma}_{Yn}\sqrt{\bar{w}'\bar{w}}).$$

For the next result let $\bar{S}_n = diag(\mu_n, S_n)$, $\tilde{X} = [\varepsilon, X]\bar{S}_n^{-1\prime}$, and $H_n = \sum_{i=1}^n (1 - P_{ii}) z_i z_i'/n$.

LEMMA A2: *If Assumptions 1-4 are satisfied and* $\sqrt{K}/\mu_n^2 \longrightarrow 0$ *then*

$$\sum_{i\neq j} \tilde{X}_i P_{ij} \tilde{X}_j' = diag(0, H_n) + o_p(1).$$

Proof: Note that

$$\tilde{X}_i = \begin{pmatrix} \mu_n^{-1}\varepsilon_i \\ S_n^{-1}X_i \end{pmatrix} = \begin{pmatrix} 0 \\ z_i/\sqrt{n} \end{pmatrix} + \begin{pmatrix} \mu_n^{-1}\varepsilon_i \\ S_n^{-1}U_i \end{pmatrix}.$$

Since $\|S_n^{-1}\| \le C\mu_n^{-1}$ we have $Var(\tilde{X}_{ik}) \le C\mu_n^{-2}$ for any element $\tilde{X}_{ik}$ of $\tilde{X}_i$. Then applying Lemma A1 to each element of $\sum_{i \neq j} \tilde{X}_i P_{ij} \tilde{X}_j'$ gives

$$\begin{aligned}\sum_{i\neq j} \tilde{X}_i P_{ij} \tilde{X}_j' &= diag(0, \sum_{i\neq j} z_i P_{ij} z_j'/n) + O_p(K^{1/2}/\mu_n^2 + \mu_n^{-1}(\sum_i \|z_i\|^2/n)^{1/2}) \\ &= diag(0, \sum_{i\neq j} z_i P_{ij} z_j'/n) + o_p(1).\end{aligned}$$

Also, note that

$$\begin{aligned}H_n - \sum_{i\neq j} z_i P_{ij} z_j'/n &= \sum_i z_i z_i'/n - \sum_i P_{ii} z_i z_i'/n - \sum_{i\neq j} z_i P_{ij} z_j'/n = z'(I-P)z/n \\ &= (z - Z\pi_{Kn}')'(I-P)(z - Z\pi_{Kn}')/n \le (z - Z\pi_{Kn}')'(z - Z\pi_{Kn}')/n \\ &\le I_G \sum_i \|z_i - \pi_{Kn} Z_i\|^2/n \longrightarrow 0,\end{aligned}$$

where the third equality follows by $PZ = Z$, the first inequality by $I - P$ idempotent, and the last inequality by $A \le tr(A)I$ for any positive semi-definite (p.s.d.) matrix A. Since this equation shows that $H_n - \sum_{i\neq j} z_i P_{ij} z_j'/n$ is p.s.d. and is less than or equal to another p.s.d. matrix that converges to zero it follows that $\sum_{i\neq j} z_i P_{ij} z_j'/n = H_n + o_p(1)$. The conclusion follows by T. Q.E.D.

In what follows it is useful to prove directly that the HLIM estimator $\tilde{\delta}$ satisfies $S_n'(\tilde{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$.

Lemma A3: *If Assumptions 1-4 are satisfied then* $S_n'(\tilde{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$.

Proof: Let $\bar{\Upsilon} = [0, \Upsilon]$, $\bar{U} = [\varepsilon, U]$, $\bar{X} = [y, X]$, so that $\bar{X} = (\bar{\Upsilon} + \bar{U})D$ for

$$D = \begin{bmatrix} 1 & 0 \\ \delta_0 & I \end{bmatrix}.$$

Let $\hat{B} = \bar{X}'\bar{X}/n$. Note that $\|S_n/\sqrt{n}\| \le C$ and by standard calculations $z'U/n \xrightarrow{p} 0$. Then

$$\|\bar{\Upsilon}'\bar{U}/n\| = \|(S_n/\sqrt{n})\, z'U/n\| \le C\,\|z'U/n\| \xrightarrow{p} 0.$$

Let $\bar{\Omega}_n = \sum_{i=1}^n E[\bar{U}_i \bar{U}_i']/n = diag(\sum_{i=1}^n \Omega_i^*/n, 0) \ge C diag(I_{G_2+1}, 0)$ by Assumption 3, where $G_2 + 1$ is the dimension of number of included endogenous variables. By M we have $\bar{U}'\bar{U}/n - \bar{\Omega}_n \xrightarrow{p} 0$, so it follows that w.p.a.1.

$$\hat{B} = (\bar{U}'\bar{U} + \bar{\Upsilon}'\bar{U} + \bar{U}'\bar{\Upsilon} + \bar{\Upsilon}'\bar{\Upsilon})/n = \bar{\Omega}_n + \bar{\Upsilon}'\bar{\Upsilon}/n + o_p(1) \geq C diag(I_{G-G_2+1}, 0).$$

Since $\bar{\Omega}_n + \bar{\Upsilon}'\bar{\Upsilon}/n$ is bounded, it follows that w.p.a.1,

$$C \leq (1, -\delta')\hat{B}(1, -\delta')' = (y - X\delta)'(y - X\delta)/n \leq C \left\|(1, -\delta')\right\|^2 = C(1 + \|\delta\|^2).$$

Next, as defined preceding Lemma A2 let $\bar{S}_n = diag(\mu_n, S_n)$ and $\tilde{X} = [\varepsilon, X]\bar{S}_n^{-1\prime}$. Note that by $P_{ii} \leq C < 1$ and uniform nonsingularity of $\sum_{i=1}^n z_i z_i'/n$ we have $H_n \geq (1-C)\sum_{i=1}^n z_i z_i'/n \geq C I_G$. Then by Lemma A2, w.p.a.1.

$$\hat{A} \stackrel{def}{=} \sum_{i \neq j} P_{ij}\tilde{X}_i\tilde{X}_j' \geq C diag(0, I_G),$$

Note that $\bar{S}_n' D(1, -\delta')' = (\mu_n, (\delta_0 - \delta)'S_n)'$ and $\bar{X}_i = D'\bar{S}_n\tilde{X}_i$. Then w.p.a.1 for all δ

$$\begin{aligned}
\mu_n^{-2}\sum_{i \neq j} P_{ij}(y_i - X_i'\delta)(y_j - X_j'\delta) &= \mu_n^{-2}(1, -\delta')\left(\sum_{i \neq j} P_{ij}\bar{X}_i\bar{X}_j'\right)(1, -\delta')' \\
&= \mu_n^{-2}(1, -\delta')D'\bar{S}_n\hat{A}\bar{S}_n' D(1, -\delta')' \geq C\left\|S_n'(\delta - \delta_0)/\mu_n\right\|^2.
\end{aligned}$$

Let $\hat{Q}(\delta) = (n/\mu_n^2)\sum_{i \neq j}(y_i - X_i'\delta)P_{ij}(y_j - X_j'\delta)/(y - X\delta)'(y - X\delta)$. Then by the upper left element of the conclusion of Lemma A2, $\mu_n^{-2}\sum_{i \neq j}\varepsilon_i P_{ij}\varepsilon_j \xrightarrow{p} 0$. Then w.p.a.1

$$\left|\hat{Q}(\delta_0)\right| = \left|\mu_n^{-2}\sum_{i \neq j}\varepsilon_i P_{ij}\varepsilon_j / \sum_{i=1}^n \varepsilon_i^2/n\right| \xrightarrow{p} 0.$$

Since $\hat{\delta} = \arg\min_\delta \hat{Q}(\delta)$, we have $\hat{Q}(\hat{\delta}) \leq \hat{Q}(\delta_0)$.Therefore w.p.a.1, by $(y - X\delta)'(y - X\delta)/n \leq C(1 + \|\delta\|^2)$, it follows that

$$0 \leq \frac{\left\|S_n'(\hat{\delta} - \delta_0)/\mu_n\right\|^2}{1 + \left\|\hat{\delta}\right\|^2} \leq C\hat{Q}(\hat{\delta}) \leq C\hat{Q}(\delta_0) \xrightarrow{p} 0,$$

implying $\left\|S_n'(\hat{\delta} - \delta_0)/\mu_n\right\|^2 / \left(1 + \left\|\hat{\delta}\right\|^2\right) \xrightarrow{p} 0$. Lemma A0 gives the conclusion. Q.E.D.

Lemma A4: *If Assumptions 1-4 are satisfied,* $\hat{\alpha} = o_p(\mu_n^2/n)$, *and* $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$ *then for* $H_n = \sum_{i=1}^n (1 - P_{ii})z_i z_i'/n$,

$$S_n^{-1}\left(\sum_{i \neq j} X_i P_{ij} X_j' - \hat{\alpha}X'X\right)S_n^{-1\prime} = H_n + o_p(1), S_n^{-1}(\sum_{i \neq j} X_i P_{ij}\hat{\varepsilon}_j - \hat{\alpha}X'\hat{\varepsilon})/\mu_n \xrightarrow{p} 0.$$

Proof: By M and standard arguments $X'X = O_p(n)$ and $X'\hat{\varepsilon} = O_p(n)$. Therefore, by $\|S_n^{-1}\| = O(\mu_n^{-1})$,

$$\hat{\alpha} S_n^{-1} X'X S_n^{-1\prime} = o_p(\mu_n^2/n) O_p(n/\mu_n^2) \xrightarrow{p} 0, \hat{\alpha} S_n^{-1} X'\hat{\varepsilon}/\mu_n = o_p(\mu_n^2/n) O_p(n/\mu_n^2) \xrightarrow{p} 0.$$

Lemma A2 (lower right hand block) and T then give the first conclusion. By Lemma A2 (off diagonal) we have $S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j/\mu_n \xrightarrow{p} 0$, so that

$$S_n^{-1} \sum_{i \neq j} X_i P_{ij} \hat{\varepsilon}_j/\mu_n = o_p(1) - \left(S_n^{-1} \sum_{i \neq j} X_i P_{ij} X_j' S_n^{-1\prime} \right) S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0. Q.E.D.$$

Lemma A5: *If Assumptions 1 - 4 are satisfied and* $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$ *then* $\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j / \hat{\varepsilon}'\hat{\varepsilon} = o_p(\mu_n^2/n)$.

Proof: Let $\hat{\beta} = S_n'(\hat{\delta} - \delta_0)/\mu_n$ and $\breve{\alpha} = \sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j/\varepsilon'\varepsilon = o_p(\mu_n^2/n)$. Note that $\hat{\sigma}_\varepsilon^2 = \hat{\varepsilon}'\hat{\varepsilon}/n$ satisfies $1/\hat{\sigma}_\varepsilon^2 = O_p(1)$ by M. By Lemma A4 with $\hat{\alpha} = \breve{\alpha}$ we have $\tilde{H}_n = S_n^{-1}(\sum_{i \neq j} X_i P_{ij} X_j' - \breve{\alpha} X'X) S_n^{-1\prime} = O_p(1)$ and $W_n = S_n^{-1}(\sum_{i \neq j} X_i P_{ij} \varepsilon_j - \breve{\alpha} X'\varepsilon)/\mu_n \xrightarrow{p} 0$, so

$$\begin{aligned} \frac{\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\hat{\varepsilon}'\hat{\varepsilon}} - \breve{\alpha} &= \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \left(\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j - \sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j - \breve{\alpha} \left(\hat{\varepsilon}'\hat{\varepsilon} - \varepsilon'\varepsilon \right) \right) \\ &= \frac{\mu_n^2}{n} \frac{1}{\hat{\sigma}_\varepsilon^2} \left(\hat{\beta}' \tilde{H}_n \hat{\beta} - 2\hat{\beta}' W_n \right) = o_p(\mu_n^2/n), \end{aligned}$$

so the conclusion follows by T. Q.E.D.

Proof of Theorem 1: First, note that if $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$ then by $\lambda_{\min}(S_n S_n'/\mu_n^2) \geq \lambda_{\min}\left(\tilde{S}\tilde{S}'\right) > 0$ we have

$$\left\| S_n'(\hat{\delta} - \delta_0)/\mu_n \right\| \geq \lambda_{\min}(S_n S_n'/\mu_n^2)^{1/2} \left\| \hat{\delta} - \delta_0 \right\| \geq C \left\| \hat{\delta} - \delta_0 \right\|,$$

implying $\hat{\delta} \xrightarrow{p} \delta_0$. Therefore, it suffices to show that $S_n'(\hat{\delta} - \delta_0)/\mu_n \xrightarrow{p} 0$. For HLIM this follows from Lemma A3. For HFUL, note that $\tilde{\alpha} = \hat{Q}(\tilde{\delta}) = \sum_{i \neq j} \tilde{\varepsilon}_i P_{ij} \tilde{\varepsilon}_j / \tilde{\varepsilon}'\tilde{\varepsilon} = o_p(\mu_n^2/n)$ by Lemma A5, so by the formula for HFUL, $\hat{\alpha} = \tilde{\alpha} + O_p(1/n) = o_p(\mu_n^2/n)$. Thus, the result for HFUL will follow from the most general result for any $\hat{\alpha}$ with $\hat{\alpha} = o_p(\mu_n^2/n)$.

For any such $\hat{\alpha}$, by Lemma A4 we have

$$\begin{aligned} S_n'(\hat{\delta}-\delta_0)/\mu_n &= S_n'(\sum_{i\neq j} X_i P_{ij} X_j' - \hat{\alpha} X'X)^{-1} \sum_{i\neq j} (X_i P_{ij}\varepsilon_j - \hat{\alpha} X'\varepsilon)/\mu_n \\ &= [S_n^{-1}(\sum_{i\neq j} X_i P_{ij} X_j' - \hat{\alpha} X'X) S_n^{-1\prime}]^{-1} S_n^{-1} \sum_{i\neq j} (X_i P_{ij}\varepsilon_j - \hat{\alpha} X'\varepsilon)/\mu_n \\ &= (H_n + o_p(1))^{-1} o_p(1) \xrightarrow{p} 0. Q.E.D. \end{aligned}$$

Now we move on to asymptotic normality results. The next result is a central limit theorem that is proven in Chao et al. (2007).

Lemma A6 (Lemma A2 of Chao et al., 2009): *If i) P is a symmetric, idempotent matrix with rank$(P) = K$, $P_{ii} \le C < 1$; ii) $(W_{1n}, U_1, \varepsilon_1)$, ..., $(W_{nn}, U_n, \varepsilon_n)$ are independent and $D_n = \sum_{i=1}^n E[W_{in}W_{in}']$ is bounded; iii) $E[W_{in}'] = 0$, $E[U_i] = 0$, $E[\varepsilon_i] = 0$ and there exists a constant C such that $E[\|U_i\|^4] \le C$, $E[\varepsilon_i^4] \le C$; iv) $\sum_{i=1}^n E[\|W_{in}\|^4] \longrightarrow 0$; v) $K \longrightarrow \infty$; then for $\bar{\Sigma}_n \stackrel{\text{def}}{=} \sum_{i\neq j} P_{ij}^2 \left(E[U_iU_i']E[\varepsilon_j^2] + E[U_i\varepsilon_i]E[\varepsilon_j U_j']\right)/K$ and for any sequence of bounded nonzero vectors c_{1n} and c_{2n} such that $\Xi_n = c_{1n}'D_n c_{1n} + c_{2n}'\bar{\Sigma}_n c_{2n} > C$, it follows that*

$$Y_n = \Xi_n^{-1/2}(\sum_{i=1}^n c_{1n}' W_{in} + c_{2n}' \sum_{i\neq j} U_i P_{ij} \varepsilon_j/\sqrt{K}) \xrightarrow{d} N(0,1).$$

Let $\tilde{\alpha}(\delta) = \sum_{i\neq j} \varepsilon_i(\delta) P_{ij} \varepsilon_j(\delta)/\varepsilon(\delta)'\varepsilon(\delta)$ and

$$\hat{D}(\delta) = -[\varepsilon(\delta)'\varepsilon(\delta)/2]\partial[\sum_{i\neq j} \varepsilon_i(\delta) P_{ij} \varepsilon_j(\delta)/]/\partial\delta = \sum_{i\neq j} X_i P_{ij} \varepsilon_j(\delta) - \tilde{\alpha}(\delta) X'\varepsilon(\delta).$$

A couple of other intermediate results are also useful.

Lemma A7: *If Assumptions 1 - 4 are satisfied and $S_n'(\bar{\delta}-\delta_0)/\mu_n \xrightarrow{p} 0$ then*

$$-S_n^{-1}[\partial \hat{D}(\bar{\delta})/\partial\delta] S_n^{-1\prime} = H_n + o_p(1).$$

Proof: Let $\bar{\varepsilon} = \varepsilon(\bar{\delta}) = y - X\bar{\delta}$, $\bar{\gamma} = X'\bar{\varepsilon}/\bar{\varepsilon}'\bar{\varepsilon}$, and $\bar{\alpha} = \tilde{\alpha}(\bar{\delta})$. Then differentiating gives

$$\begin{aligned} -\frac{\partial \hat{D}}{\partial \delta}(\bar{\delta}) &= \sum_{i\neq j} X_i P_{ij} X_j' - \bar{\alpha} X'X - \bar{\gamma} \sum_{i\neq j} \bar{\varepsilon}_i P_{ij} X_j' - \sum_{i\neq j} X_i P_{ij} \bar{\varepsilon}_j \bar{\gamma}' + 2(\bar{\varepsilon}'\bar{\varepsilon})\bar{\alpha}\bar{\gamma}\bar{\gamma}' \\ &= \sum_{i\neq j} X_i P_{ij} X_j' - \bar{\alpha} X'X + \bar{\gamma}\hat{D}(\bar{\delta})' + \hat{D}(\bar{\delta})\bar{\gamma}', \end{aligned}$$

where the second equality follows by $\hat{D}(\bar{\delta}) = \sum_{i\neq j} X_i P_{ij}\bar{\varepsilon}_j - (\bar{\varepsilon}'\bar{\varepsilon})\bar{\alpha}\bar{\gamma}$. By Lemma A5 we have $\bar{\alpha} = o_p(\mu_n^2/n)$. By standard arguments, $\bar{\gamma} = O_p(1)$ so that $S_n^{-1}\bar{\gamma} = O_p(1/\mu_n)$. Then by Lemma A4 and $\hat{D}(\bar{\delta}) = \sum_{i\neq j} X_i P_{ij}\bar{\varepsilon}_j - \bar{\alpha}X'\bar{\varepsilon}$

$$S_n^{-1}\left(\sum_{i\neq j} X_i P_{ij} X_j' - \bar{\alpha}X'X\right)S_n^{-1\prime} = H_n + o_p(1), S_n^{-1}\hat{D}(\bar{\delta})\bar{\gamma}'S_n^{-1\prime} \xrightarrow{p} 0,$$

The conclusion then follows by T. Q.E.D.

LEMMA A8: *If Assumptions 1-4 are satisfied then for* $\gamma_n = \sum_i E[U_i\varepsilon_i]/\sum_i E[\varepsilon_i^2]$ *and* $\tilde{U}_i = U_i - \gamma_n\varepsilon_i$

$$S_n^{-1}\hat{D}(\delta_0) = \sum_{i=1}^n (1 - P_{ii})z_i\varepsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}\tilde{U}_i P_{ij}\varepsilon_j + o_p(1).$$

Proof: Note that for $W = z'(P - I)\varepsilon/\sqrt{n}$ by $I - P$ idempotent and $E[\varepsilon\varepsilon'] \leq CI_n$ we have

$$\begin{aligned} E[WW'] &\leq Cz'(I - P)z/n = C(z - Z\pi_{Kn}')'(I - P)(z - Z\pi_{Kn}')/n \\ &\leq CI_G\sum_{i=1}^n \|z_i - \pi_{Kn}Z_i\|^2/n \longrightarrow 0, \end{aligned}$$

so $z'(P - I)\varepsilon/\sqrt{n} = o_p(1)$. Also, by M

$$X'\varepsilon/n = \sum_{i=1}^n E[X_i\varepsilon_i]/n + O_p(1/\sqrt{n}), \varepsilon'\varepsilon/n = \sum_{i=1}^n \sigma_i^2/n + O_p(1/\sqrt{n}).$$

Also, by Assumption 3 $\sum_{i=1}^n \sigma_i^2/n \geq C > 0$. The delta method then gives $\tilde{\gamma} = X'\varepsilon/\varepsilon'\varepsilon = \gamma_n + O_p(1/\sqrt{n})$. Therefore, it follows by Lemma A1 and $\hat{D}(\delta_0) = \sum_{i\neq j} X_i P_{ij}\varepsilon_j - \varepsilon'\varepsilon\tilde{\alpha}(\delta_0)\tilde{\gamma}$ that

$$\begin{aligned} S_n^{-1}\hat{D}(\delta_0) &= \sum_{i\neq j} z_i P_{ij}\varepsilon_j/\sqrt{n} + S_n^{-1}\sum_{i\neq j}\tilde{U}_i P_{ij}\varepsilon_i - S_n^{-1}(\tilde{\gamma} - \gamma_n)\varepsilon'\varepsilon\tilde{\alpha}(\delta_0) \\ &= z'P\varepsilon/\sqrt{n} - \sum_i P_{ii}z_i\varepsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}\tilde{U}_i P_{ij}\varepsilon_j + O_p(1/\sqrt{n}\mu_n)o_p(\mu_n^2/n) \\ &= \sum_{i=1}^n (1 - P_{ii})z_i\varepsilon_i/\sqrt{n} + S_n^{-1}\sum_{i\neq j}\tilde{U}_i P_{ij}\varepsilon_j + o_p(1). Q.E.D. \end{aligned}$$

Proof of Theorem 2: Consider first the case where $\hat{\delta}$ is HLIM. Then by Theorem 1, $\hat{\delta} \xrightarrow{p} \delta_0$. First-order conditions for LIML are $\hat{D}(\hat{\delta}) = 0$. Expanding gives

$$0 = \hat{D}(\delta_0) + \frac{\partial \hat{D}}{\partial \delta}\left(\bar{\delta}\right)(\hat{\delta} - \delta_0),$$

where $\bar{\delta}$ lies on the line joining $\hat{\delta}$ and δ_0 and hence $\bar{\beta} = \mu_n^{-1} S_n'(\bar{\delta} - \delta_0) \xrightarrow{p} 0$. Then by Lemma A7, $\bar{H}_n = S_n^{-1}[\partial \hat{D}(\bar{\delta})/\partial \delta] S_n^{-1\prime} = H_P + o_p(1)$. Then $\partial \hat{D}(\bar{\delta})/\partial \delta$ is nonsingular w.p.a.1 and solving gives

$$S_n'(\hat{\delta} - \delta) = -S_n'[\partial \hat{D}(\bar{\delta})/\partial \delta]^{-1} \hat{D}(\delta_0) = -\bar{H}_n^{-1} S_n^{-1} \hat{D}(\delta_0).$$

Next, apply Lemma A6 with $U_i = \tilde{U}_i$ and

$$W_{in} = (1 - P_{ii}) z_i \varepsilon_i / \sqrt{n},$$

By ε_i having bounded fourth moment, and $P_{ii} \leq 1$,

$$\sum_{i=1}^{n} E\left[\|W_{in}\|^4\right] \leq C \sum_{i=1}^{n} \|z_i\|^4 / n^2 \longrightarrow 0.$$

By Assumption 6, we have $\sum_{i=1}^{n} E[W_{in} W_{in}'] \longrightarrow \Sigma_P$. Let $\Gamma = \operatorname{diag}(\Sigma_P, \Psi)$ and

$$A_n = \begin{pmatrix} \sum_{i=1}^{n} W_{in} \\ \sum_{i \neq j} \tilde{U}_i P_{ij} \varepsilon_j / \sqrt{K} \end{pmatrix}.$$

Consider c such that $c'\Gamma c > 0$. Then by the conclusion of Lemma A6 we have $c'A_n \xrightarrow{d} N(0, c'\Gamma c)$. Also, if $c'\Gamma c = 0$ then it is straightforward to show that $c'A_n \xrightarrow{p} 0$. Then it follows by the Cramer-Wold device that

$$A_n = \begin{pmatrix} \sum_{i=1}^{n} W_{in} \\ \sum_{i \neq j} \tilde{U}_i P_{ij} \varepsilon_j / \sqrt{K} \end{pmatrix} \xrightarrow{d} N(0, \Gamma), \Gamma = \operatorname{diag}(\Sigma_P, \Psi).$$

Next, we consider the two cases. Case I) has K/μ_n^2 bounded. In this case $\sqrt{K} S_n^{-1} \longrightarrow S_0$, so that

$$F_n \stackrel{def}{=} [I, \sqrt{K} S_n^{-1}] \longrightarrow F_0 = [I, S_0], F_0 \Gamma F_0' = \Sigma_P + S_0 \Psi S_0'.$$

Then by Lemma A8,

$$\begin{aligned} S_n^{-1}\hat{D}(\delta_0) &= F_n A_n + o_p(1) \xrightarrow{d} N(0, \Sigma_P + S_0 \Psi S_0'), \\ S_n'(\hat{\delta} - \delta_0) &= -\bar{H}_n^{-1} S_n^{-1} \hat{D}(\delta_0) \xrightarrow{d} N(0, \Lambda_I). \end{aligned}$$

In case II we have $K/\mu_n^2 \longrightarrow \infty$. Here

$$(\mu_n/\sqrt{K})F_n \longrightarrow \bar{F}_0 = [0, \bar{S}_0], \bar{F}_0 \Gamma \bar{F}_0' = \bar{S}_0 \Psi \bar{S}_0'$$

and $(\mu_n/\sqrt{K})o_p(1) = o_p(1)$. Then by Lemma A8,

$$\begin{aligned} (\mu_n/\sqrt{K})S_n^{-1}\hat{D}(\delta_0) &= (\mu_n/\sqrt{K})F_n A_n + o_p(1) \xrightarrow{d} N(0, \bar{S}_0 \Psi \bar{S}_0'), \\ (\mu_n/\sqrt{K})S_n'(\hat{\delta} - \delta_0) &= -\bar{H}_n^{-1}(\mu_n/\sqrt{K}) S_n^{-1} \hat{D}(\delta_0) \xrightarrow{d} N(0, \Lambda_{II}). Q.E.D. \end{aligned}$$

The next two results are useful for the proof of consistency of the variance estimator are taken from Chao et al. (2007). Let $\bar{\mu}_{Wn} = \max_{i \le n} |E[W_i]|$ and $\bar{\mu}_{Yn} = \max_{i \le n} |E[Y_i]|$.

Lemma A9 (Lemma A3 of Chao et al., 2009): *If $(W_i, Y_i), (i = 1, ..., n)$ are independent, W_i and Y_i are scalars then*

$$\sum_{i \neq j} P_{ij}^2 W_i Y_j = E[\sum_{i \neq j} P_{ij}^2 W_i Y_j] + O_p(\sqrt{K}(\bar{\sigma}_{Wn}\bar{\sigma}_{Yn} + \bar{\sigma}_{Wn}\bar{\mu}_{Yn} + \bar{\mu}_{Wn}\bar{\sigma}_{Yn})).$$

Lemma A10 (Lemma A4 of Chao et al., 2009): *If W_i, Y_i, η_i, are independent across i with $E[W_i] = a_i/\sqrt{n}$, $E[Y_i] = b_i/\sqrt{n}$, $|a_i| \le C$, $|b_i| \le C$, $E[\eta_i^2] \le C$, $Var(W_i) \le C\mu_n^{-2}$, $Var(Y_i) \le C\mu_n^{-2}$, there exists π_n such that $\max_{i \le n} |a_i - Z_i'\pi_n| \longrightarrow 0$, and $\sqrt{K}/\mu_n^2 \longrightarrow 0$ then*

$$A_n = E[\sum_{i \neq j \neq k} W_i P_{ik} \eta_k P_{kj} Y_j] = O(1), \sum_{i \neq j \neq k} W_i P_{ik} \eta_k P_{kj} Y_j - A_n \xrightarrow{p} 0.$$

Next, recall that $\hat{\varepsilon}_i = Y_i - X_i'\hat{\delta}, \hat{\gamma} = X'\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon}, \gamma_n = \sum_i E[X_i \varepsilon_i] / \sum_i \sigma_i^2$ and let

$$\begin{aligned} \breve{X}_i &= S_n^{-1}(X_i - \hat{\gamma}\hat{\varepsilon}_i) = S_n^{-1}\hat{X}_i, \dot{X}_i = S_n^{-1}(X_i - \gamma_n \varepsilon_i), \\ \breve{\Sigma}_1 &= \sum_{i \neq j \neq k} \breve{X}_i P_{ik} \hat{\varepsilon}_k^2 P_{kj} \breve{X}_j', \breve{\Sigma}_2 = \sum_{i \neq j} P_{ij}^2 \left(\breve{X}_i \breve{X}_i' \hat{\varepsilon}_j^2 + \breve{X}_i \hat{\varepsilon}_i \hat{\varepsilon}_j \breve{X}_j' \right), \\ \dot{\Sigma}_1 &= \sum_{i \neq j \neq k} \dot{X}_i P_{ik} \varepsilon_k^2 P_{kj} \dot{X}_j', \dot{\Sigma}_2 = \sum_{i \neq j} P_{ij}^2 \left(\dot{X}_i \dot{X}_i' \varepsilon_j^2 + \dot{X}_i \varepsilon_i \varepsilon_j \dot{X}_j' \right). \end{aligned}$$

Note that for $\hat{\Delta} = S_n'(\hat{\delta} - \delta_0)$ we have

$$\begin{aligned}
\hat{\varepsilon}_i - \varepsilon_i &= -X_i'(\hat{\delta} - \delta_0) = -X_i' S_n^{-1\prime}\hat{\Delta}, \\
\hat{\varepsilon}_i^2 - \varepsilon_i^2 &= -2\varepsilon_i X_i'(\hat{\delta} - \delta_0) + \left[X_i'(\hat{\delta} - \delta_0)\right]^2, \\
\breve{X}_i - \dot{X}_i &= -S_n^{-1}\hat{\gamma}(\hat{\varepsilon}_i - \varepsilon_i) - S_n^{-1}(\hat{\gamma} - \gamma_n)\varepsilon_i, \\
&= S_n^{-1}\hat{\gamma} X_i' S_n^{-1\prime}\hat{\Delta} - S_n^{-1}\mu_n(\hat{\gamma} - \gamma_n)(\varepsilon_i/\mu_n), \\
\breve{X}_i\hat{\varepsilon}_i - \dot{X}_i\varepsilon_i &= X_i\hat{\varepsilon}_i - \hat{\gamma}\hat{\varepsilon}_i^2 - X_i\varepsilon_i + \gamma_n\varepsilon_i^2, \\
&= -X_i X_i'(\hat{\delta} - \delta_0) - \hat{\gamma}\left\{-2\varepsilon_i X_i'(\hat{\delta} - \delta_0) + \left[X_i'(\hat{\delta} - \delta_0)^2\right]\right\} \\
&\quad -(\hat{\gamma} - \gamma_n)\varepsilon_i^2. \\
\left\|\breve{X}_i\breve{X}_i' - \dot{X}_i\dot{X}_i'\right\| &\leq \left\|\breve{X}_i - \dot{X}_i\right\|^2 + 2\left\|\dot{X}_i\right\|\left\|\breve{X}_i - \dot{X}_i\right\|
\end{aligned}$$

Lemma A11: *If the hypotheses of Theorem 3 are satisfied then* $\breve{\Sigma}_2 - \dot{\Sigma}_2 = o_p(K/\mu_n^2)$.

Proof: Note first that $S_n/\sqrt{n}$ is bounded so by the Cauchy-Schwartz inequality, $\|\Upsilon_i\| = \|S_n z_i/\sqrt{n}\| \leq C$. Let $d_i = C + |\varepsilon_i| + \|U_i\|$. Note that $\hat{\gamma} - \gamma_n \xrightarrow{p} 0$ by standard arguments. Then for $\hat{A} = (1 + \|\hat{\gamma}\|)(1 + \left\|\hat{\delta}\right\|) = O_p(1)$, and $\hat{B} = \|\hat{\gamma} - \gamma_n\| + \left\|\hat{\delta} - \delta_0\right\| \xrightarrow{p} 0$, we have

$$\begin{aligned}
\|X_i\| &\leq C + \|U_i\| \leq d_i, |\hat{\varepsilon}_i| \leq |X_i'(\delta_0 - \hat{\delta}) + \varepsilon_i| \leq C d_i\hat{A}, \\
\left\|\dot{X}_i\right\| &= \left\|S_n^{-1}(X_i - \gamma_n\varepsilon_i)\right\| \leq C\mu_n^{-1}d_i, \left\|\breve{X}_i\right\| = \left\|S_n^{-1}(X_i - \hat{\gamma}\hat{\varepsilon}_i)\right\| \leq C\mu_n^{-1}d_i\hat{A}, \\
\left\|\breve{X}_i\breve{X}_i' - \dot{X}_i\dot{X}_i'\right\| &\leq \left(\left\|\breve{X}_i\right\| + \left\|\dot{X}_i\right\|\right)\left\|\breve{X}_i - \dot{X}_i\right\| \leq C\mu_n^{-2}d_i\hat{A}\,\|\hat{\gamma}\|\,\|\hat{\varepsilon}_i - \varepsilon_i\| + \|\hat{\gamma} - \gamma_n\|\,|\varepsilon_i| \\
&\leq C\mu_n^{-2}d_i^2\hat{A}^2\hat{B}, \\
\left|\hat{\varepsilon}_i^2 - \varepsilon_i^2\right| &\leq (|\varepsilon_i| + |\hat{\varepsilon}_i|)\,|\hat{\varepsilon}_i - \varepsilon_i| \leq C d_i^2\hat{A}\hat{B}, \\
\left\|\breve{X}_i\hat{\varepsilon}_i - \dot{X}_i\varepsilon_i\right\| &= \left\|S_n^{-1}\left(X_i\hat{\varepsilon}_i - \hat{\gamma}\hat{\varepsilon}_i^2 - X_i\varepsilon_i + \gamma_n\varepsilon_i^2\right)\right\| \\
&\leq C\mu_n^{-1}\left(\|X_i\|\,|\hat{\varepsilon}_i - \varepsilon_i| + \|\hat{\gamma}\|\,|\hat{\varepsilon}_i^2 - \varepsilon_i^2| + \left|\varepsilon_i^2\right|\,\|\hat{\gamma} - \gamma_n\|\right) \\
&\leq C\mu_n^{-1}d_i^2(\hat{B} + \hat{A}^2\hat{B} + \hat{B}) \leq C d_i^2\hat{A}^2\hat{B}, \\
\left\|\breve{X}_i\hat{\varepsilon}_i\right\| &\leq C\mu_n^{-1}d_i^2\hat{A}^2, \left\|\dot{X}_i\varepsilon_i\right\| \leq C\mu_n^{-1}d_i^2.
\end{aligned}$$

Also note that

$$E\left[\sum_{i\neq j} P_{ij}^2 d_i^2 d_j^2 \mu_n^{-2}\right] \leq C\mu_n^{-2}\sum_{i,j} P_{ij}^2 = C\mu_n^{-2}\sum_i P_{ii} = C\mu_n^{-2}K.$$

so that $\sum_{i\neq j} P_{ij}^2 d_i^2 d_j^2 \mu_n^{-2} = O_p(K/\mu_n^2)$ by the Markov inequality. Then it follows that

$$\begin{aligned}\left\|\sum_{i\neq j} P_{ij}^2 \left(\breve{X}_i\breve{X}_i'\hat{\varepsilon}_j^2 - \dot{X}_i\dot{X}_i'\varepsilon_j^2\right)\right\| &\leq \sum_{i\neq j} P_{ij}^2 \left(\left|\hat{\varepsilon}_j^2\right| \left\|\breve{X}_i\breve{X}_i' - \dot{X}_i\dot{X}_i'\right\| + \left\|\dot{X}_i\right\|^2 \left|\hat{\varepsilon}_j^2 - \varepsilon_j^2\right|\right) \\ &\leq C\mu_n^{-2}\sum_{i\neq j} P_{ij}^2 d_i^2 d_j^2 (\hat{A}^4\hat{B} + \hat{A}\hat{B}) = o_p\left(K/\mu_n^2\right).\end{aligned}$$

We also have

$$\begin{aligned}\left\|\sum_{i\neq j} P_{ij}^2 \left(\breve{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\breve{X}_j' - \dot{X}_i\varepsilon_i\varepsilon_j\dot{X}_j\right)\right\| &\leq \sum_{i\neq j} P_{ij}^2 \left(\left\|\breve{X}_i\hat{\varepsilon}_i\right\| \left\|\breve{X}_j\hat{\varepsilon}_j - \dot{X}_j\varepsilon_j\right\| + \left\|\dot{X}_j\varepsilon_j\right\| \left\|\breve{X}_i\hat{\varepsilon}_i - \dot{X}_i\varepsilon_i\right\|\right) \\ &\leq C\mu_n^{-2}\sum_{i\neq j} P_{ij}^2 d_i^2 d_j^2 (1+\hat{A}^2)\hat{A}^2\hat{B} = o_p\left(\frac{K}{\mu_n^2}\right).\end{aligned}$$

The conclusion then follows by the triangle inequality. Q.E.D.

Lemma A12: *If the hypotheses of Theorem 3 are satisfied then* $\breve{\Sigma}_1 - \dot{\Sigma}_1 = o_p(K/\mu_n^2)$.

Proof: Note first that

$$\hat{\varepsilon}_i - \varepsilon_i = -X_i'(\hat{\delta} - \delta_0) = -X_i' S_n^{-1\prime} S_n'(\hat{\delta} - \delta_0) = -\left(z_i/\sqrt{n} + S_n^{-1}U_i\right)' \hat{\Delta} = -D_i'\hat{\Delta},$$

where $D_i = z_i/\sqrt{n} + S_n^{-1}U_i$ and $\hat{\Delta} = S_n'(\hat{\delta} - \delta_0)$. Also

$$\begin{aligned}\hat{\varepsilon}_i^2 - \varepsilon_i^2 &= -2\varepsilon_i X_i'(\hat{\delta} - \delta_0) + \left[X_i'(\hat{\delta} - \delta_0)\right]^2, \\ \breve{X}_i - \dot{X}_i &= -\hat{\gamma}\hat{\varepsilon}_i + \gamma_n\varepsilon_i = S_n^{-1}\hat{\gamma}D_i'\hat{\Delta} - S_n^{-1}\mu_n\left(\hat{\gamma} - \gamma_n\right)\varepsilon_i/\mu_n.\end{aligned}$$

We now have $\breve{\Sigma}_1 - \dot{\Sigma}_1 = \sum_{r=1}^7 T_r$ where

$$\begin{aligned}T_1 &= \sum_{i\neq j\neq k}\left(\breve{X}_i - \dot{X}_i\right) P_{ik}\left(\hat{\varepsilon}_k^2 - \varepsilon_k^2\right) P_{kj}\left(\breve{X}_j - \dot{X}_j\right)', T_2 = \sum_{i\neq j\neq k}\dot{X}_i P_{ik}\left(\hat{\varepsilon}_k^2 - \varepsilon_k^2\right) P_{kj}\left(\breve{X}_j - \dot{X}_j\right)' \\ T_3 &= \sum_{i\neq j\neq k}\left(\breve{X}_i - \dot{X}_i\right) P_{ik}\varepsilon_k^2 P_{kj}\left(\breve{X}_j - \dot{X}_j\right)', T_4 = T_2', T_5 = \sum_{i\neq j\neq k}\left(\breve{X}_i - \dot{X}_i\right) P_{ik}\varepsilon_k^2 P_{kj}\dot{X}_j', \\ T_6 &= \sum_{i\neq j\neq k}\dot{X}_i P_{ik}\left(\hat{\varepsilon}_k^2 - \varepsilon_k^2\right) P_{kj}\dot{X}_j', T_7 = T_5'.\end{aligned}$$

From the above expression for $\hat{\varepsilon}_i^2 - \varepsilon_i^2$ we see that T_6 is a sum of terms of the form $\hat{B}\sum_{i\neq j\neq k}\dot{X}_i P_{ik}\eta_i P_{kj}\dot{X}_j'$ where $\hat{B} \xrightarrow{p} 0$ and η_i is either a component of $-2\varepsilon_i X_i$ or of $X_i X_i'$.

By Lemma A10 we have $\sum_{i\neq j\neq k}\dot{X}_i P_{ik}\eta_i P_{kj}\dot{X}_j' = O_p(1)$, so by the triangle inequality $T_6 \xrightarrow{p} 0$. Also, note that

$$T_5 = S_n^{-1}\hat{\gamma}\hat{\Delta}'\sum_{i\neq j\neq k} D_i P_{ik}\varepsilon_k^2 P_{kj}\dot{X}_j' + S_n^{-1}\mu_n\left(\hat{\gamma}-\gamma_n\right)\sum_{i\neq j\neq k}(\varepsilon_i/\mu_n)P_{ik}\varepsilon_k^2 P_{kj}\dot{X}_j'.$$

Note that $S_n^{-1}\hat{\gamma}\hat{\Delta}' \xrightarrow{p} 0$, $E\left[D_i\right] = z_i/\sqrt{n}$, $Var(D_i) = O(\mu_n^{-2})$, $E[\dot{X}_i] = z_i/\sqrt{n}$, and $Var(\dot{X}) = O(\mu_n^{-2})$. Then by Lemma A10 it follows that $\sum_{i\neq j\neq k} D_i P_{ik}\varepsilon_k^2 P_{kj}\dot{X}_j' = O_p(1)$ so that the $S_n^{-1}\hat{\gamma}\hat{\Delta}'\sum_{i\neq j\neq k} D_i P_{ik}\varepsilon_k^2 P_{kj}\dot{X}_j' \xrightarrow{p} 0$. A similar argument applied to the second term and the triangle inequality then give $T_5 \xrightarrow{p} 0$. Also $T_7 = T_5' \xrightarrow{p} 0$.

Next, analogous arguments apply to T_2 and T_3, except that there are four terms in each of them rather than two, and also to T_1 except there are eight terms in T_1. For brevity we omit details. Q.E.D.

LEMMA A13: *If the hypotheses of Theorem 3 are satisfied then*

$$\dot{\Sigma}_2 = \sum_{i\neq j} P_{ij}^2 z_i z_i' \sigma_j^2/n + S_n^{-1}\sum_{i\neq j} P_{ij}^2\left(E[\tilde{U}_i\tilde{U}_i']\sigma_j^2 + E[\tilde{U}_i\varepsilon_i]E[\varepsilon_j\tilde{U}_j']\right)S_n^{-1\prime} + o_p(K/\mu_n^2).$$

Proof: Note that $Var(\varepsilon_i^2) \le C$ and $\mu_n^2 \le Cn$, so that for $u_{ki} = e_k' S_n^{-1}U_i$,

$$\begin{aligned}
E[(\dot{X}_{ik}\dot{X}_{i\ell})^2] &\le CE[\dot{X}_{ik}^4 + \dot{X}_{i\ell}^4] \le C\left\{z_{ik}^4/n^2 + E[u_k^4] + z_{i\ell}^4/n^2 + E[u_\ell^4]\right\} \le C\mu_n^{-4},\\
E[(\dot{X}_{ik}\varepsilon_i)^2] &\le CE[(z_{ik}^2\varepsilon_i^2/n + u_{ki}^2\varepsilon_i^2)] \le Cn^{-1} + C\mu_n^{-2} \le C\mu_n^{-2}.
\end{aligned}$$

Also, we have, for $\tilde{\Omega}_i = E[\tilde{U}_i\tilde{U}_i']$,

$$E[\dot{X}_i\dot{X}_i'] = z_i z_i'/n + S_n^{-1}\tilde{\Omega}_i S_n^{-1\prime}, E[\dot{X}_i\varepsilon_i] = S_n^{-1}E[\tilde{U}_i\varepsilon_i].$$

Next let W_i be $e_j'\dot{X}_i\dot{X}_i' e_k$ for some j and k, so that

$$\begin{aligned}
E[W_i] &= e_j' S_n^{-1}E[\tilde{U}_i\tilde{U}_i']S_n^{-1\prime}e_k + z_{ij}z_{ik}/n, |E[W_i]| \le C\mu_n^{-2}.\\
Var(W_i) &= Var\left\{\left(e_j' S_n^{-1}U_i + z_{ij}/\sqrt{n}\right)\left(e_k' S_n^{-1}U_i + z_{ik}/\sqrt{n}\right)\right\}\\
&\le C/\mu_n^4 + C/n\mu_n^2 \le C/\mu_n^4.
\end{aligned}$$

Also let $Y_i = \varepsilon_i^2$. Then $\sqrt{K}(\bar{\sigma}_{Wn}\bar{\sigma}_{Yn} + \bar{\sigma}_{Wn}\bar{\mu}_{Yn} + \bar{\mu}_{Wn}\bar{\sigma}_{Yn}) \leq CK^{1/2}/\mu_n^2$, so applying Lemma A9 for this W_i and Y_i gives

$$\sum_{i\neq j} P_{ij}^2 \dot{X}_i \dot{X}_i' \varepsilon_j^2 = \sum_{i\neq j} P_{ij}^2 \left(z_i z_i'/n + S_n^{-1}\tilde{\Omega}_i S_n^{-1\prime} \right) \sigma_j^2 + O_p(\sqrt{K}/\mu_n^2).$$

It follows similarly from Lemma A9 with W_i and Y_i equal to elements of $\dot{X}_i\varepsilon_i$ that

$$\sum_{i\neq j} P_{ij}^2 \dot{X}_i \varepsilon_i \varepsilon_j \dot{X}_j' = S_n^{-1} \sum_{i\neq j} P_{ij}^2 E[\tilde{U}_i\varepsilon_i]E[\varepsilon_j \tilde{U}_j'] S_n^{-1\prime} + O_p(\sqrt{K}/\mu_n^2).$$

Also, by $K \longrightarrow \infty$ we have $O_p(\sqrt{K}/\mu_n^2) = o_p(K/\mu_n^2)$. The conclusion then follows by T. Q.E.D.

Lemma A14: *If the hypotheses of Theorem 3 are satisfied then*

$$\dot{\Sigma}_1 = \sum_{i\neq j\neq k} z_i P_{ik} \sigma_k^2 P_{kj} z_j'/n + o_p(1).$$

Proof: Apply Lemma A10 with W_i equal to an element of $\dot{X}_i$, Y_j equal to an element of $\dot{X}_j$, and $\eta_k = \varepsilon_k^2$. Q.E.D.

Proof of Theorem 3: Note that $\bar{X}_i = \sum_{j=1}^n P_{ij}\hat{X}_j$,

$$\begin{aligned}
&\sum_{i=1}^n (\bar{X}_i\bar{X}_i' - \hat{X}_i P_{ii}\bar{X}_i' - \bar{X}_i P_{ii}\hat{X}_i')\hat{\varepsilon}_i^2 \\
= &\sum_{i,j,k=1}^n \hat{X}_i P_{ik}\hat{\varepsilon}_k^2 P_{kj}\hat{X}_j' - \sum_{i,j=1}^n \hat{X}_i P_{ii}\hat{\varepsilon}_i^2 P_{ij}\hat{X}_j' - \sum_{i,j=1}^n \hat{X}_i P_{ij}\hat{\varepsilon}_j^2 P_{jj}\hat{X}_j' \\
= &\sum_{i,j,k=1}^n \hat{X}_i P_{ik}\hat{\varepsilon}_k^2 P_{kj}\hat{X}_j' - \sum_{i\neq j} \hat{X}_i P_{ii}\hat{\varepsilon}_i^2 P_{ij}\hat{X}_j' - \sum_{i\neq j} \hat{X}_i P_{ij}\hat{\varepsilon}_j^2 P_{jj}\hat{X}_j' - 2\sum_{i=1}^n \hat{X}_i P_{ii}^2\hat{\varepsilon}_i^2\hat{X}_i' \\
= &\sum_{i,j,k\notin\{i,j\}}^n \hat{X}_i P_{ik}\hat{\varepsilon}_k^2 P_{kj}\hat{X}_j' - \sum_{i=1}^n \hat{X}_i P_{ii}^2\hat{\varepsilon}_i^2\hat{X}_i' \\
= &\sum_{i\neq j\neq k}^n \hat{X}_i P_{ik}\hat{\varepsilon}_k^2 P_{kj}\hat{X}_j' + \sum_{i\neq j}^n P_{ij}^2 \hat{X}_i\hat{X}_i\hat{\varepsilon}_j^2 - \sum_{i=1}^n \hat{X}_i P_{ii}^2\hat{\varepsilon}_i^2\hat{X}_i'.
\end{aligned}$$

Also, for Z_i' and $\tilde{Z}_i'$ equal to the ith row of Z and $\tilde{Z} = Z(Z'Z)^{-1}$ we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{\ell=1}^{K}\left(\sum_{i=1}^{n}\tilde{Z}_{ik}\tilde{Z}_{i\ell}\hat{X}_i\hat{\varepsilon}_i\right)\left(\sum_{j=1}^{n}Z_{jk}Z_{j\ell}\hat{X}_j\hat{\varepsilon}_j\right)' \\
= &\sum_{i,j=1}^{n}\left(\sum_{k=1}^{K}\sum_{\ell=1}^{K}\tilde{Z}_{ik}Z_{jk}\tilde{Z}_{i\ell}Z_{j\ell}\right)\hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j' = \sum_{i,j=1}^{n}(\sum_{k=1}^{K}\tilde{Z}_{ik}Z_{jk})^2\hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j' \\
= &\sum_{i,j=1}^{n}(\tilde{Z}_i'Z_j)^2\hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j' = \sum_{i,j=1}^{n}P_{ij}^2\hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j'
\end{aligned}$$

Adding this equation to the previous one then gives

$$\begin{aligned}
\hat{\Sigma} &= \sum_{i\neq j\neq k}\hat{X}_iP_{ik}\hat{\varepsilon}_k^2P_{kj}\hat{X}_j' + \sum_{i\neq j}P_{ij}^2\hat{X}_i\hat{X}_i'\hat{\varepsilon}_j^2 - \sum_{i=1}^{n}\hat{X}_iP_{ii}^2\hat{\varepsilon}_i^2\hat{X}_i' + \sum_{i,j=1}^{n}P_{ij}^2\hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j' \\
&= \sum_{i\neq j\neq k}\hat{X}_iP_{ik}\hat{\varepsilon}_k^2P_{kj}\hat{X}_j' + \sum_{i\neq j}P_{ij}^2(\hat{X}_i\hat{X}_i'\hat{\varepsilon}_j^2 + \hat{X}_i\hat{\varepsilon}_i\hat{\varepsilon}_j\hat{X}_j').
\end{aligned}$$

It then follows that $S_n^{-1}\hat{\Sigma}S_n^{-1\prime} = \breve{\Sigma}_1 + \breve{\Sigma}_2$, so that

$$S_n'\hat{V}S_n = (S_n^{-1}\hat{H}S_n^{-1\prime})^{-1}S_n^{-1}\hat{\Sigma}S_n^{-1\prime}(S_n^{-1}\hat{H}S_n^{-1\prime})^{-1} = (S_n^{-1}\hat{H}S_n^{-1\prime})^{-1}(\breve{\Sigma}_1+\breve{\Sigma}_2)(S_n^{-1}\hat{H}S_n^{-1\prime})^{-1}.$$

By Lemma A4 we have $S_n^{-1}\hat{H}S_n^{-1\prime} \xrightarrow{p} H_P$. Also, note that for $\bar{z}_i = \sum_j P_{ij}z_i = e_i'Pz$,

$$\begin{aligned}
\sum_{i\neq j\neq k}z_iP_{ik}\sigma_k^2P_{kj}z_j'/n &= \sum_i\sum_{j\neq i}\sum_{k\notin\{i,j\}}z_iP_{ik}\sigma_k^2P_{kj}z_j'/n \\
&= \sum_i\sum_{j\neq i}\left(\sum_k z_iP_{ik}\sigma_k^2P_{kj}z_j' - z_iP_{ii}\sigma_i^2P_{ij}z_j' - z_iP_{ij}\sigma_j^2P_{jj}z_j'\right)/n \\
&= (\sum_k\bar{z}_k\sigma_k^2\bar{z}_k' - \sum_{i,k}P_{ik}^2z_iz_i'\sigma_k^2 - \sum_i z_iP_{ii}\sigma_i^2\bar{z}_i' + \sum_i z_iP_{ii}\sigma_i^2P_{ii}z_i' \\
&\quad -\sum_j\bar{z}_j\sigma_j^2P_{jj}z_j' + \sum_i z_jP_{jj}\sigma_j^2P_{jj}z_j')/n \\
&= \sum_i\sigma_i^2\left(\bar{z}_i\bar{z}_i' - P_{ii}z_i\bar{z}_i' - P_{ii}\bar{z}_iz_i' + P_{ii}^2z_iz_i'\right)/n - \sum_{i\neq j}P_{ij}^2z_iz_i'\sigma_j^2/n.
\end{aligned}$$

Also, it follows similarly to the proof of Lemma A8 that $\sum_i \|z_i - \bar{z}_i\|^2/n \leq z'(I -$

$P)z/n \longrightarrow 0$. Then by σ_i^2 and P_{ii} bounded we have

$$\begin{aligned}
\left\|\sum_i \sigma_i^2(\bar{z}_i\bar{z}_i' - z_iz_i')/n\right\| &\leq \sum_i \sigma_i^2(2\|z_i\|\|z_i-\bar{z}_i\| + \|z_i-\bar{z}_i\|^2)/n \\
&\leq C(\sum_i \|z_i\|^2/n)^{1/2}(\sum_i \|z_i-\bar{z}_i\|^2/n)^{1/2} + C\sum_i \|z_i-\bar{z}_i\|^2/n \longrightarrow 0, \\
\left\|\sum_i \sigma_i^2 P_{ii}(z_i\bar{z}_i' - z_iz_i')/n\right\| &\leq (\sum_i \sigma_i^4 P_{ii}^2\|z_i\|^2/n)^{1/2}(\sum_i \|z_i-\bar{z}_i\|^2/n)^{1/2} \longrightarrow 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{i\neq j\neq k} z_iP_{ik}\sigma_k^2P_{kj}z_j'/n &= \sum_i \sigma_i^2(1-P_{ii})^2 z_iz_i'/n + o(1) - \sum_{i\neq j} P_{ij}^2 z_iz_i'\sigma_j^2/n \\
&= \Sigma_P - \sum_{i\neq j} P_{ij}^2 z_iz_i'\sigma_j^2/n + o(1).
\end{aligned}$$

It then follows by Lemmas A10-A14 and the triangle inequality that

$$\begin{aligned}
\breve{\Sigma}_1 + \breve{\Sigma}_2 &= \sum_{i\neq j\neq k} z_iP_{ik}\sigma_k^2P_{kj}z_j'/n + \sum_{i\neq j} P_{ij}^2 z_iz_i'\sigma_j^2/n \\
&\quad + S_n^{-1}\sum_{i\neq j} P_{ij}^2\left(E[\tilde{U}_i\tilde{U}_i']\sigma_j^2 + E[\tilde{U}_i\varepsilon_i]E[\varepsilon_j\tilde{U}_j']\right)S_n^{-1\prime} + o_p(1) + o_p(K/\mu_n^2) \\
&= \Sigma_P + KS_n^{-1}(\Psi + o(1))S_n^{-1\prime} + o_p(1) + o_p(K/\mu_n^2) \\
&= \Sigma_P + KS_n^{-1}\Psi S_n^{-1\prime} + o_p(1) + o_p(K/\mu_n^2).
\end{aligned}$$

Then in case I) we have $o_p(K/\mu_n^2) = o_p(1)$ so that

$$S_n'\hat{V}S_n = H^{-1}\left(\Sigma_P + KS_n^{-1}\Psi S_n^{-1\prime}\right)H^{-1} + o_p(1) = \Lambda_I + o_p(1).$$

In case II) we have $(\mu_n^2/K)\,o_p(1) \xrightarrow{p} 0$, so that

$$\left(\mu_n^2/K\right)S_n'\hat{V}S_n = H^{-1}\left(\left(\mu_n^2/K\right)\Sigma_P + \mu_n^2S_n^{-1}\Psi S_n^{-1\prime}\right)H^{-1} + o_p(1) = \Lambda_{II} + o_p(1).$$

Next, consider case I) and note that $S_n'(\hat{\delta}-\delta_0) \xrightarrow{d} Y \sim N(0,\Lambda_I)$, $S_n'\hat{V}S_n \xrightarrow{p} \Lambda_I$, $c'\sqrt{K}S_n^{-1\prime} \rightarrow c'S_0'$, and $c'S_0'\Lambda_IS_0c \neq 0$. Then by the continuous mapping and Slutzky theorems,

$$\begin{aligned}
\frac{c'(\hat{\delta}-\delta_0)}{\sqrt{c'\hat{V}c}} &= \frac{c'S_n^{-1\prime}S_n'(\hat{\delta}-\delta_0)}{\sqrt{c'S_n^{-1\prime}S_n'\hat{V}S_nS_n^{-1}c}} = \frac{c'\sqrt{K}S_n^{-1\prime}S_n'(\hat{\delta}-\delta_0)}{\sqrt{c'\sqrt{K}S_n^{-1\prime}S_n'\hat{V}S_nS_n^{-1}\sqrt{K}c}} \\
\xrightarrow{d} \frac{c'S_0'Y}{\sqrt{c'S_0'\Lambda_IS_0c}} &\sim N(0,1).
\end{aligned}$$

For case II), $\left(\mu_n/\sqrt{K}\right) S_n'(\hat{\delta}-\delta_0) \xrightarrow{d} \bar{Y} \sim N(0,\Lambda_{II})$, $(\mu_n^2/K)\, S_n'\hat{V}S_n \xrightarrow{p} \Lambda_{II}$, $c'\mu_n S_n^{-1\prime} \longrightarrow c'\bar{S}_0'$, and $c'\bar{S}_0'\Lambda_{II}\bar{S}_0 c \neq 0$. Then

$$\begin{aligned}\frac{c'(\hat{\delta}-\delta_0)}{\sqrt{c'\hat{V}c}} &= \frac{c'S_n^{-1\prime}\left(\mu_n/\sqrt{K}\right) S_n'(\hat{\delta}-\delta_0)}{\sqrt{c'S_n^{-1\prime}\,(\mu_n^2/K)\,S_n'\hat{V}S_nS_n^{-1}c}} \\ &= \frac{c'\mu_n S_n^{-1\prime}\left(\mu_n/\sqrt{K}\right) S_n'(\hat{\delta}-\delta_0)}{\sqrt{c'\mu_n S_n^{-1\prime}\,(\mu_n^2/K)\,S_n'\hat{V}S_nS_n^{-1}\mu_n c}} \xrightarrow{d} \frac{c'\bar{S}_0'\bar{Y}}{\sqrt{c'\bar{S}_0'\Lambda_{II}\bar{S}_0 c}} \sim N(0,1).Q.E.D.\end{aligned}$$

References

Ackerberg, D.A. and P. Devereux (2003): "Improved JIVE Estimators for Overidentified Models with and without Heteroskedasticity," Working Paper, UCLA.

Anderson, T.W., N. Kunitomo, and T. Sawa (1982): "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator," *Econometrica* 50, 1009-1027.

Andrews, D.W.K. and J.H. Stock (2007): "Testing with Many Weak Instruments," *Journal of Econometrics* 138, 24-46.

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

Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica* 63, 657-681.

Bekker, P.A. and J. van der Ploeg (2005): "Instrumental Variable Estimation Based on Grouped Data," *Statistica Neerlandica* 59, 239-267.

Blomquist, S. and M. Dahlberg (1999): "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments," *Journal of Applied Econometrics* 14, 69-88.

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

Chao, J. and N. Swanson (2004): "Estimation and Testing Using Jackknife IV in Heteroskedastic Regressions With Many Weak Instruments," Working Paper, Rutgers University.

Chao, J. and N. Swanson (2005): "Consistent Estimation With a Large Number of Weak Instruments," *Econometrica* 73, 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, Rutgers.

Davidson, R. and J.G. MacKinnon (2006): "The Case Against Jive," (with discussion and reply), *Journal of Applied Econometrics* 21, 827-833.

Donald, S.G. and W.K Newey (2000): "A Jackknife Interpretation of the Continuous Updating Estimator," *Economics Letters* 67, 239-243.

Fuller, W.A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica* 45, 939-954.

Hahn, J. and J.A. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables", *Econometrica* 70*,* 163-189.

Hahn, J. and A. Inoue (2002): "A Monte Carlo Comparison of Various Asymptotic Approximations to the Distribution of Instrumental Variables Estimators," *Econometric Reviews* 21, 309-336.

Hahn, J., J.A. Hausman, and G.M. Kuersteiner (2004): "Estimation with Weak Instruments: Accuracy of Higher-order Bias and MSE Approximations," *Econometrics Journal, Volume 7.*

Han, C. and P.C.B. Phillips (2006): “GMM With Many Moment Conditions,” *Econometrica* 74, 147-192.

Hansen, C., J.A. Hausman, W.K. Newey (2008): “Estimation with Many Instrumental Variables,” *Journal of Business and Economic Statistics* 26, 398-422.

Hansen, L.P., J. Heaton and A. Yaron (1996): “Finite-Sample Properties of Some Alternative GMM Estimators,” *Journal of Business and Economic Statistics* 14, 262-280.

van Hasselt, M. (2000): ”The Instrumental Variables Estimator and the Effects of non-Gaussian Disturbances,” working paper.

Kunitomo, N. (1980): “Asymptotic Expansions of Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations,” *Journal of the American Statistical Association* 75, 693-700.

Nagar, A.L. ”The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations,” *Econometrica* 27, 575-595.

Newey, W.K. (1990): ”Efficient Instrumental Variables Estimation of Nonlinear Models,” *Econometrica* 58, 809-837.

Newey, W.K. and Windmeier (2009): “GMM with Many Weak Moment Conditions Many Moments” *Econometrica* 77, 687-719.

Phillips, G.D.A. and C. Hale (1977): “The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems,” *International Economic Review* 18, 219-228.

Smith, R.J. (1997): “Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation,” *Economic Journal* 107, 503-519.

Staiger, D. and J. Stock (1997): “Instrumental Variables Regression with Weak Instruments”, *Econometrica* 65, 57-586.

Stock, J. and M. Yogo (2005): “Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments,” Chapter 6, in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Andrews, D.W.K. and J.H. Stock eds., Cambridge: Cambridge University Press.

White, H. (1982): ”Instrumental Variables Regression with Independent Observations,” *Econometrica* 50, 483-500.