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
John Hopkins University

John Chao
Department of Economics
University of Maryland

Norman Swanson
Department of Economics
Rutgers University

August 2006
Revised September 2007

Abstract

It is common practice in econometrics to correct for heteroskedasticity. This paper corrects instrumental variables estimators with many instruments for heteroskedasticity. We give heteroskedasticity robust versions of the limited information maximum likelihood (LIML) and Fuller (1977, FULL) estimators; as well as heteroskedasticity consistent standard errors thereof. The estimators are based on removing the own observation terms in the numerator of the LIML variance ratio. We derive asymptotic properties of the estimators under many and many weak instruments setups. Based on a series of Monte Carlo experiments, we find that the estimators perform as well as LIML or FULL 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 A. Kowalski, R. Lewis, and K. Menzel.

1 Introduction

It is common practice in econometrics to correct standard errors for heteroskedasticity. A leading example of such correction is least squares with heteroskedasticity consistent standard errors, which is ubiquitous. Additionally, two-stage least squares (2SLS) with heteroskedasticity consistent standard errors is often used, in exactly identified models. However, such corrections seem not to be available for the Fuller (1977, FULL) and limited information maximum likelihood (LIML) estimators, in overidentified models. This perhaps surprising, given that FULL and LIML have better properties than 2SLS (see e.g. Hahn and Inoue (2002), Hahn and Hausman (2002), and Hansen, Hausman, and Newey, (2007)). The purpose of this paper is to correct these methods for heteroskedasticity under many instruments, and we shall see that it is necessary to correct both the estimators and the standard errors.

LIML and FULL are inconsistent with many instruments and heteroskedasticity, as pointed out for the case of dummy instruments and LIML by Bekker and van der Ploeg (2005), and more generally by Chao and Swanson (2004).[1] Here we give a general characterization of this inconsistency. More importantly, we propose heteroskedasticity robust versions of FULL and LIML, namely HFUL and HLIM, respectively. HLIM is a jackknife version of LIML that deletes own observation terms in the numerator of the variance ratio; and like LIML, HLIM is invariant to normalization. Also, HLIM can be interpreted as a linear combination of forward and reverse jackknife instrumental variable (JIV) estimators, analogous to Hahn and Hausman's (2002) interpretation of LIML as a linear combination of forward and reverse two-stage least squares estimators. For each estimator we also give heteroskedasticity consistent standard errors that adjust for the presence of many instruments.

We show that HLIM and HFUL are as efficient as FULL and LIML under homoskedasticity and the many weak instruments sequence of Chao and Swanson (2005). Under the many instruments sequence of Kunitomo (1980) and Bekker (1994) we show that HLIM

[1] See also Ackerberg and Devereux (2003).

may be more or less efficient than LIML. We argue that these efficiency differences will tend to be small in most applications, where the number of instrumental variables is small relative to the sample size.

The HFUL and HLIM estimators and their associated standard errors are quite simple to compute. However, similarly to least squares not being efficient under heteroskedasticity, HFUL and HLIM are also not efficient under heteroskedasticity and many instruments. Recent results of Newey and Windmeijer (2007) suggest that the continuous updating estimator (CUE) of Hansen, Heaton, and Yaron (1996) and other generalized empirical likelihood estimators (see e.g. Smith (2004)) are efficient. These estimators are quite difficult to compute, though. To address this problem, we give a linearized, jackknife version of the continuous updating estimator that is easier to compute, and for which HLIM provides simple starting values. In Monte Carlo work we do not find much advantage to using the CUE, and no advantage to using its linearized version, relative to HFUL and HLIM.

One important precedent to the research discussed in this paper is Hahn and Hausman (2002), who considered combining forward and reverse IV estimators. JIV estimators were proposed by Phillips and Hale (1977), Blomquist and Dahlberg (1999), Angrist and Imbens and Krueger (1999), and Ackerberg and Deveraux (2003). Chao and Swanson (2004) have previously given heteroskedasticity consistent standard errors and shown asymptotic normality for JIV, under many weak instruments. Newey and Windmeijer (2007) considered efficiency of IV estimators with heteroskedasticity and many weak instruments.

In a series of Monte Carlo experiments, we show that the HFUL and HLIM are approximately as efficient as LIML under homoskedasticity, unlike the JIV estimator, that was shown to perform poorly relative to LIML by Davidson and MacKinnon (2006). Also, HFUL has less bias and dispersion than FULL in most of the cases that we consider, under heteroskedasticity. These results suggest that the new estimators are promising heteroskedasticity robust and efficient alternatives to FULL, LIML, and other estimators, under many instruments.

The rest of the paper is organized as follows. In the next section, the model is outlined, and previous estimators are summarized. In Section 3, heteroskedasticity robust LIML and FULL estimators are presented; while Section 4 discusses efficiency of these estimators. Section 5 outlines how to use the same jackknifing approach used in the construction of HLIM and HFUL in order to construct a robust CUE. Asymptotic theory is gathered in Section 6, and Monte Carlo findings are presented in Section 7. All proofs are gathered in Section 8.

2 The Model and Previous Estimators

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.

This model allows for Υ to be a linear combination of Z, i.e. $\Upsilon = Z\pi$ for some $K \times G$ matrix π. Furthermore, some columns of X may be exogenous, with the corresponding column of U being zero. The model also allows for 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 define $\Upsilon_i = f_0(w_i)$ to be a vector of unknown functions of a vector w_i of underlying instruments, and $Z_i = (p_{1K}(w_i), ..., p_{KK}(w_i))'$ for 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 Donald and Newey (2001)).

To describe estimators in the extant literature, let $P = Z(Z'Z)^{-1}Z'$. The LIML estimator, $\tilde{\delta}^*$, is given by

$$\tilde{\delta}^* = \arg\min_{\delta} \hat{Q}^*(\delta), \hat{Q}^*(\delta) = \frac{(y - X\delta)'P(y - X\delta)}{(y - X\delta)'(y - X\delta)}.$$

FULL is obtained as

$$\breve{\delta}^* = (X'PX - \breve{\alpha}^*X'X)^{-1}(X'Py - \breve{\alpha}^*X'y),$$

for $\breve{\alpha}^* = [\tilde{\alpha}^* - (1 - \tilde{\alpha}^*)C/T]/[1 - (1 - \tilde{\alpha}^*)C/T]$, $\tilde{\alpha}^* = Q(\tilde{\delta}^*)$, and $C > 0$. FULL has moments of all orders, is approximately mean unbiased for $C = 1$, and is second order admissible for $C \geq 4$, under homoskedasticity and standard large sample asymptotics. Both LIML and FULL are members of a class of estimators of the form

$$\hat{\delta}^* = (X'PX - \hat{\alpha}^*X'X)^{-1}(X'Py - \hat{\alpha}^*X'y).$$

For example, LIML has this form for $\hat{\alpha}^* = \tilde{\alpha}^*$, FULL for $\hat{\alpha}^* = \breve{\alpha}^*$, and 2SLS for $\hat{\alpha}^* = 0$.

We can use the objective functions that these estimators minimize in order to characterize the problem with heteroskedasticity and many instruments. If the limit of the objective function is not minimized at the true parameter, then the estimator will not be consistent. For expository purposes, first consider 2SLS, which has the following objective function

$$\hat{Q}_{2SLS}(\delta) = (y - X\delta)'P(y - X\delta)/n = \sum_{i \neq j}(y_i - X_i'\delta)P_{ij}(y_j - X_j'\delta)/n + \sum_{i=1}^{n} P_{ii}(y_i - X_i'\delta)^2/n.$$

This objective function is a quadratic form that, like a sample average, will be close to its expectation in large samples. Its expectation is

$$E\left[\hat{Q}_{2SLS}(\delta)\right] = (\delta - \delta_0)' \sum_{i \neq j} \Upsilon_i P_{ij} \Upsilon_j'(\delta - \delta_0)/n + \sum_{i=1}^{n} P_{ii} E[(y_i - X_i'\delta)^2]/n$$

Asymptotically, the first term following the above equality will be minimized at δ_0, under certain regularity conditions. The second term is an expected squared residual that will not be minimized at δ_0 due to endogeneity. With many instruments

$$P_{ii} \nrightarrow 0,$$

so that the second term does not vanish asymptotically. Hence, with many instruments, 2SLS is not consistent, even under homoskedasticity, as pointed out by Bekker (1994).

For LIML, we can (asymptotically) replace the objective function, $\hat{Q}^*(\delta)$, with a corresponding ratio of expectations giving

$$\frac{E[(y-X\delta)'P(y-X\delta)]}{E[(y-X\delta)'(y-X\delta)]}=\frac{(\delta-\delta_0)'\sum_{i\neq j}^n P_{ij}\Upsilon_i\Upsilon_j'(\delta-\delta_0)}{\sum_{i=1}^n E[(y_i-X_i'\delta)^2]}+\frac{\sum_{i=1}^n P_{ii}E[(y_i-X_i'\delta)^2]}{\sum_{i=1}^n E[(y_i-X_i'\delta)^2]}.$$

Here, we again see that the first term following the equality will be minimized at δ_0 asymptotically. Under heteroskedasticity, 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 objective function will not have zero derivative at δ_0 asymptotically so that it is not minimized at δ_0. When this covariance does have a zero limit then it can be shown that the ratio of expectations will be minimized at δ_0 as long as for $\Omega_i=E[U_iU_i']$ the matrix

$$\left(1-\frac{\sum_{i=1}^n\sigma_i^2P_{ii}}{\sum_{i=1}^n\sigma_i^2}\right)\sum\Upsilon_i\Upsilon_i'/n+\sum_i P_{ii}\Omega_i/n-\frac{\sum_{i=1}^n\sigma_i^2P_{ii}}{\sum_{i=1}^n\sigma_i^2}\sum_{i=1}^n\Omega_i/n$$

has a positive definite limit. For the homoskedastic case it is known that LIML is consistent under many or many weak instruments (see e.g. Bekker (1994) and Chao and Swanson (2004)).

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.

LIML is inconsistent when $P_{ii} = Z_i'(Z'Z)^{-1}Z_i$ (roughly speaking this is the size of the i^{th} instrument observation) is correlated with γ_i. This can easily happen when (say) there is more heteroskedasticity in σ_i^2 than $E[X_i\varepsilon_i]$. Bekker and van der Ploeg (2005) and Chao and Swanson (2004) pointed out that LIML can be inconsistent with heteroskedasticity; but this appears to be the first statement of the critical condition that $\widehat{Cov_{\sigma^2}(P_{ii}}, \gamma_i) = 0$ for consistency of LIML.

The lack of consistency of these estimators under many instruments and heteroskedasticity can be attributed to the presence of the $i = j$ terms in their objective functions. The estimators can be made robust to heteroskedasticity by dropping these terms. Doing this for 2SLS gives

$$\bar{\delta} = \arg\min_{\delta} \sum_{i \neq j} (y_i - X_i'\delta) P_{ij} (y_j - X_j'\delta)/n$$

Solving for $\bar{\delta}$ gives

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

This is the JIV2 estimator of Angrist, Imbens, and Krueger (1994). Because the normal equations remove the $i = j$ terms, this estimator is consistent. It was pointed out by Ackerberg and Devereux (2003) and Chao and Swanson (2004) that this estimator is consistent under many weak instruments and heteroskedasticity. However, under homoskedasticity and many weak instruments, this estimator is not efficient; and Davidson and MacKinnon (2005) argued that it additionally has inferior small sample properties under homoskedasticity, when compared with LIML. The estimators that we give overcome these problems.

3 Heteroskedasticity Robust LIML and FULL

The heteroskedasticity robust LIML estimator (HLIM) is obtained by dropping the $i = j$ terms from the numerator of the LIML objective function, so that

$$\tilde{\delta} = \arg\min_{\delta} \hat{Q}(\delta), \hat{Q}(\delta) = \frac{\sum_{i \neq j} (y_i - X_i'\delta) P_{ij} (y_j - X_j'\delta)}{(y - X\delta)'(y - X\delta)}.$$

Like the jackknife IV estimator, $\tilde{\delta}$ will be consistent under heteroskedasticity because the $i = j$ terms have been removed from the numerator. In the sequel, we will show that this estimator (and an asymptotic variance estimator thereof) is consistent and asymptotically normal.

As is the case for LIML, this estimator is invariant to normalization. Let $\bar{X} = [y, X]$. Then $\tilde{d} = (1, -\tilde{\delta}')'$ solves

$$\min_{d: d_1 = 1} \frac{d' \left(\sum_{i \neq j} \bar{X}_i P_{ij} \bar{X}_j' \right) d}{d' \bar{X}' \bar{X} d}.$$

Another normalization, such as imposing that another d is equal to 1 would produce the same estimator, up to the normalization.

Also, computation of this estimator is straightforward. Similarly to LIML, $\tilde{\alpha} = \hat{Q}(\tilde{\delta})$ is the smallest eigenvalue of $(\bar{X}'\bar{X})^{-1} \sum_{i \neq j} \bar{X}_i P_{ij} \bar{X}_j'$. Also, first order conditions for $\tilde{\delta}$ are

$$0 = \sum_{i \neq j} X_i P_{ij} \left(y_j - X_j' \tilde{\delta} \right) - \tilde{\alpha} \sum_i X_i (y_i - X_i' \tilde{\delta}).$$

Solving these conditions gives

$$\tilde{\delta} = \left(\sum_{i \neq j} X_i P_{ij} X_j' - \tilde{\alpha} X'X \right)^{-1} \left(\sum_{i \neq j} X_i P_{ij} y_j - \tilde{\alpha} X'y \right).$$

This estimator has a similar form to LIML except that the $i = j$ terms have been deleted from the double sums.

It is interesting to note that LIML and HLIM coincide when P_{ii} is constant. In that case,

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

so that the LIML objective function equals the HLIM objective function plus a constant. This explains why constant P_{ii} will lead to LIML being consistent under heteroskedasticity.

HLIM is a member of a class of jackknife estimators having the form

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

The JIV estimator is obtained by setting $\hat{\alpha} = 0$. A heteroskedasticity consistent version of FULL, namely HFUL, is obtained by replacing $\tilde{\alpha}$ with $\hat{\alpha} = [\tilde{\alpha} - (1-\tilde{\alpha})C/T]/[1-(1-\tilde{\alpha})C/T]$ for some $C > 0$. The small sample properties of this estimator are unknown, but we expect its performance relative to HLIM to be similar to that of FULL relative to LIML. As pointed out by Hahn, Hausman, and Kuersteiner (2004), FULL has much smaller dispersion than LIML with weak instruments, so we expect the same for HFUL. Monte Carlo results given below confirm these properties.

An asymptotic variance estimator is useful for constructing large sample confidence intervals and tests. To describe it, let $\hat{\varepsilon}_i = y_i - X_i'\hat{\delta}$, $\hat{\gamma} = X'\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon}$, $\hat{X} = X - \hat{\varepsilon}\hat{\gamma}'$,

$$\hat{H} = \sum_{i\neq j} X_i P_{ij} X_j' - \hat{\alpha} X'X, \hat{\Sigma} = \sum_{i,j=1}^{n} \sum_{k\notin\{i,j\}} \hat{X}_i P_{ik} \hat{\varepsilon}_k^2 P_{kj} \hat{X}_j' + \sum_{i\neq j} P_{ij}^2 \hat{X}_i \hat{\varepsilon}_i \hat{\varepsilon}_j \hat{X}_j'.$$

The variance estimator is

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

We can interpret the HLIM estimator, $\tilde{\delta}$, as a combination of forward and reverse jackknife IV (JIV) estimators. For simplicity, we give this interpretation in the scalar δ case. Let $\tilde{\varepsilon}_i = y_i - X_i'\tilde{\delta}$ and $\tilde{\gamma} = \sum_i X_i \tilde{\varepsilon}_i / \sum_i \tilde{\varepsilon}_i^2$. First-order conditions for $\tilde{\delta}$ are

$$0 = -\frac{\partial \hat{Q}(\tilde{\delta})}{\partial \delta} \sum_i \tilde{\varepsilon}_i^2/2 = \sum_{i\neq j} (X_i - \tilde{\gamma}\tilde{\varepsilon}_i) P_{ij} (y_j - X_j'\tilde{\delta}) = \sum_{i\neq j} [(1+\tilde{\gamma}\tilde{\delta})X_i - \tilde{\gamma} y_i] P_{ij} (y_j - X_j'\tilde{\delta}).$$

The forward JIV estimator $\bar{\delta}$ is

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

The reverse JIV is obtained as follows. Dividing the structural equation by δ_0 gives

$$X_i = y_i/\delta_0 - \varepsilon_i/\delta_0.$$

Applying JIV to this equation in order to estimate $1/\delta_0$, and then inverting, gives the reverse JIV estimator

$$\bar{\delta}^r = \left(\sum_{i \neq j} y_i P_{ij} X_j \right)^{-1} \sum_{i \neq j} y_i P_{ij} y_j.$$

Then, collecting terms in the first-order conditions for HLIM gives

$$\begin{aligned} 0 &= (1 + \tilde{\gamma}\tilde{\delta}) \sum_{i \neq j} X_i P_{ij} (y_j - X_j' \tilde{\delta}) - \tilde{\gamma} \sum_{i \neq j} y_i P_{ij} (y_j - X_j' \tilde{\delta}) \\ &= (1 + \tilde{\gamma}\tilde{\delta}) \sum_{i \neq j} X_i P_{ij} X_j (\bar{\delta} - \tilde{\delta}) - \tilde{\gamma} \sum_{i \neq j} y_i P_{ij} X_j (\bar{\delta}^r - \tilde{\delta}). \end{aligned}$$

Dividing through by $\sum_{i \neq j} X_i P_{ij} X_j$ gives

$$0 = (1 + \tilde{\gamma}\tilde{\delta})(\bar{\delta} - \tilde{\delta}) - \tilde{\gamma}\bar{\delta}(\bar{\delta}^r - \tilde{\delta}).$$

Finally, solving for $\tilde{\delta}$ gives

$$\tilde{\delta} = \frac{(1 + \tilde{\gamma}\tilde{\delta})\bar{\delta} - (\tilde{\gamma}\bar{\delta})\, \bar{\delta}^r}{1 + \tilde{\gamma}(\tilde{\delta} - \bar{\delta})}.$$

As usual, the asymptotic variance of a linear combination of coefficients is unaffected by how the coefficients are estimated, so that a feasible version of this estimator is

$$\bar{\delta}^* = (1 + \bar{\gamma}\bar{\delta})\bar{\delta} - (\bar{\gamma}\bar{\delta})\, \bar{\delta}^r, \bar{\gamma} = \sum_{i=1}^{n} X_i (y_i - X_i' \bar{\delta}) / \sum_{i=1}^{n} (y_i - X_i' \bar{\delta})^2.$$

Because HLIM and HFUL perform so well in our Monte Carlo experiments, we do not pursue this particular estimator, however.

The above result is analogous to that of Hahn and Hausman (2002), in the sense that under homoskedasticity, LIML is an optimal combination of forward and reverse bias corrected two stage least squares estimators. We find a similar result, as HLIM is a function of forward and reverse heteroskedasticity robust JIV estimators.

4 Optimal Estimation with Heteroskedasticity

HLIM is not asymptotically efficient under heteroskedasticity and many weak instruments. In GMM terminology, it uses a nonoptimal weighting matrix, one that is not heteroskedasticity consistent for the inverse of the variance of the moments. In addition, it does not use a heteroskedasticity consistent projection of the endogenous variables on the disturbance, which leads to inefficiency in the many instruments correction term. Efficiency can be obtained by modifying the estimator so that the weight matrix and the projection are heteroskedasticity consistent. Let

$$\hat{\Omega}(\delta) = \sum_{i=1}^{n} Z_i Z_i' \varepsilon_i(\delta)^2/n, \hat{B}_k(\delta) = \left(\sum_i Z_i Z_i' \varepsilon_i(\delta) X_{ik}/n \right) \hat{\Omega}(\delta)^{-1}$$

and

$$\hat{D}_{ik}(\delta) = Z_i X_{ik} - \hat{B}_k(\delta) Z_i \varepsilon_i(\delta), \hat{D}_i(\delta) = \left[\hat{D}_{i1}(\delta), ..., \hat{D}_{iG}(\delta) \right].$$

Also, let $\bar{\delta}$ be a preliminary estimator (such as HLIM). An IV estimator that is efficient under heteroskedasticity of unknown form and many weak instruments is

$$\hat{\delta} = \left(\sum_{i \neq j} \hat{D}_i(\bar{\delta})' \hat{\Omega}(\bar{\delta})^{-1} Z_j X_j' \right)^{-1} \sum_{i \neq j} \hat{D}_i(\bar{\delta})' \hat{\Omega}(\bar{\delta})^{-1} Z_j y_j.$$

This is a jackknife IV estimator with an optimal weighting matrix, $\hat{\Omega}(\bar{\delta})^{-1}$, and where $\hat{D}_i(\bar{\delta})$ replaces $X_i Z_i'$. The use of $\hat{D}_i(\bar{\delta})$ makes the estimator as efficient as the CUE under many weak instruments.

The asymptotic variance can be estimated by

$$U = \hat{H}^{-1} \hat{\Sigma} \hat{H}^{-1}, \hat{H} = \sum_{i \neq j} X_i Z_i' \hat{\Omega}(\tilde{\delta})^{-1} Z_j X_j', \hat{\Sigma} = \sum_{i,j=1}^{n} \hat{D}_i(\tilde{\delta})' \hat{\Omega}(\tilde{\delta})^{-1} \hat{D}_j(\tilde{\delta}).$$

This estimator has a sandwich form similar to that given in Newey and Windmeijer (2007).

5 The Robust, Restricted CUE

As discussed above, HLIM has been made robust to heteroskedasticity by jackknifing, where own observation terms are removed. In general this same approach can be used to make the continuous updating estimator robust to restrictions on the weighting matrix, such as homoskedasticity. For example, LIML is a CUE, where homoskedasticity is imposed on the weighting matrix; and HLIM is its robust version.

For expository purposes, consider a general GMM setup where δ denotes a $G \times 1$ parameter vector and $g_i(\delta)$ is a $K \times 1$ vector of functions of the data and parameters satisfying $E[g_i(\delta_0)] = 0$. For example, in the linear IV environment, $g_i(\delta) = Z_i(y_i - X_i'\delta)$. Let $\tilde{\Omega}(\delta)$ denote an estimator of $\Omega(\delta) = \sum_{i=1}^n E[g_i(\delta)g_i(\delta)']/n$, where an n subscript on $\Omega(\delta)$ is suppressed for notational convenience. A CUE is given by

$$\hat{\delta} = \arg\min_{\delta} \hat{g}(\delta)'\tilde{\Omega}(\delta)^{-1}\hat{g}(\delta).$$

When $\tilde{\Omega}(\delta) = \sum_{i=1}^n g_i(\delta)g_i(\delta)'/n$ this estimator is the CUE given by Hansen, Heaton, and Yaron (1996), that places no restrictions on the estimator of the second moment matrices. In general, restrictions may be imposed on the second moment matrix. For example, in the IV setting where $g_i(\delta) = Z_i(y_i - X_i'\delta)$, we may specify $\tilde{\Omega}(\delta)$ to be only consistent under homoskedasticity,

$$\tilde{\Omega}(\delta) = (y - X\delta)'(y - X\delta)\,Z'Z/n^2.$$

In this case the CUE objective function is

$$\hat{g}(\delta)'\tilde{\Omega}(\delta)^{-1}\hat{g}(\delta) = \frac{(y - X\delta)'P(y - X\delta)}{(y - X\delta)'(y - X\delta)},$$

which is the LIML objective function, as is well known (see Hansen, Heaton, and Yaron, (1996)).

A CUE will tend to have low bias when the restrictions imposed on $\tilde{\Omega}(\delta)$ are satisfied, but may be more biased otherwise. A simple calculation can be used to explain this bias. Consider a CUE where $\tilde{\Omega}(\delta)$ is replaced by its expectation, $\bar{\Omega}(\delta) = E[\tilde{\Omega}(\delta)]$. This replacement is justified under many weak moment asymptotics. The expectation of the CUE objective function is then

$$E[\hat{g}(\delta)'\bar{\Omega}(\delta)^{-1}\hat{g}(\delta)] = (1 - n^{-1})\bar{g}(\delta)'\bar{\Omega}(\delta)^{-1}\bar{g}(\delta) + tr(\bar{\Omega}(\delta)^{-1}\Omega(\delta))/n,$$

where $\bar{g}(\delta) = E[g_i(\delta)]$ and $\Omega(\delta) = E[g_i(\delta)g_i(\delta)']$. The first term in the above expression is minimized at δ_0, where $\bar{g}(\delta_0) = 0$. When $\bar{\Omega}(\delta) = \Omega(\delta)$, then

$$tr(\bar{\Omega}(\delta)^{-1}\Omega(\delta))/n = K/n,$$

so that the second term does not depend on δ. In this case the expected value of the CUE objective function is minimized at δ_0. When $\bar{\Omega}(\delta) \neq \Omega(\delta)$, the second term will depend on δ, and so the expected value of the CUE objective function will not be minimized at δ_0. This effect will lead to bias in the CUE, because the estimator will be minimizing an objective function with expectation that is not minimized at the truth. It is also interesting to note that this bias effect will tend to increase with K. This bias was noted by Han and Phillips (2005) for two-stage GMM, who referred to the bias term as a "noise" term, and to the other term as a "signal" term.

We robustify the CUE by jackknifing (i.e. by deleting the own observation terms in the CUE quadratic form). Note that

$$E[\sum_{i \neq j} g_i(\delta)'\bar{\Omega}(\delta)^{-1}g_j(\delta)/n^2] = (1 - n^{-1})\bar{g}(\delta)'\bar{\Omega}(\delta)^{-1}\bar{g}(\delta),$$

which is always minimized at δ_0, no matter what $\bar{\Omega}(\delta)$ is. A corresponding estimator is obtained by replacing $\bar{\Omega}(\delta)$ by $\tilde{\Omega}(\delta)$ and minimizing. Namely,

$$\hat{\delta} = \arg\min_{\delta} \sum_{i \neq j} g_i(\delta)'\tilde{\Omega}(\delta)^{-1}g_j(\delta)/n^2.$$

This is a robust CUE (RCUE), that should have small bias by virtue of the jackknife form of the objective function. The HLIM estimator is precisely of this form, for $\tilde{\Omega}(\delta) = (y - X\delta)'(y - X\delta)Z'Z/n^2$.

6 Asymptotic Theory

Theoretical justification for the estimators proposed here is provided by asymptotic theory where the number of instruments grows with the sample size. Some regularity conditions are important for the results. 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, Swanson, Hausman, Newey, and Woutersen (2007)).

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 $K < n$.[2] The condition $P_{ii} \leq C < 1$ implies that $K/n \leq C$, because $K/n = \sum_{i=1}^{n} P_{ii}/n \leq C$.

Assumption 2: There is a $G \times G$ matrix, $S_n = \tilde{S}_n \operatorname{diag}(\mu_{1n}, ..., \mu_{Gn})$, and z_i such that $\Upsilon_i = S_n z_i/\sqrt{n}$, $\tilde{S}_n$ is bounded and the smallest eigenvalue of $\tilde{S}_n \tilde{S}_n'$ is bounded away from zero, 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 $\sqrt{K}/\mu_n^2 \longrightarrow 0$. Also, $\sum_{i=1}^{n} \|z_i\|^4 / n^2 \longrightarrow 0$, and $\sum_{i=1}^{n} z_i z_i'/n$ is bounded and uniformly nonsingular.

Setting $\mu_{jn} = \sqrt{n}$ leads to asymptotic theory like that in Kunitomo (1980), Morimune (1984), and Bekker (1994), where the number of instruments K can grow as fast as the sample size. In that case, the condition $\sqrt{K}/\mu_n^2 \longrightarrow 0$ would be automatically satisfied. Allowing for K to grow, and for μ_n to grow more slowly than $\sqrt{n}$, allows for many instruments without strong identification. This condition then allows for some components of the reduced form to give only weak identification (corresponding to $\mu_{jn}/\sqrt{n} \longrightarrow 0$), and other components (corresponding to $\mu_{jn} = \sqrt{n}$) to give strong identification. In particular, this condition allows for fixed constant coefficients in the reduced form.

Assumption 3: $(\varepsilon_1, U_1), ..., (\varepsilon_n, U_n)$ are independent with $E[\varepsilon_i] = 0$, $E[U_i] = 0$, $E[\varepsilon_i^4]$

[2]The observations $w_1, ..., w_T$ 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 t with $a'Z_t = a'p^K(w_i) \neq 0$, implying that $a'Z'Za > 0$.

and $E[\|U_i\|^4]$ are bounded in i, $Var((\varepsilon_i, U_i')') = diag(\Omega_i^*, 0)$, and $\sum_{i=1}^n \Omega_i^*/n$ is uniformly nonsingular.

This condition includes moment existence assumptions. It also requires the average variance of the nonzero reduced form disturbances to be nonsingular, and is useful for the proof of consistency contained in the appendix.

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 allows for an unknown reduced form that is approximated by a linear combination of the instrumental variables. It is possible to replace this assumption with the condition that $\sum_{i \neq j} z_i P_{ij} z_j'/n$ is uniformly nonsingular.

We can easily interpret all of these conditions within the context of the important example of a linear model with exogenous covariates and a possibly unknown reduced form. This example is given by

$$X_i = \begin{pmatrix} \pi_{11} Z_{1i} + \mu_n f_0(w_i)/\sqrt{n} \\ Z_{1i} \end{pmatrix} + \begin{pmatrix} v_i \\ 0 \end{pmatrix}, Z_i = \begin{pmatrix} Z_{1i} \\ p^K(w_i) \end{pmatrix},$$

where Z_{1i} is a $G_2 \times 1$ vector of included exogenous variables, $f_0(w)$ is a $G - G_2$ dimensional vector function of a fixed dimensional vector of exogenous variables, w, and $p^K(w) \stackrel{def}{=} (p_{1K}(w), ..., p_{K-G_2,K}(w))'$. The variables in X_i other than Z_{1i} are endogenous with reduced form $\pi_{11} Z_{1i} + \mu_n f_0(w_i)/\sqrt{n}$. The function $f_0(w)$ may be a linear combination of a subvector of $p^K(w)$, in which case $z_i = \pi_{Kn} Z_i$, for some π_{Kn} in Assumption 4; or it may be an unknown function that can be approximated by a linear combination of $p^K(w)$. For $\mu_n = \sqrt{n}$, this example is like the model in Donald and Newey (2001), where Z_i includes approximating functions for the optimal (asymptotic variance minimizing) instruments Υ_i, but the number of instruments can grow as fast as the sample size. When $\mu_n^2/n \longrightarrow 0$, it is a modified version where the model is more weakly identified.

To see precise conditions under which the assumptions are satisfied, let

$$z_i = \begin{pmatrix} f_0(w_i) \\ Z_{1i} \end{pmatrix}, S_n = \tilde{S}_n diag\left(\mu_n, ..., \mu_n, \sqrt{n}, ..., \sqrt{n}\right), \text{ and } \tilde{S}_n = \begin{pmatrix} I & \pi_{11} \\ 0 & I \end{pmatrix}.$$

By construction we have that $\Upsilon_i = S_n z_i/\sqrt{n}$. Assumption 2 imposes the requirements that

$$\sum_{i=1}^{n} \|z_i\|^4 /n^2 \longrightarrow 0,$$

and that $\sum_{i=1}^{n} z_i z_i'/n$ is bounded and uniformly nonsingular. The other requirements of Assumption 2 are satisfied by construction. Turning to Assumption 3, we require that $\sum_{i=1}^{n} Var(\varepsilon_i, U_i')/n$ is uniformly nonsingular. For Assumption 4, let $\pi_{Kn} = [\tilde{\pi}_{Kn}', [I_{G_2}, 0]']'$. Then Assumption 4 will be satisfied if, for each n, there exists a $\tilde{\pi}_{Kn}$ with

$$\sum_{i=1}^{n} \|z_i - \pi_{Kn}' Z_i\|^2/n = \sum_{i=1}^{n} \|f_0(w_i) - \tilde{\pi}_{Kn}' Z_i\|^2/n \longrightarrow 0.$$

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 $\pi_{11}'\hat{\delta}_1 + \hat{\delta}_2 = o_p(\mu_n/\sqrt{n})$.

The asymptotic variance of the estimator will depend on the growth rate of K relative to μ_n^2. The following condition allows for two cases.

Assumption 5: 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 $\mu_n S_n^{-1} \longrightarrow \bar{S}_0$.

To state a limiting distribution result it is helpful to also assume that certain objects converge. 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: $H_P = \lim_{n\longrightarrow\infty} \sum_{i=1}^{n}(1-P_{ii})z_i z_i'/n$, $\Sigma_p = \lim_{n\longrightarrow\infty} \sum_{i=1}^{n}(1-P_{ii})^2 z_i z_i'\sigma_i^2/n$ and $\Psi = \lim_{n\longrightarrow\infty} \sum_{i\neq j} P_{ij}^2 \left(\sigma_i^2 E[\tilde{U}_j\tilde{U}_j'] + E[\tilde{U}_i\varepsilon_i]E[\varepsilon_j\tilde{U}_j']\right)/K$.

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, Swanson, Hausman, Newey, and Woutersen (2007)).

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

$$S_n'(\hat{\delta}-\delta_0) \xrightarrow{d} N(0,\Lambda_I), \tag{6.1}$$

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

$$(\mu_n/\sqrt{K})S_n'(\hat{\delta}-\delta_0) \xrightarrow{d} N(0,\Lambda_{II}), \tag{6.2}$$

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), Morimune (1983), 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 (6.1) is satisfied, and in Case II, equation (6.2) is satisfied.*

It is interesting to compare the asymptotic variance of the HLIM estimator with that of LIML when the disturbances are homoskedastic. 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 (2006), 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 Haslett (2000)). This makes estimation of the asymptotic variance under homoskedasticity 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.

7 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

$$y = \delta_{10} + \delta_{20} x_2 + \varepsilon, x_2 = \pi z_1 + U_2$$

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 the LIML or FULL estimators, 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_{i1} and v_{i2} 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 $\gamma_i = E[X_i\varepsilon_i]/\sigma_i^2$.

We report properties of estimators and t-ratios for δ_2. We set $n = 800$ and $\rho = 0.3$ throughout and choose $K = 2, 10, 30$. We choose π so that the concentration parameter is $n\pi^2 = \mu^2 = 8, 16, 32$. 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 and the range between the .05 and .95 quantiles for LIML, HLIM, the jackknife CUE, JIV, HFUL, CUE, and FULL. Interquartile range results were similar. We find that under homoskedasticity, LIML and HFUL have quite similar properties, though LIML is slightly less 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 turn out to be important in practice. In additional tables following the references, we also find that coverage probabilities using the heteroskedasticity and many instrument consistent standard errors are quite accurate.

Median Bias $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0.00$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$HFUL\frac{1}{k}$	$JIVE$	CUE	$JCUE$
8	0	0.005	0.005	0.042	0.043	0.025	−0.034	0.005	0.005
8	8	0.024	0.023	0.057	0.057	0.027	0.053	0.025	0.032
8	28	0.065	0.065	0.086	0.091	0.067	0.164	0.071	0.092
32	0	0.002	0.002	0.011	0.011	0.007	−0.018	0.002	0.002
32	8	0.002	0.001	0.011	0.011	0.002	−0.019	0.002	0.002
32	28	0.003	0.002	0.013	0.013	0.003	−0.014	0.006	0.006

***Results based on 20,000 simulations.

Nine Decile Range: .05 to .95 $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0.00$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$HFUL\frac{1}{k}$	$JIVE$	CUE	$JCUE$
8	0	1.470	1.466	1.072	1.073	1.202	3.114	1.470	1.487
8	8	2.852	2.934	1.657	1.644	2.579	5.098	3.101	3.511
8	28	5.036	5.179	2.421	2.364	4.793	6.787	6.336	6.240
32	0	0.616	0.616	0.590	0.589	0.602	0.679	0.616	0.616
32	8	0.715	0.716	0.679	0.680	0.713	0.816	0.770	0.767
32	28	0.961	0.985	0.901	0.913	0.983	1.200	1.156	1.133

***Results based on 20,000 simulations.

Median Bias $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0.20$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$HFUL\frac{1}{k}$	$JIVE$	CUE	$JCUE$
8	0	−0.001	0.050	0.041	0.078	0.065	−0.031	−0.001	0.012
8	8	−0.623	0.094	−0.349	0.113	0.096	0.039	0.003	−0.005
8	28	−1.871	0.134	−0.937	0.146	0.134	0.148	−0.034	0.076
32	0	−0.001	0.011	0.008	0.020	0.016	−0.021	−0.001	−0.003
32	8	−0.220	0.015	−0.192	0.024	0.016	−0.021	0.000	−0.019
32	28	−1.038	0.016	−0.846	0.027	0.017	−0.016	−0.017	−0.021

***Results based on 20,000 simulations.

Nine Decile Range: .05 to .95 $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0.20$

μ^2	K	$LIML$	$HLIM$	$FULL1$	$HFUL$	$HFUL\frac{1}{k}$	$JIVE$	CUE	$JCUE$
8	0	2.219	1.868	1.675	1.494	1.653	4.381	2.219	2.582
8	8	26.169	5.611	4.776	2.664	4.738	7.781	16.218	8.586
8	28	60.512	8.191	7.145	3.332	7.510	9.975	1.5E+012	12.281
32	0	0.941	0.901	0.903	0.868	0.884	1.029	0.941	0.946
32	8	3.365	1.226	2.429	1.134	1.217	1.206	1.011	1.086
32	28	18.357	1.815	5.424	1.571	1.808	1.678	3.563	1.873

***Results based on 20,000 simulations.

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 proved in Hansen, Hausman, and Newey (2006).

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., 2007): *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* $\bar{w} = E[(W_1, ..., W_n)']$, $\bar{y} = E[(Y_1, ..., Y_n)']$, $\bar{\sigma}_{Wn} = \max_{i\le n} Var(W_i)^{1/2}$, $\bar{\sigma}_{Yn} = \max_{i\le n} Var(Y_i)^{1/2}$,

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

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 $\|S_n^{-1}\| \leq C\mu_n^{-1}$ we have $Var(\tilde{X}_{ik}) \leq C\mu_n^{-2}$ 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

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

where the third equality follows by $PZ = Z$, the first inequality by $I - P$ idempotent, and the last inequality by $A \leq tr(A)I$ 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 $\|S_n/\sqrt{n}\| \leq C$ and by standard calculations $z'U/n \xrightarrow{p} 0$. Then

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

Let $\bar{\Omega}_n = \sum_{i=1}^n E[\bar{U}_i\bar{U}_i']/n = diag(\sum_{i=1}^n \Omega_i^*/n, 0) \geq C diag(I_{G-G_2+1}, 0)$ by Assumption 3. 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 $H_n=\sum_{i=1}^n(1-P_{ii})z_iz_i'/n$,*

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

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

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

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

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

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

Proof: Let $\hat\beta=S_n'(\hat\delta-\delta_0)/\mu_n$ and $\breve\alpha=\sum_{i\neq j}\varepsilon_iP_{ij}\varepsilon_j/\varepsilon'\varepsilon=o_p(\mu_n^2/n)$. 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 $\tilde H_n=S_n^{-1}(\sum_{i\neq j}X_iP_{ij}X_j'-\breve\alpha X'X)S_n^{-1\prime}=O_p(1)$ and $W_n=S_n^{-1}(X'P\varepsilon-\breve\alpha X'\varepsilon)/\mu_n\xrightarrow{p}0$, so

$$\begin{aligned}\frac{\sum_{i\neq j}\hat\varepsilon_iP_{ij}\hat\varepsilon_j}{\hat\varepsilon'\hat\varepsilon}-\breve\alpha &= \frac{1}{\hat\varepsilon'\hat\varepsilon}\left(\sum_{i\neq j}\hat\varepsilon_iP_{ij}\hat\varepsilon_j-\sum_{i\neq j}\varepsilon_iP_{ij}\varepsilon_j-\breve\alpha\left(\hat\varepsilon'\hat\varepsilon-\varepsilon'\varepsilon\right)\right)\\ &= \frac{\mu_n^2}{n}\frac{1}{\hat\sigma_\varepsilon^2}\left(\hat\beta'\tilde H_n\hat\beta-2\hat\beta'W_n\right)=o_p(\mu_n^2/n),\end{aligned}$$

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 $\lambda_{\min}(S_nS_n'/\mu_n^2)\geq\lambda_{\min}\left(\tilde S_n\tilde S_n'\right)\geq C$ we have

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

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., 2007): *If i) P is a symmetric, idempotent matrix with rank$(P) = K$, $P_{ii} \leq C < 1$; 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 $E[\|U_i\|^4] \leq C$, $E[\varepsilon_i^4] \leq C$; iv) $\sum_{i=1}^n E[\|W_{in}\|^4] \longrightarrow 0$; v) $K \longrightarrow \infty$; then for $\bar{\Sigma}_n \stackrel{\text{def}}{=} \sum_{i\neq j} P_{ij}^2 \left(E[U_i U_i'] E[\varepsilon_j^2] + E[U_i \varepsilon_i] E[\varepsilon_j U_j']\right)/K$ 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

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

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 $E[\varepsilon\varepsilon'] \le CI_n$ we have

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

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 $\sum_{i=1}^n \sigma_i^2/n \ge C > 0$. 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, $\hat{\delta} \xrightarrow{p} \delta_0$. The 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 $U_i = U_i$ 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 $\Gamma = \mathrm{diag}\,(\Sigma_P, \Psi)$ 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

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

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,

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

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,

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

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

Lemma A9 (Lemma A3 of Chao et al., 2007): *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., 2007): *If W_i, Y_i, η_i, are independent across i with $E[W_i] = a_i/\sqrt{n}$, $E[Y_i] = b_i/\sqrt{n}$, $|a_i| \leq C$, $|b_i| \leq C$, $E[\eta_i^2] \leq C$,*

$Var(W_i) \le C\mu_n^{-2}$, $Var(Y_i) \le C\mu_n^{-2}$, *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

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

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

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

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, $\|\Upsilon_i\| = \|S_n z_i/\sqrt{n}\| \le C$. 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

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

Also note that

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

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

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

We also have

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

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

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

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

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

Note that $S_n^{-1}\hat{\gamma}\hat{\Delta}' \xrightarrow{p} 0$, $E[D_i] = z_i/\sqrt{n}$, $Var(D_i) = O(\mu_n^{-2})$, $E[\dot{X}_i] = z_i/\sqrt{n}$, 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 $Var(\varepsilon_i^2) \leq C$ and $\mu_n^2 \leq Cn$, so that for $u_{ki} = e_k' S_n^{-1}U_i$,

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

It follows that

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

It then follows by Lemmas and the triangle inequality that

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

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

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

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

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

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.

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

Angrist, J. and A. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics* 104, 979–1014.

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.

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 (2007). "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 (2001). "Choosing the Number of Instruments" *Econometrica* 69, 1161-1191.

Dufour, J.M. (1997). "Some Impossibility Theorems in Econometrics, With Applications to Structural and Dynamic Models," *Econometrica* 65, 1365 - 1388.

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

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

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.*

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.

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 (2007). "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, forthcoming.

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.

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.

Morimune, K. (1983). "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica* 51, 821-841.

Newey, W.K. and Windmeier (2007). "GMM with Many Weak Moment Conditions Many Moments," Working Paper, MIT.

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. (2004). "GEL Criteria for Moment Condition Models," Working Paper, CEMMAP, UCL.

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

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 Estimation with Independent Observations," *Econometrica* 50, 483-500.