

# Testing Overidentifying Restrictions with Many Instruments and Heteroskedasticity\*

John C. Chao, Department of Economics, University of Maryland, chao@econ.umd.edu.

Jerry A. Hausman, Department of Economics, MIT, jhausman@mit.edu.

Whitney K. Newey, Department of Economics, MIT, wnewey@mit.edu.

Norman R. Swanson, Department of Economics, Rutgers University, nswanson@econ.rutgers.edu.

Tiemen Woutersen, Department of Economics, Johns Hopkins University, woutersen@jhu.edu.

June 2010

*JEL classification:* C13, C31.

*Keywords:* heteroskedasticity, instrumental variables, jackknife estimation, many instruments, weak instruments.

---

\*Prepared for the Test 2010 Conference in Xiamen, China. Newey thanks the NSF for financial support.

Proposed Running Head: JIVE With Heteroskedasticity

Corresponding Author:

Whitney K. Newey

Department of Economics

MIT, E52-262D

Cambridge, MA 02142-1347

## Abstract

This paper gives a test of overidentifying restrictions that is robust to many instruments and heteroskedasticity. It is based on a jackknife version of the Sargan test statistic, having a numerator that is the objective function minimized by the JIVE2 estimator of Angrist, Imbens, and Krueger (1999). Correct asymptotic critical values are derived for this test when the number of instruments grows large, at a rate up to the sample size. It is also shown that the test is valid when the number instruments is fixed and there is homoskedasticity. This test improves on recently proposed tests by allowing for heteroskedasticity and by avoiding assumptions on the instrument projection matrix. The asymptotics is based on the heteroskedasticity robust many instrument asymptotics of Chao et. al. (2010).

## 1 Introduction

The Sargan (1958) and Hansen (1982) tests of instrument validity are known to be quite sensitive to the number of restrictions being tested. This paper proposes an alternative test that is robust to many instruments and to heteroskedasticity. It is based on subtracting out the diagonal terms in the numerator of the Sargan statistic and normalizing appropriately. This test has a jackknife interpretation, being based on the objective function of the JIVE2 estimator of Angrist, Imbens, and Krueger (1999). We show that the test is valid under homoskedasticity with fixed number of instruments and with number of instruments going to infinity at any rate up to the sample size itself.

Recently Anatolyev and Gospodinov (2009) and Lee and Okui (2010) have given tests that allow for many instruments but impose homoskedasticity. Our test is valid under their conditions and also with heteroskedasticity. Also, we do not impose side conditions on the instrument projection matrix. The asymptotic theory is based on the results of Chao et. al. (2010) and Hausman et. al. (2010), including a central limit theorem that imposes no side conditions on the instrumental variable projection matrix.

In Section 2 we describe the model and test statistic. In Section 3 we give the asymptotic theory.

## 2 The Model and Test Statistic

We adopt the same model and notation as in Hausman et. al. (2010) and Chao et. al. (2010). 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,  $\Upsilon$  is the reduced form matrix, and  $U$  is the disturbance matrix. For the asymptotic approximations, the elements of  $\Upsilon$  will be implicitly allowed to depend on  $n$ , although we suppress dependence of  $\Upsilon$  on  $n$ , for notational convenience. Estimation of  $\delta_0$  will be based on an  $n \times K$  matrix,  $Z$ , of instrumental variable observations with  $\text{rank}(Z) = K$ . Here we will treat  $Z$  and  $\Upsilon$  as nonrandom for simplicity

though it is possible to do asymptotic theory conditional on these as in Chao et. al. (2010). We will assume that  $E[\varepsilon] = 0$  and  $E[U] = 0$ .

This model allows for  $\Upsilon$  to be a linear combination of  $Z$  (i.e.  $\Upsilon = Z\pi$ , for some  $K \times G$  matrix  $\pi$ ). 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  $X'_i$ ,  $\Upsilon'_i$ , and  $Z'_i$  denote the  $i^{th}$  row (observation) for  $X$ ,  $\Upsilon$ , 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 let  $Z_i = (p_{1K}(w_i), \dots, p_{KK}(w_i))'$ , 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.

For estimation of  $\delta$  we consider heteroskedasticity robust version of the Fuller (1977) estimator of Hausman et. al. (2010), referred to as HFUL. Other heteroskedasticity and many instrument robust estimators could also be used, such as jackknife instrumental variable (IV) estimators of Angrist, Imbens, and Krueger (1999) or the continuously updated GMM estimator (CUE). We focus on HFUL because of its high efficiency relative to jackknife IV, because it has moments, and because it is computationally simple relative to CUE. 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 the matrix in this expression is not symmetric it has real eigenvalues because it is a product of symmetric, positive semi-definite matrices. Let

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

HFUL is given by

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

Thus, HFUL can be computed by finding the smallest eigenvalue of a matrix and then using this explicit formulae. Motivation for HFUL is further discussed in Hausman et. al. (2010).

To describe the overidentification statistic, let  $\hat{\varepsilon}_i = y_i - X'_i\hat{\delta}$ ,  $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$ ,  $\hat{\varepsilon}(2) = (\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)'$ , and  $P(2)$  be the n-dimensional square matrix with  $ij^{th}$  component equal to  $P_{ij}^2$ . Also, let  $\sum_{i \neq j}$

denote the double sum over all  $i$  not equal to  $j$ . The test statistic is

$$\hat{T} = \frac{\hat{\varepsilon}' P \hat{\varepsilon} - \sum_{i=1}^n P_{ii} \hat{\varepsilon}_i^2}{\sqrt{\hat{V}}} + K, \hat{V} = \frac{\hat{\varepsilon}(2)' P(2) \hat{\varepsilon}(2) - \sum_i P_{ii}^2 \hat{\varepsilon}_i^4}{K} = \frac{\sum_{i \neq j} \hat{\varepsilon}_i^2 P_{ij}^2 \hat{\varepsilon}_j^2}{K}$$

Treating  $\hat{T}$  as if it is chi-squared with  $K - G$  degrees of freedom will be asymptotically correct if  $K \rightarrow \infty$  no faster than  $n$  and when  $K$  is fixed and  $\varepsilon_i$  is homoskedastic. Let  $q_r(\tau)$  be the  $\tau^{th}$  quantile of the chi-squared distribution with  $r$  degrees of freedom. A test with asymptotic rejection frequency  $\alpha$  is to reject the null hypothesis if

$$\hat{T} \geq q_{K-G}(1 - \alpha).$$

We will show that the test with this critical region has a probability of rejection that converges to  $\alpha$ .

To explain the form of this test statistic, note that the numerator is

$$\hat{\varepsilon}' P \hat{\varepsilon} - \sum_{i=1}^n P_{ii} \hat{\varepsilon}_i^2 = \sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j.$$

This object is the numerator of the Sargan (1958) statistic with the own observation terms subtracted out. It has a jackknife form, in the sense that it is the sum of sums where the own observations have been deleted. If  $\hat{\delta}$  were chosen to minimize this expression, it would be the JIVE2 estimator of Angrist, Imbens, and Krueger (1999).

One effect of removing the own observations is that  $\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j$  would be mean zero if  $\hat{\varepsilon}_i$  were replaced by  $\varepsilon_i$ . In fact,  $\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j$  has a martingale difference structure that leads to it being asymptotically normal as  $K \rightarrow \infty$ , e.g. as in Lemma A2 of Chao et. al. (2010). The denominator incorporates a heteroskedasticity consistent estimator of the variance of  $\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j$ . By dropping terms that have zero expectation, similarly to Chao et. al. (2010), it follows that for  $\sigma_i^2 = E[\varepsilon_i^2]$ ,

$$\begin{aligned} E[(\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j)^2] &= E[\sum_{i,j} \sum_{k \notin \{i,j\}} P_{ik} P_{jk} \varepsilon_i \varepsilon'_j \varepsilon_k^2 + \sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2] \\ &= E[2 \sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2] = 2 \sum_{i \neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2 \end{aligned}$$

Similarly to White (1980) the variances are replaced by squared residuals to obtain  $\hat{V}$ . Also, 2 is replaced by  $1/K$  and  $K$  is added to normalize the statistic to be chi-squared with  $K$  fixed and  $\varepsilon_i$  homoskedastic. Unfortunately, it does not appear possible to normalize the statistic to be chi-squared if there is heteroskedasticity when  $K$  is fixed.

### 3 Many Instrument Asymptotics

The asymptotic theory we give combines the many instrument asymptotics of Kunitomo (1980), Morimune (1983), and Bekker (1994) with the many weak instrument asymptotics of Chao and Swanson (2005). Some regularity conditions are important for this theory. Let  $Z'_i, \varepsilon_i, U'_i$ , and  $\Upsilon'_i$  denote the  $i^{th}$  row of  $Z, \varepsilon, U$ , and  $\Upsilon$  respectively.

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

The restriction that  $\text{rank}(Z) = K$  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  $K < n$ .<sup>1</sup> The condition  $P_{ii} \leq C < 1$  implies that  $K/n \leq C$ , because  $K/n = \sum_{i=1}^n P_{ii}/n \leq C$ .

The next condition specifies that the reduced form  $\Upsilon_i$  is a linear combination of a set of variables  $z_i$  having certain properties.

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

This condition is similar to Assumption 2 of Hansen, Hausman, and Newey (2008). It accommodates linear models 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, as further discussed in Hausman et. al. (2010). The  $\mu_n^2$  can be thought of as a version of the concentration parameter, determining the convergence rate of estimators of  $\delta_{0G}$ , just as the concentration parameter does in other settings. For  $\mu_n^2 = n$ , the convergence rate will be  $\sqrt{n}$ , where Assumptions 1 and 2 permit  $K$  to grow as fast as the sample size, corresponding to a many instrument asymptotic approximation like Kunitomo (1980), Morimune (1983), and Bekker (1994). For  $\mu_n^2$  growing slower than  $n$  the convergence rate will be slower than  $1/\sqrt{n}$ , leading to an asymptotic approximation like that of Chao and Swanson (2005).

---

<sup>1</sup>The observations  $w_1, \dots, w_n$  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  $a' Z' Z a > 0$ .

**Assumption 3:** There is a constant  $C > 0$  such that  $(\varepsilon_1, U_1), \dots, (\varepsilon_n, U_n)$  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  $\lambda_{\min}(\sum_{i=1}^n \Omega_i^*/n) \geq 1/C$ .

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 stay away from zero.

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

This condition and  $P_{ii} \leq C < 1$  will imply that for a large enough sample

$$\begin{aligned} \sum_{i \neq j} P_{ij} \Upsilon_i \Upsilon_j' / n &= \Upsilon' P \Upsilon / n - \sum_{i=1}^n P_{ii} \Upsilon_i \Upsilon_i' / n = \sum_{i=1}^n (1 - P_{ii}) \Upsilon_i \Upsilon_i' / n - \Upsilon' (I - P) \Upsilon / n \\ &= \sum_{i=1}^n (1 - P_{ii}) \Upsilon_i \Upsilon_i' / n + o(1) \geq (1 - C) \sum_{i=1}^n \Upsilon_i \Upsilon_i' / n, \end{aligned}$$

so that the structurally parameters are identified asymptotically. Also, Assumption 4 is not very restrictive because flexibility is allowed in the specification of  $\Upsilon_i$ . If we simply make  $\Upsilon_i$  the expectation of  $X_i$  given the instrumental variables then Assumption 4 holds automatically.

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

It simplifies the asymptotic theory to assume that certain objects converge and to allow for two cases of growth rates of  $K$  relative to  $\mu_n^2$ . These conditions could be relaxed at the expense of further notation and detail, as in Chao et. al.. 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:**  $\mu_n S_n^{-1} \rightarrow S_0$  and either I)  $K/\mu_n^2 \rightarrow \alpha$  for finite  $\alpha$  or; II)  $K/\mu_n^2 \rightarrow \infty$ .

Also, each of the following exist:

$$\begin{aligned} H_P &= \lim_{n \rightarrow \infty} \sum_{i=1}^n (1 - P_{ii}) z_i z_i' / n, \Sigma_P = \lim_{n \rightarrow \infty} \sum_{i=1}^n (1 - P_{ii})^2 z_i z_i' \sigma_i^2 / n, \\ \Psi &= \lim_{n \rightarrow \infty} \sum_{i \neq j} P_{ij}^2 \left( \sigma_i^2 E[\tilde{U}_j \tilde{U}_j'] + E[\tilde{U}_i \varepsilon_i] E[\varepsilon_j \tilde{U}_j'] \right) / K. \end{aligned}$$

The first result shows that the chi-square approximation is asymptotically correct when  $K$  grows with  $n$ .

**THEOREM 1:** *If Assumptions 1-6 are satisfied then  $\Pr(\hat{T} \geq q_{K-G}(1-\alpha)) \rightarrow \alpha$ .*

The next result shows asymptotic validity of the chi-squared approximation when  $K$  is fixed.

**THEOREM 2:** *If  $E[\varepsilon_i^2] = \sigma^2$ ,  $K$  is fixed,  $Z'Z/n \rightarrow Q$  nonsingular,  $Z'\Upsilon/n \rightarrow D$  with  $\text{rank}(D) = G$ ,  $E[\varepsilon_i^4] \leq C$ ,  $\|\Upsilon_i\| \leq C$ , and Assumption 3 is satisfied, then  $\Pr(\hat{T} \geq q_{K-G}(1-\alpha)) \rightarrow \alpha$ .*

This test should have power against some forms of misspecification. Under misspecification  $\hat{V}$  will still be bounded and bounded away from zero. Also, for  $\bar{\varepsilon}_i = E[y_i - X'_i p\lim(\hat{\delta})]$ , the normalized numerator  $\sum_{i \neq j} P_{ij} \hat{\varepsilon}_i \hat{\varepsilon}_j / \sqrt{K}$  will be centered at

$$\left( \bar{\varepsilon}' P \bar{\varepsilon} - \sum_i P_{ii} \bar{\varepsilon}_i^2 \right) / \sqrt{K}.$$

Assuming a linear combination of  $Z$  approximates  $\bar{\varepsilon}$  this is close

$$\sum_i \bar{\varepsilon}_i^2 (1 - P_{ii}) / \sqrt{K}$$

This will increase at rate  $n/\sqrt{K}$  by  $P_{ii}$  bounded away from one.

$\hat{T}$  provides specification check for many instrument estimator  $\hat{\delta}$ . Note however that it may not be optimal as a test of the null hypothesis that  $E[\bar{\varepsilon}_i] = 0$ . The magnitude of the test statistic under the alternative grows faster when  $K$  grows slower. Thus, for higher power it would be better to use fewer instruments.

## 4 Appendix A - Proofs of Theorems

We will define a number of notation and abbreviations.  $C$  denotes a generic positive constant that may be different in different uses and let M, CS, and T denote the Markov inequality, the Cauchy-Schwartz inequality, and the Triangle inequality respectively. Also, for random variables  $W_i$ ,  $Y_i$ , and  $\eta_i$ , let  $\bar{w}_i = E[W_i]$ ,  $\tilde{W}_i = W_i - \bar{w}_i$ ,  $\bar{y}_i = E[Y_i]$ ,  $\tilde{Y}_i = Y_i - \bar{y}_i$ ,  $\bar{\eta}_i = E[\eta_i]$ ,  $\tilde{\eta}_i = \eta_i - \bar{\eta}_i$ ,  $\bar{y} = (\bar{y}_1, \dots, \bar{y}_n)'$ ,  $\bar{w} = (\bar{w}_1, \dots, \bar{w}_n)'$ ,

$$\begin{aligned}\bar{\mu}_W &= \max_{1 \leq i \leq n} |\bar{w}_i|, \quad \bar{\mu}_Y = \max_{1 \leq i \leq n} |\bar{y}_i|, \quad \bar{\mu}_\eta = \max_{1 \leq i \leq n} |\bar{\eta}_i|, \\ \bar{\sigma}_W^2 &= \max_{i \leq n} \text{Var}[W_i], \quad \bar{\sigma}_Y^2 = \max_{i \leq n} \text{Var}[Y_i], \quad \bar{\sigma}_\eta^2 = \max_{i \leq n} \text{Var}[\eta_i];\end{aligned}$$

The following Lemmas are special cases of results in Chao et. al. (2010) but are given here for exposition:

**LEMMA A1:** Suppose that the following conditions hold: i)  $P$  is a symmetric, idempotent matrix with  $\text{rank}(P) = K$ ,  $P_{ii} \leq C < 1$ ; ii)  $(W_{1n}, U_1, \varepsilon_1), \dots, (W_{nn}, U_n, \varepsilon_n)$  are independent and  $D_n = \sum_{i=1}^n E[W_{in}W_{in}']$  satisfies  $\|D_n\| \leq C$ ; 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)  $\sum_{i=1}^n E[\|W_{in}\|^4] \rightarrow 0$ ; v)  $K \rightarrow \infty$  as  $n \rightarrow \infty$ . 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 any sequences  $c_{1n}$  and  $c_{2n}$  depending on  $Z$  conformable vectors with  $\|c_{1n}\| \leq C$ ,  $\|c_{2n}\| \leq C$ ,  $\Xi_n = c_{1n}' D_n c_{1n} + c_{2n}' \bar{\Sigma}_n c_{2n} > 1/C$ , it follows that

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

Proof: This is Lemma A2 of Chao et. al. (2010) when  $Z$  and  $\Upsilon$  are not random. Q.E.D.

**LEMMA A2:** If Assumptions 1-3 are satisfied then

$$S_n^{-1} \sum_{i \neq j} X_i P_{ij} X_j' S_n^{-1} = O_p(1), \quad S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j = O_p(1 + \sqrt{K/r_n}).$$

Proof: The second conclusion holds by Lemma A5 of Chao et al. (2010), and by that same result,

$$S_n^{-1} \sum_{i \neq j} X_i P_{ij} X_j' S_n^{-1} = \sum_{i \neq j} z_i P_{ij} z_j' / n + o_p(1).$$

We also have

$$\sum_{i \neq j} z_i P_{ij} z'_j / n = z' P z / n - \sum_i P_{ii} z_i z_i / n$$

and both  $z' P z / n \leq z' z / n$  and  $\sum_i P_{ii} z_i z'_i / n \leq z' z / n$  are bounded, giving the first conclusion. Q.E.D.

LEMMA A3: If  $\hat{\delta} \rightarrow \delta$ ,  $E[\|X_i\|^2] \leq C$ ,  $E[\varepsilon_i^4] \leq C$ ,  $\varepsilon_1, \dots, \varepsilon_n$  are mutually independent, and either  $K \rightarrow \infty$  or  $\max_{i \leq n} P_{ii} \rightarrow 0$  then

$$\frac{\sum_{i \neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2}{K} - \frac{\sum_{i \neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2}{K} \xrightarrow{p} 0.$$

Proof: Hence by  $\hat{\delta} \xrightarrow{p} \delta$  we have  $\|\hat{\delta} - \delta\|^2 \leq \|\hat{\delta} - \delta\|$  with probability approaching one (w.p.a.1).

Hence w.p.a.1, for  $d_i = 3(1 + \|X_i\|^2)$ ,

$$|\hat{\varepsilon}_i^2 - \varepsilon_i^2| \leq 2 \|X_i\| \|\hat{\delta} - \delta\| + \|X_i\|^2 \|\hat{\delta} - \delta\|^2 \leq d_i \|\hat{\delta} - \delta\|.$$

Also by  $\sum_{i,j} P_{ij}^2 = \sum_i P_{ii} = K$ ,

$$E[\sum_{i \neq j} P_{ij}^2 d_i d_j] / K \leq C \sum_{i \neq j} P_{ij}^2 / K \leq C, E[\sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 d_j] / K \leq C.$$

Then by M,

$$\sum_{i \neq j} P_{ij}^2 d_i d_j / K = O_p(1), \sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 d_j / K = O_p(1).$$

Therefore, for  $\hat{V}_n = \sum_{i \neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2 / K$ ,  $\tilde{V}_n = \sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2 / K$  we have

$$\begin{aligned} |\hat{V}_n - \tilde{V}_n| &\leq \sum_{i \neq j} P_{ij}^2 |\hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2 - \varepsilon_i^2 \varepsilon_j^2| / K \\ &\leq \|\hat{\delta} - \delta\|^2 \sum_{i \neq j} P_{ij}^2 d_i d_j / K + 2 \|\hat{\delta} - \delta\| \sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 d_j / K \xrightarrow{p} 0. \end{aligned}$$

Let  $V_n = \sum_{i \neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2 / K$  and  $v_i = \varepsilon_i^2 - \sigma_i^2$ . Note that by  $P_{ij} = P_{ji}$ ,

$$\sum_{i \neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2 - \sum_{i \neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2 = 2 \sum_{i \neq j} P_{ij}^2 v_i \sigma_j^2 + \sum_{i \neq j} P_{ij}^2 v_i v_j.$$

Note that  $E[v_i^2] \leq E[\varepsilon_i^4] \leq C$ , so we have

$$\begin{aligned} E[(\sum_{i \neq j} P_{ij}^2 v_i \sigma_j^2 / K)^2] &= K^{-2} \sum_i \sum_{j \neq i} \sum_{k \neq i} P_{ij}^2 P_{ik}^2 E[v_i^2] \sigma_j^2 \sigma_k^2 \\ &\leq CK^{-2} \sum_i \sum_j P_{ij}^2 \sum_k P_{ik}^2 = CK^{-2} \sum_i P_{ii}^2 \\ &\leq CK^{-1} \max_{i \leq n} P_{ii} \sum_i P_{ii} / K \leq CK^{-1} \max_{i \leq n} P_{ii} \rightarrow 0. \end{aligned}$$

Also, by CS,  $\max_{i,j \leq n} P_{ij}^2 \leq \max_{i \leq n} P_{ii}^2$ , so that

$$\begin{aligned} E[(\sum_{i \neq j} P_{ij}^2 v_i v_j / K)^2] &= 2K^{-2} \sum_{i \neq j} P_{ij}^4 E[v_i^2] E[v_j^2] \leq CK^{-2} \sum_{i,j} P_{ij}^4 \\ &\leq CK^{-1} \max_{i \leq n} P_{ii}^2 \sum_{i,j} P_{ij}^2 / K = CK^{-1} \max_{i \leq n} P_{ii}^2 \longrightarrow 0. \end{aligned}$$

Then by T and M we have  $\tilde{V}_n - V_n \xrightarrow{P} 0$ . The conclusion then follows by T. Q.E.D.

PROOF OF THEOREM 1: Note that

$$\begin{aligned} \frac{\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{K}} &= \sum_{i \neq j} [\varepsilon_i - X'_i(\hat{\delta} - \delta)] P_{ij} [\varepsilon_j - X'_j(\hat{\delta} - \delta)] / \sqrt{K} \\ &= \frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{K}} + (\hat{\delta} - \delta)' S_n \left[ S_n^{-1} \sum_{i \neq j} X_i P_{ij} X'_j S_n^{-1} \right] S'_n(\hat{\delta} - \delta) / \sqrt{K} \\ &\quad + 2(\hat{\delta} - \delta)' S_n \left[ S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j \right] / \sqrt{K}. \end{aligned}$$

If  $K/\mu_n^2 \longrightarrow \alpha < \infty$  (case I of Assumption 6) then by Theorem 2 of Hausman et. al. (2010) we have  $S'_n(\hat{\delta} - \delta) = O_p(1)$ . Then by Lemma A2 we have

$$\frac{\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{K}} = \frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{K}} + o_p(1). \quad (1)$$

If  $K/\mu_n^2 \longrightarrow \infty$  (case II of Assumption 6) then by Theorem 2 of Hausman et. al. (2010),  $(\mu_n/\sqrt{K})S'_n(\hat{\delta} - \delta_0) = O_p(1)$ , so that by  $\sqrt{K}/\mu_n^2 \longrightarrow 0$ ,

$$\begin{aligned} (\hat{\delta} - \delta)' S_n \left[ S_n^{-1} \sum_{i \neq j} X_i P_{ij} X'_j S_n^{-1} \right] S'_n(\hat{\delta} - \delta) / \sqrt{K} &= O_p(1) \left( K/\mu_n^2 \right) / \sqrt{K} = o_p(1), \\ (\hat{\delta} - \delta)' S_n \left[ S_n^{-1} \sum_{i \neq j} X_i P_{ij} \varepsilon_j \right] / \sqrt{K} &= O_p(1)(\sqrt{K}/\mu_n)O_p(1 + \sqrt{K}/\mu_n) / \sqrt{K} \\ &= O_p(1/\mu_n + \sqrt{K}/\mu_n^2) = o_p(1). \end{aligned}$$

Therefore, eq. (1) is also satisfied when  $K/\mu_n^2 \longrightarrow \infty$ .

Next, note that  $\sigma_i^2 \geq C$  by Assumption 3 and  $P_{ii} \leq C < 1$  by Assumption 1, so that

$$V_n = \frac{\sum_{i \neq j} \sigma_i^2 P_{ij}^2 \sigma_j^2}{K} > C \left( \frac{\sum_{i,j} P_{ij}^2}{K} - \frac{\sum_i P_{ii}^2}{K} \right) = C \frac{\sum_i P_{ii}(1 - P_{ii})}{K} > C > 0.$$

Also,  $E[\varepsilon_i^4] \leq C$  and as shown above,  $E[\sum_{i \neq j} (\varepsilon_i P_{ij} \varepsilon_j)^2] = 2KV_n$ . Now apply Lemma A1 with  $W_{in} = 0$ ,  $c_{1n} = 0$ , and  $c_{2n} = 1$ . It follows by the conclusion of Lemma A1 that

$$\frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2KV_n}} \xrightarrow{d} N(0, 1).$$

Next, by Theorem 1 of Hausman et. al. (2010) we have  $\hat{\delta} \xrightarrow{p} \delta$ , so that by Lemma A3,  $\hat{V}_n - V_n \xrightarrow{p} 0$ . Then by  $V_n$  bounded and bounded away from zero,  $\sqrt{V_n/\hat{V}_n} \xrightarrow{p} 1$ . Therefore by the Slutsky theorem,

$$\frac{\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{2K\hat{V}_n}} = \frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2K\hat{V}_n}} + \frac{o_p(1)}{\sqrt{2\hat{V}_n}} = \sqrt{\frac{V_n}{\hat{V}_n}} \frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2KV_n}} + o_p(1) \xrightarrow{d} N(0, 1).$$

Next, note that  $\hat{T} \geq q_{K-G}(1-\alpha)$  if and only if

$$\frac{\sum_{i \neq j} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{2K\hat{V}_n}} \geq \frac{q_{K-G}(1-\alpha) - K}{\sqrt{2K}}.$$

It is known that as  $K \rightarrow \infty$ ,  $[q_{K-G}(1-\alpha) - (K-G)]/\sqrt{2(K-G)} \rightarrow q(1-\alpha)$ , where  $q(1-\alpha)$  is the  $1-\alpha$  quantile of the standard normal distribution. Also, we have

$$= \sqrt{\frac{K-G}{K}} \left( \frac{q_{K-G}(1-\alpha) - (K-G)}{\sqrt{2(K-G)}} \right) - \frac{G}{\sqrt{2K}} \rightarrow q(1-\alpha).$$

The conclusion now follows. Q.E.D.

**PROOF OF THEOREM 2:** It follows in the usual way from the conditions that

$$\sqrt{n}(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, \sigma^2(D'Q^{-1}D)^{-1}).$$

In addition, it is straightforward to show that  $Z'Z/n \rightarrow Q$  nonsingular implies that  $\max_{i \leq n} P_{ii} \rightarrow 0$ ; e.g. see McFadden (1982). Furthermore, note for  $d_i = 3(1 + \|X_i\|^2)$  from the proof of Lemma A3 that

$$E[\sum_i P_{ii} d_i] \leq \sum_i P_{ii} E[d_i] \leq C,$$

so  $\sum_i P_{ii} d_i = O_p(1)$ . Then similarly to the proof of Lemma A3, by  $P_{ii} \geq 0$ ,

$$\left| \sum_i P_{ii} (\hat{\varepsilon}_i^2 - \varepsilon_i^2) \right| \leq \sum_i P_{ii} |\hat{\varepsilon}_i^2 - \varepsilon_i^2| \leq \sum_i P_{ii} d_i \|\hat{\delta} - \delta\| = O_p(1)o_p(1) \xrightarrow{p} 0.$$

Also, we have

$$\begin{aligned} E[(\sum_i P_{ii}\varepsilon_i^2 - K\sigma^2)^2] &= E[(\sum_i P_{ii}\{\varepsilon_i^2 - \sigma^2\})^2] = \sum_i P_{ii}^2 Var(\varepsilon_i^4) \\ &\leq C \max_{i \leq n} P_{ii} \sum_i P_{ii} \longrightarrow 0. \end{aligned}$$

Then by the Markov and Triangle inequalities,

$$\sum_i P_{ii}\varepsilon_i^2 \xrightarrow{p} K\sigma^2.$$

Also, since (as just shown)  $\sum_i P_{ii}^2 \longrightarrow 0$  it follows by Lemma A3 and  $\sigma_i^2 = \sigma^2$  that

$$\hat{V}_n - \sigma^4 = \frac{\sum_{i \neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2}{K} - \sigma^4 \frac{\sum_{i \neq j} P_{ij}^2}{K} - \sigma^4 \frac{\sum_i P_{ii}^2}{K} = o_p(1) + o(1) \xrightarrow{p} 0.$$

Therefore,

$$\hat{T} = \frac{\sigma^2}{\sqrt{\hat{V}}} \frac{\hat{\varepsilon}' P \hat{\varepsilon}}{\sigma^2} + K - \frac{\sum_{i=1}^n P_{ii} \hat{\varepsilon}_i^2}{\sqrt{\hat{V}}} = [1 + o_p(1)] \frac{\hat{\varepsilon}' P \hat{\varepsilon}}{\sigma^2} + o_p(1).$$

It follows by standard arguments that  $\hat{\varepsilon}' P \hat{\varepsilon} / \sigma^2 \xrightarrow{d} \chi^2(K - G)$ , so the conclusion follows by the Slutsky Lemma. Q.E.D.

## 5 References

- Ackerberg, D.A. and P. Devereaux (2003) Improved JIVE estimators for overidentified models with and without heteroskedasticity. Working Paper, University of California, Los Angeles.
- Anatolyev, S. and Gospodinov, N. (2009). Specification testing in models with many instruments, *Econometric Theory*, forthcoming.
- 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 variable estimators. *Econometrica* 62, 657-681.

- Bekker, P. A. and J. van der Ploeg (2005) Instrumental variable estimation based on grouped data. *Statistica Neerlandica* 59, 506-508.
- 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.C. and N.R. Swanson (2005) Consistent estimation with a large number of weak instruments. *Econometrica* 73, 1673-1692.
- Chao, J.C., N.R. Swanson, J.A. Hausman, W.K. Newey, and T. Woutersen.(2010): Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments, working paper.
- Fuller, W.A. (1977) Some properties of a modification of the limited information estimator,” *Econometrica* 45, 939-954.
- Hansen, L.P. (1982): Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029-1054.
- Hansen, C., J.A. Hausman, and W.K. Newey (2008) Estimation with many instrumental variables. *Journal of Business and Economic Statistics* 26, 398-422..
- Hausman, J.A., W.K. Newey, T.M. Woutersen, J. Chao, and N.R. Swanson (2010) IV estimation with heteroskedasticity and many instruments. Working Paper, MIT.
- 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.
- Lee, Y. and R. Okui (2010) A specification test for instrumental variables regression with many instrument, working paper.
- McFadden, D. (1982) Large sample properties of least squares, Lecture notes, MIT.
- 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 F. Windmeijer (2009) GMM with many weak moment conditions. *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.

Sargan, J.D. (1958), The estimation of economic relationships using instrumental variables, Econometrica, 26, pp.393–415.

White, H. (1980) A heteroskedasticity-consistent variance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817-38.