Supplementary Materials for

Conditional Linear Combination Tests for Weakly Identified Models

Isaiah Andrews

June 8, 2016

Section A gives empirical results for the Angrist and Krueger (1991) data and discusses implementation of PI tests and confidence sets. Section B provides further details on the derivation of the limit problems discussed in Section 2 of the paper. Section C shows that general nonlinear GMM models which are weakly identified in the sense of Stock and Wright (2000) give rise to limiting problems of the form (2). Section D concerns our linear IV simulations, gives power plots for PI tests in linear IV with homoskedastic errors, provides further information on our simulation design, and discusses our implementation of the MM1-SU, MM2-SU, QLR, and PI tests. Section E provides additional details on our implementation of PI and QLR tests in Section A. Finally, Section F discusses simulation results and derivations in a nonlinear new Keynesian Phillips curve model.

Supplementary Materials A: Application and Implementation

To illustrate the proposed procedures we calculate confidence sets using data from Angrist and Krueger (1991) and a range of specifications. We detail the steps required to construct joint plug-in confidence sets, and discuss computational considerations.

A.1 Data and Specifications

Angrist and Krueger (1991) study the relationship between years of schooling and labor market earnings. They note that children born later in the calendar year face a longer period of compulsory schooling than do those born earlier in the calendar year, and argue that a worker's quarter of birth thus provides a valid instrument from years of schooling. The baseline specifications of Angrist and Krueger (1991) take log wages to be linear in years of schooling, but a number of empirical studies (see Heckman et al. (2006), Heckman et al. (2008), and references therein) have found evidence of nonlinearity in this relationship. Here we consider a range of specifications which allow nonlinear effects.[20]

Angrist and Krueger (1991) assume a linear structural equation

$$\widetilde{Y}_t = \beta \widetilde{E}_t + \widetilde{W}_t' \delta + \varepsilon_t$$

where $\widetilde{Y}_t$ is log weekly wages, $\widetilde{E}_t$ is years of schooling, and $\widetilde{W}_t$ is a vector of control variables (including a constant). To relax the assumption that $\widetilde{Y}_t$ is linear in $\widetilde{E}_t$, here we consider specifications of the form

$$\widetilde{Y}_t = \beta_1 \widetilde{E}_t + \sum_{i=2}^{p} \beta_i 1\left\{\widetilde{E}_t \geq c_i\right\} + \widetilde{W}_t' \delta + \varepsilon_t$$

where c_i is a constant. We consider $c_i \in \{12, 14, 16, 18\}$, leading to the five nested specifications listed in Table 3. Following Angrist and Krueger (1991), Table I, we interpret 12 years of schooling as a high school degree, 14 years of schooling as two years of post-secondary education, 16 years of schooling as a college degree, and 18 years of schooling as a masters degree.

We focus on data from the 1930-1939 cohort. For this data, Angrist and Krueger (1991) show that quarter of birth has a statistically significant relationship (at the 5% level) to all the levels of education considered except for the masters degree. However, by including dummies for multiple levels of education in the same specification (as well as the linear term in $\widetilde{E}_t$) we are increasing the demands made of the instruments, and concerns about weak identification are relevant here.

[20] Considering multiple specifications for the Angrist and Krueger (1991) setting also allows us illustrate how the computational burden of the plug-in test scales with the number of endogenous regressors, holding all other features constant.

Specification	c_i	Highest c_i corresponds to
A	None	None
B	12	High school degree
C	12, 14	Two years of postsecondary education
D	12, 14, 16	College degree
E	12, 14, 16, 18	Masters degree

Table 3: Specifications for Angrist and Krueger (1991) data. Specification A is the linear specification considered by Angrist and Krueger (1991). Specification B adds a dummy for completing 12 years of schooling, which following Angrist and Krueger (1991) we interpret as completing high school. Specification C adds a dummy for completing 14 years of schooling, or two years of post-secondary education. Specification D adds a dummy for completing 16 years of schooling, which we interpret as completing college. Finally, Specification E adds a dummy for completing 18 years of education, which we interpret as completing a masters degree.

We choose controls $\widetilde{W}_t$ and instruments $\widetilde{Z}_t$ as in specification II of Staiger and Stock (1997) for the Angrist and Krueger (1991) data.[21] For each specification, we construct $\widetilde{X}_t = (E_t, 1\{E_t \geq c_1\}, ..., 1\{E_t \geq c_n\})'$. We eliminate the control variables $\widetilde{W}_t$ by projection, and for $\widetilde{X}$, $\widetilde{W}$, and $\widetilde{Z}$ matrices with rows $\widetilde{X}_t'$, $\widetilde{W}_t'$, and $\widetilde{Z}_t'$, respectively, we let $M_{\widetilde{W}} = I - \widetilde{W}\left(\widetilde{W}'\widetilde{W}\right)^{-1}\widetilde{W}'$ and define $X = M_{\widetilde{W}}\widetilde{X}$, $Y = M_{\widetilde{W}}\widetilde{Y}$, and $Z = M_{\widetilde{W}}\widetilde{Z}$.

To construct confidence sets it will be helpful to have a bounded parameter space for β. Here we consider $\beta_1 \in [-0.5, 0.5]$, which allows the marginal effect for each additional year of schooling to range from an almost 40% decrease in wages to an over 60% increase. For $i \geq 2$ we consider $\beta_i \in [-2, 2]$, which allows an additional almost 90% decrease to over 600% increase in wages for completing year c_i. In both cases this more than suffices to cover the economically plausible range of parameter values.

A.2 Implementation

We calculate joint confidence sets for the parameter vector β in each specification by inverting identification-robust tests. In particular, for a given test ϕ the corresponding confidence set for β will be $\{\beta_0 : \phi \text{ does not reject } H_0 : \beta = \beta_0\}$. We report PI, QCLR, S, K, and QLR confidence sets. We do not report confidence sets based on the MM procedures, since these procedures depend on a choice of weight function and MM only propose weights for the case with a single endogenous regressor, so their suggested tests

[21] This specification has 30 instruments, formed by interacting quarter of birth dummies with year of birth dummies.

are not directly applicable to specifications B-E.

The next subsection gives a detailed discussion of the implementation of the PI test at a given point. We discuss the computational demands of the PI test and the tradeoffs involved in particular implementation choices. Finally, we discuss how we invert the PI test (and the other robust tests) to construct confidence sets.

A.2.1 Calculating the PI Test

To evaluate the the PI test of $H_0 : \beta = \beta_0$ for a particular value β_0 in specifications A-E, we introduce finite-sample analogs S_T, D_T, K_T, and J_T to the limiting random variables S, D, K, and J, where $S_T = g_T' g_T$ and so on. We compute the PI test in four steps:

1. Calculate the statistics S_T, K_T, D_T, and $\hat{\gamma}$
2. Calculate $\hat{\mu}_D$
3. Calculate the plug-in weight $a_{PI}(D_T) = a_{MMRU}(\hat{\mu}_D)$
4. Calculate the critical value $c_\alpha(a_{MMRU}(\hat{\mu}_D))$ and evaluate the test

1: Calculate statistics For X_t, Y_t, and Z_t' the rows of X, Y, and Z respectively, we define $f_t(\beta)$ as in (4) and let $\hat{\Omega}(\beta)$ be the usual estimator for $Var\left(f_t(\beta)', vec\left(\frac{\partial}{\partial \beta'} f_t(\beta)\right)'\right)$. We take $\hat{\gamma} = vec\left(\hat{\Omega}\right)$. Decomposing $\hat{\Omega}$ as in (5), we then let

$$g_T(\beta) = \frac{1}{\sqrt{T}} \hat{\Omega}_{ff}(\beta)^{-\frac{1}{2}} \sum_{t=1}^{T} f_t(\beta)$$
$$\Delta g_T(\beta) = \frac{1}{\sqrt{T}} \hat{\Omega}_{ff}(\beta)^{-\frac{1}{2}} \sum_{t=1}^{T} \frac{\partial}{\partial \beta'} f_t(\beta)$$

and define $\hat{\Sigma}_{g\theta}(\beta)$, $\hat{\Sigma}_{\theta\theta}(\beta)$ as in Example I. Note that these definitions are just as in Example I, save where necessary to accommodate the presence of multiple endogenous regressors. Given g_T and Δg_T we then calculate S_T, K_T, and D_T in the same manner as S, K, and D, replacing all variance matrices by their estimates.

2: Calculate $\hat{\mu}_D$ A natural estimator for $\hat{\mu}_D$ is D_T. As we saw in Section 7.1, however, this choice yields sub-optimal power in linear IV with homoskedastic errors and a single endogenous regressor. Thus we instead use a generalization of the positive-part estimator considered in Sections 7.1 and 7.2. In particular, denoting column i of

D_T by $[D_T]_i$ we begin with an unbiased estimator of $R = \mu_D' \left(\sum_i Var\left([D_T]_i\right)\right)^{-1} \mu_D$ in the limit problem, restrict the eigenvalues of this estimator to be weakly positive, and choose $\hat{\mu}_D$ such that $\hat{\mu}_D' \left(\sum_i Var\left([D_T]_i\right)\right)^{-1} \hat{\mu}_D$ yields the chosen estimator of R. In cases with a single endogenous regressor this approach recovers the positive-part estimator used above. For more on $\hat{\mu}_D$, see Section E of the supplementary materials.

3: Calculate the plug-in weight We calculate $a_{MMRU}(\mu_D)$ using a discrete approximation to the MMRU problem. In particular, we restrict a to $A = \{0, 0.01, 0.02, ..., 1\}$ and for each element β_i of the vector β consider a uniform grid of J points,

$$\beta_i \in B_{i,J} = \left\{\beta_{i,L}, \beta_{i,L} + \frac{\beta_{i,U} - \beta_{i,L}}{J-1}, ..., \beta_{i,U}\right\}$$

where $\beta_{i,L}$ and $\beta_{i,U}$ are the lower and upper bounds of the parameter space for β_i. We then define B_J as the Cartesian product $B_{1,J} \times ... \times B_{n,J}$. For a given value of μ_D, we approximate $\mathcal{M}_D(\mu_D)$ by

$$M_D(\mu_D) = \left\{vec^{-1}\left(\left(I_{pk} - (\beta - \beta_0)' \otimes \left(\hat{\Sigma}_{\theta g}\hat{\Sigma}_{gg}^{-1}\right)\right)^{-1} vec(\mu_D)\right)(\beta - \beta_0) : \beta \in B_J\right\},$$

the set of values m consistent with μ_D and $\beta \in B_J$ (for $\hat{\Sigma}_{\theta g} = \Sigma_{\theta g}$).[22] We then calculate

$$a_{MMRU}(\hat{\mu}_D) = \arg\min_{a \in A} \sup_{m \in M_D(\hat{\mu}_D)} \left(\sup_{\tilde{a} \in A} E_{m,\hat{\mu}_D}[\phi_{\tilde{a}}] - E_{m,\hat{\mu}_D}[\phi_a]\right) \tag{26}$$

by evaluating $E_{m,\hat{\mu}_D}[\phi_a]$ for all $(a, m) \in A \times M$, which can be done by simulation.[23] To further speed this step, note that $E_{m,\hat{\mu}_D}[\phi_a] = \int E_{\tau_J(D),\tau_K(D)}[\phi_a] \, dF_D(\hat{\mu}_D)$ so we can tabulate $E_{\tau_J,\tau_K}[\phi_a]$ in advance and calculate the integral by simulation. Indeed, this is the approach we take in practice - see Section E for details.

4: Calculate the critical value Given $a_{MMRU}(\hat{\mu}_D)$, the conditional critical value $c_\alpha(a_{MMRU}(\hat{\mu}_D))$ is simply the $1-\alpha$ quantile of a $(1 - a_{MMRU}(\hat{\mu}_D))\chi_p^2 + a_{MMRU}(\hat{\mu}_D)\chi_{k-p}^2$ distribution for $a_{MMRU}(\hat{\mu}_D)$ fixed. By virtue of the discrete approximation to the

[22] Here we take $vec^{-1}(\cdot)$ to denote the inverse of the vectorization operator in (9). Thus, $vec^{-1}(\cdot)$ maps $kp \times 1$ vectors to $k \times p$ matrices.

[23] To reduce simulation noise, we in fact take $a_{MMRU}(\hat{\mu}_D)$ to be the largest value which comes within 10^{-5} of minimizing (26).

p	1	2	3	4	5
$J = 41$	1.38	28.66			
$J = 21$	1.09	7.90	157.84		
$J = 11$	0.97	3.21	21.63	275.91	
$J = 5$	0.89	1.60	3.84	10.61	52.70

Table 4: Runtime in seconds for computing PI tests in Matlab, averaged over 100 runs, using a laptop computer with an Intel i5 1.7 gigahertz processor and 4 GB of RAM**.** Vertical axis lists the number of grid points used, while horizontal axis lists the number of endogenous regressors. For empty cells, memory limits precluded efficient computation on laptop.

MMRU problem, we know that $a_{MMRU}(\hat{\mu}_D) \in A$. Thus, to save time we calculate the $1-\alpha$ quantile of $(1-a)\chi^2_p + a \cdot \chi^2_{k-p}$ for $a \in A$, based on 1 million simulations, before we begin and save the results. We then simply look up the appropriate critical value each time we evaluate the test.[24] Finally, given S_T, K_T, $a_{MMRU}(\hat{\mu}_D)$, and $c_\alpha(a_{MMRU}(\hat{\mu}_D))$ we evaluate the test

$$\phi_{PI} = 1\left\{K_T + a_{MMRU}(\hat{\mu}_D) \cdot J_T > c_\alpha(a_{MMRU}(\hat{\mu}_D))\right\}.$$

A.2.2 Computation Time

By taking J and A large, the discrete problem (26) can approximate the non-discretized minimax regret problem arbitrarily well. In step 3 above, however, we must evaluate the power $E_{m,\hat{\mu}_D}[\phi_a]$ at least $|M| \cdot |A| = J^p \cdot |A|$ times. Thus, there is a curse of dimensionality in the parameter β. When the dimension p is small we can take a fine grid of values β (that is, a large J) at little cost, but for p large the cost of increasing J can be high. To illustrate this, Table 4 reports average runtimes for computing PI tests in Matlab using an Intel i5 1.7 gigahertz processor and 4 GB of RAM.[25]

As we can see from Table 4, even on a relatively slow computer computing PI tests is reasonably fast when p is equal to one, taking less than 1.5 seconds for all values J considered. Even for higher-dimensional p, computing PI tests is reasonably fast when $J = 5$ (taking less than 15 seconds for $p < 5$). However, when we increase J and p together the computational burden of evaluating PI tests increases rapidly.

[24]Without the discrete approximation to the MMRU problem, we could instead calculate the critical value by simulation for each test evaluation.

[25]The final confidence sets were computed in parallel on a server, however.

Managing computational cost for large p: To reduce the computational demands of the PI tests in higher dimensional problems, we need to reduce the number of evaluations of $E_{m,\hat{\mu}_D}[\phi_a]$. If we are interested in power against a known subset of alternatives, restricting attention to this set naturally implies such a reduction. Absent such restrictions, we can reduce $|M|$ by limiting attention to deviations from the null along specific directions. For instance, while above we define B by taking the Cartesian product of the parameter grids $B_{i,J}$ for each parameter, we could instead restrict attention to alternatives which differ from the null only in one parameter at a time, holding the other parameters at their null values. This results in the MMRU test against the class of alternatives $\mathcal{M}_D$ which differ from the null only in one structural parameter. Calculating the discrete approximation to this problem requires only $|A| \cdot p \cdot J$ evaluations of $E_{m,\hat{\mu}_D}[\phi_a]$, which is an enormous reduction when p and J are large. The MMRU property against a smaller set $\mathcal{M}_D$ is of course weaker, but the computational savings may be essential when p is large. One can likewise redefine $\mathcal{M}_D$ to be the set of alternatives which differ in two or three parameters, as desired.

Computational Choices and Size Control It is important to emphasize that plug-in tests continue to control size even if we use a very crude algorithm to calculate $a_{PI}(D)$, since the resulting test remains a CLC test, and so controls size by Theorem 3. Thus, in cases where computing $a_{PI}(D)$ is challenging, doing a poor job on this step may result in a test with inferior power but will not lead to size distortion.

A.2.3 Computing Confidence Sets

As discussed above, the PI confidence set is obtained by inverting the PI test, yielding $\{\beta_0 : \phi_{PI} \text{ does not reject } H_0 : \beta = \beta_0\}$. We construct an approximate confidence set by drawing a million points β_0 uniformly at random from the parameter space for β.[26] For each draw β_0, we test the null that β_0 is the true value and keep β_0 only if the null is not rejected.[27] In the present application sampling uniformly from the

[26]The one exception are the results for the conditional QLR test of I. Andrews and Mikusheva (2016*a*) reported in columns C-E of Table 6, which due to computational cost are based on one hundred thousand draws in specification C and ten thousand draws in specifications D and E. See Section E for further details on the implementation of this test.

[27]The reported confidence sets are then constructed using the non-rejected points, for example taking the min and max to obtain a confidence interval, and including a neighborhood of all non-rejected points to construct a two-dimensional confidence set.

	QCLR	PI	S	K	QLR
95% Confidence Set	[0.046,0.128]	[0.046,0.128]	[-0.002,0.186]	[0.047,0.126]	[0.045,0.128]
CI Length	0.082	0.083	0.188	0.079	0.083

Table 5: Confidence sets in specification A for Angrist and Krueger (1991) data, constructed as discussed in Section A.2. PI confidence set is constructed using $J = 41$ grid points.

parameter space performs well, but one could also consider more refined MCMC-based approaches as in Chernozhukov et al. (2009).

Since we compute the PI test using the discretized problem (26), note that a point β_0 can be included in the PI confidence set only if it is included in the confidence set for some fixed-weight linear combination test which takes $a(D) = a \in A$, that is, if

$$\min_{a \in A} \left(1\left\{(1-a) \cdot K_T + a \cdot S_T > c_\alpha(a)\right\}\right) = 0. \tag{27}$$

Once we've calculated the statistics K_T and S_T, however, it is very fast to check whether (27) holds. Thus we only calculate the PI test when (27) holds, which substantially reduces the number of times we need to evaluate the PI test, the exact degree of reduction depending on the specification considered.

A.3 Empirical Results:

We now report the identification-robust confidence sets obtained using the Angrist and Krueger (1991) data. Table 5 gives confidence sets in specification A, where the PI confidence set is computed using $J = 41$ grid points. Here we see that the QCLR, PI, K, and QLR confidence sets are all quite similar, with the PI and QLR confidence sets being slightly longer that the QCLR confidence set, while the K confidence set is somewhat shorter than QCLR. Finally, the S confidence set is over twice as long as the others, and is the only confidence set considered which includes zero.

Figure 4 plots confidence sets for specification B. Here the QCLR, PI, S, K, and QLR confidence sets cover 6.8%, 6.4%, 15.8%, 5.1%, and 6.7% of the parameter space, respectively, and the ordering of the confidence sets in terms of volume is much as in specification A, save that the PI confidence set is now smaller than the QCLR and QLR confidence sets, and the QLR confidence set is smaller than the QCLR. When projected on the axes no confidence set excludes zero for either parameter, though

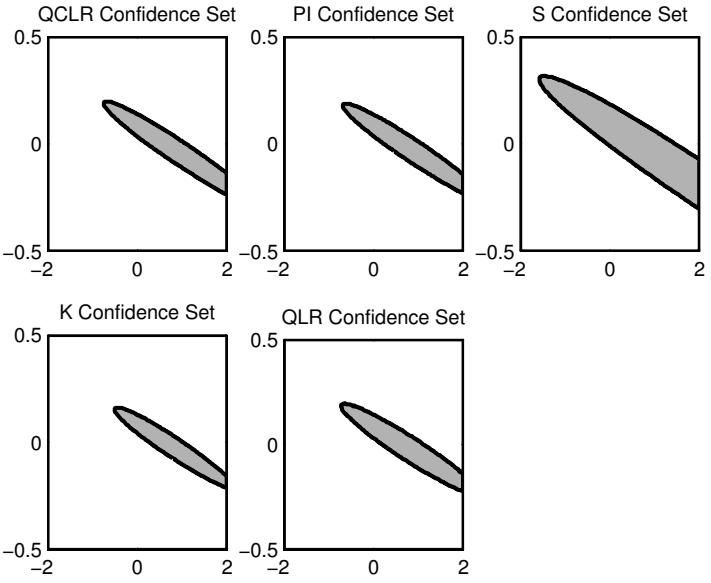

Figure 4: Confidence sets in specification B for Angrist and Krueger (1991) data, constructed as discussed in Section A.2. PI confidence set based on $J = 41$ grid points. Vertical axis corresponds to coefficient on E, while horizontal axis corresponds to coefficient on $1\{E_t > 12\}$.

all confidence sets other than S exclude (0,0) so the corresponding tests reject the joint hypothesis of zero coefficients for both endogenous regressors. Moreover, as is intuitively reasonable all confidence sets suggest some substitution between the two endogenous regressors: the larger the linear effect of education, the smaller the effect of graduating high school must be to justify the observed data.

It is difficult to report results for specifications C-E since the joint confidence sets are of dimension three or more, and we must project on lower-dimensional subspaces to obtain easily reported objects.[28] Table 6 reports one-dimensional confidence sets obtained by projecting the joint confidence sets on the individual parameters, where we use $J = 5$ grid points to calculate PI tests in all cases for consistency. From this table, several points become clear. First, for specification A reducing from $J = 41$ to $J = 5$ grid points in the calculation of the PI test has only a small effect on the PI

[28]Note that we may also be interested in confidence sets for individual parameters for their own sake. For this purpose, projection-based intervals of the type considered here will be valid but typically conservative. Unfortunately, eliminating this conservativeness is a open problem, and PI tests constructed by plugging in estimates for poorly identified nuisance parameters will not in general have correct size. For related results, see Guggenberger et al. (2012).

confidence set, while this change has a somewhat larger effect in specification B. As we increase the number of endogenous regressors the one-dimensional confidence sets become less informative, and in specifications D and E cover the full parameter space for each of the parameters. At the same time, the fraction of the joint parameter space covered by the confidence sets does not increase nearly so dramatically, so the large size of marginal confidence sets reflects in part conservativeness resulting from projection. We see that the S confidence set has the largest volume in all specifications other than D, while the PI confidence set is smaller than the QCLR confidence set in all specifications save A, often by a substantial margin. The QLR confidence set is small in specifications A and B, but its volume increases in specifications C-E. While the K confidence set has the smallest volume in most specifications studied, this is the result of the bounds chosen for the parameter space. Even in specification A, if we double the bounds of the parameter space (results not shown) we find that the K confidence set has two disjoint components while the PI, QCLR, and QLR confidence sets remain connected and have less than half the volume of the K confidence set.

Supplementary Materials B: Derivation of Limit Problems for Examples

In this section, we provide additional details on the derivation of the limit problems in examples I and II.

Example I: Weak IV Re-writing our moment condition we have that

$$f_T(\beta_0) = f_T(\beta_0) - f_T(\beta) + f_T(\beta) = \frac{1}{T}\sum \left(X_t\beta - X_t\beta_0\right) Z_t + f_T(\beta).$$

Note that the expectation of $f_T(\beta)$ under true parameter value β is zero by our identifying assumption, so $E_\beta\left[f_T(\beta_0)\right] = E\left[\frac{1}{T}\sum X_t Z_t\right](\beta - \beta_0)$. Since

$$E\left[\frac{1}{T}\sum X_t Z_t\right] = E\left[\frac{1}{T}\sum Z_t\left(Z_t'\pi + V_{2,t}\right)\right] = E\left[\frac{1}{T}\sum Z_t Z_t'\right]\pi,$$

we can see that provided that $\frac{1}{T}\sum Z_t Z_t' \to_p Q_Z$ for Q_Z positive definite and $\frac{1}{\sqrt{T}}\sum Z_t V_{1,t}$ and $\frac{1}{\sqrt{T}}\sum Z_t V_{2,t}$ converge in distribution to jointly normal random vectors, the weak-

Specification		A	B	C	D	E
E_t	QCLR	[0.046,0.128]	[-0.232,0.198]	[-0.481,0.5]	[-0.5,0.5]	[-0.5,0.5]
	PI	[0.046,0.130]	[-0.213,0.181]	[-0.464,0.5]	[-0.5,0.5]	[-0.5,0.5]
	S	[-0.002,0.185]	[-0.296,0.316]	[-0.5,0.5]	[-0.5,0.5]	[-0.5,0.5]
	K	[0.047,0.126]	[-0.206,0.163]	[-0.451,0.5]	[-0.5,0.5]	[-0.5,0.5]
	QLR	[0.045,0.128]	[-0.217,0.194]	[-0.481,0.5]	[-0.5,0.5]	[-0.5,0.5]
$1\{E_t \geq 12\}$	QCLR		[-0.733,2]	[-2,2]	[-2,2]	[-2,2]
	PI		[-0.601,2]	[-2,2]	[-2,2]	[-2,2]
	S		[-1.555,2]	[-2,2]	[-2,2]	[-2,2]
	K		[-0.488,2]	[-2,2]	[-2,2]	[-2,2]
	QLR		[-0.705,2]	[-2,2]	[-2,2]	[-2,2]
$1\{E_t \geq 14\}$	QCLR			[-2,2]	[-2,2]	[-2,2]
	PI			[-2,2]	[-2,2]	[-2,2]
	S			[-2,2]	[-2,2]	[-2,2]
	K			[-2,2]	[-2,2]	[-2,2]
	QLR			[-2,2]	[-2,2]	[-2,2]
$1\{E_t \geq 16\}$	QCLR				[-2,2]	[-2,2]
	PI				[-2,2]	[-2,2]
	S				[-2,2]	[-2,2]
	K				[-2,2]	[-2,2]
	QLR				[-2,2]	[-2,2]
$1\{E_t \geq 18\}$	QCLR					[-2,2]
	PI					[-2,2]
	S					[-2,2]
	K					[-2,2]
	QLR					[-2,2]
Volume	QCLR	8.23%	6.82%	11.92%	11.21%	13.14%
	PI	8.39%	6.02%	8.50%	8.56%	10.8%
	S	18.75%	15.75%	16.77%	15.13%	16.34%
	K	7.91%	5.09%	7.51%	8.28%	11.64%
	QLR	8.33%	6.69%	14.22%	15.25%	15.94%

Table 6: One-dimensional projections and volume of 95% joint confidence sets in specifications A-E for Angrist and Krueger (1991) data, constructed as discussed in Section A.2. PI confidence sets are computed using $J = 5$ grid points. Volume is the percent of parameter space covered by the joint confidence set.

instruments sequence $\pi_T = \frac{c}{\sqrt{T}}$ implies that under true parameter value β,

$$\sqrt{T}\begin{pmatrix} f_T(\beta_0) \\ -\frac{\partial}{\partial\beta} f_T(\beta_0) \end{pmatrix} \to_d N\left(\begin{pmatrix} Q_Z c(\beta - \beta_0) \\ Q_Z c \end{pmatrix}, \Omega(\beta_0)\right).$$

Combined with the consistency of $\hat{\Omega}_{ff}$, this immediately yields (6).

Example II: Minimum Distance The identifying assumption for the minimum distance model imposes $\eta = f(\theta)$. Note, however, that

$$g_T(\theta_0) = \hat{\Omega}_\eta^{-\frac{1}{2}}(\hat{\eta} - f(\theta_0)) = \hat{\Omega}_\eta^{-\frac{1}{2}}(\hat{\eta} - f(\theta)) + \hat{\Omega}_\eta^{-\frac{1}{2}}(f(\theta) - f(\theta_0))$$

where by assumption the first term converges to a $N(0, I_k)$ distribution and the second term converges to $\Omega_\eta^{-\frac{1}{2}}(f(\theta) - f(\theta_0))$ by the Continuous Mapping Theorem and the assumed consistency of $\hat{\Omega}_\eta$. The consistency of $\Delta g_T(\theta)$ for $\Omega^{-\frac{1}{2}}\frac{\partial}{\partial\theta} f(\theta_0)$ follows similarly, immediately implying (7).

Supplementary Materials C: Limit Problem for Weak GMM Models

In this section, we prove some additional results for GMM models which are weakly identified in the sense of Stock and Wright (2000). Suppose we begin with a moment function $f_t(\psi)$ which is differentiable in the parameter ψ and satisfies the usual GMM identifying assumption that $E_\psi[f_t(\psi)] = 0$, and are interested in testing $H_0 : \psi = \psi_0$. Suppose that, much like in Stock and Wright (2000), our parameter vector $\psi = (\psi_1, \psi_2)$ is such that ψ_1 is weakly identified while ψ_2 is strongly identified, and that the expectation of $f_t(\psi_0)$ under alternative ψ is

$$E_\psi[f_t(\psi_0)] = \tilde{h}_1(\psi_1) + \frac{1}{\sqrt{T}}\tilde{h}_2(\psi_1, \psi_2)$$

for $\tilde{h}_1$, $\tilde{h}_2$ continuously differentiable. Letting ψ denote the true parameter value, for sample size T let us reparametrize in terms of $\theta = \theta_T = \left(\sqrt{T}(\psi_1 - \psi_{1,0}), \psi_2\right)$ and note that the null can now be written $H_0 : \theta = \theta_0 = (0, \theta_{2,0})$. This reparameterization is infeasible as it demands knowledge of the unknown true value ψ_1, but this is irrelevant

provided we use a test which is invariant to linear reparameterizations. Let

$$g_t(\theta) = f\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}, \theta_2\right)$$

denote the moment function under this new parametrization, and note that the expectation of $g_t(\theta_0)$ under alternative θ is

$$\begin{aligned} E_\theta\left[g_t(\theta_0)\right] &= \tilde{h}_1\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}\right) + \frac{1}{\sqrt{T}}\tilde{h}_2\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}, \theta_2\right) \\ &= \tilde{h}_1\left(\psi_{1,0} - \frac{\theta_1}{\sqrt{T}}\right) + \frac{1}{\sqrt{T}}\tilde{h}_2\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}, \theta_2\right) \\ = \tilde{h}_1\left(\psi_{1,0}\right) + \frac{1}{\sqrt{T}}\frac{\partial}{\partial\psi_1'}\tilde{h}_1\left(\psi_{1,0} + \frac{\bar{\theta}_1}{\sqrt{T}}\right)\theta_1 &+ \frac{1}{\sqrt{T}}\tilde{h}_2\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}, \theta_2\right) \end{aligned}$$

where in the last step we have taken a mean value expansion in θ_1 with intermediate value $\bar{\theta}_1$. Note, however, that the identifying assumption for GMM implies that $\tilde{h}_1\left(\psi_{1,0}\right) = 0$ while under our continuity assumptions

$$\frac{\partial}{\partial\psi_1'}\tilde{h}_1\left(\psi_{1,0} + \frac{\bar{\theta}_1}{\sqrt{T}}\right) \to \frac{\partial}{\partial\psi_1'}\tilde{h}_1\left(\psi_{1,0}\right)$$

and

$$\tilde{h}_2\left(\psi_{1,0} + \frac{\theta_1}{\sqrt{T}}, \theta_2\right) \to \tilde{h}_2\left(\psi_{1,0}, \theta_2\right).$$

Hence, $E_\theta\left[g_t(\theta_0)\right] = h(\theta) + o\left(\frac{1}{\sqrt{T}}\right)$ where

$$h(\theta) = \frac{1}{\sqrt{T}}\frac{\partial}{\partial\psi_1'}\tilde{h}_1\left(\psi_{1,0}\right)\theta_1 + \frac{1}{\sqrt{T}}\tilde{h}_2\left(\psi_{1,0}, \theta_2\right). \tag{28}$$

Note that the strongly identified parameters θ_1 enter $h(\theta)$ linearly while the weakly identified parameters θ_2 may enter non-linearly. Suppose that for our original moment functions $f_t(\theta)$, we have that under the sequence of alternatives $\psi_T = \left(\psi_{1,0} - \frac{1}{\sqrt{T}}\theta_1, \psi_2\right)$

$$\frac{1}{\sqrt{T}}\begin{pmatrix} \sum f_t(\psi_0) - E_{\psi_T}\left[f_t(\psi_0)\right] \\ vec\left(\sum \frac{\partial}{\partial\psi'} f_t(\psi_0) - E_{\psi_T}\left[\frac{\partial}{\partial\psi'} f_t(\psi_0)\right]\right) \end{pmatrix} \to_d N\left(0, \Omega_f\right)$$

where Ω_f is consistently estimable and Ω_{ff}, the upper-left block of Ω_f, is full rank. Since alternative θ in the new parametrization corresponds to this sequence of alternatives in the original parametrization, this implies that under θ we have

$$\frac{1}{\sqrt{T}}\begin{pmatrix} \sum g_t(\theta_0) - E_\theta\left[g_t(\theta_0)\right] \\ vec\left(\sum \frac{\partial}{\partial\theta'} g_t(\theta_0) - E_\theta\left[\frac{\partial}{\partial\theta'} g_t(\theta_0)\right]\right) \end{pmatrix} \to_d N(0,\Omega)$$

for $\Omega = \begin{pmatrix} \Omega_{gg} & \Omega_{g\theta} \\ \Omega_{\theta g} & \Omega_{\theta\theta} \end{pmatrix}$ consistently estimable and $\Omega_{gg} = \Omega_{ff}$ full-rank. Letting

$$g_T(\theta_0) = \frac{1}{\sqrt{T}}\hat{\Omega}_{gg}^{-\frac{1}{2}} \sum_t g_t(\theta_0)$$

and

$$\Delta g_T(\theta_0) = \frac{1}{\sqrt{T}}\hat{\Omega}_{gg}^{-\frac{1}{2}} \sum_t \frac{\partial}{\partial\theta} g_t(\theta_0)$$

note that

$$\begin{pmatrix} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{pmatrix} \to_d N\left(\begin{pmatrix} m \\ \mu \end{pmatrix}, \begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix}\right)$$

where $\mu = \lim_{T\to\infty} E_\theta\left[\frac{1}{\sqrt{T}}\sum \frac{\partial}{\partial\theta'} g_t(\theta_0)\right]$ provided this limit exists and $m = h(\theta) \in \mathcal{M}(\mu,\gamma)$, where $\mathcal{M}(\mu,\gamma)$ will depend on the structure of the problem at hand: in some cases it may be that without additional structure we cannot restrict the set of possible values m and have $\mathcal{M}(\mu) = \mathbb{R}^k$ while in others, like Example I, we may be able to obtain further restrictions. Note further, that while we framed the analysis here using reparameterization in terms of local alternatives for strongly identified parameters, we could equivalently have formulated Δg_T using the Jacobian of the original moment function, $\frac{\partial}{\partial\psi'} f_t(\psi_0)$, post-multiplied by an appropriate sequence of normalizing matrices A_T, as in Appendix 1.

We can say a bit more regarding the strongly identified parameters θ_1. Note that by the definition of θ, $\frac{\partial}{\partial\theta_1'} g_t(\theta_0) = \frac{1}{\sqrt{T}}\frac{\partial}{\partial\psi_1'} f_t(\psi_0)$. Hence, $\frac{1}{\sqrt{T}}\sum \frac{\partial}{\partial\theta_1'} g_t(\theta_0) = \frac{1}{T}\sum \frac{\partial}{\partial\psi_1'} f_t(\psi_0)$ and we can re-write μ as $\lim_{T\to\infty} E_\theta\left[\begin{array}{cc} \frac{1}{T}\sum \frac{\partial}{\partial\psi_1'} f_t(\psi_0) & \frac{1}{\sqrt{T}}\sum \frac{\partial}{\partial\psi_2'} f_t(\psi_0) \end{array}\right]$. Further, the central limit theorem we have assumed for $\frac{1}{\sqrt{T}}\sum \frac{\partial}{\partial\psi_1'} f_t(\psi_0)$ implies that $\frac{1}{\sqrt{T}}\sum \frac{\partial}{\partial\theta_1'} g_t(\theta_0) \to_p \mu_1 = \lim_{T\to\infty} E_\theta\left[\frac{1}{T}\sum \frac{\partial}{\partial\psi_1'} f_t(\psi_0)\right]$. Together with (28) this implies that under standard regularity conditions (see e.g. Newey and McFadden (1994)) $h(\theta_1,\theta_{2,0}) = \mu_1\cdot\theta_1$ and hence that in the special case where all parameters are strongly identified we obtain

the Gaussian shift limiting problem

$$\begin{pmatrix} g \\ \Delta g \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \cdot \theta \\ \mu \end{pmatrix}, \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\right).$$

Supplementary Materials D: Further Details of Weak IV Simulations

In Section 7.2 of the paper we discuss simulation results in weak IV limit problems calibrated to match parameters estimated using data from Yogo (2004). This section details the estimates, simulation design, and implementation of CLR, MM, QLR, and PI tests underlying these results. We also give some additional results on the linear IV model with homoskedastic errors, and the performance of MM tests with tuning parameters as in Moreira and Moreira (2013).

D.1: Results for Homoskedastic IV Model

We directly compare the weight function $a_{CLR}(D)$ implied by the CLR test to the plug-in weight functions $a_{PI}(D)$ for plug-in tests in the homoskedastic IV model. We then plot the power curves of all tests considered.

D.1.1 Weight Function Comparison

The task of comparing the weight functions implied by PI tests for the various estimators of r is considerably simplified by the following lemma:

Lemma 2 *For A and B symmetric positive-definite matrices of dimension 2×2 and $k \times k$, respectively, the function $a_{MMRU}(\mu_D)$ in the limit problem (6) with $\Omega = A \otimes B$ can be taken to depend on μ_D only through $r = \mu_D' \Sigma_D^{-1} \mu_D$.*

Proof: Note that $\Omega = A \otimes B$ implies that $\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix} \otimes I$. To prove the result, it is easier to work with the formulation of the problem discussed in AMS. In particular, consider $k \times 1$ random vectors $\widetilde{S}$ and $\widetilde{T}$ (denoted by S and T in AMS) with

$$\begin{pmatrix} \widetilde{S} \\ \widetilde{T} \end{pmatrix} \sim N\left(\begin{pmatrix} c_\beta \mu_\pi \\ d_\beta \mu_\pi \end{pmatrix}, I\right),$$

were c_β ranges over $\mathbb{R}$ for different true values of β. AMS (Theorem 1) show that the maximal invariant to rotations of the instruments is $\left(\widetilde{S}'\widetilde{S}, \widetilde{S}'\widetilde{T}, \widetilde{T}'\widetilde{T}\right)$, and note that the S statistic can be written $S = \widetilde{S}'\widetilde{S}$, while the K statistic is $K = \frac{\left(\widetilde{S}'\widetilde{T}\right)^2}{\widetilde{T}'\widetilde{T}}$. Kleibergen (2007) considers a finite-sample Gaussian IV model with a known covariance matrix for the structural errors, and his Theorem 3 establishes that (in our notation) $\widetilde{T}'\widetilde{T} = D'\Sigma_D^{-1}D'$, where $\Sigma_D = I\left(\frac{A_{22}}{A_{11}} - \left(\frac{A_{12}}{A_{11}}\right)^2\right)$. Hence, in the limit problem (6) with $\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix} \otimes I$, the maximal invariant under rotations of the instruments $\left(\widetilde{S}'\widetilde{S}, \widetilde{S}'\widetilde{T}, \widetilde{T}'\widetilde{T}\right)$ is a one-to-one transformation of $(J, K, D'\Sigma_D D)$.

By the imposed invariance to rotations of the instruments, it is without loss of generality to assume that $d_\beta\mu_\pi = e_1 \cdot \sqrt{r}$, where $e_1 \in \mathbb{R}^k$ has a one in its first entry and zeros everywhere else. Hence, $\tilde{T}'\tilde{T} = D'\Sigma_D D \sim \chi^2_k(r)$. For fixed r, the distribution of $(J, K, D'\Sigma_D D)$ depends only on $c_\beta\mu_\pi = ||m||e_1$ and on consistently estimable parameters. The value of r imposes no restrictions on the value of $||m||$. Hence, the power of any unconditional linear combination test ϕ_a can be written as a function of $||m||$ and r, the power envelope for unconditional linear combination tests is defined by $\beta^u_{||m||,r} = \sup_{a\in[0,1]} E_{||m||,r}[\phi_a]$, and the maximum regret for any unconditional linear combination test (taking μ_D and hence r to be known) is

$$\sup_{||m||\in\mathbb{R}_+} \left(\beta^u_{||m||,r} - E_{||m||,r}[\phi_a]\right)$$

which depends only on r. We can thus take the MMRU test $\phi_{MMRU}(\mu_D)$ to depend on μ_D only through $r = \mu_D'\Sigma^{-1}\mu_D$.□

Since the weights of both the CLR test and the plug-in approaches discussed in Section 6.1 depend on $\hat{r}$ alone, in Figure 5 we plot the values of $a_{CLR}(\hat{r})$, $a_{MMRU}(\hat{r})$, $a_{MMRU}(\hat{r}_{MLE})$, $a_{MMRU}(\hat{r}_{PP})$, and $a_{MMRU}(\hat{r}_{KRS})$ as functions of $\hat{r}$ for $k = 5$. All the weight functions exhibit similar qualitative behavior, placing large weight on S for small values of $\hat{r}$ and increasing the weight on K as $\hat{r}$ grows, but there are some notable differences. Perhaps most pronounced, $a_{MMRU}(\hat{r})$ is lower than any of the other functions, as is intuitively reasonable given that $\hat{r}$ tends to overestimate r. As previously noted both $\hat{r}_{MLE}$ and $\hat{r}_{PP}$ are zero for a range of strictly positive values $\hat{r}$.

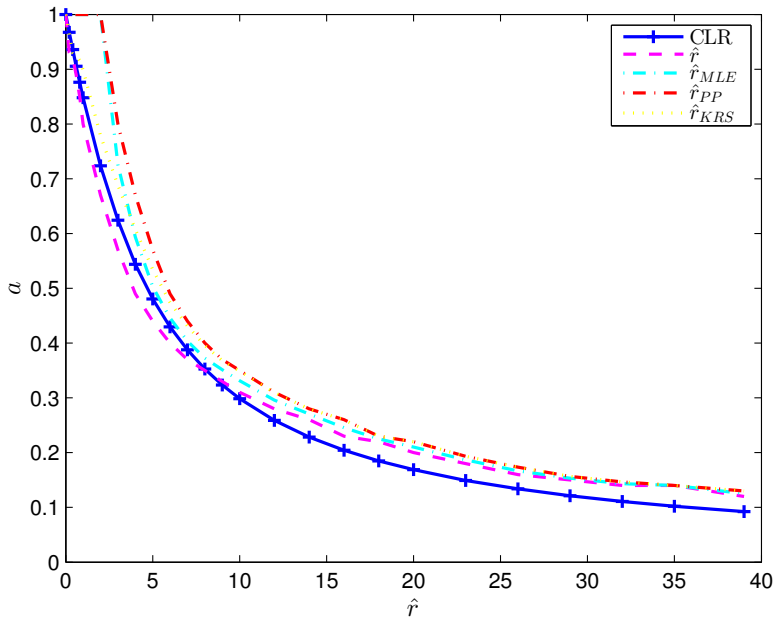

Figure 5: Weight functions $a_{CLR}(\hat{r})$ for CLR and $a_{MMRU}\left(\tilde{r}\left(\hat{r}\right)\right)$ for PI tests with different estimators $\tilde{r}$ of r discussed in Section 6.1 for linear IV with five instruments and homoskedastic errors.

D.1.2 Power Curves for Homoskedastic IV Model

Figures 6 and 7 plot power curves for the CLR, K, AR, and PI tests in the linear IV calibrations discussed in Section 7.1 of the paper.

D.2: Estimation of Parameters for the Limit Problem

The behavior of $(g, \Delta g)$ in the weak IV limit problem (6) is determined entirely by (m, μ, Ω). The set $\mathcal{M}(\mu)$ of possible values m given μ is $\mathcal{M}(\mu) = \{b \cdot \mu : b \in \mathbb{R}\}$, so to simulate the power properties of different tests in the limit problem all we require are values of μ and Ω.

To obtain values for these parameters, as noted in the text we use data from Yogo's (2004) paper on weak instrument-robust inference on the elasticity of inter-temporal substitution. For all countries we use quarterly data for a (country-specific) period beginning in the 1970's and ending in the late 1990's. We focus on estimation based

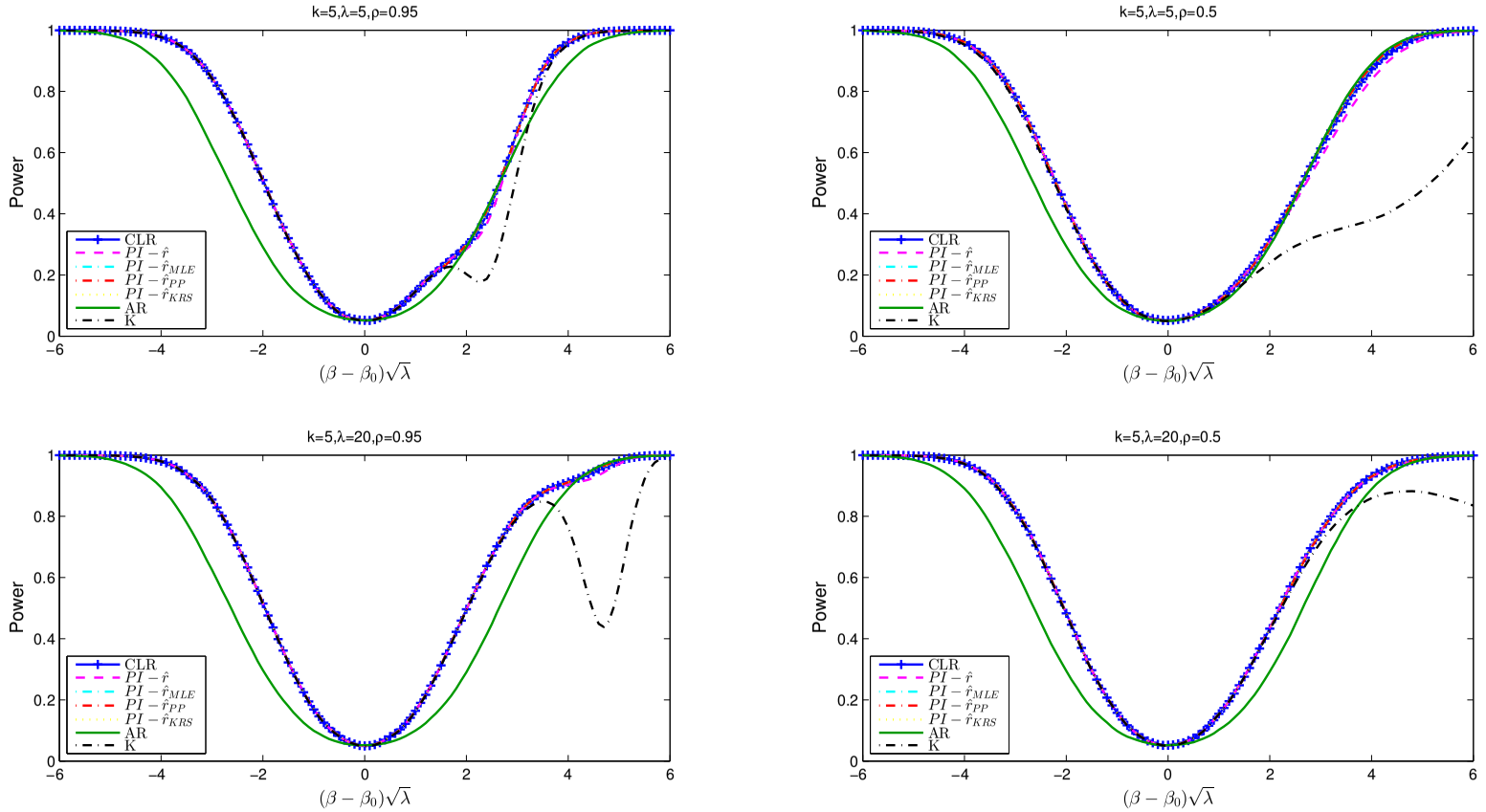

Figure 6: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with five instruments, discussed in Section 7.1.

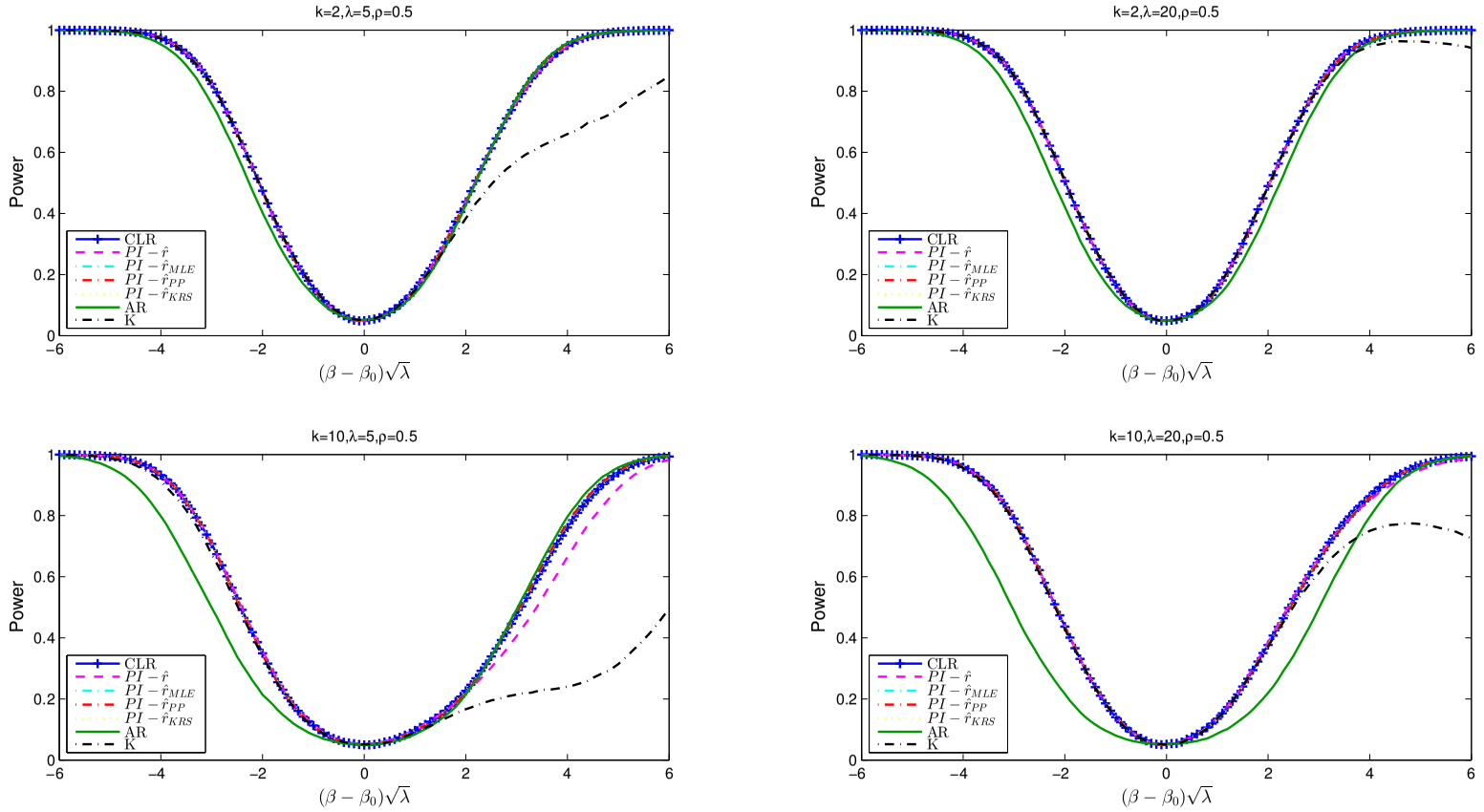

Figure 7: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with two instruments and ten instruments, discussed in Section 7.1.

on the linear IV moment condition

$$f_t(\beta) = Z_t \left(Y_t - X_t \beta \right)$$

where Y_t is the change in consumption (Yogo's Δc), X_t is the real interest rate, and Z_t is a 4×1 vector of instruments which following Yogo we take to be lagged values of the nominal interest rate, inflation, consumption growth, and the log dividend-price ratio. We focus on the case with X_t the risk-free rate since this is the case for which Yogo finds the strongest relationship between the instruments and the endogenous regressor (see Table 1 of Yogo (2004)). All data is de-meaned prior to beginning the analysis.

For country i we estimate μ by $\hat{\mu}_i = \frac{1}{\sqrt{T}} \sum Z_t X_t$, take $\hat{\beta}_i$ to be the two-step GMM estimate of β, and let $\hat{\Omega}_i$ be the Newey-West covariance estimator for $Var\left(\frac{1}{\sqrt{T}} \sum \left(f_t(\hat{\beta}_i)', Z_t' X_t \right)' \right)$ based on 3 lags of all variables. These estimates will not in general be consistent for the parameters of the limit problem under weak-instrument asymptotics, but give us empirically reasonable values for our simulations.

D.3: Simulation Design

For each country i we consider the problem of testing $H_0 : \beta = \beta_0$ in the limit problem. For true parameter value β and $\tilde{\mu}_i = \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \hat{\mu}_i$, in simulation runs $b = 1, ..., B$ we draw

$$\begin{pmatrix} g_b \\ \Delta g_b \end{pmatrix} \sim N\left(\begin{pmatrix} \tilde{\mu}_i \left(\beta - \beta_0 \right) \\ \tilde{\mu}_i \end{pmatrix}, \hat{\Sigma}_i \right)$$

where

$$\hat{\Sigma}_i = \begin{bmatrix} I & \hat{\Sigma}_{g\theta,i} \\ \hat{\Sigma}_{\theta g,i} & \hat{\Sigma}_{\theta\theta,i} \end{bmatrix} = \begin{bmatrix} I & \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \hat{\Omega}_{f\beta,i} \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \\ \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \hat{\Omega}_{\beta f,i} \hat{\Omega}_{ff,i}^{-\frac{1}{2}} & \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \hat{\Omega}_{\beta\beta,i} \hat{\Omega}_{ff,i}^{-\frac{1}{2}} \end{bmatrix}.$$

Note that this is the limiting distribution (6) of the normalized moment condition and Jacobian $(g_T, \Delta g_T)$ in a weak IV problem with true parameters β, $\Omega = \hat{\Omega}_i$, and $\mu = \hat{\mu}_i$. We then calculate the S and K tests $\phi_{S,b}$, $\phi_{K,b}$ as in (17) and (16). We define the QCLR test as in Theorem 4 and, following Kleibergen (2005), take $r = D_b' \hat{\Sigma}_{D,i}^{-1} D_b$ for $\hat{\Sigma}_{D,i} = \hat{\Sigma}_{\theta\theta,i} - \hat{\Sigma}_{\theta g,i} \hat{\Sigma}_{g\theta,i}$. Details on the implementation of the MM-SU tests are

discussed in the next section. Finally, to calculate the PI test we take

$$\hat{\mu}_{D,b} = D_b \cdot \frac{\sqrt{\max\left\{D_b'\hat{\Sigma}_{D,i}^{-1}D_b - k, 0\right\}}}{\sqrt{D_b'\hat{\Sigma}_{D,i}^{-1}D_b}}$$

which is a generalization of the positive-part estimator $\hat{r}_{PP}$ to the non-Kronecker case, and consider $\phi_{PI,b} = \phi_{MMRU\left(\hat{\mu}_{D,b}\right)}$. Details on calculation of the PI test are given in the Section D.7.

For each value β in a grid we estimate the power of each test against this alternative by averaging over B, e.g. estimating the power of ϕ_K by $\frac{1}{B}\sum \phi_{K,b}$ (where we take $B = 5,000$), and repeat this exercise for each of the eleven countries considered.

D.4: The MM1-SU and MM2-SU Tests

The MM1-SU and MM2-SU procedures of MM maximize weighted average power against particular weights on (β, μ) over a class of tests satisfying a sufficient condition for local unbiasedness. To apply the results of MM in our context, we can take $(Z'Z)^{-\frac{1}{2}}Z'Y$ (in the notation of MM) to equal $\begin{pmatrix} g & \Delta g \end{pmatrix}$ and then derive their statistics S and T as they describe, noting that S as defined in MM is, up to rotation, equal to g as defined here. MM calculate their weights using the 2×2 and $k \times k$ symmetric positive definite matrices Ω^* and Φ^* solving $\min \|\Sigma - \Omega \otimes \Phi\|_F$ (see MM for the weights). To choose the scalings for (Ω^*, Φ^*), we follow van Loan and Pitsianis (1993) and normalize the Frobeinus norm of Φ^* to one. Thus, since we estimate different covariance matrices for each of the 11 countries in the Yogo data, the MM tests use different weight functions for each country. For each pair (Ω^*, Φ^*) MM consider two different weight functions, which they label MM1 and MM2 respectively. Each of these weight functions features a tuning parameter (which MM call σ and ζ for the MM1 and MM2 weights, respectively). Following MM we set both σ and ς equal to one tenth of the sample size. Results based on an alternative choice of tuning parameters, as in Moreira and Moreira (2013), are reported below.

MM consider several different tests based on their weights. They find in simulation that weighted average power optimal conditionally similar tests can have some undesirable power properties in non-homoskedastic linear IV models, in particular exhibiting substantial bias. To remedy this they impose further restrictions on the class of tests, considering first locally unbiased (LU) tests, which satisfy $\frac{\partial}{\partial \beta} E_{\beta_0,\mu}[\phi] = 0$ for

all $\mu \in \mathbb{M}$. They then consider the class of strongly unbiased (SU) tests which satisfy the condition $E_{\beta_0,\mu}[\phi g] = 0$ for all $\mu \in \mathbb{M}$, and which they show are a subset of the LU tests. They find that weighted average power optimal tests within this class (based on the MM1 and MM2 weights) have good power in their simulations, and it is these tests which we consider here.

As noted in the main text, the class of CLC tests is a subset of the SU tests. To see this, note that (for fixed D) S and K are both invariant to switching the sign of g. Since $g \sim N(0, I_k)$ conditional on $D = d$ under $\beta = \beta_0$, we can see that for any conditional linear combination test $\phi_{a(D)}$,

$$E_{\beta_0,\mu}\left[\phi_{a(D)} g | D = d\right] = E_{\beta_0,\mu}\left[-\phi_{a(D)} g | D = d\right] = 0$$

and thus that all CLC tests satisfy $E_{\beta_0,\mu}\left[\phi_{a(D)} g\right] = 0$ and are SU tests. Since the MM1-SU and MM2-SU tests are weighted average power optimal in the class of SU tests, it follows that their weighted average power must be at least as high as that of any CLC test (for their respective weights).

The MM-SU tests are not available in closed form. However, as discussed in MM, approximating these tests numerically is fairly straightforward. We implement the MM-SU tests using the linear programing approach discussed by MM, using 5,000 draws for S ($J = 5,000$ in MM's notation). Evaluating the MM-SU tests also involves an integral over a function of β, which we likewise approximate via Monte-Carlo (based on 1,000 draws). One issue we encountered at this stage is that some of the matrices used in the construction of the MM weights are near-degenerate, leading to negative eigenvalues when evaluated numerically. The issue appears to be purely numerical, but despite extensive exploration, as well as consultation with Marcelo Moreira, we have not succeeded in fully eliminating this issue. It seems unlikely to have a substantial impact on the simulated performance of the MM procedures, but it is possible it could have some effect.

D.5: Results for Alternative MM Tuning Parameters

As noted in the text, Moreira and Moreira (2013) took the tuning parameters in the MM1-SU and MM2-SU tests both equal to one, rather than setting them to one-tenth of the sample size as do MM. Since we previously followed Moreira and Moreira (2013) in implementation of these tests, for comparability with previous versions here

	QCLR	AR	K	PI	MM1-SU	MM2-SU	QLR
Australia	0.60%	17.78%	1.58%	4.86%	8.12%	3.02%	3.98%
Canada	4.00%	20.98%	4.92%	10.06%	9.00%	6.66%	11.62%
France	0.66%	20.44%	0.76%	6.92%	3.12%	2.64%	4.76%
Germany	4.18%	20.90%	10.42%	6.94%	8.80%	8.84%	18.72%
Italy	5.16%	14.72%	9.08%	5.54%	11.92%	6.54%	4.14%
Japan	40.12%	16.56%	85.08%	6.52%	9.70%	14.84%	8.48%
Netherlands	2.22%	19.16%	8.24%	7.82%	2.76%	2.26%	3.20%
Sweden	1.96%	19.72%	2.52%	5.02%	7.36%	2.16%	2.24%
Switzerland	4.02%	21.36%	4.10%	7.86%	8.76%	7.20%	2.60%
United Kingdom	26.68%	18.86%	37.40%	11.18%	14.48%	5.66%	10.98%
United States	13.70%	17.22%	16.10%	8.74%	24.10%	13.54%	3.66%

Table 7: Maximal point-wise power shortfall relative to other tests considered, for simulations calibrated to match data in Yogo (2004). QCLR denotes the quasi-CLR test of Kleibergen (2005) while PI is the plug-in test discussed in Section 7.2.1 of the paper. AR is the Anderson Rubin (or S) test, K is Kleibergen (2005)'s K test, and MM1-SU and MM2-SU are the weighted average power optimal SU tests of Moreira and Moreira (2013). Note that here we set the tuning parameters in these tests as in Moreira and Moreira (2013), rather than Moreira and Moreira (2015).

we report results for the same calibrations to Yogo data considered in the paper where we now select the tuning parameters as in Moreira and Moreira (2013). Figures 8-10 plot power curves for the eleven simulation designs considered, while Table 7 reports the maximal power shortfall for each test relative to the other tests considered in each simulation design. As noted in the text, for this choice of tuning parameters the PI test has the smallest maximal power shortfall of all the tests considered.

D.6: Implementation of QLR Test

To evaluate the QLR test of I. Andrews and Mikusheva (2016*a*) we need to evaluate a conditional critical value function, which requires simulating the conditional distribution of a QLR statistic under the null. To accelerate the required calculations we follow Andrews and Mikusheva (2016*a*) and take a discrete approximation to the parameter space and evaluate both the QLR statistic and critical values via grid search. Critical values are then based on 1,000 simulation draws.

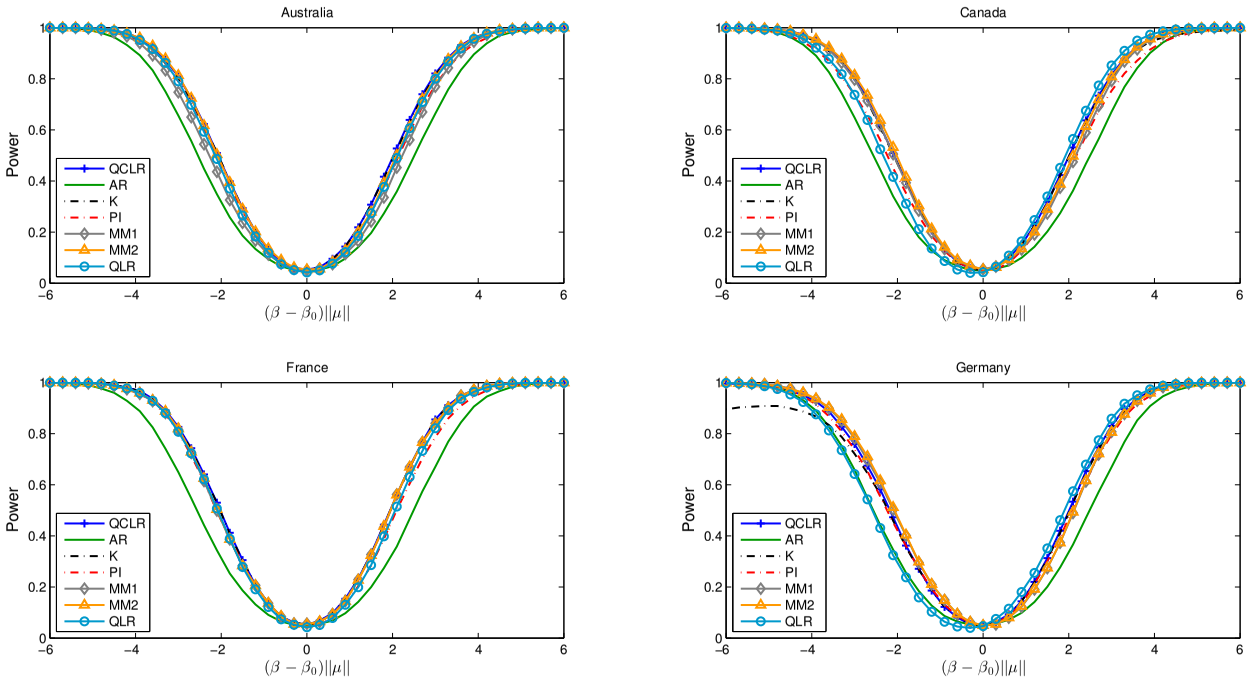

Figure 8: Power functions for QCLR, AR (or S), K, PI, MM1-SU, MM2-SU, and QLR tests in simulation calibrated to Yogo (2004) data with four instruments, selecting MM tuning parameters as in Moreira and Moreira (2013).

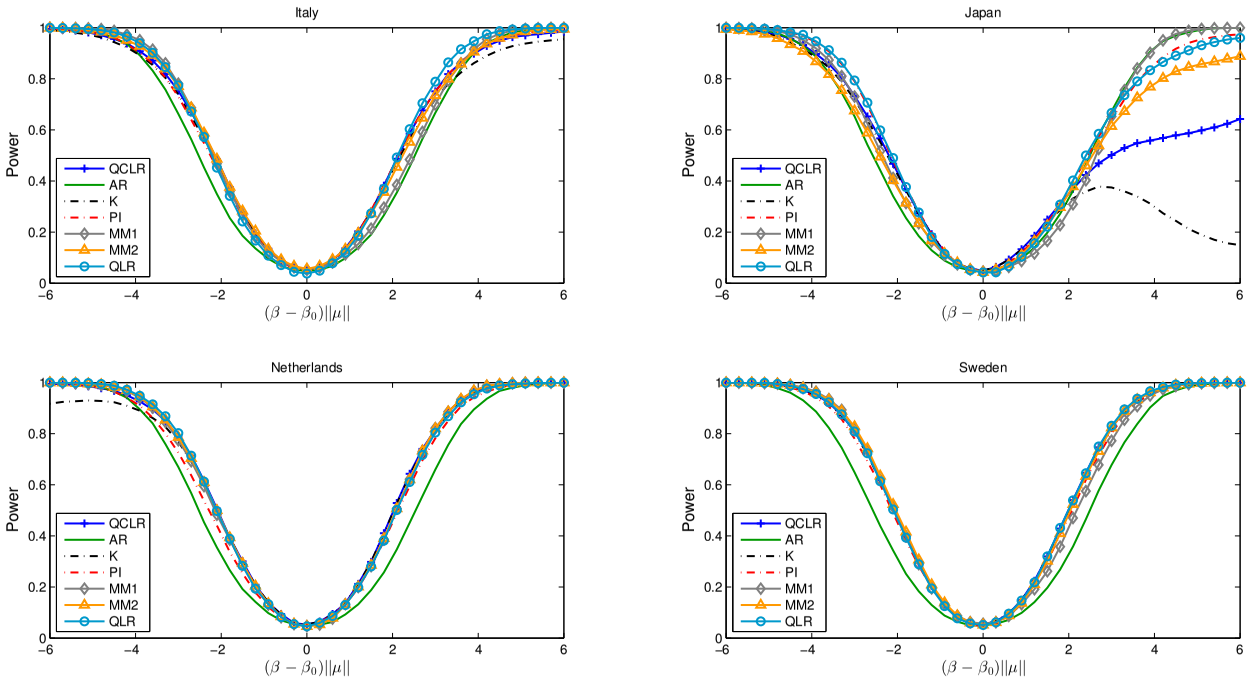

Figure 9: Power functions for QCLR, AR (or S), K, PI, MM1-SU, MM2-SU, and QLR tests in simulation calibrated to Yogo (2004) data with four instruments, selecting MM tuning parameters as in Moreira and Moreira (2013).

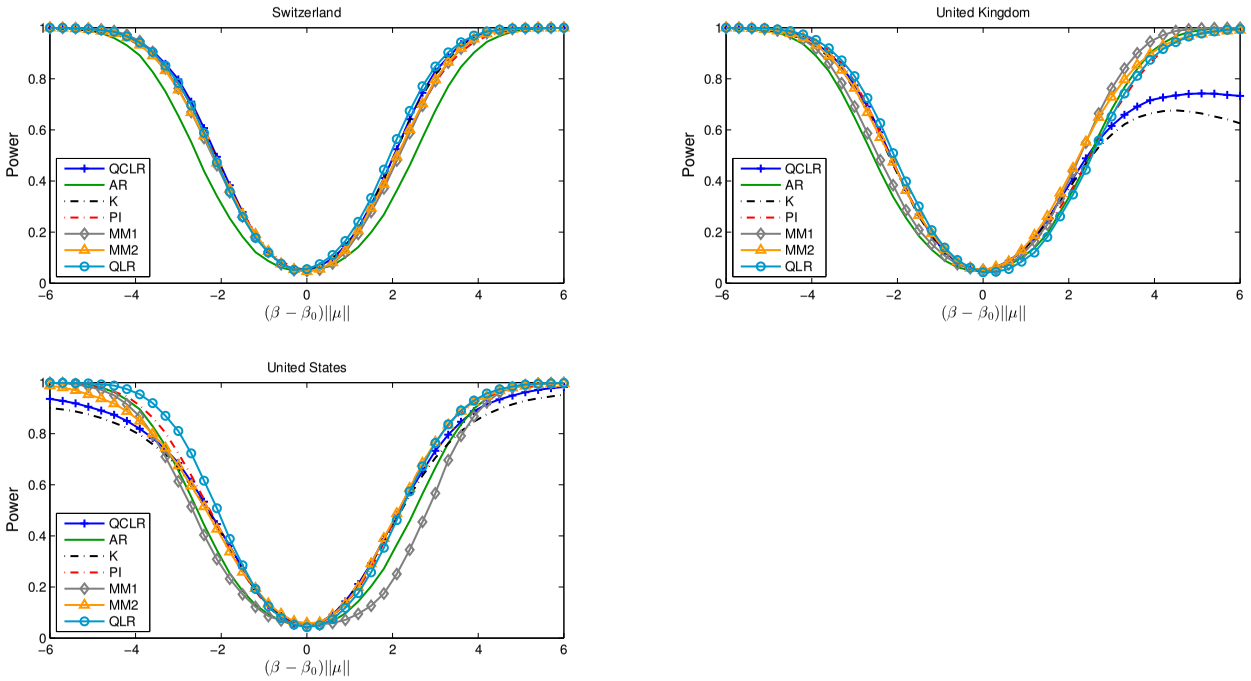

Figure 10: Power functions for QCLR, AR (or S), K, PI, MM1-SU, MM2-SU, and QLR tests in simulation calibrated to Yogo (2004) data with four instruments, selecting MM tuning parameters as in Moreira and Moreira (2013).

D.7: Implementation of PI Test

To implement the PI test, we need to calculate the MMRU test

$$\phi_{MMRU\left(\hat{\mu}_{D,b}\right)} = 1\left\{\left(1 - a_{MMRU\left(\hat{\mu}_{D,b}\right)}\right) K_r + a_{MMRU\left(\hat{\mu}_{D,b}\right)} S_r \geq c_\alpha\left(a_{MMRU\left(\hat{\mu}_{D,b}\right)}\right)\right\}$$

so the critical task is evaluating $a_{MMRU\left(\hat{\mu}_{D,b}\right)}$. As discussed in Section 6 above, $a_{MMRU\left(\hat{\mu}_{D,b}\right)}$ solves a minimization problem which depends on $\hat{\mu}_{D,b}$ and on $\hat{\Sigma}_i$.

As in Section A, we approximate a_{MMRU} by considering grids of values in a and β. We first simulate the critical values $c_\alpha(a)$ for linear combination tests based on $K + a \cdot J$ for $a \in A = \{0, 0.01, ..., 1\}$, which are simply the $1-\alpha$ quantiles of $\chi^2_p + a \cdot \chi^2_{k-p}$ distributions, and store these values for later use. To speed up power simulations, for each $a \in A$ and (τ_J, τ_K) values in a grid we calculate

$$Pr\left\{\chi^2_p(\tau_K) + a \cdot \chi^2_{k-p}(\tau_J) > c_\alpha(a)\right\}$$

based on 10^6 simulations and store the results as well.

We next consider a grid $B = \{-5, -4.9, ..., 5\}$ of 101 values for the alternative β_h. For each value β_h we solve for

$$\hat{\mu}_{b,h} = \left(I - \hat{\Sigma}_{\theta g,i}(\beta_h - \beta_0)\right)^{-1} \hat{\mu}_{D,b}$$

which gives us the value μ for which D would have mean $\hat{\mu}_{D,b}$ under alternative β_h. Note that the mean m of g under β_h is then $m_{b,h} = \hat{\mu}_{b,h}(\beta_h - \beta_0)$. We take draws $l = 1, ..., L = 10,000$ from

$$D_{b,l} \sim N\left(\hat{\mu}_{D,b}, \hat{\Sigma}_{D,i}\right)$$

and for each (h, l) pair we calculate $\tau_{K,b,h,l} = m'_{b,h} P_{D_{b,l}} m_{b,h}$ and $\tau_{J,b,h,l} = m'_{b,h} M_{D_{b,l}} m_{b,h}$.

We could estimate the power of the linear combination test with weight a against alternative β_h by

$$\hat{E}\left[\phi_a | \beta = \beta_h\right] = \frac{1}{L} \sum Pr\left\{\chi^2_p(\tau_{K,b,h,l}) + a \cdot \chi^2_{k-p}(\tau_{J,b,h,l}) > c_\alpha(a)\right\}.$$

Instead, to reduce the amount of required computation we note that for (b, h) fixed, $\tau_{K,b,h,l} + \tau_{J,b,h,l} = m'_{b,h} m_{b,h}$ and thus for fixed (b, h) the power of the linear combination test with weight a can be written as a function of $\tau_{K,b,h,l}$ alone. Using this observation,

we group the ten smallest values of $\tau_{K,b,h,l}$, the next ten smallest, etc. and assign each cell the (τ_K, τ_J) values given the by average of its endpoints. This gives us pairs $(\bar{\tau}_{K,q}, \bar{\tau}_{J,q})$ for $q \in \{1, ..., 1000\}$, and we estimate

$$\hat{E}\left[\phi_a | \beta = \beta_h\right] = \frac{1}{1000} \sum Pr\left\{\chi^2_p\left(\bar{\tau}_{K,q}\right) + a \cdot \chi^2_{k-p}\left(\bar{\tau}_{J,q}\right) > c_\alpha(a)\right\}$$

where by using $(\bar{\tau}_{K,q}, \bar{\tau}_{J,q})$ we need only calculate power 1000 times rather than 10000. To further speed computation, we approximate $Pr\left\{\chi^2_p\left(\bar{\tau}_{K,q}\right) + a \cdot \chi^2_{k-p}\left(\bar{\tau}_{J,q}\right) > c_\alpha(a)\right\}$ by interpolating using our stored values for $Pr\left\{\chi^2_p\left(\tau_K\right) + a \cdot \chi^2_{k-p}\left(\tau_J\right) > c_\alpha(a)\right\}$.

For each $a \in A$ we estimate the maximum regret by

$$\sup_{\beta_h \in B}\left(\max_{\tilde{a} \in A} \hat{E}\left[\phi_{\tilde{a}} | \beta = \beta_h\right] - \hat{E}\left[\phi_a | \beta = \beta_h\right]\right)$$

and pick $a_{MMRU}\left(\hat{\mu}_{D,b}\right)$ as the largest value $a \in A$ which comes within 10^{-5} of minimizing this quantity- we do this instead of taking $a_{MMRU}\left(\hat{\mu}_{D,b}\right)$ to be the true minimizing value in order to slightly reduce simulation noise in $a_{MMRU}\left(\hat{\mu}_{D,b}\right)$.

Supplementary Materials E: Additional Implementation Details

This section provides further details on our implementation of the PI and QLR tests and confidence sets in Section A.

E.1 Estimator $\hat{\mu}_D$

Note that for $V = \left(\sum_i vec\left([D]_i\right)\right)^{-1}$ and $[A]_{ij}$ the element in row i and column j of A,

$$E\left[D'VD\right]_{ij} = E\left[[D]_i' V [D]_j\right] = E\left[[\mu_D]_i' V [\mu_D]_j\right] + E\left[[D - \mu_D]_i' V [D - \mu_D]_j\right]$$

$$= [\mu_D' V \mu_D]_{ij} + E\left[tr\left([D - \mu_D]_i' V [D - \mu_D]_j\right)\right]$$

$$= [\mu_D' V \mu_D]_{ij} + tr\left(V \cdot E\left[[D - \mu_D]_j [D - \mu_D]_i'\right]\right)$$

$$= [\mu_D' V \mu_D]_{ij} + tr\left(V \cdot Cov\left([D]_j, [D]_i\right)\right).$$

Thus, an unbiased estimator of $[\mu_D' V \mu_D]_{ij}$ is $[D'VD]_{ij} - tr\left(V \cdot Cov\left([D]_j, [D]_i\right)\right)$. Let $\hat{R} = D'VD$ and define $\hat{R}_U$ to be the corresponding unbiased estimator for $R = \mu_D' V \mu_D$. Unlike R, $\hat{R}_U$ may have negative eigenvalues. To address this possibility, let $\hat{R}_P$ be matrix with the same eigenvectors as $\hat{R}_U$ which sets any negative eigenvalues to zero. Our estimator of μ_D is

$$\hat{\mu}_D = D\hat{R}^{-\frac{1}{2}} R_P^{\frac{1}{2}}$$

which can be seen to imply estimate $\hat{R}_P$ for R. Moreover, in the case with a single endogenous regressor, this choice of $\hat{\mu}_D$ can be seen to yield the positive-part estimator used in Section 7.2 and described in footnote 15. To obtain a feasible estimator $\hat{\mu}_D$, we simply plug in a consistent estimator for V, based on $\hat{\Sigma}_D = \hat{\Sigma}_{\theta\theta} - \hat{\Sigma}_{\theta g}\hat{\Sigma}_{g\theta}$.

E.2: Calculation of $a_{PI}(D) = a_{MMRU}(\hat{\mu}_D)$

After approximating the MMRU problem as described in Section D.7, to calculate the plug-in weight $a_{MMRU}(\hat{\mu}_D)$ we need to repeatedly evaluate $E_{m,\hat{\mu}_D}[\phi_a]$ for different values a. As in D.7 above, we reduce the computational burden at this step by recognizing that $E_{m,\hat{\mu}_D}[\phi_a] = \int E_{\tau_J(D),\tau_K(D)}[\phi_a]\, dF_D(\hat{\mu}_D)$ and tabulating $E_{\tau_J,\tau_K}[\phi_a]$ in advance. In particular, for each $a \in A$ and (τ_J, τ_K) values in a grid we calculate

$$Pr\left\{\chi_p^2(\tau_K) + a \cdot \chi_{k-p}^2(\tau_J) > c_\alpha(a)\right\}$$

based on a million simulations and store the results. Note that the resulting values depend only on k, p, and A. Thus, in computing e.g. confidence sets, we only have to do this tabulation once per specification.

To evaluate $E_{m,\hat{\mu}_D}[\phi_a]$, for each $(\hat{\mu}_D, m)$ pair, we take draws $l = 1, ..., L = 10,000$ from

$$vec(D_l) \sim N\left(vec(\hat{\mu}_D), \hat{\Sigma}_D\right)$$

for $\hat{\Sigma}_D = \hat{\Sigma}_{\theta\theta} - \hat{\Sigma}_{\theta g}\hat{\Sigma}_{g\theta}$. For each l we calculate $\tau_{K,l} = m' P_{D_l} m$ and $\tau_{K,l} = m' M_{D_l} m$. To approximate $\int E_{\tau_J(D),\tau_K(D)}[\phi_a]\, dF_D(\hat{\mu}_D)$, we group the twenty smallest values of $\tau_{K,l}$, the next twenty smallest, etc. and assign each cell the (τ_K, τ_J) values given the by average of its endpoints. This gives us pairs $(\bar{\tau}_{K,q}, \bar{\tau}_{J,q})$ for $q \in \{1, ..., 500\}$, and we estimate

$$\int E_{\tau_J(D),\tau_K(D)}[\phi_a]\, dF_D(\hat{\mu}_D) = \frac{1}{500}\sum Pr\left\{\chi_p^2(\bar{\tau}_{K,q}) + a \cdot \chi_p^2(\bar{\tau}_{J,q}) > c_\alpha(a)\right\}$$

where by using $(\bar{\tau}_{K,q}, \bar{\tau}_{J,q})$ we need only calculate power 500 times rather than 10,000. To compute $Pr\left\{\chi^2_p(\bar{\tau}_{K,q}) + a \cdot \chi^2_{k-p}(\bar{\tau}_{J,q}) > c_\alpha(a)\right\}$ we then simply interpolate using our stored values.

E.3: Evaluation of QLR Critical Values

To evaluate the conditional QLR test of I. Andrews and Mikusheva (2016*a*) we need to repeatedly simulate from the conditional distribution of the QLR statistic (the difference between the continuously updating GMM objective evaluated at the the null and at the continuously updating GMM estimator) given D under the null. This requires repeated numerical optimization, which becomes more challenging as we increase the dimension of the parameter and is very demanding for specifications D and E for the Angrist and Krueger (1991) data. For each parameter β_i we construct a grid of five values evenly spaced between $\beta_{i,L}$ and $\beta_{i,U}$, and then construct a grid of values of β by taking the Cartesian product of these grids for the individual parameters. Note that the resulting grid is the same as the grid of values B_J used to evaluate the PI test for $J = 5$.

In each simulation run we evaluate the continuously updating GMM objective function at each point in B_J. We then take the point with the lowest value of the objective function as the starting value for numerical (simplex) minimization. Critical values are based on 500 simulation draws.

Supplementary Materials F: Inference on the New Keynesian Phillips Curve

To illustrate the application of PI tests to a nonlinear example, we study the performance of robust minimum distance inference on new Keynesian Phillips curve (NKPC) parameters. There is considerable evidence that some NKPC parameters are weakly identified: Mavroeidis et al. (2014) review the empirical literature on the role of expectations in the NKPC and find that parameter estimates are extremely sensitive to model specification and, conditional on correct specification, suffer from weak identification. To address these weak identification issues Magnusson and Mavroeidis (2010) (henceforth MM) propose identification-robust S and K statistics for testing hypotheses on NKPC parameters using a minimum distance approach. These statistics will

form the basis for our analysis.

MM study a simple new Keynesian Phillips curve model

$$\pi_t = \frac{(1-\nu)^2}{\nu(1+\rho)} x_t + \frac{1}{1+\rho} E[\pi_{t+1}|\mathcal{I}_t] + \frac{\rho}{1+\rho}\pi_{t-1} + \varepsilon_t \tag{29}$$

where π_t is inflation, x_t is a measure of marginal costs, $E[\cdot|\mathcal{I}_t]$ denotes an expectation conditional on information available at time t, ε_t is an exogenous shock with $E[\varepsilon_{t+1}|\mathcal{I}_t] = 0$, and the parameters ν and ρ denote the degree of price stickiness and price indexation, respectively. Following Sbordone (2005), MM further assume that (π_t, x_t) follows a nth order vector auto-regressive (VAR) process, which can be written in companion form as

$$z_t = A(\varphi) z_{t-1} + \epsilon_t$$

where $z_t = (\pi_t, x_t, ..., \pi_{t-n+1}, x_{t-n+1})'$ is a $2n \times 1$ vector, $A(\varphi)$ is a $2n \times 2n$ matrix, φ is the vector of $4n$ unknown VAR parameters, and ϵ_t are VAR innovations with $E[\epsilon_{t+1}|\mathcal{I}_t] = 0$. For e_π and e_x unit vectors such that $e_\pi' z_t = \pi_t$, $e_x' z_t = x_t$ and $\theta = (\nu, \rho)$, define the $2n$-dimensional distance function $f(\varphi, \theta)$ as

$$f(\varphi,\theta) = A(\varphi)' \left\{ \left[I - \frac{1}{1+\rho} A(\varphi)' \right] e_\pi - \frac{(1-\nu)^2}{\nu(1+\rho)} e_x \right\} - \frac{\rho}{1+\rho} e_\pi.$$

MM show that the NKPC model (29) implies that the true parameter values φ and θ satisfy $f(\varphi, \theta) = 0$, and propose testing $H_0 : \theta = \theta_0$ using an identification-robust minimum distance approach.

To model weak identification in this context, suppose the data is generated by a sequence of models with drifting true VAR coefficients $\varphi_T = \varphi + \frac{1}{\sqrt{T}} c_\varphi + o\left(\frac{1}{\sqrt{T}}\right)$. We assume that the usual OLS estimates for the VAR coefficients are consistent and asymptotically normal

$$\sqrt{T}(\hat{\varphi} - \varphi_T) \to_d N(0, \Sigma_{\varphi\varphi})$$

where we have a consistent estimator $\hat{\Sigma}_{\varphi\varphi}$ for $\Sigma_{\varphi\varphi}$. The Δ-method (Theorem 3.1 in Van der Vaart (2000)) then yields that

$$\sqrt{T}(f(\hat{\varphi},\theta) - f(\varphi_T,\theta)) \to_d N\left(0, \frac{\partial}{\partial\varphi'} f(\varphi,\theta) \Sigma_{\varphi\varphi} \frac{\partial}{\partial\varphi'} f(\varphi,\theta)'\right).$$

To model weak identification in this context MM assume that $\frac{\partial}{\partial\theta'} f(\varphi_T, \theta) = \frac{1}{\sqrt{T}} C$

for a fixed matrix C, with the result that $\sqrt{T}\frac{\partial}{\partial\theta'}f(\varphi_T,\theta)$ is constant across T. This leads to the usual issues associated with weak identification, including nonstandard limiting distributions for non-robust test statistics. Here, we will take a more flexible approach and assume only that $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta)$ drifts towards some, potentially reduced-rank, matrix as the sample size grows.

To apply our robust testing approach in this context, define

$$\hat{\Omega}_{ff}=\frac{\partial}{\partial\varphi'}f(\hat{\varphi},\theta_0)\hat{\Sigma}_{\varphi\varphi}\frac{\partial}{\partial\varphi'}f(\hat{\varphi},\theta_0)'$$

which is a consistent, Δ-method-based estimator for $\Omega_{ff}=\lim_{T\to\infty}T\cdot Var\left(f(\hat{\varphi},\theta_0)'\right)$. We can then define

$$g_T(\theta)=\sqrt{T}\hat{\Omega}_{ff}^{-\frac{1}{2}}f(\hat{\varphi},\theta)$$

$$\Delta g_T(\theta)=\hat{\Omega}_{ff}^{-\frac{1}{2}}\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta)A_T.$$

for A_T a sequence of full-rank normalizing matrices which may depend on the sequence of true VAR parameters φ_T. Under sequences of true parameter values θ_T such that $g_T(\theta_0)$ converges in distribution, corresponding to local alternatives for strongly identified parameters and fixed alternatives for weakly identified ones, arguments discussed in Section F.3 below yield the weak convergence

$$\begin{pmatrix} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{pmatrix}\to_d\begin{pmatrix} g \\ \Delta g \end{pmatrix}\sim N\left(\begin{pmatrix} m \\ \mu \end{pmatrix},\begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix}\right). \tag{30}$$

where Δg is full rank almost surely, $m\in\mathcal{M}(\mu,\gamma)$ for $\mathcal{M}(\mu,\gamma)$ appropriately defined, Σ is consistently estimable, and details on all terms may be found below. Hence, this model falls into the class considered in the paper. While Δg_T depends on the (generally unknown) sequence of normalizing matrices A_T, provided we restrict attention to postmultiplication-invariant CLC tests we can instead conduct tests based on the feasible statistics $(\tilde{g}_T,\Delta\tilde{g}_T,\tilde{\gamma})=h\left(g_T,\Delta g_T,\hat{\gamma};A_T^{-1}\right)$. For $\hat{\gamma}$ as defined below the statistics S_T and K_T based on $(\tilde{g}_T,\Delta\tilde{g}_T,\tilde{\gamma})$ are equivalent to the $MD-AR$ and $MD-K$ statistics discussed in MM.

F.1 Coverage Simulations

After assuming that (π_t, x_t) follows a VAR(3), MM apply their approach to create confidence sets for the parameter θ based on quarterly US data from 1984 to 2008 and show that their robust minimum distance approach yields smaller confidence sets than an identification-robust GMM approach. MM suggested using S and JK tests $\phi_{S_T} = 1\left\{S_T > \chi^2_{6,1-\alpha}\right\}$ and

$$\phi_{JK_T} = \max\left\{1\left\{K_T > \chi^2_{2,1-\alpha_K}\right\}, 1\left\{J_T > \chi^2_{4,1-\alpha_J}\right\}\right\},$$

where $\alpha_K = 0.8 \cdot \alpha$ and $\alpha_J = 0.2 \cdot \alpha$. They use the JK test rather than the K test ϕ_{K_T} to address spurious power declines for the K test. We take these tests, together with the K test ϕ_{K_T}, as the benchmarks against which we compare the performance of the PI test. In particular, we consider the plug-in test

$$\phi_{PI_T} = 1\left\{(1 - a_{PI}\left(D_T, \tilde{\gamma}\right)) \cdot K_T + a_{PI}\left(D_T, \tilde{\gamma}\right) \cdot S_T > c_\alpha(a_{PI}(D_T, \tilde{\gamma}))\right\}$$

where as before

$$a_{PI}(D, \gamma) = \arg\min_{a \in [0,1]} \sup_{m \in \mathcal{M}_D(\hat{\mu}_D, \gamma)} (\beta^u_m - E_{m, \hat{\mu}_D, \gamma}\left[\phi_a\right])$$

and for simplicity we take $\hat{\mu}_D = D$.

To compare the performance of the PI test to the tests discussed by MM, we calibrate a simulation example based on the empirical application of MM. In particular, we estimate structural and reduced-form parameters using the data studied by MM and generate samples of 100 observations based on these estimates together with the assumption of Gaussian errors ϵ_t (see below for details).[29] We calculate the true size of nominal 5% tests, based on 10,000 simulations, and report the results in Table 8. We find that all the tests over-reject, which is unsurprising given the non-linearity of the model together with the small sample size, but that only the JK and S tests has true size exceeding 10%.

Next, we simulate false coverage probabilities for confidence sets formed by inverting these tests. In particular we calculate the rejection rates for PI, JK, S, and K

[29] We simulate samples of size 100 because MM use a dataset with 99 observations in their empirical application.

	PI	JK	S	K
Size	9.34%	10.52%	12.28%	8.74%

Table 8: Size of nominal 5% tests in NKPC simulation example based on 10,000 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

	PI	JK	S	K
PI	*	3.0%	4.2%	1.0%
JK	6.4%	*	2.0%	6.0%
S	31.2%	26.6%	*	30.0%
K	17.6%	17.6%	17.4%	*

Table 9: Maximal point-wise differences in false coverage probability of nominal 5% tests in NKPC example. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test falls short relative to the JK test is 3%. Based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

tests of hypotheses $H_0 : \theta = \theta_0$ for θ_0 not equal to the true parameter value.[30] Table 9 reports the maximal difference in point-wise false coverage probability across tests, based on 500 simulations. For each test we report the largest margin by which the rejection probability of that test falls short relative to that of the other tests considered over $\theta_0 \in (0,1)^2$, which is the parameter space for the model.[31] For example, the second entry of the first row of Table 9 reports

$$\sup_{\theta_0 \in (0,1)^2} E_{\tilde{\theta}} \left[\phi_{JK_T,\theta_0} - \phi_{PI_T,\theta_0} \right]$$

where ϕ_{PI_T,θ_0} and ϕ_{JK_T,θ_0} denote the PI and JK tests of $H_0 : \theta = \theta_0$, respectively, and $\tilde{\theta} = (\tilde{\nu}, \tilde{\rho}) = (0.96, 0.48)$ is the true parameter value in the simulations. As these

[30] We focus on calculating false coverage probabilities rather than power because there are many reduced-form parameter values φ compatible with a given structural parameter value θ^*, and the power of tests of $H_0 : \theta = \theta_0$ against θ^* will generally depend on φ. Hence, to simulate the power function we must either adopt some rule to pick φ based on θ^* or calculate power on a 12-dimensional space, whereas to calculate false coverage probabilities it suffices to consider a 2-dimensional space of values θ.

[31] For computational reasons, our simulations use a discretized version of this parameter space- see below.

	PI	JK	S	K
PI	*	1.6%	2.0%	8.2%
JK	11.4%	*	1.4%	16.4%
S	42.0%	33.2%	*	46.0%
K	20.4%	21.6%	21.8%	*

Table 10: Maximal point-wise differences in false coverage probability of **size corrected** 5% tests in NKPC example. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test falls short relative to the JK test is 1.6%. Based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

	PI	JK	S	K
Expected Area: feasible confidence sets	0.084	0.0873	0.110	0.094
Expected Area: corrected confidence sets	0.131	0.141	0.169	0.138

Table 11: Expected area of 95% confidence sets formed by inverting tests in NKPC example, based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

results make clear, the PI test outperforms the other tests studied and has the smallest maximal rejection rate shortfall. The JK test also performs reasonably well, with a much smaller maximal rejection rate shortfall than the S and K tests. Interpreting these results is complicated by the fact that, while all the tests considered have correct asymptotic size under weak identification, their finite sample size differs substantially. To account for such size differences, Table 10 reports results analogous to those of Table 9 based on (infeasible) size-corrected versions of all four tests. As in Table 9, we can see that the PI test offers the best performance, followed by the JK test.[32]

After simulating false coverage probabilities, it is easy to calculate the expected

[32]To size-correct the S and K tests, we simply take their critical values to be the 95th percentiles of their respective distributions for testing $H_0 : \theta = \tilde{\theta}$. To size-correct the PI test we consider

$$\phi^*_{PI_T} = 1\left\{(1 - a_{PI}\left(D_T, \tilde{\gamma}\right)) \cdot K_T + a_{PI}\left(D_T, \tilde{\gamma}\right) \cdot S_T - c_\alpha(a_{PI}(D_T, \tilde{\gamma})) > c^*\right\}$$

where c^* is chosen to give correct size when testing $H_0 : \theta = \tilde{\theta}$. Likewise, the size-corrected JK test is

$$\phi^*_{JK_T} = 1\left\{\max\left\{K_T - \chi^2_{2,1-\alpha_K}, J_T - \chi^2_{4,1-\alpha_J}\right\} > c^*\right\}$$

for c^* chosen to ensure correct size for testing $H_0 : \theta = \tilde{\theta}$. Note that if we instead take $c^* = 0$, these coincide with the non-size-corrected PI and JK tests.

area of confidence sets obtained by inverting the PI, JK, S, and K tests. The expected area for confidence sets formed by inverting both the feasible and size-corrected tests is reported in Table 11. As we can see, using size-corrected tests increases the area of all confidence sets. In each case the PI test produces confidence sets with the smallest expected area, while the S test yields confidence sets with the largest expected area. The feasible JK test yields smaller confidence sets than the feasible K test, but size correction reveals that this is due in part to finite-sample size distortions for the JK test: when we invert size-corrected tests, we find that JK confidence sets have larger expected area than K confidence sets. A further advantage of the PI-test-based confidence sets is that, like K-test-based confidence sets, they are non-empty in all 500 simulations, whereas confidence sets formed by inverting the JK and S tests are empty in 3.2% and 4.8% of simulations, respectively.[33] These results confirm that the PI test outperforms the other tests considered.

F.2 Details of NKPC Example

Define the infeasible estimator $\hat{\Omega}$ by

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{ff} & \hat{\Omega}_{f\theta} \\ \hat{\Omega}_{\theta f} & \hat{\Omega}_{\theta\theta} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{\partial}{\partial\varphi'} f(\hat{\varphi},\theta_0)\hat{\Sigma}_{\varphi\varphi}\frac{\partial}{\partial\varphi'} f(\hat{\varphi},\theta_0)' & \cdots \\ \frac{\partial}{\partial\varphi'} vec\left(\frac{\partial}{\partial\theta'} f(\hat{\varphi},\theta_0)A_T/\sqrt{T}\right)\hat{\Sigma}_{\varphi\varphi}\frac{\partial}{\partial\varphi'} f(\hat{\varphi},\theta)' & \\ \frac{\partial}{\partial\varphi'} f(\hat{\varphi},\theta_0)\hat{\Sigma}_{\varphi\varphi}\frac{\partial}{\partial\varphi'} vec\left(\frac{\partial}{\partial\theta'} f(\hat{\varphi},\theta_0)A_T/\sqrt{T}\right)' & \\ \frac{\partial}{\partial\varphi'} vec\left(\frac{\partial}{\partial\theta'} f(\hat{\varphi},\theta_0)A_T/\sqrt{T}\right)\hat{\Sigma}_{\varphi\varphi}\frac{\partial}{\partial\varphi'} vec\left(\frac{\partial}{\partial\theta'} f(\hat{\varphi},\theta_0)A_T/\sqrt{T}\right)' & \end{pmatrix}$$

and note that given our assumptions this will be consistent for

$$\Omega = \lim_{T\to\infty} Var\left(\sqrt{T} f(\hat{\varphi},\theta_0)', vec\left(\frac{\partial}{\partial\theta'} f(\hat{\varphi},\theta_0)A_T\right)'\right).$$

To derive the weak convergence (30), as well as the form of the matrices A_T, note that since we have assumed $\varphi_T \to \varphi$ and $\sqrt{T}(\hat{\varphi} - \varphi_T) \to_d N(0, \Sigma_{\varphi\varphi})$, the Δ-method

[33] Note that there is no guarantee that confidence sets formed by inverting the PI test will be non-empty in general.

(Theorem 3.1 in Van der Vaart (2000)) yields that

$$\sqrt{T}\left(f_T(\hat{\varphi},\theta_0)-f_T(\varphi_T,\theta_0)\right)\to_d N\left(0,\frac{\partial}{\partial\varphi'}f(\varphi,\theta_0)\Sigma_{\varphi\varphi}\frac{\partial}{\partial\varphi'}f(\varphi,\theta_0)'\right)$$

$$\sqrt{T}\cdot vec\left(\tfrac{\partial}{\partial\theta'}f(\hat{\varphi},\theta_0)-\tfrac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right)\to_d$$
$$\tfrac{\partial}{\partial\varphi'}\left(vec\left(\tfrac{\partial}{\partial\theta'}f(\hat{\varphi},\theta_0)\right)\right)\Sigma_{\varphi\varphi}\tfrac{\partial}{\partial\varphi'}\left(vec\left(\tfrac{\partial}{\partial\theta'}f(\hat{\varphi},\theta_0)\right)\right)'.$$

We can see that the assumed convergence of $g_T(\theta_0)=\sqrt{T}\hat{\Omega}_{gg}^{-\frac{1}{2}}f(\hat{\varphi},\theta_0)$ thus holds only if $\sqrt{T}f(\varphi_T,\theta_0)$ converges. To obtain convergence in distribution for $\Delta g_T(\theta_0)$, we will need to choose an appropriate sequence of normalizing matrices A_T, which may in turn depend on the sequence of true VAR parameters φ_T. To examine this issue in more detail, in the next subsection we briefly discuss two ways in which identification could fail in this model, one resulting in weak identification for ν and the other in weak identification for ρ.

F.2.1 Possible Sources of Weak Identification

Since we have assumed that (π_t, z_t) follow a VAR(3), we have that φ is 12-dimensional and can take

$$A(\varphi)=\begin{bmatrix}\varphi_{11}&\varphi_{12}&\varphi_{13}&\varphi_{14}&\varphi_{15}&\varphi_{16}\\ \varphi_{21}&\varphi_{22}&\varphi_{23}&\varphi_{24}&\varphi_{25}&\varphi_{26}\\ 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\end{bmatrix}.$$

Note that $e_\pi=(1,0,0,0,0,0)'$ and $e_x=(0,1,0,0,0,0)'$.

Fix a true parameter value θ. Identification of ν fails if $\varphi_{2i}=0$ for all $i\in\{1,...,6\}$. In this case we have that $A(\varphi)'e_x=0$, with the consequence that ν does not enter the distance function $f(\varphi,\theta)$ and $\frac{\partial}{\partial\nu}f(\varphi,\theta)=0$. To model ν as weakly identified, fix $\varphi_{1i,T}=\varphi_{1i}$ for $i\in\{1,...,6\}$ at values such that $f(\varphi_T,\theta)=0$ when $\varphi_{2i}=0$ for $i\in\{1,...,6\}$. We can take sequences of true VAR parameter values φ_T such that $\varphi_{1i,T}=\frac{1}{\sqrt{T}}c_{1,i}+o\left(\frac{1}{\sqrt{T}}\right)$,$\varphi_{2i,T}=\frac{1}{\sqrt{T}}c_{2,i}+o\left(\frac{1}{\sqrt{T}}\right)$ and $f(\varphi_T,\theta)=0\ \forall T$, which will imply that $\sqrt{T}\frac{\partial}{\partial\nu}f(\varphi_T,\theta)\to C_\nu$ for a 6×1 vector C_ν. Thus, if we take $A_T=\begin{bmatrix}\sqrt{T}&0\\0&1\end{bmatrix}$ we will have that the first column of $\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta)A_T$ converges in distribution to a non-

degenerate random vector. Provided the values $\varphi_{1i,T}$ are such that $\frac{\partial}{\partial\rho}f\left(\varphi_T,\theta\right)\to C_\rho$ for a non-zero vector C_ρ, then $\hat{\Omega}_{gg}^{-\frac{1}{2}}\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta)A_T\to_d\Delta g$ for a matrix Δg which is full rank almost surely, as we assumed.

The parameter ρ may also be weakly identified. In particular, note that

$$\frac{\partial}{\partial\rho}f(\varphi,\theta)=\ A\left(\varphi\right)'\left\{\left[\tfrac{1}{(1+\rho)^2}A\left(\varphi\right)'\right]e_\pi+\tfrac{(1-\nu)^2}{\nu(1+\rho)^2}e_x\right\}-\tfrac{1}{(1+\rho)^2}e_\pi$$

so if

$$\left(I-A\left(\varphi\right)'A\left(\varphi\right)'\right)e_\pi=A\left(\varphi\right)'e_x\frac{\left(1-\nu\right)^2}{\nu}$$

then $\frac{\partial}{\partial\rho}f(\varphi,\theta)=0$ for all values of ρ, so ρ is unidentified. In the same manner as above, for any pair (φ,ν) satisfying this restriction we can take ν fixed and construct a sequence φ_T converging to φ at a $\sqrt{T}$ rate such that $\Omega_{gg}^{-\frac{1}{2}}\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta)A_T\to_d\Delta g$ for Δg full rank almost-surely with $A_T=\begin{bmatrix}1&0\\0&\sqrt{T}\end{bmatrix}$.

F.2.2 Derivation of the Limit Problem

To derive the form of the limit problem (30) we need to understand the behavior of g_T and Δg_T under alternatives. Note that for alternative θ_T and true reduced-form parameter value φ_T, we have that since $f(\varphi_T,\theta_T)=0$,

$$f(\varphi_T,\theta_0)=f(\varphi_T,\theta_0)-f(\varphi_T,\theta_T).$$

Define

$$m(\varphi_T,\theta_T)=f(\varphi_T,\theta_0)-f(\varphi_T,\theta_T)=$$
$$A\left(\varphi_T\right)'\left\{\left(\tfrac{1}{1+\rho_T}-\tfrac{1}{1+\rho_0}\right)A\left(\varphi_T\right)'e_\pi+\left(\tfrac{(1-\nu_T)^2}{\nu_T(1+\rho_T)}-\tfrac{(1-\nu_0)^2}{\nu_0(1+\rho_0)}\right)e_x\right\}+\left(\tfrac{\rho_T}{1+\rho_T}-\tfrac{\rho_0}{1+\rho_0}\right)e_\pi,$$

and note that the assumed convergence for g_T implies that $\sqrt{T}m(\varphi_T,\theta_T)$ converges to m. To determine the form of the set $\mathcal{M}\left(\mu\right)$, which characterizes the behavior of m under various alternatives, note that

$$\frac{\partial}{\partial\theta}f(\varphi_T,\theta_0)=\left[\ A\left(\varphi_T\right)'\left(\tfrac{1-\nu_0^2}{\nu_0^2(1+\rho_0)}\right)e_x\quad A\left(\varphi_T\right)'\left\{\left[\tfrac{1}{(1+\rho_0)^2}A\left(\varphi_T\right)'\right]e_\pi+\tfrac{(1-\nu_0)^2}{\nu_0(1+\rho_0)^2}e_x\right\}-\tfrac{1}{(1+\rho_0)^2}e_\pi\ \right]$$

and hence

$$A\left(\varphi_T\right) e_x = \left(\frac{1-\nu_0^2}{\nu_0^2\left(1+\rho_0\right)}\right)^{-1}\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1 =$$

$$h_1(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1$$

for $h_1\left(\theta_0\right) = \left(\frac{1-\nu_0^2}{\nu_0^2(1+\rho_0)}\right)^{-1}$, and

$$A\left(\varphi_T\right)' A\left(\varphi_T\right)' e_\pi = \left(1+\rho_0\right)^2\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_2 - \frac{\left(1-\nu_0\right)^2}{\nu_0} A\left(\varphi_T\right)' e_x + e_\pi =$$

$$h_2(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_2 - h_3(\theta_0)h_1(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1 + e_\pi,$$

for $h_2\left(\theta_0\right) = \left(1+\rho_0\right)^2$ and $h_3\left(\theta_0\right) = \frac{\left(1-\nu_0\right)^2}{\nu_0}$. For $m\left(\varphi_T,\theta_T\right)$ as defined above, this implies that

$$\begin{gathered} m\left(\varphi_T,\theta_T\right) = \\ \left(\frac{1}{1+\rho_T}-\frac{1}{1+\rho_0}\right)\left(h_2(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_2 - h_3(\theta_0)h_1(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1 + e_\pi\right) + \\ \left(\frac{\left(1-\nu_T\right)^2}{\nu_T(1+\rho_T)}-\frac{\left(1-\nu_0\right)^2}{\nu_0(1+\rho_0)}\right) h_1(\theta_0)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1 + \left(\frac{\rho_T}{1+\rho_T}-\frac{\rho_0}{1+\rho_0}\right) e_\pi = \\ h_4(\theta_0,\theta_T)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_1 + h_5(\theta_0,\theta_T)\left[\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)\right]_2 \end{gathered}$$

for

$$h_4\left(\theta_0,\theta_T\right) = -\left(\frac{1}{1+\rho_T}-\frac{1}{1+\rho_0}\right) h_3(\theta_0)h_1(\theta_0) + \left(\frac{\left(1-\nu_T\right)^2}{\nu_T\left(1+\rho_T\right)}-\frac{\left(1-\nu_0\right)^2}{\nu_0\left(1+\rho_0\right)}\right) h_1(\theta_0)$$

and

$$h_5\left(\theta_0,\theta_T\right) = \left(\frac{1}{1+\rho_T}-\frac{1}{1+\rho_0}\right) h_2(\theta_0).$$

Thus, knowledge of $\frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0)$ suffices to let us calculate $m(\varphi_T,\theta)$ for any alternative θ in the sample of size T. Consequently, in each sample size T, an estimate of $\mu_T = \frac{\partial}{\partial\theta'} f(\varphi_T,\theta_0) A_T$ implies a corresponding set $\mathcal{M}_T(\mu_T) = \left\{\sqrt{T} m(\varphi_T,\theta) : \theta\in\Theta\right\}$. For a given convergent sequence φ_T, we can then define $\mathcal{M}$ in the limit problem as $\mathcal{M}(\mu) = \lim_T\left(\mathcal{M}_T(\mu)\cap C\right)$ for any compact set C: the restriction to the set C ensures convergence, and has the effect of restricting attention to a particular neighborhood of fixed alternatives for weakly identified parameters and local alternatives for strongly identified parameters. Note that in any given sample size we need not know A_T to

calculate $\mathcal{M}_T(\mu_T)$ once given an estimate of $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)$, so if we just treat $\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta_0)$ as a Gaussian random matrix with mean $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)$ and proceed accordingly, this will (asymptotically) correspond to using the correct $\mathcal{M}(\mu)$ under all sequences yielding limit problems in this class. Indeed, this is the approach we adopt to calculate plug-in tests in our simulations.

F.2.3 NKPC Simulation Details

The assumption that (π_t, x_t) follows a 3rd order VAR means that, once we make a distributional assumption on the driving shocks ϵ_t, we can simulate data from the NKPC model discussed above for any combination of parameters (φ,θ) such that $f(\varphi,\theta)=0$. For $\hat{\varphi}$ the VAR coefficients estimated from the data used by MM with estimated variance matrix $\hat{\Sigma}_{\varphi\varphi}$, we find the coefficients $\left(\tilde{\varphi},\tilde{\theta}\right)$ solving

$$\left(\tilde{\varphi},\tilde{\theta}\right)=\arg\min_{(\varphi,\theta):f(\varphi,\theta)=0}\left(\hat{\varphi}-\varphi\right)'\hat{\Sigma}_{\varphi\varphi}^{-1}\left(\hat{\varphi}-\varphi\right).$$

This yields the pair of reduced form and structural coefficients consistent with the NKPC model which, in a covariance-weighted sense, are as close as possible to the estimated VAR coefficient $\hat{\varphi}$. Using the residuals $\hat{\epsilon}_t$ from calculating the VAR coefficients $\hat{\varphi}$, we estimate the covariance matrix of the driving shocks by

$$\hat{V}_\epsilon=\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\epsilon}_t-\frac{1}{T}\sum_{s=1}^{T}\hat{\epsilon}_s\right)\left(\hat{\epsilon}_t-\frac{1}{T}\sum_{s=1}^{T}\hat{\epsilon}_s\right)'.$$

Taking ϵ_t to be normally distributed, to conduct our simulations we then generate samples of 100 observations from the the model with true parameter values $\left(\tilde{\varphi},\tilde{\theta}\right)$ and true covariance matrix $\hat{V}_\epsilon$ for ϵ_t.

For computational purposes, when calculating PI tests and simulating coverage probabilities we discretize the parameter space, considering grids of values in both ν and ρ. For both parameters we consider grids ranging from 0.005 to 0.995, with grid points spaced 0.03 apart.

Additional References

J.D. Angrist and A.B. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106:979–1014, 1991.

V. Chernozhukov, C. Hansen, and M. Jansson. Finite sample inference for quantile regression models. *Journal of Econometrics*, 152:93–103, 2009.

Patrik Guggenberger, Frank Kleibergen, Sophocles Mavroeidis, and Linchun Chen. On the asymptotic size of subset anderson-rubin and largrange multiplier tests in linear instrumental variables regression. *Econometrica*, 80:2649–2666, 2012.

James J. Heckman, Lance J. Lochner, and Petra E. Todd. *Handbook of the Economics of Education*, chapter Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond, pages 307–458. Elsevier, 2006.

James J. Heckman, Lance J. Lochner, and Petra E. Todd. Earnings functions and rates of return. *Journal of Human Capital*, 2:1–31, 2008.

C.F. Van Loan and N. Pitsianis. *Linear Algebra for Large-Scale and Real-Time Applications*, chapter Approximation with Kronecker Products, pages 293–314. Kluwer Academic Publishers, 1993.

S. Mavroeidis, M. Plagborg-Moller, and J. Stock. Empirical evidence on inflation expectations in the new keynesian phillips curve. Forthcoming in the Journal of Economic Literature, 2014.

W.K Newey and D.L. McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, chapter 36. Elsever, 1994.

A.M. Sbordone. Do expected future marginal costs drive inflation dynamics? *Journal of Monetary Economics*, 52:1183–1197, 2005.

A.W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.