Supplementary Appendix to

"Conditional Inference with a Functional Nuisance Parameter"

By Isaiah Andrews[1] and Anna Mikusheva [2]

Abstract

This Supplementary Appendix contains additional results concerning the interpretation of our conditional critical values, the bounded completeness of our sufficient statistics, the derivation of the conditioning process $h_T(\cdot)$ in homoscedastic linear IV, the power of tests in a simple Gaussian model, the power of the conditional QLR tests in linear IV with non-homoscedastic errors, proofs of asymptotic results stated in the paper, a theoretical analysis and additional simulation results for the quantile IV model, and additional results for Stock and Wright (2000)'s setting.

Key words: weak identification, similar test, conditional inference

First draft: September 2014. This draft: February, 2016.

S1 Interpretation of conditional critical values

The are many ways to represent a given test using data-dependent critical values. In particular, for any statistic $S(g_T)$ such that the test that rejects when $S(g_T) > 0$ has correct size, the test that rejects when the statistic $R = R(g_T, \Sigma)$ exceeds the random critical value $R - S(g_T)$ has correct size as well, and indeed will be the same test in that it rejects for precisely the same realizations of the data. The goal of this section is to point out the sense in which the test that rejects when $R > c_\alpha(h_T)$, for $c_\alpha(h_T)$ the critical value we propose, is naturally connected to the test statistic R. An interesting corollary of this result is that this test can be viewed as a best approximation (within the class of size-α conditionally similar tests) to any test based on R that uses a fixed critical value.

Let $\Phi_{C,\alpha}$ denote the class of size-α tests of $H_0 : m_T(\theta_0) = 0$ which are conditionally similar given h_T. For a given realization of the data a test in this class rejects the null with

[1]Harvard Society of Fellows. Harvard Department of Economics, Littauer Center 124, Cambridge, MA 02138. Email iandrews@fas.harvard.edu. NSF Graduate Research Fellowship support under grant number 1122374 is gratefully acknowledged.

[2]Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02139. Email: amikushe@mit.edu. Financial support from the Castle-Krob Career Development Chair and the Sloan Research Fellowship is gratefully acknowledged. We thank Alex Belloni, Victor Chernozhukov, Kirill Evdokimov, and Martin Spindler for helpful discussions.

probability $\tilde{\phi} \in [0,1]$, while also satisfying the conditional size restriction $E[\tilde{\phi}|h_T] = \alpha$ under any $m_T \in \mathcal{M}_0$.[3]

Lemma S1.1 *Suppose that the conditional distribution of R given h_T is almost surely continuous. For any non-decreasing function $F(\cdot)$ and any $m_T \in \mathcal{M}_0$ the test $\phi = \mathbb{I}\{R > c_\alpha(h_T)\}$ solves the problem*

$$\max_{\tilde{\phi} \in \Phi_{C,\alpha}} E\left[\tilde{\phi} F(R)\right]. \tag{1}$$

If $F(\cdot)$ is strictly increasing then ϕ is the almost-everywhere unique solution, in the sense that any other test $\tilde{\phi}$ solving optimization problem (1) above is equal to ϕ with probability one.

Proof: Let $f(h_T) = F(c_\alpha(h_T))$. Note that for any (potentially randomized) test $\tilde{\phi} \in [0,1]$ the following inequality holds almost surely:

$$(\phi - \tilde{\phi})(F(R) - f(h_T)) \geq 0.$$

Indeed, if $F(R) > f(h_T)$ then $R > c_\alpha(h_T)$, and thus $\phi = 1 \geq \tilde{\phi}$, and if $F(R) < f(h_T)$, then $R < c_\alpha(h_T)$ and thus $\phi = 0 \leq \tilde{\phi}$. As a result,

$$\begin{aligned} 0 \leq E\left[(\phi - \tilde{\phi})(F(R) - f(h_T))\right] = \\ = E\left[f(h_T)\left(E\left[\tilde{\phi}|h_T\right] - E\left[\phi|h_T\right]\right)\right] + E[\phi F(R)] - E[\tilde{\phi} F(R)]. \end{aligned}$$

If $\tilde{\phi} \in \Phi_{C,\alpha}$ then the first term equals zero, and we have $E[\tilde{\phi} F(R)] \leq E[\phi F(R)]$.

To establish the second statement of the Lemma, assume that $\tilde{\phi} \in \Phi_{C,\alpha}$ is such that $E[\tilde{\phi} F(R)] = E[\phi F(R)]$. Then

$$E\left[(\phi - \tilde{\phi})(F(R) - f(h_T))\right] = 0.$$

The integral of an almost-surely-non-negative function is equal to zero only if the function itself is equal to zero almost surely. We assumed that the conditional distribution of R

[3] We may equally well define $\tilde{\phi} \in \{0,1\}$ to be a realized outcome of the test, which may depend on an auxiliary randomization as in the paper. This distinction is unimportant for Lemma S1.1, though we use outcome notation for Corollary S1.1.

given h_T is almost surely continuous and F is strictly increasing. Thus the probability of the event $\{F(R) = f(h_T)\}$ is zero, so $\phi = \tilde{\phi}$ almost surely. $\square$

Lemma S1.1 establishes that the test ϕ can be interpreted as a maximizer of $E[\phi F(R)]$ over the class of conditionally similar tests for any distribution consistent with the null and any non-decreasing function F. This property makes precise the sense in which ϕ is the conditionally similar test most associated with large values of R. Note further that if the family of distributions for $h_T(\cdot)$ consistent with the null is complete, so that all similar tests are conditionally similar, then the conclusion of Lemma S1.1 continues to hold when we replace $\Phi_{C,\alpha}$ with $\Phi_{S,\alpha}$, the class of level-α similar tests.

A particularly interesting consequence of Lemma S1.1 is to relate the test ϕ to the test $\phi^* = \mathbb{I}\{R > c^*\}$ which is also based on R but, unlike ϕ, uses a fixed critical value. In particular:

Corollary S1.1 *If the conditional distribution of R given h_T is almost surely continuous, then for any $m_T \in \mathcal{M}_0$ the test ϕ solves*

$$\min_{\tilde{\phi} \in \Phi_{C,\alpha}} E\left[\left(\tilde{\phi} - \phi^*\right)^2\right] \tag{2}$$

where for a randomized test $\tilde{\phi}$ we use the final outcome in evaluating (2).

Proof: As noted in e.g. Section 3.5 of Lehmann and Romano, any randomized test $\tilde{\phi}(g_T)$ based on g_T can be represened as a non-randomized test $\tilde{\phi}(g_T, U)$ based on g_T and a uniform random variable U independent of the data. Using this representation, note that

$$E\left[\left(\tilde{\phi} - \phi^*\right)^2\right] = E\left[\tilde{\phi}\right] - 2E\left[\tilde{\phi}\phi^*\right] + E\left[\phi^*\right].$$

Conditional similarity of $\tilde{\phi}$ at level α implies that $E\left[\tilde{\phi}\right] = \alpha$, while $E[\phi^*]$ is unaffected by the choice of $\tilde{\phi}$. Thus, equation (2) is equivalent to

$$\max_{\tilde{\phi} \in \Phi_{C,\alpha}} E\left[\tilde{\phi}\phi^*\right].$$

Since $\phi^* = \mathbb{I}\{R > c^*\}$, the result follows immediately from Lemma S1.1. $\square$

Thus, if one would like to use a test ϕ^* but for its lack of size control, our approach yields the conditionally similar test that best approximates ϕ^* under any distribution

consistent with the null. Thus, analogous to the size-correction approaches discussed in the literature on non-standard tests (e.g. D. Andrews and Guggenberger (2009)), it seems reasonable to interpret ϕ as a size-corrected version of ϕ^* or, alternatively, as the best "conditional-similarity-corrected" version of ϕ^*, where the best is taken to mean minimizing squared approximation error under to any distribution consistent with the null.

S2 Boundedly complete families

Conditional similarity of a test is a very strong restriction that can be hard to justify because it significantly reduces the class of possible tests. If, however, one wishes to create a similar test (a test that has exactly correct rejection probability under the null for all values of the nuisance parameter), all similar tests will automatically be conditionally similar for a given sufficient statistic if the family of distributions for that sufficient statistic under the null is boundedly complete.

Definition 1 *The family of distributions $\mathcal{P}$ for a statistic h (or a σ-field $\sigma(h)$) is called boundedly complete if for all bounded $\sigma(h)$-measurable functions $f(h)$, the property that $E_P[f(h)] = 0$ for all $P \in \mathcal{P}$ implies that $f(h) = 0$ almost surely for all $P \in \mathcal{P}$.*

Lemma S2.1 *(Lehmann and Romano (2005)) If the family of distributions for the sufficient statistic h under the null is boundedly complete, then all similar tests of $H_0 : m \in \mathcal{M}_0$ are conditionally similar given h. In particular, any random variable ϕ with values in $[0,1]$ satisfying the similarity condition:*

$$E_P[\phi] = \alpha \text{ for any } P \in \mathcal{P}_0$$

also satisfies the conditional similarity condition

$$E_P[\phi \mid h] = \alpha \text{ a.s. for any } P \in \mathcal{P}_0.$$

Consider the exact Gaussian problem discussed in Section 3.2 of the paper in which we observe the process $g(\theta) = m(\theta) + G(\theta)$ for an unknown deterministic mean function $m(\cdot) \in \mathcal{M}$ and a mean-zero Gaussian process $G(\cdot)$ with known covariance $\Sigma(\theta, \tilde{\theta}) = EG(\theta)G(\tilde{\theta})'$. We again assume that $\mathcal{M}$ is the set of potential mean functions, which is

in general infinite-dimensional, and wish to test the hypothesis $H_0 : m(\theta_0) = 0$. Let $\mathcal{M}_0$ be the subset of $\mathcal{M}$ containing all functions m satisfying $m(\theta_0) = 0$. Let $\mathcal{P}_0$ be the set of distributions for g corresponding to mean functions in $\mathcal{M}_0$. Define $h(\cdot) = H(g, \Sigma)$ as in equation (3) from the paper. As shown in Lemma 1 of the paper, h is a sufficient statistic for $m \in \mathcal{M}_0$.

If the parameter space for θ is finite ($\Theta = \{\theta_0, \theta_1, ..., \theta_n\}$) the conditions for bounded completeness are well known and easy to check. In particular, in this case our problem reduces to that of observing a $k(n+1)$-dimensional Gaussian vector $g = (g(\theta_0)', ..., g(\theta_n)')'$ with unknown mean $(0, m_1' = m(\theta_1)', ..., m_n' = m(\theta_n)')'$ and known covariance. If the set $\mathcal{M}$ of possible values for the nuisance parameter $(m_1', ..., m_n')'$ contains a rectangle with a non-empty interior then the family of distributions for h under the null is boundedly complete, and all similar tests are conditionally similar given h. To generalize this statement to an infinite-dimensional nuisance parameter we can use Lemma 3.3 of Janssen and Ostrovski (2005) to obtain a sufficient condition for completeness.

Lemma S2.2 *(Janssen and Ostrovski (2005)) Consider an increasing sequence $\mathcal{M}_n \subseteq \mathcal{M}_0$ of subsets of the parameter space and let $\mathcal{P}_n = \{P_m, m \in \mathcal{M}_n\}$ be an increasing set of families of distributions with $\mathcal{P}_\infty = \cup_n \mathcal{P}_n$. Let $\mathcal{A}_n$ denote an increasing sequence of σ-algebras with $\mathcal{A}_\infty = \cup_n \mathcal{A}_n$. Assume that for each $n \in \mathbb{N}$ the σ-algebra $\mathcal{A}_n$ is sufficient for $m \in \mathcal{M}_n$ while $\mathcal{P}_n$ is boundedly complete for $\mathcal{A}_n$. Assume that $\mathcal{A}_\infty$ is sufficient for $\mathcal{P}_0$ and that for any $P \in \mathcal{P}_0$ there exists $P_1 \in \mathcal{P}_\infty$ such that P is absolutely continuous with respect to P_1. Then $\mathcal{P}_0$ is boundedly complete for $\mathcal{A}_\infty$.*

Consider the exact Gaussian problem and assume that the covariance function $\Sigma(\cdot, \cdot)$ is continuous while Θ is a separable metric space. Consider the corresponding process $h(\cdot)$, which has mean $m(\cdot)$ and covariance function $\tilde{\Sigma}(\cdot, \cdot)$:

$$\tilde{\Sigma}(\theta, \theta_1) = \Sigma(\theta, \theta_1) - \Sigma(\theta, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}\Sigma(\theta_0, \theta_1).$$

Note that $\tilde{\Sigma}(\theta_0, \theta) = 0$ for all θ. Let $\mathcal{H}$ be a Hilbert space with reproducing kernel $\tilde{\Sigma}(\cdot, \cdot)$. This is defined as the closure of the set of functions $\phi(\cdot) = \sum_{j=1}^n \alpha_j \tilde{\Sigma}(\cdot, \theta_j)$ with respect to the inner product

$$\langle \sum_{j=1}^n \alpha_j \tilde{\Sigma}(\cdot, \theta_j), \sum_{i=1}^m \beta_i \tilde{\Sigma}(\cdot, s_i) \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_j \beta_i \tilde{\Sigma}(\theta_j, s_i).$$

Assume that $\mathcal{M}_0 \subseteq \mathcal{H}$. This assumption can be justified using Theorem 57 of Berlinet and Thomas-Agnan (2004), which states that two Gaussian processes with common covariance function Σ and potentially different mean functions $m_X(\cdot)$ and $m_Y(\cdot)$ are mutually absolutely continuous if and only if $m_X - m_Y \in \mathcal{H}$.

Theorem S2.1 *Consider the exact Gaussian problem. Assume the covariance function $\Sigma(\cdot,\cdot)$ is continuous, and Θ is a separable metric space. Assume that $\mathcal{M}_0 \subseteq \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with reproducing kernel $\tilde{\Sigma}(\cdot,\cdot)$, and $(\phi_j)_{j=1}^{\infty}$ is a basis in $\mathcal{H}$. Assume that for any n there exists an n-dimensional closed rectangle B_n with a non-empty interior such that for any $\alpha = (\alpha_1, ..., \alpha_n) \in B_n$ we have $\sum_{j=1}^{n} \alpha_j \phi_j \in \mathcal{M}_0$. Then h is boundedly complete for $\mathcal{P}_0$.*

Proof of Theorem S2.1. Given the assumptions, the Hilbert space $\mathcal{H}$ is separable. Without loss of generality we can assume that ϕ_j is an orthonormal basis in $\mathcal{H}$. According to Theorem 14 in Berlinet and Thomas-Agnan (2004) we have $\tilde{\Sigma}(\theta_1, \theta_2) = \sum_{j=1}^{\infty} \phi_j(\theta_1)\phi_j(\theta_2)'$. For any $m \in \mathcal{M}_0 \subseteq \mathcal{H}_R$ we have

$$m(\theta) = \sum_{j=1}^{\infty} \langle m, \phi_j \rangle_{\mathcal{H}} \phi_j(\theta) = \sum_{j=1}^{\infty} a_j \phi_j(\theta).$$

Consider the canonical congruence ψ, which is a one-to-one correspondence between the space $\mathcal{H}$ and the L^2 closure of the space of linear combinations of the process $h(\theta) - m(\theta)$ (we refer the interested reader to Berlinet and Thomas-Agnan (2004) for details). A key property of the canonical congruence is that $\psi(\tilde{\Sigma}(\cdot, \theta)) = h(\theta) - m(\theta)$ for all θ. Let $\xi_i = \psi(\phi_i)$ be the random variables corresponding to the basis functions, which are independent (for different i), standard normal variables since $E(\psi(\phi_i)\psi(\phi_j)) = \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \mathbb{I}\{i = j\}$. There exists a canonical expansion of the mean-zero process $h(\cdot) - m(\cdot)$:

$$h(\theta) - m(\theta) = \psi(\tilde{\Sigma}(\cdot, \theta)) = \sum_{j=1}^{\infty} \phi_j(\theta)\xi_j.$$

Observing the process $h(\cdot)$ is therefore equivalent to observing the sequence of random variables

$$Y_j = \langle h, \phi_j \rangle_{\mathcal{H}} = \langle m, \phi_j \rangle_{\mathcal{H}} + \xi_j.$$

Denote by $\mathcal{M}_n$ a subset of $\mathcal{M}_0$ consisting of functions $m(\cdot)$ such that $\langle m, \phi_j \rangle = 0$ for all $j > n$, thus $m(\theta) = \sum_{j=1}^{n} a_j \phi_j(\theta)$ for $m \in \mathcal{M}_n$. Notice that for $m \in \mathcal{M}_n$ we have $Y_j = a_j + \xi_j$ for $j \leq n$ and $Y_j = \xi_j$ for $j > n$. To apply Lemma S2.2 we can take $\mathcal{A}_n$ to be the σ-algebra generated by variables $Y_1, ..., Y_n$. Note that $\mathcal{A}_n$ is sufficient for $\mathcal{P}_n$. Our assumption about rectangular B_n ensures that $\mathcal{P}_n$ is complete for $\mathcal{A}_n$. All the other conditions of Lemma S2.2 are satisfied, so the result follows. $\square$

S3 Homoscedastic Linear IV

Let us consider Example 2 from the paper with the additional assumption of homoscedastic errors. The moment condition considered is $g_T(\theta) = \frac{1}{\sqrt{T}} Z'(Y - D\theta)$, where Y and D are $T \times 1$ and $T \times p$ matrices of endogenous variables, Z is a $T \times k$ non-random matrix of instruments, and θ is a $p \times 1$ parameter of interest. Let Ω be the $(p+1) \times (p+1)$ covariance matrix of the reduced form errors, which are assumed to be homoscedastic. Then $\Sigma(\theta, \theta_0) = \frac{1}{T} Z'Z(1, -\theta')\Omega(1, -\theta_0')'$. As a result we have

$$h_T(\theta) = g_T(\theta) - \frac{(1, -\theta')\Omega(1, -\theta_0')'}{(1, -\theta_0')\Omega(1, -\theta_0')'} g_T(\theta_0).$$

Below we prove that

$$h_T(\theta) = \frac{1}{\sqrt{T}(1, -\theta_0')\Omega(1, -\theta_0')'} [Z'Y, Z'D] \, \Omega^{-1} \begin{pmatrix} \theta_0' \\ I_p \end{pmatrix} B(\theta - \theta_0), \tag{3}$$

where B is a full rank $p \times p$ matrix. Thus $h_T(\theta)$ is linear in θ, so conditioning on $h_T(\cdot)$ is equivalent to conditioning on the $k \times p$ matrix $[Z'Y, Z'D] \, \Omega^{-1} \begin{pmatrix} \theta_0' \\ I_p \end{pmatrix}$, which is the T statistic of Moreira (2003).

Proof of statement (3): Plugging $g_T(\theta)$ into the formula for h_T:

$$\begin{aligned} h_T(\theta) &= \frac{1}{\sqrt{T}(1, -\theta_0')\Omega(1, -\theta_0')'} [Z'Y, Z'D] \begin{pmatrix} 0 & \theta' - \theta_0' \\ \theta_0 - \theta & \theta\theta_0' - \theta_0\theta' \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\theta_0 \end{pmatrix} \\ &= \frac{1}{\sqrt{T}(1, -\theta_0')\Omega(1, -\theta_0')'} [Z'Y, Z'D] \begin{pmatrix} 0 & x' \\ -x & x\theta_0' - \theta_0 x' \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\theta_0 \end{pmatrix}, \end{aligned}$$

for $\theta = \theta_0 + x$. We decompose Ω as $\Omega = \begin{pmatrix} \omega & w' \\ w & \Omega_2 \end{pmatrix}$, where ω is a scalar, w is $p \times 1$, and Ω_2 is $p \times p$. Consider the last three terms in the previous formula:

$$\begin{pmatrix} 0 & x' \\ -x & x\theta_0' - \theta_0 x' \end{pmatrix} \begin{pmatrix} \omega & w' \\ w & \Omega_2 \end{pmatrix} \begin{pmatrix} 1 \\ -\theta_0 \end{pmatrix}$$

$$= \begin{pmatrix} x'w - x'\Omega_2\theta_0 \\ -\omega x + xw'\theta_0 + x\theta_0'w - x\theta_0'\Omega_2\theta_0 - \theta_0 x'w + \theta_0 x'\Omega_2\theta_0 \end{pmatrix}$$

$$= \begin{pmatrix} (w' - \theta_0'\Omega_2) \\ [(2w'\theta_0 - \omega - \theta_0'\Omega_2\theta_0)I_p - \theta_0 w' + \theta_0\theta_0'\Omega_2] \end{pmatrix} x.$$

Next, note that

$$\begin{pmatrix} w' - \theta_0'\Omega_2 \\ (2w'\theta_0 - \omega - \theta_0'\Omega_2\theta_0)I_p - \theta_0 w' + \theta_0\theta_0'\Omega_2 \end{pmatrix} = \Omega^{-1} \begin{pmatrix} \theta_0' \\ I_p \end{pmatrix} B.$$

To prove this statement, it suffices to note that

$$\Omega \begin{pmatrix} w' - \theta_0'\Omega_2 \\ (2w'\theta_0 - \omega - \theta_0'\Omega_2\theta_0)I_p - \theta_0 w' + \theta_0\theta_0'\Omega_2 \end{pmatrix} = \begin{pmatrix} \theta_0' B \\ B \end{pmatrix}$$

for

$$B = ww' - w\theta_0'\Omega_2 + (2w'\theta_0 - \omega - \theta_0'\Omega_2\theta_0)\Omega_2 - \Omega_2\theta_0 w' + \Omega_2\theta_0\theta_0'\Omega_2.$$

All that remains is to show that B is full rank. To this end, note that

$$B = yy' - (y'\Omega_2^{-1}y)\Omega_2 - (\omega - w'\Omega_2^{-1}w)\Omega_2$$

for $y = w - \Omega_2\theta_0$. Thus B is symmetric. We show that $-B$ is positive-definite and thus is full rank. First, note that for any $p \times 1$ vector a we have

$$a' \left(yy' - (y'\Omega_2^{-1}y)\Omega_2\right) a \leq 0$$

as follows from the Cauchy-Schwarz inequality applied to vectors $\Omega^{1/2}a$ and $\Omega^{-1/2}y$. Thus

$-yy' + (y'\Omega_2^{-1}y)\Omega_2$ is a positive semi-definite matrix. We also note that $\omega - w'\Omega_2^{-1}w$ is strictly positive by the positive definiteness of Ω. Indeed,

$$(1, -w'\Omega_2^{-1}) \begin{pmatrix} \omega & w' \\ w & \Omega_2 \end{pmatrix} \begin{pmatrix} 1 \\ -\Omega_2^{-1}w \end{pmatrix} = \omega - w'\Omega_2^{-1}w > 0.$$

Thus, B is full rank.

S4 Power considerations

In this section we provide derivations for statements made in Section 3.4 of the paper, as well as an extension to a case with multiple alternatives. The setting considered in Section 3.4 can be described as follows. We observe two k-dimensional random vectors $\xi = g_T(\theta_0)$ and $\eta = g_T(\theta^*)$:

$$(\xi', \eta')' \sim N((\mu', \lambda')', \Sigma),$$

where the $2k \times 2k$- covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ is known, while the means μ and λ are unknown and satisfy the restriction $\|\mu\| \cdot \|\lambda\| = 0$. We wish to test the null $H_0 : \mu = 0, \lambda \in \mathbb{R}^k$ against the alternative $H_1 : \mu \neq 0, \lambda = 0$.

S4.1 Tests based on ξ alone

We derive the power envelope for tests based only on ξ by finding the optimal test against a specific alternative μ (noting that under this restriction we can ignore the parameter λ, since it affects only the distribution of η). By the Neyman-Pearson lemma, the optimal test rejects when $\xi'\Sigma_{11}^{-1}\xi - (\xi - \mu)'\Sigma_{11}^{-1}(\xi - \mu)$ exceeds its $1 - \alpha$-quantile under $\mu = 0$ or, equivalently, when

$$\xi'\Sigma_{11}^{-1}\mu / \sqrt{\mu'\Sigma_{11}^{-1}\mu} > z_{1-\alpha}, \tag{4}$$

where $z_{1-\alpha}$ is the $1 - \alpha$-quantile of the standard normal distribution. As is immediately clear this most powerful test is one-sided and is biased as a test of the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. The power of this test (evaluated using the true μ) yields power envelope PE-1 in the paper.

To construct the power envelope for the class of *unbiased* tests based on ξ, consider the sub-problem which assumes the direction of μ known, that is, $\mu \propto \mu^*$ for a known vector μ^*. In this sub-problem the only unknown parameter is the coefficient of proportionality between μ and μ^*, which can be written as $a = \mu'\Sigma_{11}^{-1}\mu^*(\mu^{*\prime}\Sigma_{11}^{-1}\mu^*)^{-1}$, and the testing problem (based on ξ) becomes $H_0 : a = 0$ against $H_1 : a \neq 0$. A sufficient statistic for the unknown parameter a is $(\mu^{*\prime}\Sigma_{11}^{-1}\mu^*)^{-1}\mu^{*\prime}\Sigma_{11}\xi \sim N(a, 1/(\mu^{*\prime}\Sigma_{11}^{-1}\mu^*))$. The uniformly most powerful unbiased test in this sub-problem rejects if the square of the statistic in equation (4) above exceeds the $1-\alpha$-quantile of the χ_1^2 distribution. It is easy to see that this test is unbiased for the initial problem of testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ as well, and as such provides the best unbiased test against alternative μ. By changing the value of μ we can construct the power envelope, labeled PE-2 in the paper, for the class of unbiased tests based on ξ alone.

Note that the optimal unbiased test depends in a significant way on the alternative tested, and thus there is no uniformly most powerful unbiased test. In practice it is difficult to choose the particular alternative on which to focus, and for a given choice of μ the tests above will have very low level power against some alternatives. Kleibergen's K test can be viewed as a plug-in version of the optimal unbiased test with an orthogonalized derivative estimator substituting for μ. As noted in I. Andrews (2015), this derivative estimator may perform quite well for alternatives close to the null but its performance can deteriorate against more distant alternatives. Further, as noted in Section 3.5 of the paper, replacing μ by a derivative estimator seems to make sense when the unknown function $m_T(\cdot)$ is approximately linear in θ, but may not work as well when $m_T(\cdot)$ is substantially nonlinear.

Another way of eliminating the dependence of the optimal test on μ is to consider only tests invariant to a full-rank $k \times k$ linear transformation of the data, noting that this transformation preserves both the null and alternative. One can easily see that a maximal invariant under this transformation is the S statistic $\xi'\Sigma_{11}^{-1}\xi$ and that the uniformly most powerful invariant test rejects when S is large. Thus, the S (or Anderson-Rubin) test is the uniformly most powerful invariant test based on ξ alone, and its power function yields power envelope PE-3.

S4.1.1 Power Bound for Tests in Previous Literature

In the paper we note that the power envelope PE-2 for unbiased tests based on ξ gives an upper bound on the power of most of the tests studied in the previous weak identification literature, including the S, K, JK, and GMM-M tests of Kleibergen (2005). The power envelope PE-2 likewise bounds the power of the CLC tests studied by I. Andrews (2015), and (in the linear IV model) of the SU tests studied by Moreira and Moreira (2015). This follows from the fact that all of these tests satisfy a version of the SU condition of Moreira and Moreira (2015).

Specifically, consider tests ϕ which depend on the data only through $g_T(\theta_0)$ and D_T, where consistent with the exact Gaussian model studied in Section 3 of the paper we assume that $g_T(\theta_0) \sim N(\mu, \Sigma(\theta_0, \theta_0))$ and $D_T \sim N(\mu_D, \Sigma_D)$. Once we restrict attention to $(g_T(\theta_0), D_T)$, D_T is sufficient for μ_D and all the tests discussed above are conditionally similar given D_T in that

$$E_{\mu=0}[\phi | D_T = d] = \alpha \tag{5}$$

for almost every d. These tests further satisfy the restriction

$$E_{\mu=0}[\phi g_T(\theta_0) | D_T = d] = 0 \tag{6}$$

for almost every d. Conditions (5) and (6) together are sufficient for the SU condition of Moreira and Moreira (2015).

As noted in the supplement to I. Andrews (2015), a sufficient condition for condition (6) is that $\phi(g_T(\theta_0), D_T) = \phi(-g_T(\theta_0), D_T)$, which holds for all the tests discussed above save for the SU tests of Moreira and Moreira (2015), which instead impose restriction (6) by construction. Given equations (5) and (6), however, standard arguments imply that the most powerful test against alternative μ rejects when $\left(\mu' \Sigma(\theta_0, \theta_0)^{-1} g_T(\theta_0)\right)^2$ is large, and thus coincides with the test used to construct PE-2, which in the linear IV model also coincides with the power envelope for SU tests derived in Moreira and Moreira (2015).

S4.2 Tests based on ξ and η

Assume that under the null the parameter λ can take any value in some k-dimensional rectangle (which implies that similar tests must be conditionally similar given h). If we

let $\mu_h = -\Sigma_{21}\Sigma_{11}^{-1}\mu$ and denote the variance of h by Σ_h, then following Montiel Olea (2013) we know that the point-optimal similar test against μ will reject when

$$\frac{\exp(-(\xi-\mu)'\Sigma_{11}^{-1}(\xi-\mu)-(h-\mu_h)'\Sigma_h^{-1}(h-\mu_h))}{\exp(-\xi'\Sigma_{11}^{-1}\xi)}$$

exceeds its $1-\alpha$ quantile under the null conditional on h. However, we can see that once we condition on h, the only random component of the ratio is ξ. Moreover, for fixed h this statistic is a strictly increasing function of $\xi'\Sigma_{11}^{-1}\mu$. Thus, the test will reject when $\xi'\Sigma_{11}^{-1}\mu$ exceeds its $1-\alpha$ conditional quantile under the null. However, $\xi'\Sigma_{11}^{-1}\mu$ is independent of h, so we obtain that the most powerful similar test against μ based on (ξ,η) is the same as the most powerful test based on ξ alone, described in equation (4).

It is without loss of generality to normalize $\Sigma_{11} = \Sigma_{22} = I_k$. Under this normalization the conditional pQLR test rejects the null when

$$\xi'\xi - \eta'\eta > c_\alpha(h), \text{ where } h = \eta - \Sigma_{21}\xi.$$

In what follows we discuss optimality properties of the conditional pQLR test in two cases: first, under the assumption that the covariance matrix Σ of variables $(\xi,\eta) = (g_T(\theta_0)', g_T(\theta^*)')'$ can be written as the Kronecker product of a 2×2 matrix with a $k\times k$ matrix, and then in the general case.

S4.2.1 Kronecker structure case

In this section we show that under the Kronecker structure assumption discussed in the paper, which imposes that $\Sigma_{21} = \rho I_k$ for $|\rho| < 1$, and thus that $h = \eta - \rho\xi$, the conditional pQLR test is the uniformly most powerful invariant similar test based on ξ and η and is unbiased.

First, note that the testing problem is invariant to rotations of ξ and η. Specifically, for F a $k\times k$ orthonormal matrix ($F'F = I_k$),

$$((F\xi)', (F\eta)')' \sim N(((F\mu)', (F\lambda)')', \Sigma),$$

and the null hypothesis is unchanged. A test $\phi(\xi,\eta)$ is said to be invariant if for any orthonormal F we have $\phi(F\xi, F\eta) = \phi(\xi,\eta)$. The power function $\beta(\mu)$ of an invariant

test ϕ satisfies

$$\beta(F\mu) = \beta(\mu) = \beta(\|\mu\|).$$

Consider an arbitrary weighting function $\pi(\mu)$ on the set of alternatives. Then the weighted average power (WAP) of the invariant test ϕ is

$$\int \beta(\mu)\pi(\mu)d\mu = \int_0^\infty \int_{\|\mu\|=x} \beta(\mu)\pi(\mu)d\mu dx =$$
$$= \int_0^\infty \beta(x) \left[\int_{\|\mu\|=x} \pi(\mu)d\mu \right] dx = \int_0^\infty \beta(x) \left[q(x) \int_{\|\mu\|=x} d\mu \right] dx,$$

where $q(x) = \frac{\int_{\|\mu\|=x} \pi(\mu)d\mu}{\int_{\|\mu\|=x} d\mu}$. Let us define a new weighting function $\tilde{\pi}(\mu) = q(\|\mu\|)$. We see that for any invariant test, WAP with respect to π is the same as WAP with respect to $\tilde{\pi}$.

Now let us find the invariant similar test with the highest WAP for an arbitrary weight $\pi(\mu)$. By the argument above this test also maximizes WAP for weight $\tilde{\pi}(\mu)$. We construct the WAP optimal similar test with respect to $\tilde{\pi}$ (without assuming invariance). According to the results in Montiel Olea (2013), this test will reject when

$$\frac{\int \tilde{\pi}(\mu) \exp\left\{ -1/2(\xi' - \mu', \eta') \begin{pmatrix} I_k & \rho I_k \\ \rho I_k & I_k \end{pmatrix}^{-1} (\xi' - \mu', \eta')' \right\} d\mu}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h),$$

where the conditional critical value $c_\alpha(h)$ is chosen to control size conditionally on h. Given the symmetry of $\tilde{\pi}(\mu) = q(\|\mu\|)$, this is equivalent to

$$\frac{\int_0^\infty q(x) \int_{\|\mu\|=x} \exp\left\{ -\frac{1}{2(1-\rho^2)}\|\xi - \mu - \rho\eta\|^2 - \frac{1}{2}\eta'\eta \right\} d\mu dx}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h).$$

Letting $u = \xi - \rho\eta$, the optimal test rejects when

$$G(u) \frac{\exp\left\{ -\frac{1}{2(1-\rho^2)}\|u\|^2 - \frac{1}{2}\eta'\eta \right\}}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h),$$

where

$$G(u) = \int_0^\infty q(x) \exp\left\{ -\frac{x^2}{2(1-\rho^2)} \right\} \int_{\|\mu\|=x} \exp\left\{ \frac{1}{2(1-\rho^2)}\mu' u \right\} d\mu dx.$$

We can see that function $G(u)$ depends on u only through $\|u\|$. So, the optimal test statistic has the form

$$\widetilde{G}(\|u\|)\exp\left\{-\frac{1}{2}(\eta'\eta-\xi'\xi)\right\}$$

and the test has to control size conditionally on h. We can notice that conditional on h the statistic

$$\|u\|^2=\|\xi-\rho(h+\rho\xi)\|^2=(1-\rho^2)\left((1-\rho^2)\xi'\xi-2\rho h'\xi+\frac{\rho^2h'h}{1-\rho^2}\right)$$

is a positive affine transformation of the statistic

$$\xi'\xi-\eta'\eta=\xi'\xi-(h+\rho\xi)'(h+\rho\xi)=(1-\rho^2)\xi'\xi-2\rho h'\xi-h'h.$$

We arrive at the following conclusions:

1. The WAP optimal test statistic with respect to $\tilde{\pi}(\mu)$ is invariant to transformation of (ξ,η) to $(F\xi,F\eta)$ (or, equivalently, transformation of (ξ,h) to $(F\xi,Fh)$) and the conditional critical value function $c_\alpha(h)$ is likewise invariant. Thus, the WAP optimal similar test with respect to $\tilde{\pi}(\mu)$ is automatically invariant, and thus is the WAP optimal invariant similar test against both $\pi(\mu)$ and $\tilde{\pi}(\mu)$.

2. Conditional on h the WAP optimal test statistic depends on data only through $\|u\|$.

3. Since $G(u)$ is increasing in $\|u\|$, the WAP optimal invariant similar test rejects the null for large values of $\|u\|$ no matter what the weight functions π or $\tilde{\pi}$ are. Thus it is uniformly most powerful in the class of invariant similar tests.

4. This optimal test is the conditional pQLR test.

5. The S test belongs to the class of the invariant similar tests, as such its power is weakly dominated by the power of the conditional pQLR test at all points.

6. Since the S test is unbiased, we conclude that the conditional pQLR is unbiased as well.

While we have derived this optimality result directly from the form of the weighted average power optimal similar tests for this problem, note that under the restrictions

considered here (using only the values on the moment process at two points, and the Kronecker product structure for Σ), the testing problem here is equivalent to the problem of testing one structural parameter value against another in linear IV with homoscedastic errors. Thus, the optimality of the pQLR test can instead be derived as a special case of Corollary 1 in Mills et al. (2014).

S4.2.2 General case

We next relax the Kronecker structure assumption and consider the general case, where Σ_{21} is not necessarily proportional to the identity matrix. In this context we show that the conditional pQLR test is WAP optimal within the class of similar tests for a particular class of weight functions.

We consider a weight function $\pi(\mu)$ with the property that $\pi(\mu) = \pi(\tilde{\mu})$ for all μ and $\tilde{\mu}$ such that $\tilde{\mu}'(I - \Sigma_{12}\Sigma_{21})^{-1}\tilde{\mu} = \mu'(I - \Sigma_{12}\Sigma_{21})^{-1}\mu$. The corresponding WAP optimal similar test rejects when

$$\frac{\int \pi(\mu) \exp\left\{-\frac{1}{2}(\xi' - \mu', \eta') \begin{pmatrix} I_k & \Sigma_{12} \\ \Sigma_{21} & I_k \end{pmatrix}^{-1} (\xi' - \mu', \eta')'\right\} d\mu}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h)$$

or, equivalently,

$$\frac{\int \pi(\mu) \exp\left\{-\frac{1}{2}(\xi - \mu - \Sigma_{12}\eta)'(1 - \Sigma_{12}\Sigma_{21})^{-1}(\xi - \mu - \Sigma_{12}\eta) - \frac{1}{2}\eta'\eta\right\} d\mu}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h).$$

Letting $u = \xi - \Sigma_{12}\eta$, the WAP optimal test rejects when

$$\frac{e^{-\frac{1}{2}u'(1-\Sigma_{12}\Sigma_{21})^{-1}u - \frac{1}{2}\eta'\eta} G(u)}{\exp\{-\frac{1}{2}\xi'\xi\}} > c_\alpha(h)$$

for

$$G(u) = \int \pi(\mu) \exp\left\{u'(1 - \Sigma_{12}\Sigma_{21})^{-1}\mu - \frac{1}{2}\mu'(1 - \Sigma_{12}\Sigma_{21})^{-1}\mu\right\} d\mu.$$

Note that for the weight function we consider, the function $G(u)$ depends on u only through $u'(1 - \Sigma_{12}\Sigma_{21})^{-1}u = \|u\|^2_{\Sigma_{12}}$. Indeed, consider the scalar product $\langle v, w\rangle_{\Sigma_{12}} = v'(1 - \Sigma_{12}\Sigma_{21}')^{-1}w$ and the group of orthonormal matrices with respect to this scalar product, namely all matrices F such that $F(1 - \Sigma_{12}\Sigma_{21}')^{-1}F' = (1 - \Sigma_{12}\Sigma_{21}')^{-1}$. Then

for any $\tilde{u}$ such that $u'(1-\Sigma_{12}\Sigma_{21}')^{-1}u = \tilde{u}'(1-\Sigma_{12}\Sigma_{21}')^{-1}\tilde{u}$, we have that there exists a matrix F from this group such that $u = F\tilde{u}$. Thus in the formula for $G(\tilde{u})$ we can change the integration from μ to $F\mu$, and obtain that $G(\tilde{u})$ is equal to $G(u)$.

Thus far we have shown that the optimal WAP similar test with respect to weight π rejects when

$$G(\|u\|_{\Sigma_{12}}) \exp\left\{-\frac{1}{2}\|u\|_{\Sigma_{12}}^2\right\} \exp\left\{-\frac{1}{2}(\eta'\eta - \xi'\xi)\right\} > c_\alpha(h)$$

where the function G depends on π and the test is conditionally similar given h.

Finally, we note that for a given h the statistic $\|u\|_{\Sigma_{12}}^2$ is equal to $\xi'\xi - \eta'\eta$ plus a constant depending only on h. Indeed;

$$\|u\|_{\Sigma_{12}}^2 = (\xi - \Sigma_{12}(h + \Sigma_{21}\xi))'(I - \Sigma_{12}\Sigma_{21})^{-1}(\xi - \Sigma_{12}(h + \Sigma_{21}\xi))$$

$$= \xi'(I - \Sigma_{12}\Sigma_{21})\xi - 2h'\Sigma_{21}\xi + h'\Sigma_{21}(I - \Sigma_{12}\Sigma_{21})^{-1}\Sigma_{12}h$$

and

$$\xi'\xi - \eta'\eta = \xi'\xi - (h + \Sigma_{21}\xi)'(h + \Sigma_{21}\xi) = \xi'(I - \Sigma_{12}\Sigma_{21})\xi - 2h'\Sigma_{21}\xi - h'h.$$

Thus, we see that the two statistics differ by a summand that depends only on h. Thus, the WAP optimal test rejects when $G(\|u\|_{\Sigma_{12}})$ is large. Since $G(\|u\|_{\Sigma_{12}})$ is increasing in $\|u\|_{\Sigma_{12}}$, this implies that the conditional WAP test is equal to the conditional pQLR test.

S4.3 Extension to multiple alternatives

In the simulations discussed in Section 3.4 of the paper, we consider the problem of testing the null that the true parameter value is θ_0 against the alternative that it is θ^*. In applications, however, we are typically interested in the composite alternative that $\theta \neq \theta_0$ and so use the QLR or other tests rather than the pQLR test. To explore the effect of testing against such alternatives, in this section we consider a generalization of the stylized model studied in Section 3.4 of the paper which adds additional points to the parameter space.

Specifically, we consider a case where the GMM parameter space is finite, $\Theta_N = \{\theta_0, \theta_1, ..., \theta_N\}$. As in Section 3.4 we consider $k = 5$ moments. Since the GMM parameter

space is finite, observing the process $g_T(\cdot)$ reduces to observing

$$\begin{pmatrix} g_T(\theta_0) \\ g_T(\theta_1) \\ \vdots \\ g_T(\theta_N) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix}, \begin{pmatrix} \Sigma(\theta_0,\theta_0) & \Sigma(\theta_0,\theta_1) & \cdots & \Sigma(\theta_0,\theta_N) \\ \Sigma(\theta_1,\theta_0) & \Sigma(\theta_1,\theta_1) & \cdots & \Sigma(\theta_1,\theta_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(\theta_N,\theta_0) & \Sigma(\theta_N,\theta_1) & \cdots & \Sigma(\theta_N,\theta_N) \end{pmatrix} \right).$$

To obtain a simple parametrization for the covariance structure, we consider the case with

$$\begin{pmatrix} \Sigma(\theta_0,\theta_0) & \Sigma(\theta_0,\theta_1) & \cdots & \Sigma(\theta_0,\theta_N) \\ \Sigma(\theta_1,\theta_0) & \Sigma(\theta_1,\theta_1) & \cdots & \Sigma(\theta_1,\theta_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(\theta_N,\theta_0) & \Sigma(\theta_N,\theta_1) & \cdots & \Sigma(\theta_N,\theta_N) \end{pmatrix} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \otimes I_k.$$

Without futher loss of generality consider $\mu = \|\mu\| e_1$ and $\lambda_i = \|\lambda_i\| e_1$, for e_1 the first standard basis vector. Under the null we have that $\|\mu\| = 0$, while under the alternative correct specification of the model implies that $\prod_{i=1}^{N} \|\lambda_i\| = 0$. Since alternatives $\theta_1, ..., \theta_N$ play a symmetric role here, we focus on the case where θ_1 is the true parameter value, so under the alternative we have $\|\lambda_1\| = 0$. For simplicity we impose that $\lambda_2 = ... = \lambda_N$.

We ultimately obtain a problem indexed by two known parameters (the number of alternatives N, and the correlation ρ) and two unknown parameters (μ_1 and $\lambda_{N,1}$). For our power simulations, we consider $N \in \{2, 3, 4, 5, 10, 20\}$, $\rho \in \{0.3, 0.5, 0.9, 0.99\}$, and take both μ_1 and $\lambda_{N,1}$ to vary over the grid $\{-6, -5.7, ..., 5.7, 6\}$. As in the paper we study power envelopes PE-1 and PE-2, as well as the power of the S and pQLR tests. Unlike Section 3.4 of the paper we consider the conditional QLR test, based on the statistic

$$g_T(\theta_0)' \Sigma(\theta_0,\theta_0)^{-1} g_T(\theta_0) - \min_{i \in \{1,...,N\}} g_T(\theta_i)' \Sigma(\theta_i,\theta_i)^{-1} g_T(\theta_i).$$

Since the space of unknown alternatives is now two-dimensional it becomes more difficult to plot the results. Instead, for each (N, ρ) pair we consider the largest amount by which the power of each test falls short of the power of each of the other tests, as well as the average rejection probability with respect to uniform weights on each point of the $(\mu_1, \lambda_{N,1})$ grid.

The results of this exercise, based on 10,000 simulation draws, are reported in Tables 1-6. As these results make clear, even as we add more points to the parameter space the QLR test performs very well (and, indeed, outperforms the power envelope PE-2 over much of the parameter space) when ρ is large. For $N > 2$, however, this dominance is no longer uniform over the parameter space for $(\mu_1, \lambda_{N,1})$, and there are some alternatives where the power of QLR falls below PE-2. Unsurprisingly, the power of the conditional QLR test is typically (though not everywhere) exceeded by that of the pQLR test. More surprisingly, the power of the QLR test sometimes falls below that of the S test, albeit only by a small amount. In all of the simulation designs we consider the average power of the QLR test (averaged over all alternatives considered) exceeds the power of the S test, and the degree of out-performance is increasing in ρ.

S5 The conditional QLR test in linear IV

Our results apply to linear IV with both homoscedastic and non-homoscedastic (heteroscedastic, serially correlated, or clustered) errors. As discussed in Section 3.3 of the paper, the linear-in-parameters structure of the IV moment condition means that conditioning on $h_T(\cdot)$ in linear IV with homoscedastic errors is equivalent to conditioning on Moreira (2003)'s T statistic, so in that case our conditioning coincides with that of Moreira (2003) and the conditional QLR test becomes Moreira's CLR test. The conditioning process $h_T(\cdot)$ in the linear IV model remains linear in the parameter θ even when errors are non-homoscedastic, however, and conditioning on $h_T(\cdot)$ is more generally equivalent to conditioning on its Jacobian at θ_0. As noted in Section 3.3, this Jacobian is the negative of Kleibergen (2005)'s conditioning statistic D_T. Thus, in linear IV with non-homoscedastic errors conditioning on $h_T(\cdot)$ is equivalent to conditioning on Kleibergen's D_T.

One implication of this equivalence is that the conditioning statistic introduced in Kleibergen (2005) is already enough to allow the use of conditional QLR tests in non-homoscedastic IV. This fact does not appear to have been noted in the previous literature, nor does the performance of the conditional QLR test in non-homoscedastic models appear to have been explored. Thus, while the results of our paper are not needed to allow the use of conditional QLR tests in a non-homoscedastic IV setting, for completeness we compare the performance of the conditional QLR test to alternative procedures that

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	0%	68%
S	36%	26%	*	6%	6%	58%
pQLR	32%	22%	0%	*	0%	60%
QLR	32%	22%	0%	1%	*	60%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	0%	0%	68%
S	36%	26%	*	18%	18%	58%
pQLR	25%	13%	0%	*	1%	64%
QLR	25%	13%	0%	0%	*	64%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	8%	9%	68%
S	36%	25%	*	33%	33%	58%
pQLR	11%	1%	0%	*	1%	70%
QLR	11%	1%	0%	0%	*	70%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	5%	5%	73%
PE-2	13%	*	0%	18%	17%	68%
S	35%	25%	*	40%	40%	58%
pQLR	3%	0%	0%	*	1%	73%
QLR	3%	0%	0%	0%	*	73%

Table 1: Power comparisons for $N = 2$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	1%	68%
S	36%	26%	*	6%	7%	58%
pQLR	31%	20%	0%	*	1%	60%
QLR	34%	23%	0%	4%	*	60%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	0%	68%
S	36%	25%	*	17%	19%	58%
pQLR	24%	12%	0%	*	3%	64%
QLR	29%	18%	0%	9%	*	64%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	8%	10%	68%
S	36%	25%	*	33%	34%	58%
pQLR	9%	0%	0%	*	3%	71%
QLR	16%	5%	1%	12%	*	69%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	3%	4%	73%
PE-2	13%	*	0%	16%	16%	68%
S	36%	25%	*	38%	38%	58%
pQLR	1%	0%	0%	*	3%	74%
QLR	11%	4%	2%	14%	*	72%

Table 2: Power comparisons for $N = 3$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	0%	0%	68%
S	35%	24%	*	3%	3%	58%
pQLR	34%	23%	1%	*	2%	59%
QLR	38%	28%	4%	5%	*	58%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	0%	68%
S	35%	25%	*	14%	16%	58%
pQLR	27%	15%	1%	*	3%	63%
QLR	35%	24%	2%	12%	*	62%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	4%	5%	68%
S	36%	25%	*	29%	30%	58%
pQLR	13%	1%	0%	*	3%	69%
QLR	23%	11%	2%	14%	*	67%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	11%	11%	68%
S	36%	25%	*	35%	35%	58%
pQLR	5%	0%	0%	*	2%	72%
QLR	18%	7%	4%	16%	*	70%

Table 3: Power comparisons for $N = 4$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	0%	68%
S	36%	26%	*	4%	5%	58%
pQLR	34%	23%	0%	*	2%	59%
QLR	39%	28%	3%	7%	*	59%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	0%	0%	68%
S	36%	26%	*	16%	19%	58%
pQLR	25%	12%	0%	*	4%	64%
QLR	35%	24%	1%	14%	*	63%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	6%	8%	68%
S	36%	26%	*	31%	32%	58%
pQLR	9%	0%	0%	*	5%	70%
QLR	20%	8%	1%	13%	*	69%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	1%	2%	73%
PE-2	13%	*	0%	14%	14%	68%
S	36%	25%	*	37%	37%	58%
pQLR	2%	0%	0%	*	3%	73%
QLR	14%	6%	3%	15%	*	71%

Table 4: Power comparisons for $N = 5$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	1%	73%
PE-2	13%	*	0%	0%	1%	68%
S	36%	26%	*	5%	7%	58%
pQLR	32%	21%	0%	*	3%	60%
QLR	39%	28%	3%	7%	*	59%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	0%	1%	68%
S	36%	26%	*	16%	20%	58%
pQLR	24%	12%	0%	*	5%	64%
QLR	38%	27%	2%	16%	*	62%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	12%	*	0%	6%	9%	68%
S	36%	26%	*	32%	34%	58%
pQLR	9%	1%	0%	*	5%	70%
QLR	24%	12%	2%	17%	*	68%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	1%	73%
PE-2	12%	*	0%	12%	13%	68%
S	36%	25%	*	36%	36%	58%
pQLR	1%	1%	0%	*	2%	73%
QLR	15%	6%	3%	14%	*	71%

Table 5: Power comparisons for $N = 10$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

$\rho = 0.3$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	0%	68%
S	36%	26%	*	5%	8%	58%
pQLR	33%	22%	0%	*	3%	60%
QLR	40%	30%	5%	9%	*	59%

$\rho = 0.5$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	0%	73%
PE-2	13%	*	0%	0%	1%	68%
S	36%	26%	*	17%	21%	58%
pQLR	23%	11%	0%	*	5%	64%
QLR	39%	29%	4%	20%	*	62%

$\rho = 0.9$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	0%	1%	73%
PE-2	12%	*	0%	7%	9%	68%
S	36%	25%	*	32%	34%	58%
pQLR	8%	0%	0%	*	5%	70%
QLR	24%	12%	2%	18%	*	68%

$\rho = 0.99$						
	PE-1	PE-2	S	pQLR	QLR	Average
PE-1	*	0%	0%	1%	2%	73%
PE-2	13%	*	0%	14%	14%	68%
S	36%	26%	*	37%	37%	58%
pQLR	1%	0%	0%	*	2%	73%
QLR	14%	6%	3%	15%	*	72%

Table 6: Power comparisons for $N = 20$ alternatives in a stylized model, based on 10,000 simulation draws and rounded to the nearest percent.. The entry in row i, column j reports the largest amount by which the power of test i falls short relative to the power of test j over the grid of alternatives considered. The "average" column reports the average power of each test, using uniform weights over $(\mu_1, \lambda_{N,1}) \in \{-6, -5.7, ..., 5.7, 6\}^2$.

have been proposed for the linear IV model with non-homoscedastic errors. A further reason this comparison may be of interest is that, as noted by a referee, unlike in general nonlinear GMM models the QLR statistic is a true likelihood ratio statistic in linear IV with non-homoscedastic Gaussian errors.[4]

To asses the performance of the QLR test we compare it to the AR test and the K and GMM-M tests of Kleibergen (2005). We also include the power function of the infeasible "oracle" pQLR test which tests the null against the true alternative at each point. All the tests considered control size and, indeed, are conditionally similar given Kleibergen (2005)'s conditioning statistic D_T. To calculate the QLR statistic we minimize the GMM objective via grid search. To check that this grid search does not affect the performance of the QLR test, in unreported results we simulate the performance of an infeasible QLR test which tunes the grid used in optimization based on the unknown data generating process, and find results nearly identical to those reported here. Critical values for both the QLR and pQLR tests are obtained using 1,000 simulation draws.

We adopt the simulation design of I. Andrews (2015), which is calibrated to data used by Yogo (2004) to study the effect of weak instruments on estimation of the elasticity of intertemporal substitution in a linear Euler Equation model. Yogo uses data from eleven different countries and considers multiple specifications, taking either the risk free interest rate or an equity return to be the endogenous regressor. As noted by Moreira and Moreira (2015) the instruments are more strongly correlated with the risk free rate than with the equity return, and there is a correspondingly more severe weak instruments problem when we use equity returns as the endogenous regressor. Since we are interested in behavior for a range of different identification scenarios we follow I. Andrews (2015) and calibrate our simulations to specifications which use the risk free rate as the endogenous regressor.

Figures 1-3 plot simulated power, based on 5,000 simulation draws, for all the tests we consider. Notably, the QLR test is more powerful than the GMM-M test against many (though not all) alternatives, and avoids the declines in power which affect the GMM-M test in some specifications (for example the calibrations to data from Japan and the UK). While there are some simulation designs, for example the calibration to German data,

[4]Kleibergen (2007) shows that the Anderson-Rubin statistic in non-homoscedastic instrumental variables models with Gaussian errors and known variance can be interpreted as a concentrated likelihood, from which the interpretation of the QLR statistic as a likelihood ratio statistic follows immediately.

where the QLR test has less power against some alternatives than do the other feasible tests considered and also displays a small level of bias, there are other cases, for example the calibration to US data, where the power of the QLR test exceeds that of the other feasible tests considered.

The power of the oracle pQLR test typically exceeds that of the other tests considered. Even in well-identified models we would expect this to be the case, since under strong instrument asymptotics one can show that the pQLR test will be (locally asymptotically) equivalent to a one-sided t-test, while the other tests considered will be locally asymptotically equivalent to a two-sided t-test. While the power of the pQLR test need not exceed that of the QLR test in general, we find that its power is indeed higher in these simulations.

I. Andrews (2015) and Moreira and Moreira (2015) have recently used the additional structure imposed by the linear IV model to propose tests for the non-homoscedastic IV setting motivated by optimality considerations. Results comparing the performance of these tests to the QLR test are reported in a note on I. Andrews's website, and will be incorporated into the next revision of I. Andrews (2015). There, one can see that the QLR test is competitive with these tests, in the sense that its power neither uniformly dominates, nor is uniformly dominated by, these alternative procedures.

S6 Proofs of results stated in the paper

Proof of Lemma 2 of the paper. Consider a $k \times 1$ vector ξ, a k-dimensional function $h(\cdot)$ of the q-dimensional argument θ satisfying $h(\theta_0) = 0$, and a covariance function $\Sigma(\cdot,\cdot)$ satisfying Assumption 2 of the paper. Let $\Sigma_\theta = \Sigma(\theta,\theta)$, $\Sigma_0 = \Sigma(\theta_0,\theta_0)$, $V(\theta) = \Sigma(\theta,\theta_0)\Sigma(\theta_0,\theta_0)^{-1}$. Recall the definition of the QLR statistic:

$$R(\xi, h(\cdot), \Sigma(\cdot,\cdot)) = \xi'\Sigma_0^{-1}\xi - \inf_\theta \left(h(\theta) + V(\theta)\xi\right)' \Sigma_\theta^{-1} \left(h(\theta) + V(\theta)\xi\right). \tag{7}$$

We now restrict our attention to those values of ξ and Σ for which $\xi'\Sigma_0^{-1}\xi \leq C$ for a fixed constant $C > 0$.

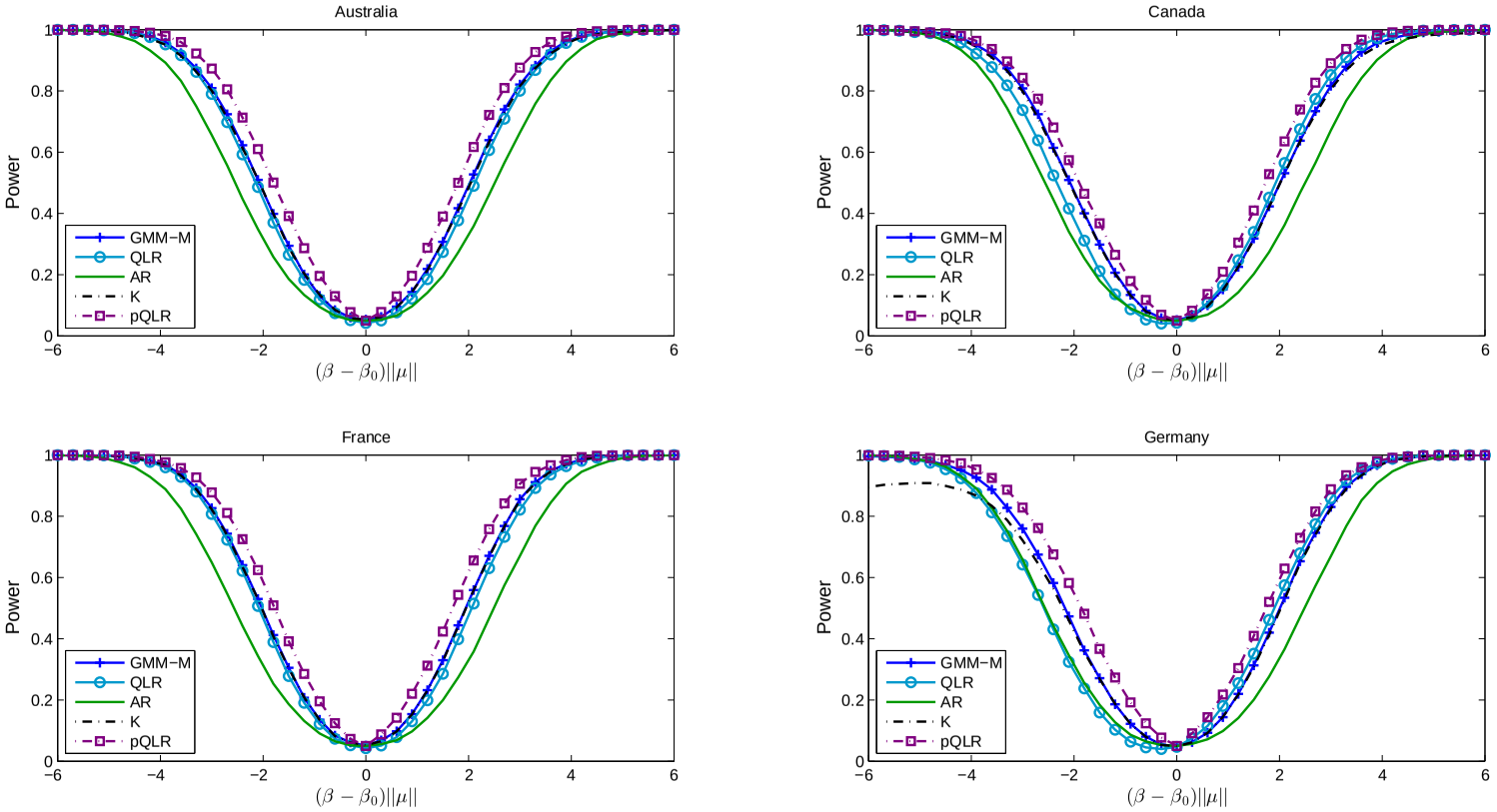

Figure 1: Power functions for the GMM-M, QLR, AR (or S), K, and pQLR tests in simulation calibrated to Yogo (2004) data.

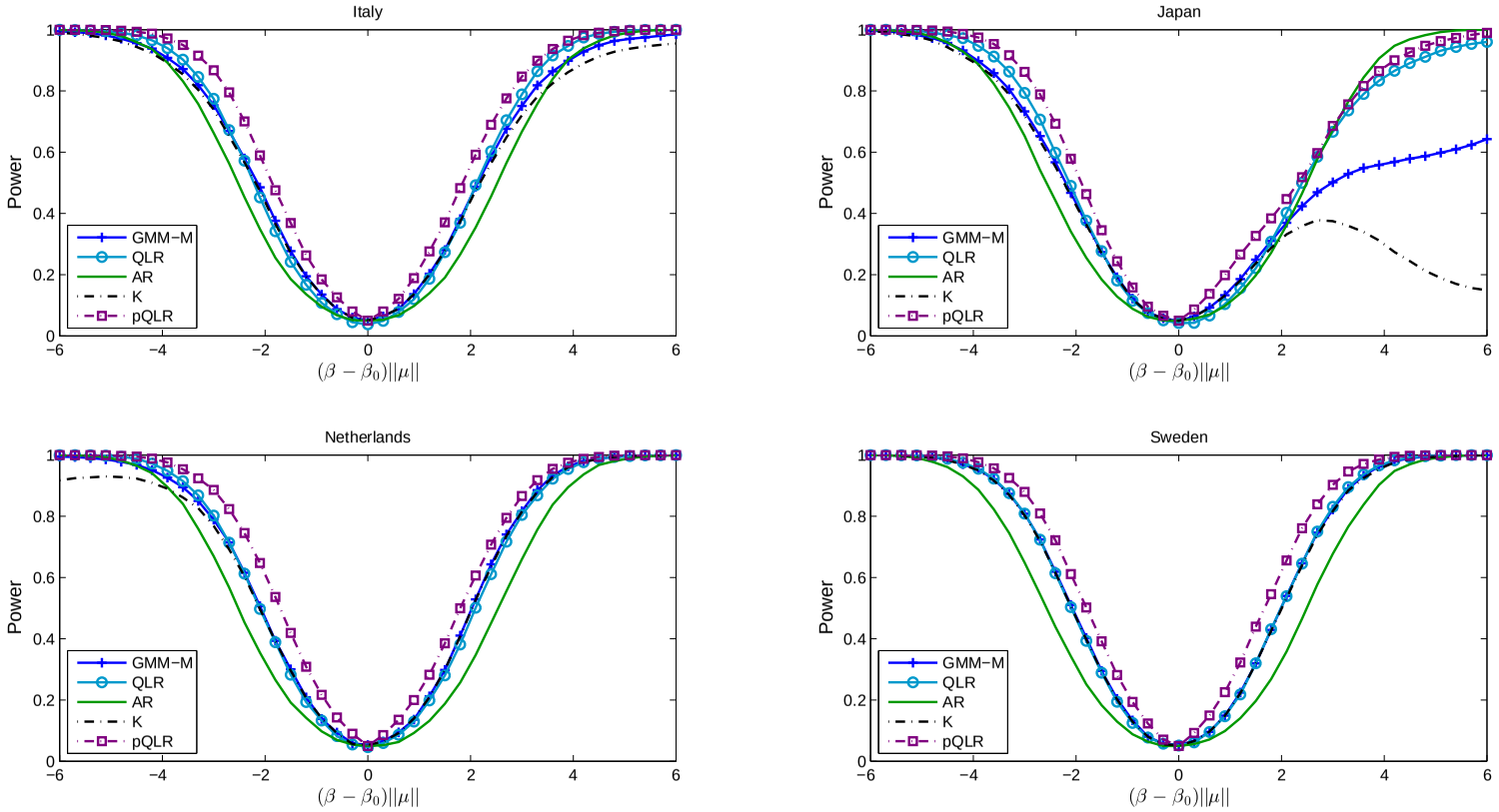

Figure 2: Power functions for the GMM-M, QLR, AR (or S), K, and pQLR tests in simulation calibrated to Yogo (2004) data.

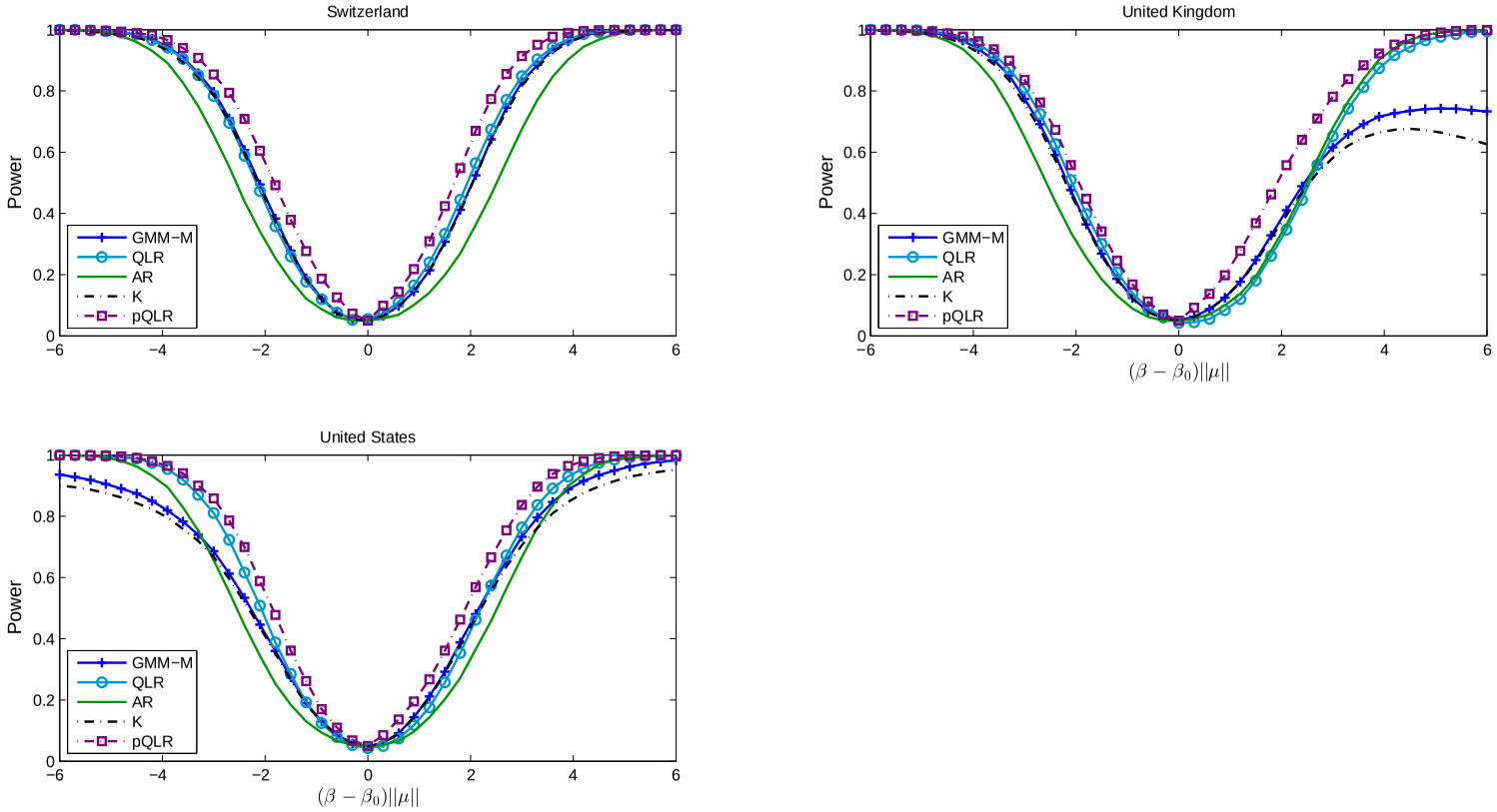

Figure 3: Power functions for the GMM-M, QLR, AR (or S), K, and pQLR tests in simulation calibrated to Yogo (2004) data.

If the optimum in equation (7) is attained[5] at point θ^*, then

$$\frac{1}{\bar{\lambda}} \left\| h\left(\theta^*\right) + V(\theta^*)\xi \right\|^2 \leq \left(h(\theta^*) + V(\theta^*)\xi\right)' \Sigma_{\theta^*}^{-1} \left(h(\theta^*) + V(\theta^*)\xi\right) \leq \xi' \Sigma_0^{-1} \xi \leq C.$$

Consider some ξ, Σ and two functions h_1 and h_2 that satisfy our assumptions. For function h_i let the optimum in equation (7) be achieved at θ_i. Then

$$|R(\xi, h_1, \Sigma) - R(\xi, h_2, \Sigma)|$$

$$\leq \max_{\theta \in \{\theta_1, \theta_2\}} \left| (h_1(\theta) + V(\theta)\xi)' \Sigma_\theta^{-1} (h_1(\theta) + V(\theta)\xi) - (h_2(\theta) + V(\theta)\xi)' \Sigma_\theta^{-1} (h_2(\theta) + V(\theta)\xi) \right|$$

$$= \max_{\theta \in \{\theta_1, \theta_2\}} 2 \left| \left(\frac{h_1(\theta) + h_2(\theta)}{2} + V(\theta)\xi \right)' \Sigma_\theta^{-1} (h_2(\theta) - h_1(\theta)) \right|$$

$$\leq \max_{\theta \in \{\theta_1, \theta_2\}} 2 \left\| \frac{h_1(\theta) + h_2(\theta)}{2} + V(\theta)\xi \right\| \bar{\lambda} d(h_1, h_2).$$

Here we used that $\|\Sigma_\theta^{-1}\| \leq \bar{\lambda}$.

For $\theta = \theta_1$ we have

$$\left\| \frac{h_1(\theta) + h_2(\theta)}{2} + V(\theta)\xi \right\| \leq \|h_1(\theta) + V(\theta)\xi\| + \left\| \frac{h_1(\theta) - h_2(\theta)}{2} \right\| \leq \sqrt{\bar{\lambda} C} + d(h_1, h_2).$$

A similar statement holds for $\theta = \theta_2$. Thus

$$|R(\xi, h_1, \Sigma) - R(\xi, h_2, \Sigma)| \leq 2\bar{\lambda} d(h_1, h_2) \left(\sqrt{\bar{\lambda} C} + d(h_1, h_2) \right). \tag{8}$$

Equation (8) implies that $R(\xi, h, \Sigma)$ is Lipschitz in h. Indeed, for all h_1 and h_2 such that $d(h_1, h_2) < \bar{\lambda}^{1/2}$ we have

$$|R(\xi, h_1, \Sigma) - R(\xi, h_2, \Sigma)| \leq 2\bar{\lambda}^{3/2} (\sqrt{C} + 1) d(h_1, h_2).$$

For all ξ such that $\xi' \Sigma_0^{-1} \xi \leq C$ we have $0 \leq R(\xi, h_1, \Sigma) \leq C$, thus

$$|R(\xi, h_1, \Sigma) - R(\xi, h_2, \Sigma)| \leq 2C \leq 2C \bar{\lambda}^{-1/2} d(h_1, h_2)$$

[5] In cases where the optimum cannot be attained we consider θ^* such that the function at this point is within $\delta > 0$ of the infimum. We can choose δ based on the bound we wish to obtain. In such cases all inequalities must be corrected by an additional δ term.

for all $d(h_1, h_2) \geq \bar{\lambda}^{1/2}$. Thus, $R(\xi, h, \Sigma)$ is Lipschitz for all h.

We can similarly show that R is Lipschitz in ξ for $\xi'\Sigma_0^{-1}\xi < C$. Fix h, Σ and consider ξ_1 and ξ_2. Again let the corresponding optima be achieved at θ_i. Then

$$\begin{aligned}
&|R(\xi_1, h, \Sigma) - R(\xi_2, h, \Sigma)| \\
&\leq \left|(\xi_1 + \xi_2)'\Sigma_0^{-1}(\xi_1 - \xi_2)\right| + \max_{\theta \in \{\theta_1, \theta_2\}} 2 \left| \left(h(\theta) + V(\theta)\frac{\xi_1 + \xi_2}{2} \right)' \Sigma_\theta^{-1} V(\theta)(\xi_1 - \xi_2) \right| \\
&\leq 2\sqrt{\bar{\lambda} C}\|\xi_1 - \xi_2\| + \max_{\theta \in \{\theta_1, \theta_2\}} 2 \left\| h(\theta) + V(\theta)\frac{\xi_1 + \xi_2}{2} \right\| \bar{\lambda}\|\xi_1 - \xi_2\| \\
&\leq 2\sqrt{\bar{\lambda} C}\|\xi_1 - \xi_2\| + 2\bar{\lambda}(\sqrt{\bar{\lambda} C} + \bar{\lambda}^{3/2}\sqrt{C})\|\xi_1 - \xi_2\|,
\end{aligned}$$

where between the second and third lines we used the fact that

$$\|\Sigma_\theta^{-1} V_\theta\| = \|\Sigma_\theta^{-1}\Sigma(\theta, \theta_0)\Sigma_0^{-1}\| \leq \|\Sigma_\theta^{-1/2}\Sigma_0^{-1/2}\| \leq \bar{\lambda}.$$

Finally, let us prove that R is Lipschitz with respect to Σ. Fix ξ, h and consider two covariance functions Σ_1 and Σ_2. Again let the corresponding optima be achieved at θ_1 and θ_2. Then

$$\begin{aligned}
&|R(\xi, h, \Sigma_1) - R(\xi, h, \Sigma_2)| \leq \left|\xi'\Sigma_{1,0}^{-1}(\Sigma_{1,0} - \Sigma_{2,0})\Sigma_{2,0}^{-1}\xi\right| \\
&+ \max_{\theta \in \{\theta_1, \theta_2\}} \left|(h(\theta) + V_1(\theta)\xi)' \Sigma_{1,\theta}^{-1} (h(\theta) + V_1(\theta)\xi) - (h(\theta) + V_2(\theta)\xi) \Sigma_{2,\theta}^{-1} (h(\theta) + V_2(\theta)\xi)\right|.
\end{aligned}$$

The first term on the right-hand side is bounded by $\bar{\lambda}^3 C d(\Sigma_1, \Sigma_2)$. Consider now the second term on the right-hand side for $\theta = \theta_1$. It is no greater than

$$\begin{aligned}
&\left|(h(\theta) + V_1(\theta)\xi)' \Sigma_{1,\theta}^{-1}(\Sigma_{1,\theta} - \Sigma_{2,\theta})\Sigma_{2,\theta}^{-1} (h(\theta) + V_1(\theta)\xi)\right| \\
&+2 \left| \left(h(\theta) + \frac{V_1(\theta) + V_2(\theta)}{2}\xi \right)' \Sigma_{2,\theta}^{-1}(V_1(\theta) - V_2(\theta))\xi \right| \\
&\leq \bar{\lambda}^3 C\|\Sigma_{1,\theta} - \Sigma_{2,\theta}\| + 2\bar{\lambda}\,\|h(\theta) + V_1(\theta)\xi\|\,\|V_1(\theta) - V_2(\theta)\|\|\xi\| \\
&+2 \left|\xi'(V_1(\theta) - V_2(\theta))'\Sigma_{2,\theta}^{-1}(V_1(\theta) - V_2(\theta))\xi\right| \\
&\leq \bar{\lambda}^3 C\|\Sigma_{1,\theta} - \Sigma_{2,\theta}\| + 2\bar{\lambda}^2 C\|V_1(\theta) - V_2(\theta)\| + 2\bar{\lambda}^2 C\|V_1(\theta) - V_2(\theta)\|^2.
\end{aligned}$$

A similar argument applies for $\theta = \theta_2$. Now note that

$$\begin{aligned} \|V_1(\theta) - V_2(\theta)\| &= \left\|\Sigma_1(\theta,\theta_0)\Sigma_{1,0}^{-1} - \Sigma_2(\theta,\theta_0)\Sigma_{2,0}^{-1}\right\| \\ &\leq \|\Sigma_1(\theta,\theta_0) - \Sigma_2(\theta,\theta_0)\|\|\Sigma_{1,0}^{-1}\| + \|\Sigma_2(\theta,\theta_0)\|\|\Sigma_{1,0} - \Sigma_{2,0}\| \\ &\leq 2\bar{\lambda} d(\Sigma_1,\Sigma_2) \end{aligned}$$

and $\|\Sigma_{1,\theta} - \Sigma_{2,\theta}\| \leq d(\Sigma_1,\Sigma_2)$. By arguments like those used to establish the Lipschitz property in h above, this implies that R is Lipschitz in Σ. □

Proof of Theorem 3 of the paper. We first verify Assumption 1.

$$\begin{aligned} g_T(\theta) - m_T(\theta) &= g_T^{(L)}(\widehat{\beta}(\theta),\theta) - m_T^{(L)}(\beta(\theta),\theta) \\ &= G_T^{(L)}(\beta(\theta),\theta) + \left(m_T^{(L)}(\widehat{\beta}(\theta),\theta) - m_T^{(L)}(\beta(\theta),\theta)\right) + \left(G_T^{(L)}(\widehat{\beta}(\theta),\theta) - G_T^{(L)}(\beta(\theta),\theta)\right) \\ &= G_T^{(L)}(\beta(\theta),\theta) + M_T(\theta)\sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) + \left(G_T^{(L)}(\widehat{\beta}(\theta),\theta) - G_T^{(L)}(\beta(\theta),\theta)\right) + r_T(\theta), \end{aligned}$$

where $G_T^{(L)}(\beta,\theta) = g_T^{(L)}(\beta,\theta) - m_T^{(L)}(\beta,\theta)$ and

$$r_T(\theta) = m_T^{(L)}(\widehat{\beta}(\theta),\theta) - m_T^{(L)}(\beta(\theta),\theta) - M_T(\theta)\sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)).$$

Take an arbitrarily small $\varepsilon > 0$. By Assumption 6 there exist constants $C > 0$ and $\delta > 0$ such that for all large T the event

$$A = \left\{ \sqrt{T}\sup_\theta \|\widehat{\beta}(\theta) - \beta(\theta)\| < C \text{ and } \sup_\theta \sup_{|\beta - \beta(\theta)| \leq \delta} \|G_T^{(L)}(\beta,\theta) - G_T^{(L)}(\beta(\theta),\theta)\| < \varepsilon \right\}$$

occurs with high probability, $P\{A\} > 1 - \varepsilon$. For all realizations of the event A for large enough T we therefore have that

$$\sup_\theta \left\| G_T^{(L)}(\widehat{\beta}(\theta),\theta) - G_T^{(L)}(\beta(\theta),\theta) \right\| < \varepsilon.$$

At the same time for all realizations in A, Assumption 9 implies that for large enough T (such that $\delta_T > C$) we have that $\sup_\theta \|r_T(\theta)\| < \varepsilon$.

Now let us take any functional $f \in BL_1$, which is defined on a set of k-dimensional

functions of θ. For realizations that belong to the event A we have

$$\begin{aligned}\left\| f\left(g_T(\theta) - m_T(\theta)\right) - f\left(G_T^{(L)}(\beta(\theta),\theta) + M_T(\theta)\sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta))\right)\right\| \\ \leq \sup_\theta \|r_T(\theta)\| + \sup_\theta \left\| G_T^{(L)}(\widehat{\beta}(\theta),\theta) - G_T^{(L)}(\beta(\theta),\theta)\right\| < 2\varepsilon. \qquad (9)\end{aligned}$$

But for any realization that does not belong to the event A, the left-hand side of equation (9) is bounded by 2. Thus

$$\begin{aligned}\left\| Ef\left(g_T(\theta) - m_T(\theta)\right) - Ef\left(G_T^{(L)}(\beta(\theta),\theta) + M_T(\theta)\sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta))\right)\right\| \\ < 2\varepsilon P(A) + 2P(\text{ not } A) < 4\varepsilon.\end{aligned}$$

Finally we notice that according to Assumption 6, process $G_T^{(L)}(\beta(\theta),\theta)+M_T(\theta)\sqrt{T}(\widehat{\beta}(\theta)-\beta(\theta))$ uniformly converges to a mean zero Gaussian process with the covariance function specified in Theorem 3 of the paper so long as $M_T(\theta)$ is uniformly bounded, which is assumed in Assumption 9. Thus Assumption 1 follows.

Next we check that Assumption 2 holds for the covariance function stated in Theorem 3 of the paper.

$$\begin{aligned}\lambda_{\min}(\Sigma(\theta,\theta)) &= \inf_{x\in\mathbb{R}^k} \frac{x'(I_k, M_T(\theta))\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)(I_k, M_T(\theta))'x}{x'x} \\ \geq \inf_{x\in\mathbb{R}^k} &\frac{x'(I_k, M_T(\theta))\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)(I_k, M_T(\theta))'x}{\|(I_k, M_T(\theta))'x\|^2} \inf_{x\in\mathbb{R}^k} \frac{\|(I_k, M_T(\theta))'x\|^2}{x'x} \\ &\geq \inf_{y\in\mathbb{R}^{k+p}} \frac{y'\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)y}{y'y} \inf_{x\in\mathbb{R}^k} \frac{x'(I_k + M_T(\theta)M_T(\theta)')x}{x'x} \geq 1/\bar{\lambda}\end{aligned}$$

Similarly

$$\begin{aligned}\lambda_{\max}(\Sigma(\theta,\theta)) &= \sup_{x\in\mathbb{R}^k} \frac{x'(I_k, M_T(\theta))\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)(I_k, M_T(\theta))'x}{x'x} \\ \leq \sup_{x\in\mathbb{R}^k} &\frac{x'(I_k, M_T(\theta))\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)(I_k, M_T(\theta))'x}{\|(I_k, M_T(\theta))'x\|^2} \sup_{x\in\mathbb{R}^k} \frac{\|(I_k, M_T(\theta))'x\|^2}{x'x} \\ &\leq \sup_{y\in\mathbb{R}^{k+p}} \frac{y'\Sigma_L(\beta(\theta),\theta,\beta(\theta),\theta)y}{y'y} \sup_{x\in\mathbb{R}^k} \frac{x'(I_k + M_T(\theta)M_T(\theta)')x}{x'x} \\ &\leq \bar{\lambda}(1 + \sup_\theta \|M_T(\theta)\|^2)\end{aligned}$$

where we use the Frobenius norm for the potentially non-square matrix M_T.

Finally, we check that the estimator of the covariance function provided in Theorem

3 is uniformly consistent.

$$
\begin{aligned}
&\sup_{\theta,\theta_1} \|\widehat{\Sigma}(\theta,\theta_1) - \Sigma(\theta,\theta_1)\| \\
&\leq \sup_{\theta,\theta_1} \left\| (I_k, M_T(\theta)) \left(\widehat{\Sigma}_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1) - \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1) \right) (I_k, M_T(\theta_1))' \right\| \\
&\quad + \sup_{\theta,\theta_1} \left\| (0, \widehat{M}_T(\theta) - M_T(\theta)) \widehat{\Sigma}_L(\widehat{\beta}(\theta),\theta,\widehat{\beta}(\theta_1),\theta_1) (I_k, M_T(\theta_1))' \right\| \\
&\quad + \sup_{\theta,\theta_1} \left\| (I_k, \widehat{M}_T(\theta)) \widehat{\Sigma}_L(\widehat{\beta}(\theta),\theta,\widehat{\beta}(\theta_1),\theta_1) (0, \widehat{M}_T(\theta_1) - M_T(\theta_1))' \right\| \\
&\leq (1 + \sup_{\theta} \|M_T(\theta)\|)^2 \sup_{\theta,\theta_1} \|\widehat{\Sigma}_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1) - \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1)\| \\
&+ 2 \sup_{\beta,\theta} \|\widehat{\Sigma}_L(\beta,\theta,\beta,\theta)\| (1 + \sup_{\theta} \|M_T(\theta)\| + \sup_{\theta} \|\widehat{M}_T(\theta) - M_T(\theta)\|) \sup_{\theta} \|\widehat{M}_T(\theta) - M_T(\theta)\|
\end{aligned}
$$

Assumption 9 implies that the last term uniformly converges to zero. We also notice that for $\sup_\theta \left\| \sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) \right\| \leq \delta_T$ we have

$$
\begin{aligned}
&\sup_{\theta,\theta_1} \|\widehat{\Sigma}_L(\widehat{\beta}(\theta),\theta,\widehat{\beta}(\theta_1),\theta_1) - \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1)\| \\
&\qquad \leq \sup_{\beta,\theta,\beta_1,\theta_1} \|\widehat{\Sigma}_L(\beta,\theta,\beta_1,\theta_1) - \Sigma_L(\beta,\theta,\beta_1,\theta_1)\| \\
&+ \sup_{\theta,\theta_1} \sup_{\|\beta-\beta(\theta)\| \leq \frac{\delta_T}{\sqrt{T}}} \sup_{\|\beta_1-\beta(\theta_1)\| \leq \frac{\delta_T}{\sqrt{T}}} \|\Sigma_L(\beta,\theta,\beta_1,\theta_1) - \Sigma_L(\beta(\theta),\theta,\beta(\theta_1),\theta_1)\|. \qquad (10)
\end{aligned}
$$

Since $\delta_T \to \infty$ while $\frac{\delta_T}{\sqrt{T}} \to 0$, Assumption 6 implies that $\sup_\theta \left\| \sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) \right\| \leq \delta_T$ holds with probability approaching one, while Assumption 7 implies that the last term in equation (10) converges uniformly to zero. Finally, the first term on the right hand side of equation (10) converges uniformly to zero due to Assumption 8. Putting everything together we conclude that the estimator $\widehat{\Sigma}$ satisfies Assumption 3. □

S7 Quantile IV regression

S7.1 Mean function for quantile IV

This section derives the mean function $m_T(\cdot)$ for Example 3 in the paper. Suppose we observe i.i.d. data consisting of an outcome variable Y_t, an almost-surely positive endogenous regressor D_t, and instruments Z_t. For U_t a zero-median shock independent of Z_t, suppose Y_t follows $Y_t = \gamma D_t + (D_t + 1)U_t$. These variables obey the Quantile IV

model of Chernozhukov and Hansen (2005) for all quantiles, and satisfy

$$E\left[\left(\mathbb{I}\{Y_t - \theta_0 D_t \leq 0\} - 1/2\right) Z_t\right] = 0$$

for $\theta_0 = \gamma$, so we can use this moment condition for inference. This moment restriction holds for arbitrary joint distributions of (D_t, Z_t, U_t) provided that U_t and Z_t are independent and U_t has median zero. However, different distributions produce different mean functions.

Consider a weakly identified example with $Z_t = \frac{1}{\sqrt{T}} F(Z_t^*) + (1 - \frac{1}{\sqrt{T}})\eta_t$ and $D_t = \exp\{Z_t^* - U_t\}$, where Z_t^*, U_t, η_t are mutually independent and $EF(Z_t^*) = E\eta_t = 0$. We use the $\frac{1}{\sqrt{T}}$ scaling to ensure only a weak relationship between the instruments Z_t and the endogenous regressor D_t. Since we consider a simplified model with no intercept, we impose $EZ_t = 0$ to avoid drawing identifying power from a misspecified intercept under $\theta \neq \theta_0$. Sections 5.1 and 6.1 of the paper consider the more general (and realistic) case, which treats the intercept as a nuisance parameter.

Note that

$$\begin{aligned} m(\theta_0 + \delta) &= \sqrt{T} E\left[Z_t \mathbb{I}\{-\delta D_t + (D_t + 1)U_t \leq 0\}\right] \\ &= E\left[F(Z_t^*)\mathbb{I}\{-\delta D_t + (D_t + 1)U_t \leq 0\}\right] + \sqrt{T}(1 - \frac{1}{\sqrt{T}}) E\left[\eta_t \mathbb{I}\{-\delta D_t + (D_t + 1)U_t \leq 0\}\right] \\ &= E\left[F(Z_t^*)\mathbb{I}\{-\delta D_t + (D_t + 1)U_t \leq 0\}\right] = E\left[F(Z_t^*)\mathbb{I}\left\{\frac{(D_t(Z_t^*, U_t) + 1)U_t}{D_t(Z_t^*, U_t)} \leq \delta\right\}\right]. \end{aligned}$$

Function $f(x) = (1 + be^x)x$ is monotonically increasing for each $b > 0$ and thus has an inverse, so for any y there is a solution to equation $(1 + be^x)x = y$. Denoting this solution by $x(y, b)$,

$$\mathbb{I}\left\{\frac{(D(Z^*, U) + 1)U}{D(Z^*, U)} \leq \delta\right\} = \mathbb{I}\{(1 + e^{-Z^*}e^U)U \leq \delta\} = \mathbb{I}\{U \leq x(\delta, e^{-Z^*})\}.$$

So we have

$$\begin{aligned} m(\theta_0 + \delta) &= E\left[F(Z^*)\mathbb{I}\left\{\frac{(D(Z^*, U) + 1)U}{D(Z^*, U)} \leq \delta\right\}\right] \\ &= E\left[F(Z^*)E\left(\mathbb{I}\left\{\frac{(D(Z^*, U) + 1)U}{D(Z^*, U)} \leq \delta\right\}\middle| Z^*\right)\right] = E\left[F(Z^*)F_U(x(\delta, e^{-Z^*}))\right], \end{aligned}$$

where $F_U(\cdot)$ is the cdf of U. Depending on F and the marginal distributions of U and Z^*

one can get a wide variety of mean functions in this setting, many of which are highly non-linear.

S7.2 Asymptotics for quantile IV

Lemma S7.1 *For the quantile IV model, Assumption 10 from the paper implies the validity of Assumptions 6-9 and thus implies that Theorem 3 holds for the concentrated moment function in this context.*

Proof of Lemma S7.1. Here as in the paper we use the notation $\varepsilon_t(\beta,\theta) = Y_t - D_t'\theta - C_t'\beta$ and $\varepsilon_t(\theta) = \varepsilon_t(\beta(\theta),\theta)$. Let us also introduce $\phi_\tau(u) = \tau - \mathbb{I}\{u<0\}$.

According to Proposition 1 of Chernozhukov and Hansen (2008), Assumption 10 guarantees that

$$\sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) = -J(\theta)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_\tau(\varepsilon_t(\theta))C_t + o_p(1),$$

where the $o_p(1)$ term is uniform in θ.

Consider the process

$$\begin{pmatrix} g_T^{(L)}(\beta,\theta) - Eg_T^{(L)}(\beta,\theta) \\ \sqrt{T}(\widehat{\beta}(\theta) - \beta(\theta)) \end{pmatrix} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix} \phi_\tau(\varepsilon_t(\beta,\theta))Z_t - E[\phi_\tau(\varepsilon_t(\beta,\theta))Z_t] \\ -J^{-1}(\theta)\phi_\tau(\varepsilon_t(\theta))C_t \end{pmatrix}.$$

Angrist, Chernozhukov and Fernandez-Val (2006) establish a functional central limit theorem for this process. In particular, they argue that the function class $\{\mathbb{I}\{Y - D'\theta - C'\beta\}\}$ is a VC subgraph class and thus is bounded Donsker. Consequently $\{\mathbb{I}\{Y - D'\theta - C'\beta\}(Z,C)\}$ is Donsker with square-integrable envelope function $2(\|Z\| + \|C\|)$. This implies equicontinuity of the above process which, together with the finite-dimensional Central Limit Theorem, establishes Assumption 6.

The first part of Assumption 7 bounds the eigenvalues of the matrix

$$\begin{pmatrix} E[\phi_\tau(\varepsilon(\theta))^2 ZZ'] - E[\phi_\tau(\varepsilon(\theta))Z]E[\phi_\tau(\varepsilon(\theta))Z]' & E[\phi_\tau(\varepsilon(\theta))^2 ZC'] \\ E[\phi_\tau(\varepsilon(\theta))^2 CZ'] & E[\phi_\tau(\varepsilon(\theta))^2 CC'] \end{pmatrix}.$$

Since $\phi_\tau(\varepsilon(\theta))$ is a binary variable taking values τ and $-(1-\tau)$, the required bounds trivially follow from Assumption 10 (i). The second part of Assumption 7, namely the

continuity of Σ_L in β along $\beta(\theta)$, comes from the fact that

$$E\left(|\phi_\tau(\varepsilon_t(\beta,\theta)) - \phi_\tau(\varepsilon_t(\theta))| \mid C, Z\right)$$
$$\leq E\left(\mathbb{I}\{|\varepsilon_t(\theta)| \leq |C'(\beta - \beta(\theta))|\} \mid C, Z\right) \leq const\|C\|\|\beta - \beta(\theta)\|,$$

where we used Assumption 10 (ii).

The validity of Assumption 8 – the uniform consistency of the estimator $\widehat{\Sigma}$ – follows from standard consistency arguments.

Finally, we examine Assumption 9. We have $m_T^{(L)}(\beta,\theta) = E\left[\phi_\tau(\varepsilon(\beta,\theta))Z'\right]$. Consider

$$M_T(\theta) = \frac{\partial m_T^{(L)}(\beta,\theta)}{\partial \beta} \mid_{\beta=\beta(\theta)} = E\left[f_{\varepsilon(\theta)}(0)CZ'\right],$$

which is bounded due to Assumption 10 (i) and (ii). Assumption 10 also implies that $m_T^{(L)}(\beta,\theta)$ has a uniformly bounded second derivative in β along $\beta(\theta)$. Thus, Taylor expansion implies the validity of the first part of Assumption 9 (linearizability of $m_T^{(L)}$). Uniform consistency of the estimator $\widehat{M_T}(\theta)$ follows by standard arguments for kernel estimators. □

S7.3 Additional Quantile IV Simulation Results

This section reports additional simulation results for the quantile IV simulation designs discussed in the paper. In particular, it gives simulated size for all tests considered and power under additional parameter values. We first report results for $k = 10$ instruments (as in the paper), and then report results for $k = 5$ instruments.

S7.3.1 Results for $k = 10$ Instruments

Simulated Size. Tables 7 and 8 report the simulated size of AR, K, JK, GMM-M, and QLR tests for a variety of parameter values in the symmetric and asymmetric simulation designs respectively. Note that we do not report size for the pQLR test since the simulations consider the infeasible pQLR test which tests θ_0 against the true value θ at each point. The pQLR statistic for testing θ_0 against θ_0 is identically zero. For ease of reading, however, the power plots below and in the paper set the rejection probability of the pQLR test to 5% at the null θ_0. Note, further, that there is an implicit restriction on possible values of ρ_S and π_S due to the fact that the covariance matrix

ρ_S	0.25				0.5				0.9	
π_S	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1
AR	5.9%	5.8%	5.1%	5.0%	5.6%	5.5%	5.6%	5.7%	5.4%	6.4%
K	6.7%	6.0%	6.1%	7.0%	5.4%	5.9%	6.0%	6.3%	9.3%	7.2%
JK	6.5%	5.6%	5.8%	6.1%	5.3%	5.7%	5.7%	6.0%	9.0%	7.0%
GMM-M	5.7%	6.0%	5.5%	6.2%	5.5%	5.2%	5.6%	5.9%	5.9%	6.4%
QLR	5.0%	5.8%	5.8%	5.4%	5.0%	5.5%	6.0%	5.4%	5.4%	5.7%

Table 7: Simulated size of nominal 5% tests in symmetric quantile IV simulation design with ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

ρ_A	0.25				0.5				0.9	
π_A	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1
AR	4.6%	4.7%	4.5%	5.0%	5.3%	3.9%	3.9%	3.8%	4.9%	4.7%
K	4.9%	6.5%	6.7%	6.6%	4.8%	5.7%	6.4%	6.7%	6.1%	5.3%
JK	4.5%	5.6%	6.5%	5.8%	4.4%	5.2%	5.4%	5.9%	5.2%	5.0%
GMM-M	4.6%	5.9%	6.5%	6.7%	4.9%	5.5%	5.1%	6.8%	4.6%	4.9%
QLR	5.5%	5.2%	5.5%	5.8%	6.6%	5.9%	5.2%	4.8%	4.4%	4.3%

Table 8: Simulated size of nominal 5% tests in asymmetric quantle IV simulation design with ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

of $(\xi_{U,t}, \xi_{D,t}, \xi_{Z_1,t}, ..., \xi_{Z_{10},t})$ must positive semi-definite, which precludes consideration of $\pi_S = 0.15$ and $\pi_S = 0.2$ when $\rho_S = 0.9$. Likewise, when $\rho_A = 0.9$ we cannot consider $\pi_A = 0.6$ and $\pi_A = 0.8$.

As these tables make clear, all tests considered have simulated size within 5% of their nominal size over the designs considered. The largest deviations of simulated size from nominal size arise for the K and JK tests in the symmetric design with $\rho_S = 0.9$. While these deviations are still not large, one might wonder to what extent power comparisons across tests would change if we took these distortions into account. To investigate this question we calculated size-corrected power curves, and found them qualitatively very similar to the raw results. These results are available upon request.

Power Simulations Figures 4-6 plot power curves for symmetric simulation designs as described in the paper, while Figures 7-9 do the same for asymmetric simulation designs. For completeness these plots repeat some of the results reported in the paper, while in each case also reporting results for designs with instruments stronger than the cases considered in the text. As expected given the local asymptotic efficiency of the K, GMM-M, and QLR tests in the well-identified case, when we increase the strength of the instruments the power curves for these tests tend to converge. From a theoretical

ρ_S	0.25				0.5				0.9	
π_S	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1
AR	4.9%	5.1%	5.0%	4.8%	4.8%	5.2%	5.1%	5.0%	5.5%	5.2%
K	4.6%	4.1%	4.2%	4.4%	5.1%	4.2%	4.1%	4.2%	6.2%	4.9%
JK	5.0%	4.6%	4.5%	4.7%	5.1%	4.4%	4.2%	4.3%	6.7%	5.4%
GMM-M	4.7%	4.7%	4.7%	4.5%	4.9%	4.7%	4.4%	4.3%	5.6%	4.9%
QLR	4.4%	4.1%	4.0%	3.6%	4.3%	4.0%	4.0%	4.0%	4.6%	4.8%

Table 9: Simulated size of nominal 5% tests in symmetric quantile IV simulation design with five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

perspective, note that the AR and JK tests are locally asymptotically inefficient in the well-identified case (though the degree of inefficiency for the JK test is small), while the pQLR test is locally asymptotically equivalent to a one-sided test in this case. Thus the pQLR power envelope considered here does not converge to the power functions of the other tests considered when we increase the strength of the instruments.

S7.3.2 Results for $k = 5$ Instruments

The results reported above and in the paper all focus on designs with $k = 10$ instruments. To examine the effect of changing the number of instruments, here we report results where we reduce the number of instruments to $k = 5$ while holding the other parameters constant. This change has a different effect in the symmetric and asymmetric simulation designs. In the symmetric designs the instruments are independent and equally informative about the endogenous regressor, and reducing the number of instruments leads to a decline in power for all tests considered. By contrast, in the asymmetric simulation design reducing the number of polynomials considered increases the power of many tests, suggesting that the sixth to tenth order polynomials included in the simulation design with $k = 10$ were not particularly informative.

Simulated Size Tables 9 and 10 report the simulated size of all tests considered under the symmetric and asymmetric simulation designs, respectively. We find that all tests have simulated size reasonably close to nominal size. In particular, unlike in the design with $k = 10$ the simulated size of the K and JK tests never exceeds 7%.

Power Simulations Figures 10-12 report power curves for the symmetric simulation designs with five instruments, while Figures 13-15 report results for the asymmetric

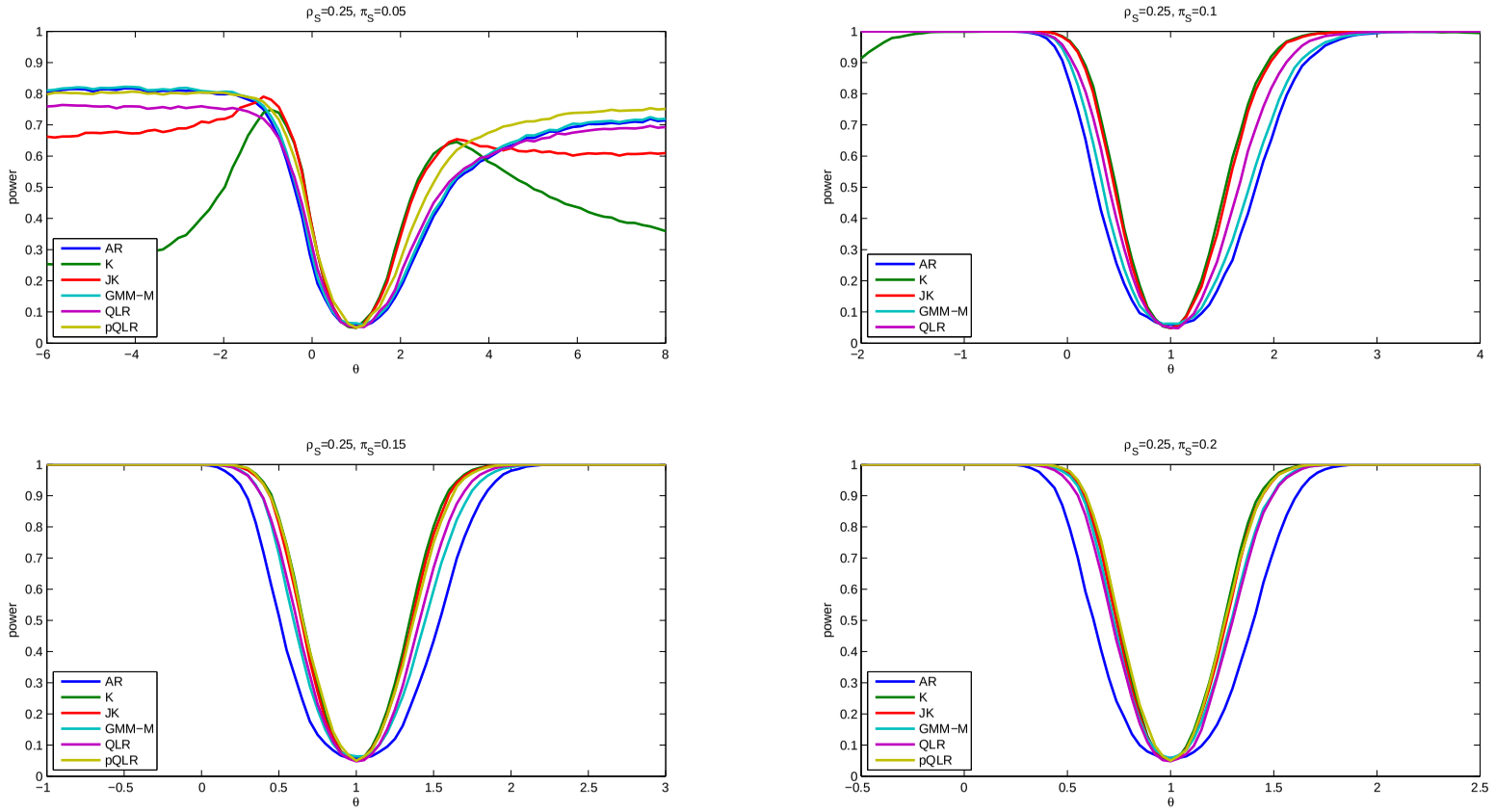

Figure 4: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.25$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

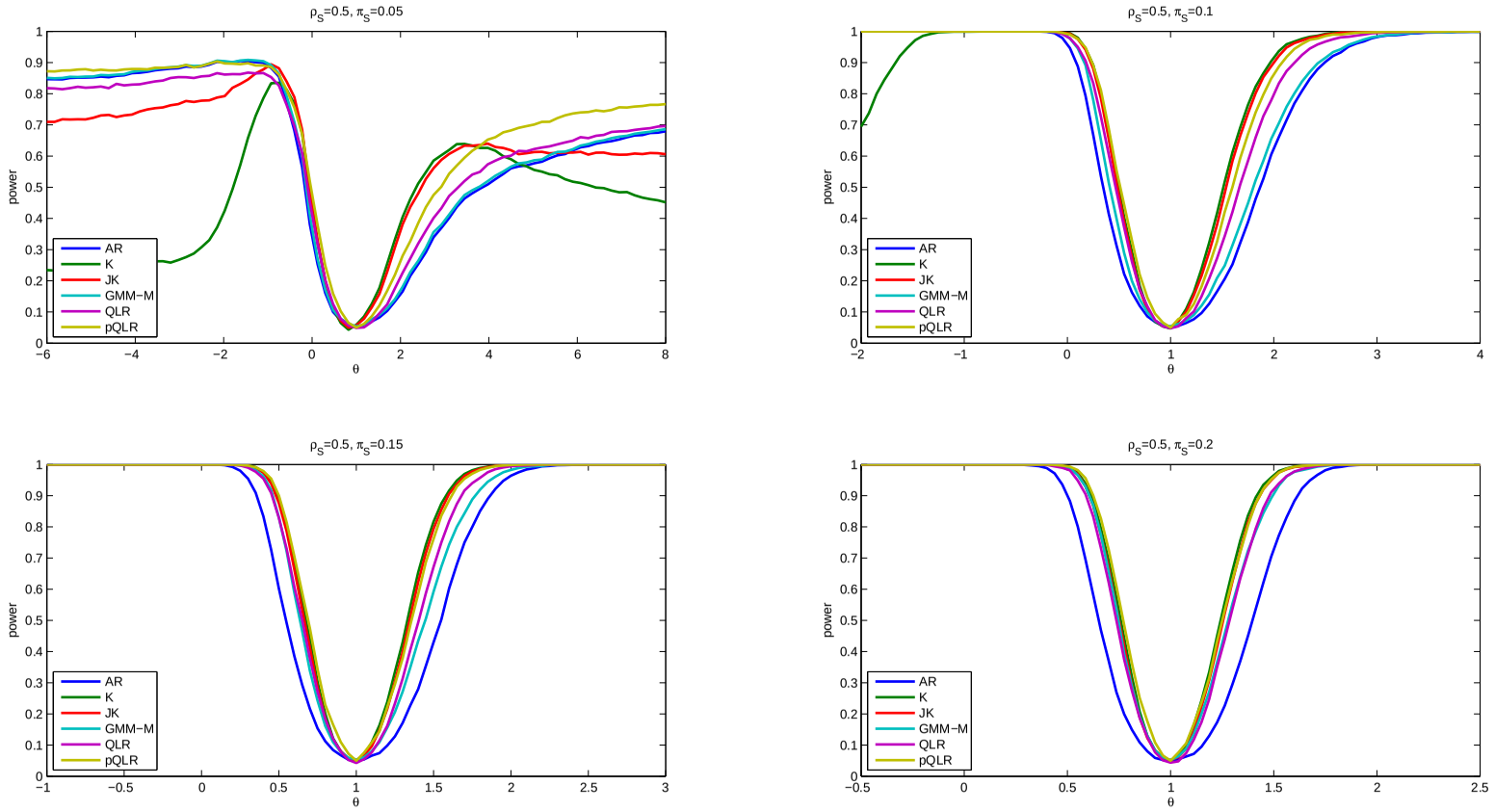

Figure 5: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.5$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

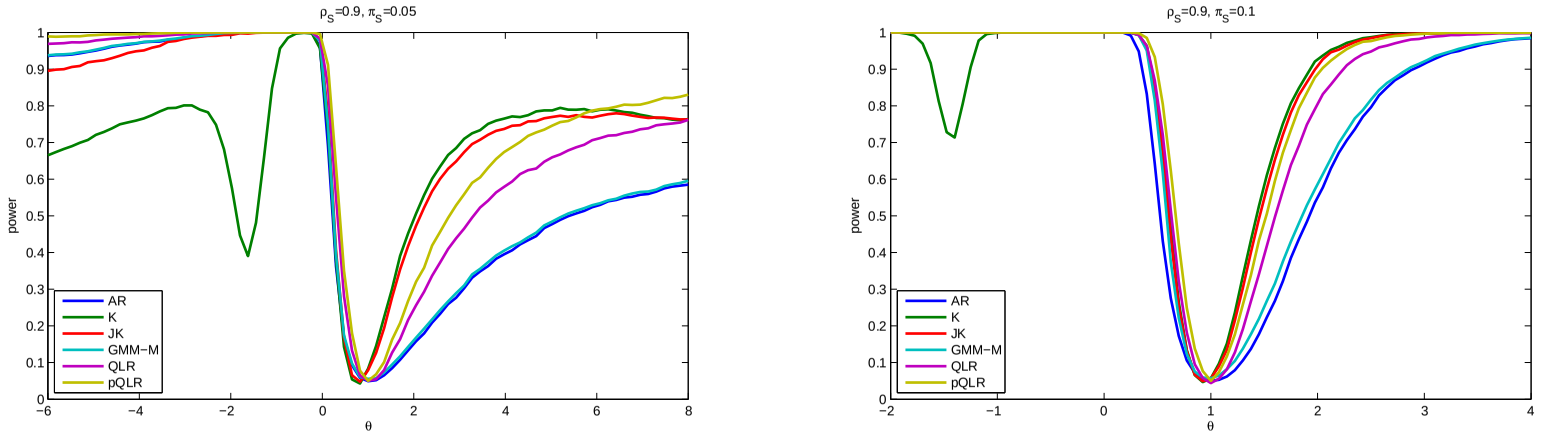

Figure 6: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.9$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

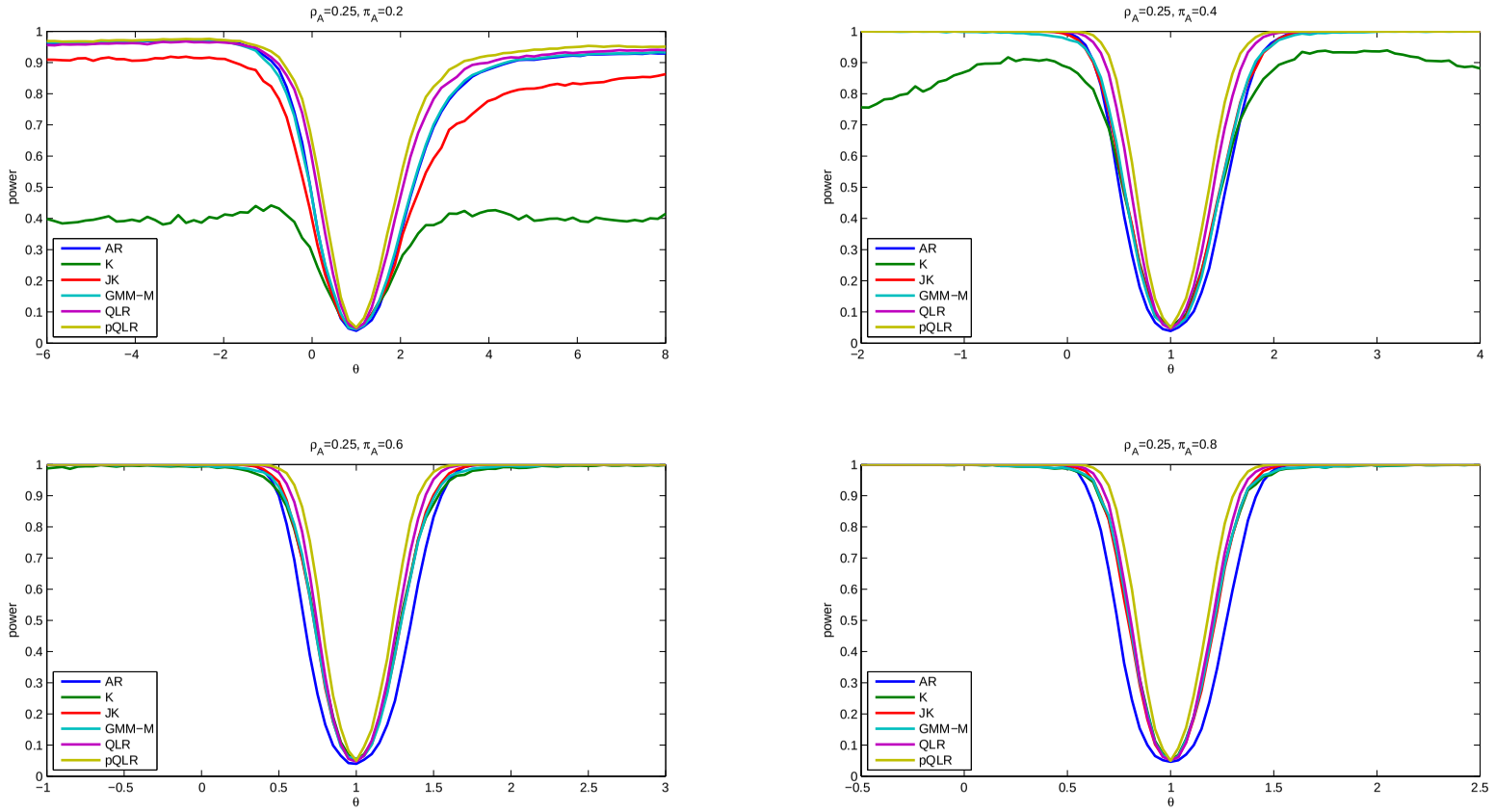

Figure 7: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.25$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

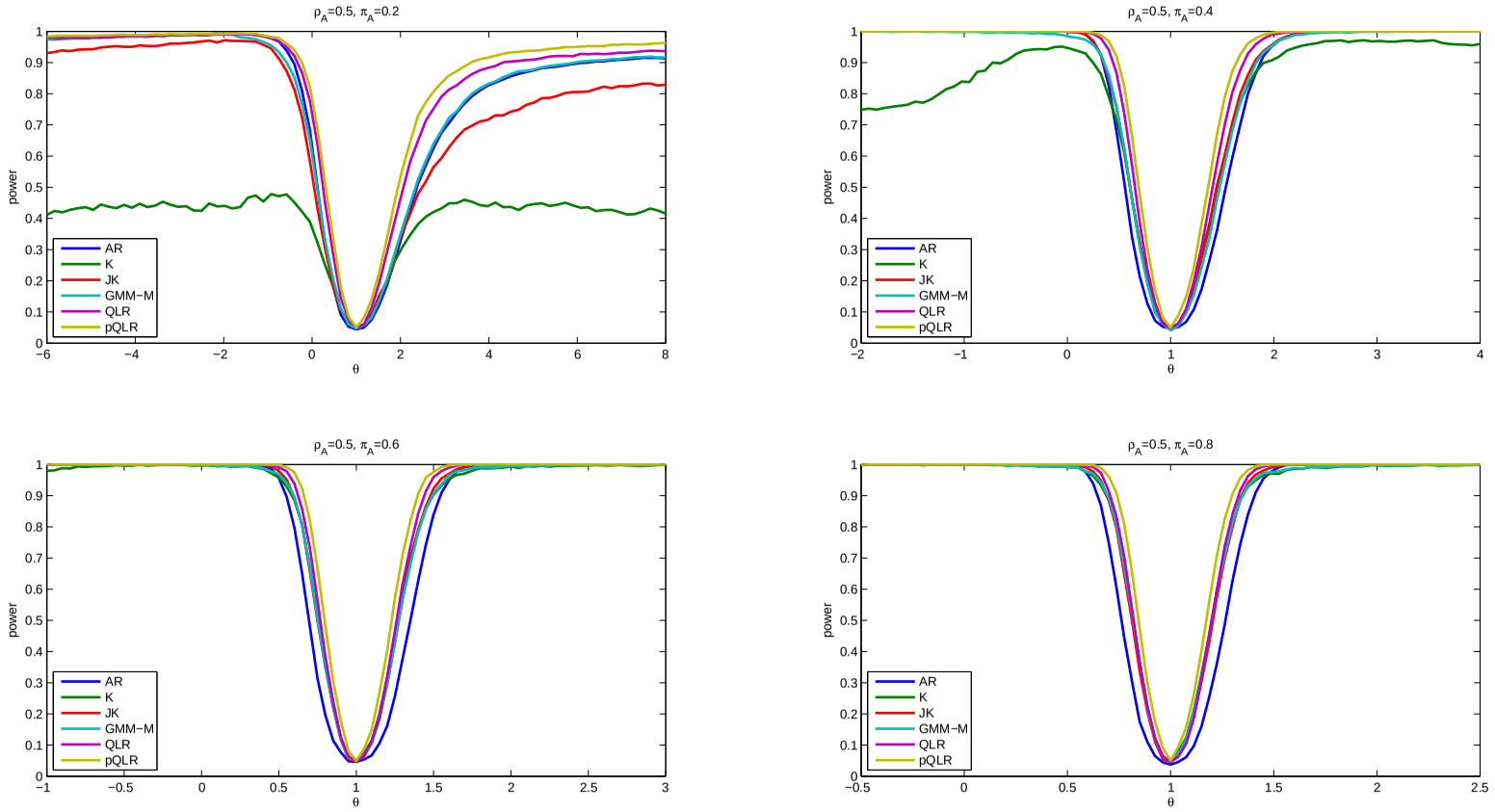

Figure 8: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.5$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

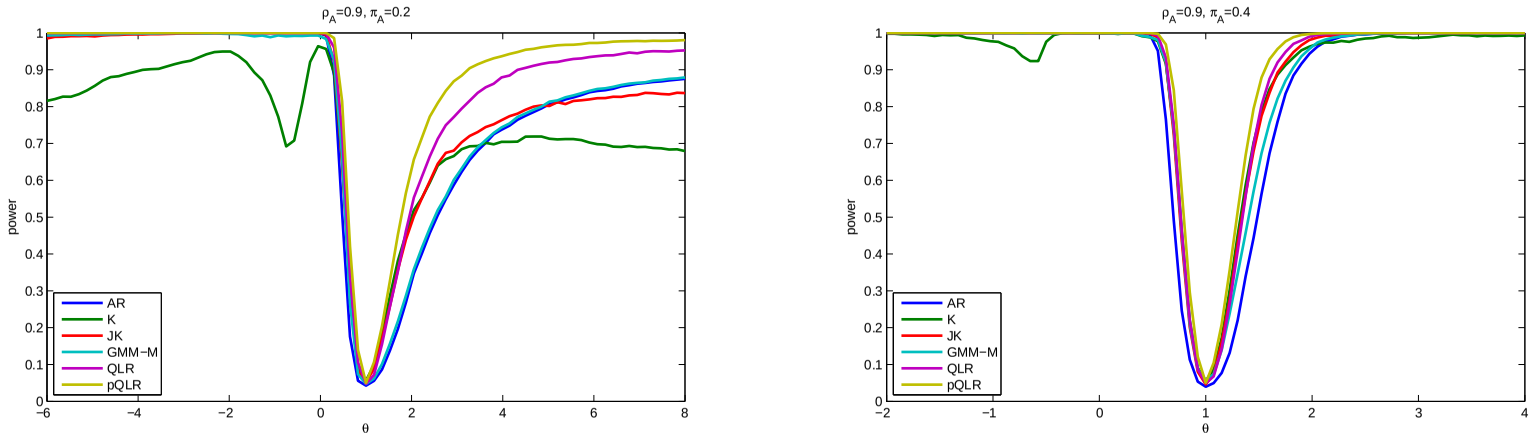

Figure 9: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.9$, ten instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

ρ_A	0.25				0.5				0.9	
π_A	0.05	0.1	0.15	0.2	0.05	0.1	0.15	0.2	0.05	0.1
AR	4.4%	4.4%	3.8%	4.0%	3.9%	3.9%	3.7%	4.1%	4.3%	3.8%
K	5.2%	4.9%	4.6%	4.4%	4.2%	4.8%	5.0%	5.5%	4.9%	5.5%
JK	4.7%	4.8%	4.6%	4.1%	4.2%	4.5%	5.2%	5.2%	4.5%	5.3%
GMM-M	3.7%	4.4%	4.2%	4.1%	4.1%	4.4%	5.0%	5.7%	4.7%	5.2%
QLR	4.1%	4.0%	3.6%	4.0%	4.4%	3.9%	3.8%	4.1%	4.0%	4.3%

Table 10: Simulated size of nominal 5% tests in asymmetric quantle IV simulation design with five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

simulation designs. Relative to designs with ten instruments, we see that the tests have substantially less power in the symmetric designs, but similar or higher power in the asymmetric simulation designs. As noted above, this stems from the fact that the number of instruments plays a different role in the symmetric and asymmetric designs, with each instrument bringing equal and independent information in the symmetric design but not in the asymmetric design. Qualitatively the results are quite similar to those in the ten instrument case: the AR test is inefficient in strongly identified case, while it performs reasonably well in weakly identified cases. The K and JK suffer power declines at distant alternatives in weakly identified cases, as well as substantial power losses in the asymmetric simulation design when the derivative of the moments is not a reliable guide to behavior under the alternative. The GMM-M test in general shows stable performance, though in most cases its power is exceeded by that of the conditional QLR test, which seems to be a desirable option among those considered.

S8 Stock and Wright setting

In this section, we demonstrate that the results of Section 5 of the paper can also be applied to the weak GMM models studied in Stock and Wright (2000) when the nuisance parameter is strongly identified and the parameter under test is weakly identified. Assume that we again begin with a moment function $g_T^{(SW)}$ whose mean function can be written as

$$Eg_T^{(SW)}(\beta,\theta) = m_T^{(SW)}(\beta,\theta) = \sqrt{T}m_1(\beta) + m_2(\beta,\theta),$$

for $(\beta,\theta) \in B \times \Theta$. We impose the following assumptions:

SW1 $g_T^{(SW)}(\beta,\theta) - m_T^{(SW)}(\beta,\theta) \Rightarrow G^{(SW)}(\beta,\theta)$ uniformly over $\mathcal{P}_0$ where $G^{(SW)}(\beta,\theta)$ is a

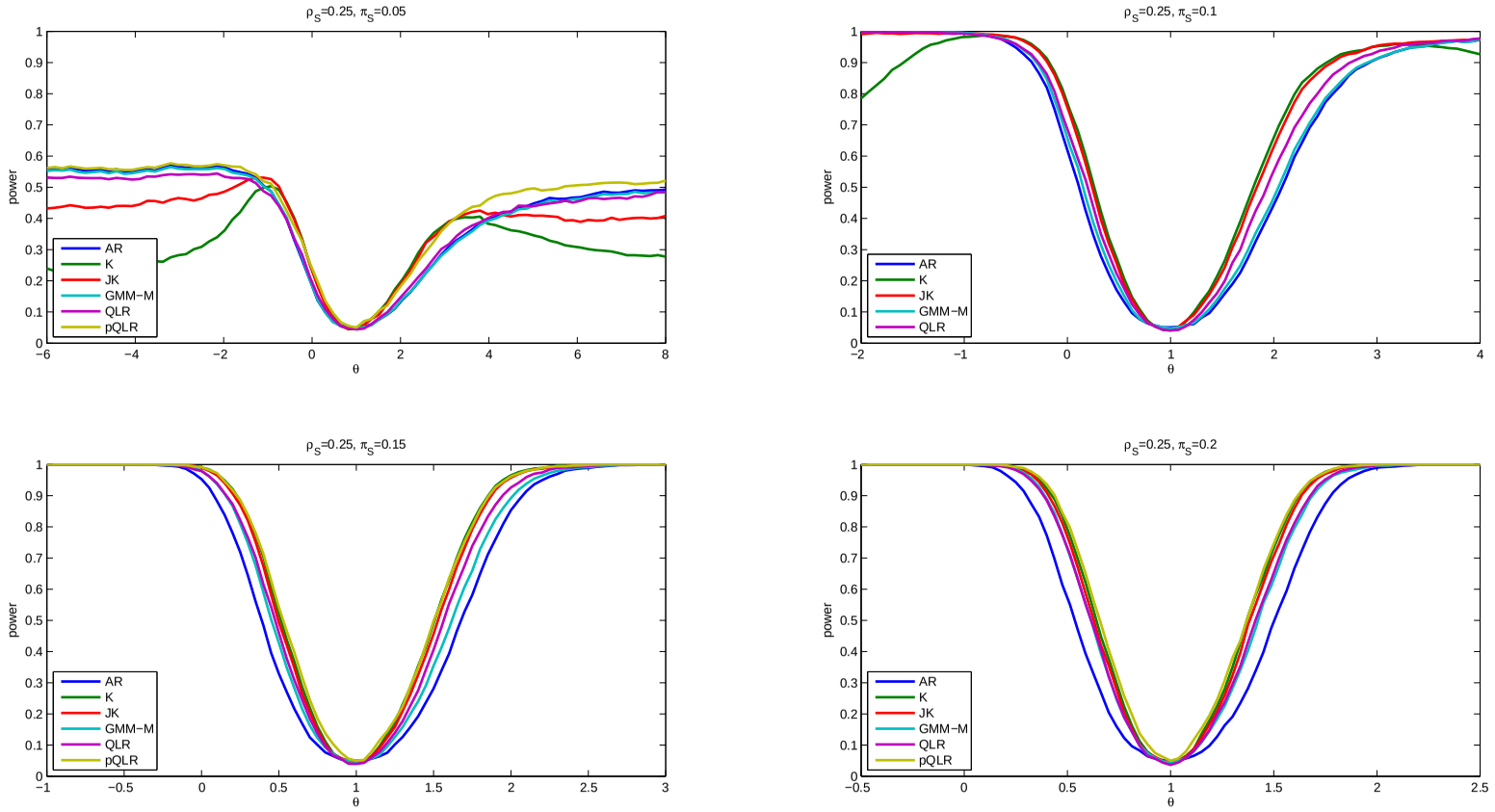

Figure 10: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.25$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

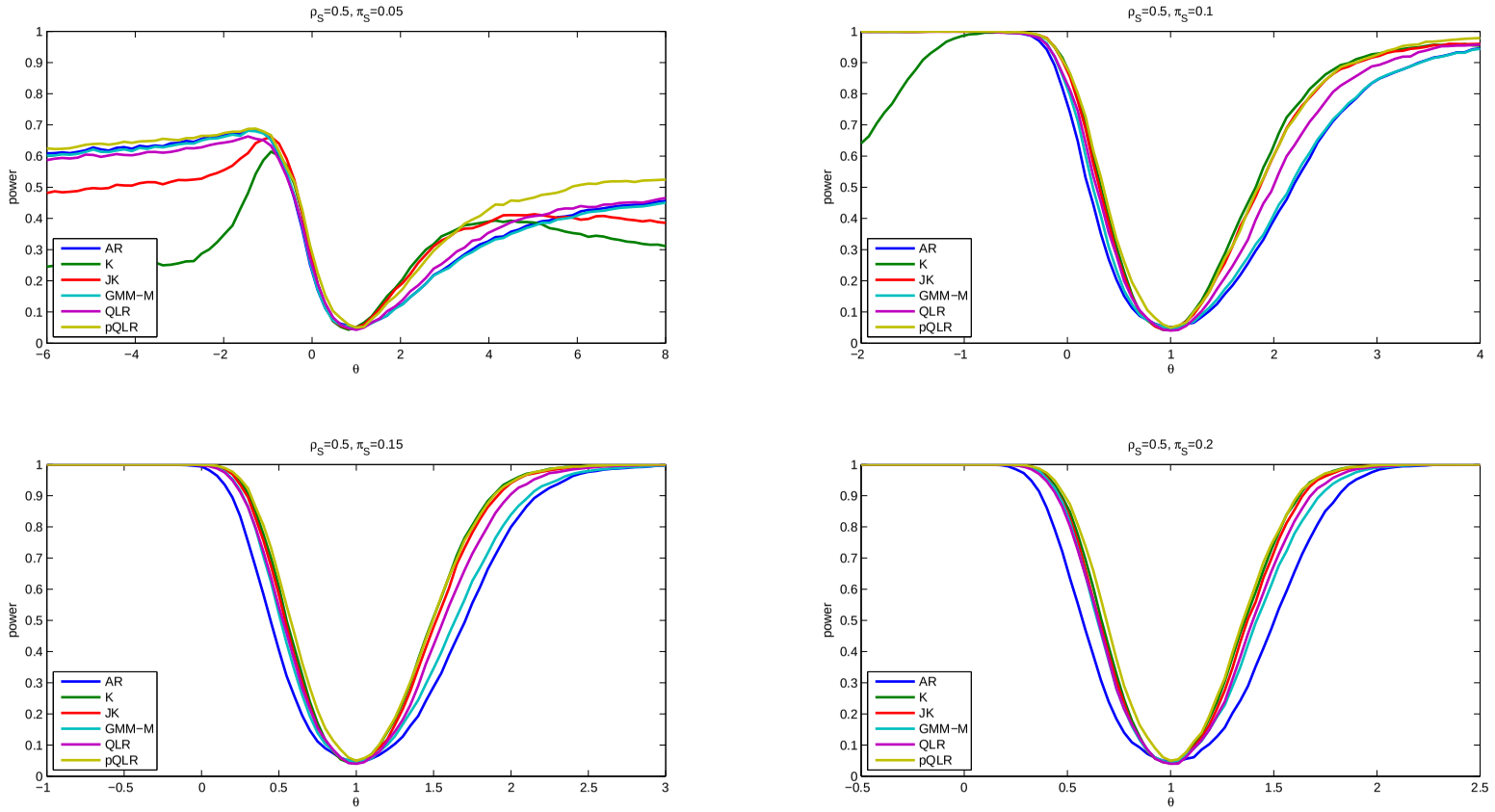

Figure 11: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.5$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

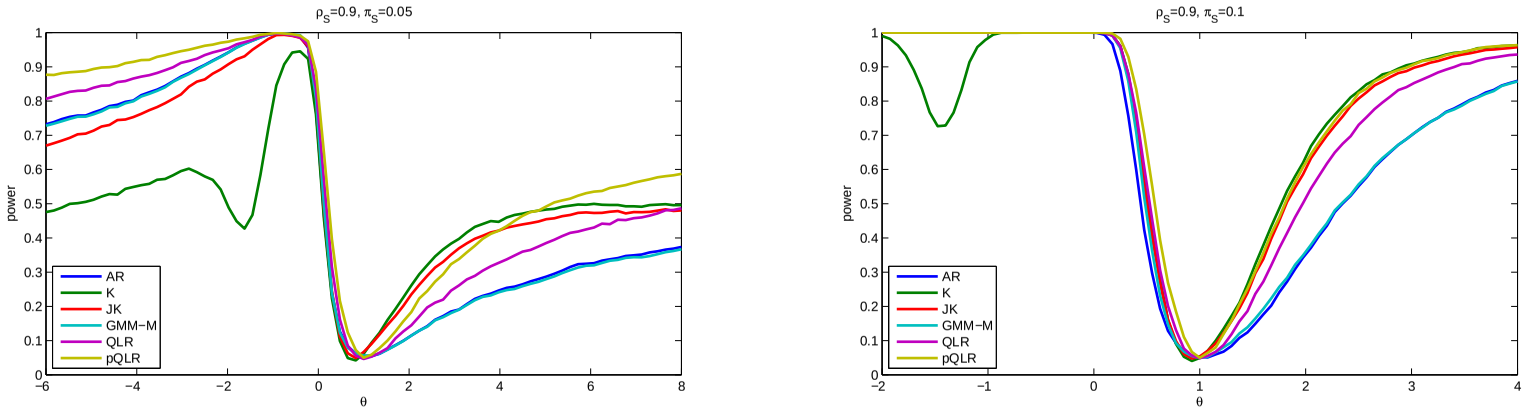

Figure 12: Simulated power of nominal 5% tests in symmetric quantile IV simulation design with $\rho_S = 0.9$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

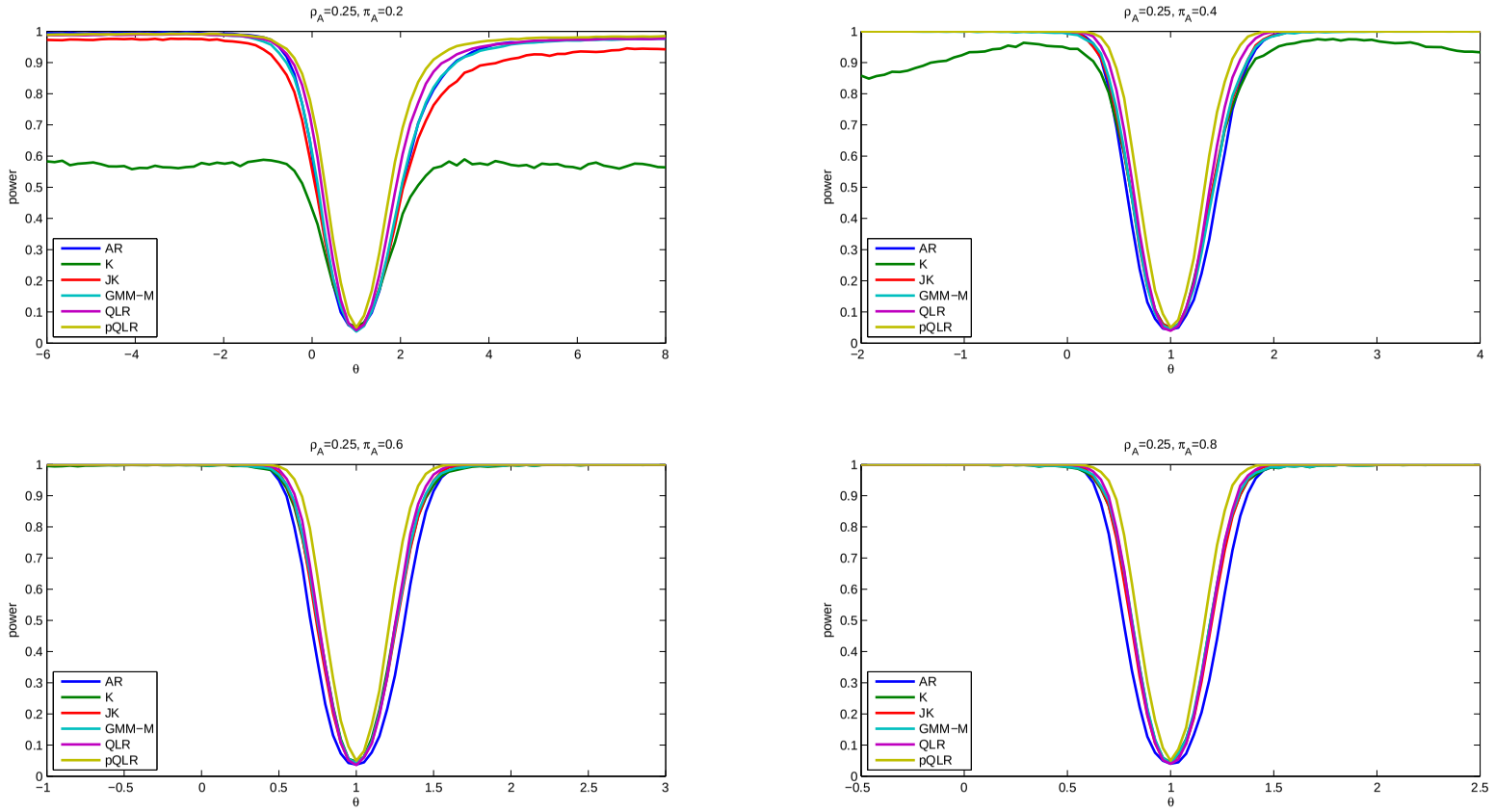

Figure 13: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.25$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

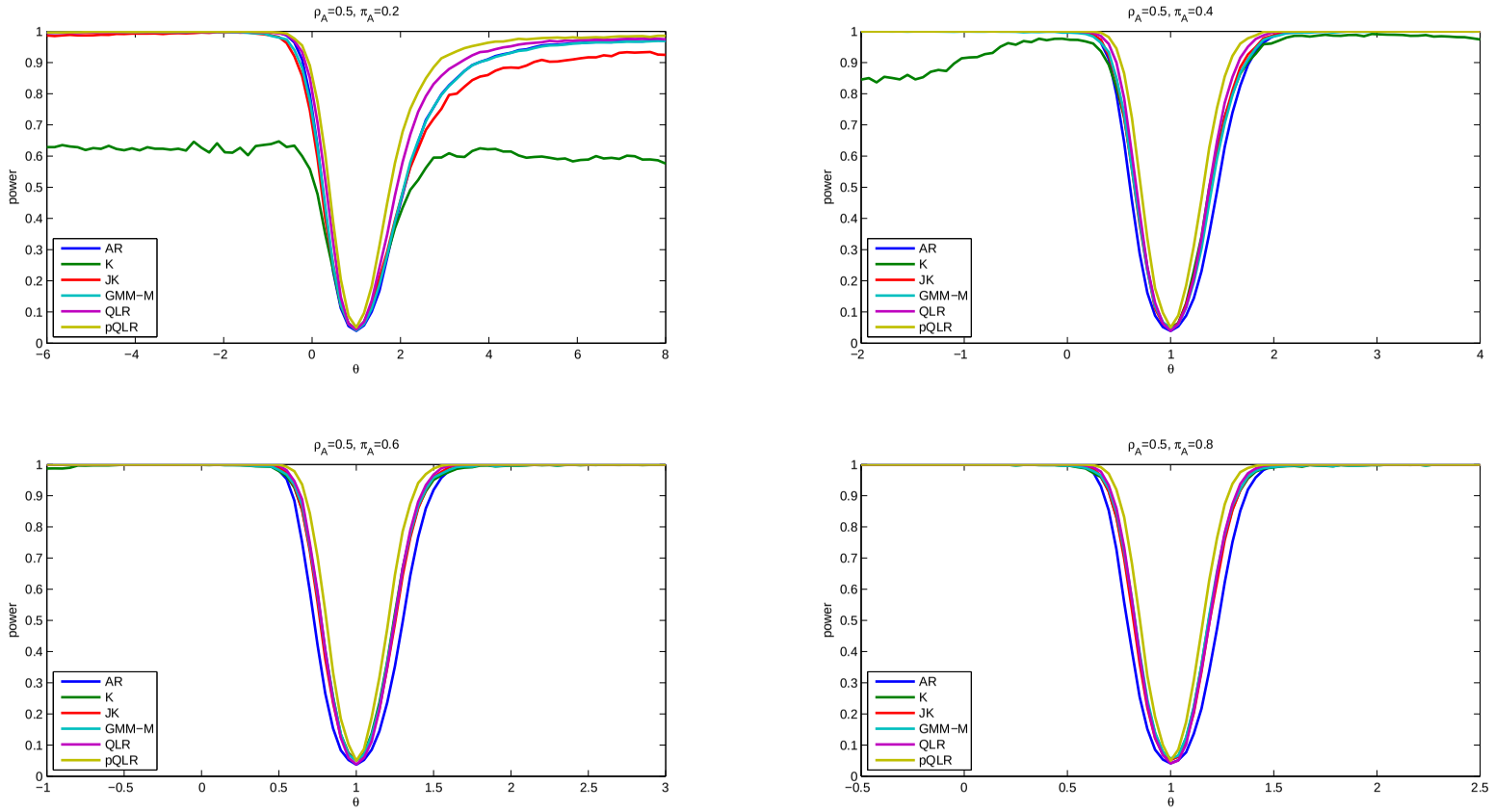

Figure 14: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.5$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

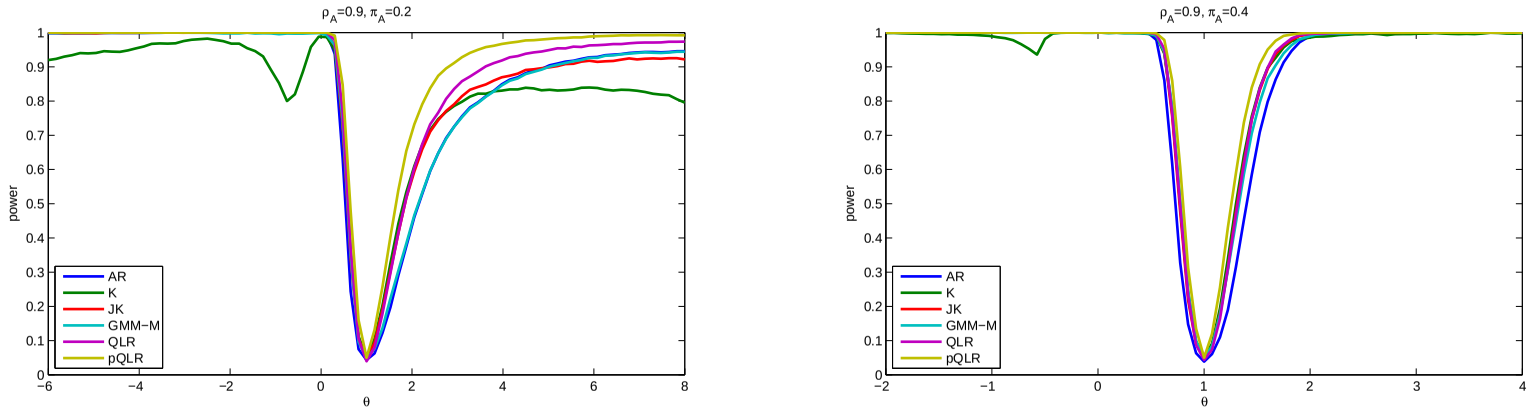

Figure 15: Simulated power of nominal 5% tests in asymmetric quantile IV simulation design with $\rho_A = 0.9$, five instruments and 1,000 observations. Based on 2,500 simulation replications and 1,000 draws of conditional critical values.

Gaussian process with mean zero and covariance function $\Sigma^{(SW)}(\beta, \theta, \beta_1, \theta_1)$. Further, $G^{(SW)}(\beta, \theta)$ is uniformly equicontinuous and uniformly bounded over $\mathcal{P}_0$.

SW2 Assumption 2 holds for $\Sigma^{(SW)}(\psi, \psi_1)$, $\psi = (\theta, \beta)$. Further, $\Sigma^{(SW)}(\psi, \psi_1)$ is uniformly continuous in ψ, ψ_1 uniformly over $\mathcal{P}_0$.

SW3 We have an estimator $\widehat{\Sigma}^{(SW)}$ for $\Sigma^{(SW)}$ which satisfies Assumption 3.

SW4 $m_1(\beta_0) = 0$ for an interior point $\beta_0 \in \mathcal{B}$, and for all $\delta > 0$ there exists $\varepsilon > 0$ such that $\|m_1(\beta)\| < \varepsilon$ implies $\|\beta - \beta_0\| < \delta$ for all $P \in \mathcal{P}_0$.

SW5 $m_1(\cdot)$ is continuously differentiable, and

$$\lambda_{\min}\left(\left(\left.\frac{\partial m_1(\beta)}{\partial \beta}\right|_{\beta=\beta_0}\right)'\left(\left.\frac{\partial m_1(\beta)}{\partial \beta}\right|_{\beta=\beta_0}\right)\right) > 1/\bar{c}$$

for some positive constant $\bar{c}$ and all $P \in \mathcal{P}_0$. Likewise, the maximal eigenvalue of the above matrix is uniformly bounded above by $\bar{c}$. $m_2(\beta, \theta)$ is continuously differentiable, and $m_2(\theta_0, \beta_0) = 0$. Further, both $m_2(\beta, \theta)$ and $\left.\frac{\partial m_2(\theta_0, \beta)}{\partial \beta}\right|_{\beta=\beta_0}$ are uniformly bounded over β, θ, and $\mathcal{P}_0$.

SW6 $\frac{1}{\sqrt{T}}\left(\frac{\partial}{\partial \beta} g_T^{(SW)}(\beta, \theta) - \frac{\partial}{\partial \beta} m_T^{(SW)}(\beta, \theta)\right) \to_p 0$ uniformly in (β, θ) and uniformly over $\mathcal{P}_0$.

We consider a sequence of (possibly parameter- and data-dependent) weighting matrices $W_T(\beta, \theta)$ which we will assume converge uniformly in probability to some positive-definite limit $W(\beta, \theta)$ which may depend on P but is uniformly bounded and positive-definite over $\mathcal{P}_0$. For example, we might take $W_T(\beta, \theta) = \widehat{\Sigma}^{(SW)}(\beta, \theta, \beta, \theta)^{-1}$. We will also assume that $\frac{\partial}{\partial \beta} W_T(\beta, \theta)$ is uniformly $O_p(1)$. Given these weighting matrices, define $\beta_T(\theta)$ to be the pseudo-true value of β given θ in the sample of size T:

$$\beta_T(\theta) = \arg\min_{\beta} m_T^{(SW)}(\beta, \theta)' W(\beta, \theta) m_T^{(SW)}(\beta, \theta).$$

It is important to note that if $\theta \neq \theta_0$ in general we have a misspecified moment condition model for β, in the sense that there does not exist a value of β such that the initial moment conditions are satisfied. The fact that the model is misspecified leads to a pseudo-true value $\beta_T(\theta)$ which depends on the choice of weighting matrix W_T and on the sample size.

However, the mis-specification in this setting is mild (uniformly of order $\frac{1}{\sqrt{T}}$), with the result that $\beta_T(\theta)$ converges to β_0 at rate $\frac{1}{\sqrt{T}}$ uniformly in θ uniformly over $\mathcal{P}_0$.

We estimate the structural nuisance parameter by

$$\hat{\beta}(\theta) = \arg\min_{\beta} g_T^{(SW)}(\beta,\theta)' W_T(\beta,\theta) g_T^{(SW)}(\beta,\theta).$$

The assumptions above guarantee that $\sqrt{T}(\widehat{\beta}(\theta) - \beta_T(\theta))$ converges to a Gaussian process.

Once we plug in the estimator of the nuisance parameter, the effective dimension of the process $g_T^{(SW)}(\widehat{\beta}(\theta),\theta)$ is reduced by the dimension of β, which is p. To avoid degeneracy we consider the linearly transformed moment condition $g_T^{(L)}(\beta,\theta) = (\hat{R}^{\perp})' g_T^{(SW)}(\beta,\theta)$, where $\hat{R}^{\perp}$ is a full-rank $k \times (k-p)$ dimensional matrix orthogonal to $\hat{R} = \frac{1}{\sqrt{T}} \left. \frac{\partial g_T^{(SW)}(\beta,\theta_0)}{\partial \beta} \right|_{\beta=\hat{\beta}(\theta_0)}$. We argue that the transformed $g_T^{(L)}$ together with the nuisance parameter estimator satisfy Assumptions 6-9 of the paper. The argument is straightforward, so for brevity we merely sketch it here.

Standard arguments show that uniformly over θ, for $R = \left. \frac{\partial m_1(\beta)}{\partial \beta} \right|_{\beta=\beta_0}$,

$$\begin{aligned} &\sqrt{T}(\widehat{\beta}(\theta) - \beta_T(\theta)) \\ &= \left(R' W(\beta,\theta_0) R\right)^{-1} R' W(\beta,\theta_0) \left(g_T^{(SW)}(\beta(\theta),\theta) - m_T^{(SW)}(\beta(\theta),\theta) \right) + o_p(1). \end{aligned}$$

Assumption SW6, together with the consistency of $\hat{\beta}(\theta_0)$, implies the uniform consistency of $\hat{R}$. Together with Assumption SW5, this implies that we can take $\hat{R}^{\perp}$ to be uniformly consistent for a full-rank $k \times (k-p)$ matrix $R^{\perp}$ orthogonal to R. The continuous mapping theorem and Assumption SW1 then imply that if we replace the $\hat{R}^{\perp}$ in the definition of $g_T^{(L)}$ with $R^{\perp}$, the error from this substitution is uniformly negligible.

This substitution results in the moment function

$$\left(R^{\perp}\right)' g_T^{(SW)}(\theta,\hat{\beta}(\theta)) = \left(R^{\perp}\right)' \left(g_T^{(SW)}(\beta,\theta(\theta)) + R\sqrt{T}(\hat{\beta}(\theta) - \beta(\theta)) \right) + o_p(1).$$

Since $R'\ R^{\perp} = 0$ we obtain that

$$g_T^{(L)}(\theta) = \left(R^{\perp}\right)' g_T^{(SW)}(\beta(\theta),\theta) + o_p(1)$$

and thus that Assumption 6 holds. Assumption 7 follows from SW2, while Assumption 8 follows from SW3. Finally, taking $M_T(\theta) = \left(R^{\perp}\right)' \frac{1}{\sqrt{T}} \frac{\partial m_T^{(SW)}(\beta,\theta)}{\partial \beta}\Big|_{\beta=\beta(\theta)}$, Assumption 9 can be shown to follow from SW5 and SW6.

S9 References

Andrews, I. (2015): "Conditional Linear Combination Tests for Weakly Identified Models," *mimeo*.

Andrews, D.W.K. and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721-762.

Angrist, J., V. Chernozhukov, and I. Fernandez-Val (2006): "Quantile Regression Under Misspecification with an Application to the US Wage Structure," *Econometrica*, 74 (2), 539-563.

Berlinet, A. and C. Thomas-Agnan (2004): *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, New York: Springer.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.

Chernozhukov, V. and C. Hansen (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132(2), 491-525.

Chernozhukov, V. and C. Hansen (2008): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142 (1), 379-398.

Janssen, A. and V. Ostrovski (2005): "The Convolution Theorem of Hajek and Le Cam - Revisited," *Statistics & Decisions*, 1

Kleibergen, F. (2005): "Testing Parameters in GMM without Assuming that They are Identified," *Econometrica*, 73(4), 1103-1124.

Kleibergen, F. (2007): "Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics," *Journal of Econometrics*, 139, 181-216.

Lehmann, E.L. and J.P. Romano (2005): *Testing Statistical Hypotheses*, New York: Springer; 3rd edition.

Mills, B., M. Moreira, and L. Vilela (2014): "Tests Based on t-Statistics for IV Regression with Weak Instruments," *Journal of Econometrics*, 182 (2), 351-363.

Moreira, H. and M. Moreira (2015): "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," *Unpublished manuscript*.

Montiel Olea, J.L. (2013): "Efficient Conditionally Similar Tests: Finite-Sample Theory and Large-Sample Applications," *unpublished manuscript*.

Stock, J. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68 (5), 1055-96.

Yogo, M. (2004), "Estimating the Elasticity of Intertemporal Substitution when Instruments are Weak," *Review of Economics and Statistics*, 86 (3), 797-810.