Conditional Linear Combination Tests for Weakly Identified Models

Isaiah Andrews*

January 29, 2014

Abstract

This paper constructs powerful tests applicable in a wide range of weakly identified contexts, including linear instrumental variables and nonlinear generalized method of moments (GMM) models. Our approach proceeds in two steps. First, we introduce the class of conditional linear combination tests, which reject the null hypothesis when a data-dependent convex combination of two identification-robust statistics is large. These tests control size under weak identification and are admissible, locally most powerful, and weighted average power maximizing in a conditional testing problem. In instrumental variables models with one endogenous regressor the conditional likelihood ratio test of Moreira (2003) is a conditional linear combination test, and in general models the class of conditional linear combination tests is equivalent to a class of quasi-conditional likelihood ratio tests. Second, we suggest using minimax regret conditional linear combination tests and propose a computationally tractable class of tests that plug in an estimator for a nuisance parameter. These plug-in tests offer substantially higher power than alternative approaches, matching the near-optimal performance of the conditional likelihood ratio test in homoskedastic weak instrumental variables models and substantially out-performing alternative procedures in a non-homoskedastic weak instrumental variables model and a non-linear new Keynesian Phillips curve model. These tests have optimal power in many strongly identified models, and so allow powerful identification-robust inference in a wide range of linear and non-linear models without sacrificing efficiency if identification is strong.

JEL Classification: C12, C18, C44

Keywords: Instrumental variables, nonlinear models, power, size, test, weak identification

*Department of Economics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E19-750, Cambridge, MA 02139 USA. Email: iandrews@mit.edu. The author is grateful to Anna Mikusheva, Whitney Newey, and Jerry Hausman for their guidance and support, and to Arun Chandrasekhar, Victor Chernozhukov, Denis Chetverikov, Kirill Evdokimov, Benjamin Feigenberg, Patricia Gomez-Gonzalez, Sally Hudson, Peter Hull, Conrad Miller, Scott Nelson, Jose Montiel Olea, Miikka Rokkanen, Adam Sacarny, Annalisa Scognamiglio, Brad Shapiro, Ashish Shenoy, Stefanie Stantcheva and the participants of the MIT Econometrics Lunch for helpful comments. NSF Graduate Research Fellowship support under grant number 1122374 is gratefully acknowledged.

1 Introduction

Researchers in economics are frequently interested in inference on causal or structural parameters. Unfortunately, in cases where the data contains only limited information useful for estimating these parameters, commonly used approaches to estimation and inference can break down and researchers who rely on such techniques risk drawing highly misleading inferences. Models where the usual approaches to inference fail due to limited information about model parameters are referred to as *weakly identified*. A large and growing literature develops identification-robust hypothesis tests, which control size regardless of identification strength and so limit the probability of rejecting true hypotheses in weakly identified contexts. The results to date on the power of identification-robust tests, that is their probability of rejecting false hypotheses, are, however, quite limited. In this paper we develop powerful identification-robust tests applicable to a wide range of models. Our approach relies on two innovations. First, we introduce a novel class of procedures, the class of conditional linear combination tests, which includes many known robust tests. Second, we suggest choosing conditional linear combination tests that minimize maximum regret, which is an intuitive optimality criterion not previously applied in this setting.

Our first step in constructing powerful tests is to introduce the class of *conditional linear combination* (CLC) tests. These tests depend on a convex combination of the generalized Anderson-Rubin (S) statistic introduced by Stock and Wright (2000) and the score (K) statistic introduced by Kleibergen (2005) for GMM models (or their analogs for generalized minimum distance, generalized empirical likelihood, or other settings), where the weight assigned to each depends on a conditioning statistic D also introduced by Kleibergen (2005). Tests based on S have stable power but are inefficient under strong identification, while tests based on K are efficient when identification is strong but can have low power when identification is weak. In many models D can be viewed as measuring identification strength, and its behavior governs the performance of tests based on K. CLC tests use information from D to determine how to weight the S and K statistics, and select critical values based on D in such a way that all tests in this class have correct size.

The class of conditional linear combination tests is quite large, and includes the S test of Stock and Wright (2000) and K test of Kleibergen (2005) for GMM and the conditional likelihood ratio (CLR) test of Moreira (2003) for linear instrumental variables (IV) models with a single endogenous regressor. More generally, we prove that the class of CLC tests is equivalent to a suitably defined class of quasi-CLR tests. CLC tests enjoy a number of optimality properties in

a testing problem which arises after conditioning on D, where we show that they are admissible, locally most powerful against particular sequences of alternatives, and weighted average power maximizing for a continuum of different weight functions.

Our second innovation is to use minimax regret CLC tests. This approach yields CLC tests with good overall power properties, and which in particular have power functions that lie as close as possible to the power envelope for this class in a uniform sense. By construction, these tests minimize the largest margin by which the power of the test selected might fall short relative to any other CLC test the researcher might have picked, thus minimizing the extent to which a researcher might regret their choice. Minimax regret has recently seen use in other areas of economics and econometrics (see Stoye (2009) for references) but, while natural for this context, has not to our knowledge been applied to the problem of selecting powerful tests for weakly identified models. Minimax regret tests must be obtained numerically which, while quite straightforward for some models, can be computationally daunting for others. In contexts where calculating true minimax regret tests is infeasible, we suggest a class of computationally simple plug-in minimax regret tests that plug in an estimate for a nuisance parameter.

To demonstrate the advantages of our approach, we consider several simulation examples. In linear IV with homoskedastic Gaussian errors and one endogenous regressor, Andrews, Moreira, and Stock (AMS, 2006) show that the CLR test of Moreira (2003) is "nearly" uniformly most powerful in a class of invariant two-sided tests. We show that plug-in minimax regret tests using reasonable plug-in estimators match the near-optimal performance of the CLR test in this case. Given that much of the data encountered in econometric practice is dependent (serially or spatially correlated, clustered), heteroskedastic, or both, however, it is of considerable interest to examine the performance of weak instrument-robust tests more broadly. To this end we calibrate a simulation to match heteroskedastic time-series data used by Yogo (2004) and find that our plug-in minimax regret test substantially outperforms Kleibergen's (2005) quasi-CLR test for general GMM models. The under-performance of Kleibergen's quasi-CLR test can be traced to the fact that the K statistic may perform especially poorly in non-homoskedastic IV. Kleibergen's test uses the CLR weight function, which is optimal under homoskedasticity but does not account for deterioration in the performance of the K statistic when we move away from the homoskedastic case. In contrast, the plug-in test proposed in this paper successfully accounts for the covariance structure of the data and delivers powerful, stable performance in both the homoskedastic and non-homoskedastic cases. To illustrate the application of our approach to a non-linear example, we also consider a

identification-robust generalized minimum distance approach to inference on new Keynesian Phillips curve parameters studied in Magnusson and Mavroeidis (2010). In a simulation calibrated to match the empirical application of that paper, we find that confidence sets based on a plug-in minimax regret test again outperform the other available approaches.

To develop intuition and illustrate results, we consider inference on parameters in linear IV and minimum distance models as recurring examples. Similarly to Mueller (2011) we assume that certain functions of the data converge in distribution to random variables in a limit problem and use this limit problem to study the performance of different procedures. To formally justify this approach we derive a number of asymptotic results, showing that the asymptotic size and power of CLC tests under the assumed convergence are simply their size and power in the limit problem. We further show that a large class of CLC tests control size uniformly in heteroskedastic linear IV with a single endogenous regressor. Moreover, we give conditions under which CLC tests, and plug-in minimax regret tests in particular, will be asymptotically efficient under strong identification, in the sense of being asymptotically uniformly most powerful in the class of tests depending on (S, K, D). Applying these results to our examples, we show that the tests we propose are asymptotically efficient in linear IV and minimum distance models when identification is strong. As a side result, we derive an essentially complete class of tests based on (S, K, D).

Before proceeding it is worth relating the approach taken in this paper to the recent econometric literature on optimal testing in non-standard models, including Mueller (2011), Elliott, Mueller, and Watson (2012), Olea (2012), and Moreira and Moreira (2013). The approaches studied in those papers apply under a weak convergence condition like the one we assume, and in each case the authors derive tests maximizing weighted average power. If a researcher has a well-defined weight function over the alternative with respect to which they want to maximize average power these approaches deliver optimal tests, either over the class of all tests or over the class of tests satisfying some auxiliary restrictions, and have a great deal to recommend them. In general, these tests are not available in closed form and will depend on the weight function chosen, however, and the nature of this dependence in a given context can be quite opaque. Consequently, in cases where the researcher has no particular weight function in mind, it can be unclear what a given choice of weight function will imply for the power of the resulting test. In contrast, the minimax regret approach suggested in this paper obviates the need to chose a weight function, instead attempting to pick tests which lie as close as possible to the power envelope for the class of CLC tests. Relative to the papers discussed above, the approach of this paper greatly restricts the class of

tests considered, first in confining attention to tests that depend only on S, K, and D, and then in further focusing on CLC tests. While this restriction reduces the strength of optimality statements, it renders the resulting tests much more transparent: conditional on D, the procedures discussed in this paper are simply tests based on a known convex combination of the S and K statistics, making it simple to understand their behavior. This transparency has other advantages, and it is relatively straightforward to give conditions under which CLC tests will be efficient under strong identification. This is particularly true of plug-in minimax regret tests which, while not generally optimal from a minimax regret perspective, often yield easy-to-characterize behavior under strong-identification asymptotics. In contrast, while Elliott, Mueller, and Watson (2012) discuss tests that "switch" to standard procedures with high probability on certain parts of the parameter space, and that can thus yield standard tests under strong asymptotics, implementing their procedures is extremely computationally costly in many empirically relevant contexts.

In the next section we outline the weak convergence assumption that will form the basis of our analysis and illustrate this assumption using our IV and minimum distance examples. In Section 3 we define several statistics including S, K, and D, and discuss a number of tests which have been proposed based on these statistics. Section 4 defines CLC tests, shows that CLR tests are CLC tests, and proves the equivalence of the class of CLC tests and a class of quasi-CLR tests. Section 5 defines a testing problem which arises conditional on D and shows that CLC tests are admissible, locally most powerful, and weighted average power maximizing in this problem. As a side result, we derive an essentially complete class of tests. Section 6 defines minimax regret CLC tests, shows that such tests exist, defines plug-in tests, and discusses implementation of these procedures. Section 7 focuses on the linear IV model and shows that suitably defined plug-in minimax regret tests match the near-optimal performance of the CLR test under homoskedasticity and substantially outperform existing alternatives under heteroskedasticity. Section 8 derives a number of asymptotic results, including uniform size control for CLC tests in linear IV and efficiency of plug-in tests in our examples under strong asymptotics. Finally, Section 9 simulates the performance of identification-robust confidence sets for new Keynesian Phillips curve parameters.

2 Weakly Identified Limit Problems

In this section we describe a class of limit problems that arise in many weakly identified contexts and illustrate this class by deriving the limit problems for two examples. We assume a sequence

of models indexed by sample size T, where sample T has distribution $F_T(\theta, \gamma)$ where $\theta \in \Theta$ is a p-dimensional parameter of interest and $\gamma \in \Gamma$ is an l-dimensional consistently estimable nuisance parameter. We will be concerned with testing $H_0 : \theta = \theta_0$ and assume we observe three objects: a $k \times 1$ vector $g_T(\theta_0)$ which will in many cases be an appropriately scaled moment vector, distance function, or analog thereof, a $k \times p$ matrix $\Delta g_T(\theta_0)$ which will often be some transformation of the Jacobian of $g_T(\theta)$ with respect to θ, and an estimate $\hat{\gamma}$ for γ. We assume that for all fixed $(\theta, \gamma) \in \Theta \times \Gamma$ we have that

$$\begin{pmatrix} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{pmatrix} \to_d \begin{pmatrix} g \\ \Delta g \end{pmatrix} \tag{1}$$

and $\hat{\gamma} \to_p \gamma$ under the sequence of data-generating processes $F_T(\theta, \gamma)$, where

$$\begin{pmatrix} g \\ vec(\Delta g) \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ vec(\mu) \end{pmatrix}, \begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix} \right), \tag{2}$$

and $m = m(\theta, \theta_0, \gamma) \in \mathcal{M}(\mu, \gamma)$ for a set $\mathcal{M}(\mu, \gamma) \subseteq \mathbb{R}^k$ which may depend on $\mu \in \mathbb{M}$ and γ. Here we use $vec(A)$ to denote vectorization, which maps $k \times p$ matrix A to a $kp \times 1$ vector. We further assume that $\Sigma_{\theta g}$ and $\Sigma_{\theta\theta}$ are continuous functions of γ and are thus consistently estimable. We will generally suppress the dependence of the terms in the limit problem on the parameters (θ, γ) when there is no loss of clarity from doing so, writing simply m, μ, and so forth. We are interested in problems where the null hypothesis $\theta = \theta_0$ implies $m = 0$, and will focus on testing $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m \in \mathcal{M}(\mu) \backslash \{0\}, \mu \in \mathbb{M}$.

Limit problems of the form (2) arise in a wide variety of weakly identified models. In the remainder of this section we show that weakly identified instrumental variables and minimum distance models generate limit problems of this form, deferring some derivations to Appendix A. General results showing that weakly identified GMM models give rise to limit problems of the form (2) are given in Appendix B.

Example I: Weak IV The linear instrumental variables model with a single endogenous regressor, written in reduced form, is

$$\begin{aligned} Y &= Z\pi\beta + V_1 \\ X &= Z\pi + V_2 \end{aligned} \tag{3}$$

for Z a $T \times k$ matrix of instruments, X a $T \times 1$ vector of endogenous regressors, Y a $T \times 1$ vector of outcome variables, and V_1 and V_2 both $T \times 1$ vectors of residuals. In such models we are generally

interested in testing a hypothesis $H_0 : \beta = \beta_0$ about the scalar coefficient β, treating the $k \times 1$ vector of first-stage parameters π as nuisance parameters. As elsewhere in the literature (see e.g. AMS) we can accommodate additional exogenous regressors, but omit such exogenous variables here to simplify the exposition.

The usual identifying assumption in IV models is that $E[V_{1,t}Z_t] = E[V_{2,t}Z_t] = 0$ for Z_t the transpose of row t of Z, which allows us to write linear IV as a special case of GMM with moment condition

$$f_t(\beta) = (Y_t - X_t\beta) Z_t \tag{4}$$

and identifying assumption $E_\beta[f_t(\beta)] = 0$ (where $E_\theta[X]$ denotes the expectation of X under true parameter value θ). Provided that $\frac{1}{T}\sum Z_t Z_t' \to_p Q_Z$ for Q_Z positive definite and $\frac{1}{\sqrt{T}}\sum Z_t V_{1,t}$ and $\frac{1}{\sqrt{T}}\sum Z_t V_{2,t}$ converge in distribution to jointly normal random vectors with a consistently estimable covariance matrix, for a fixed π with $||\pi|| > 0$ it is straightforward to construct consistent, asymptotically normal GMM estimates based on (4) and to use these estimates to test hypotheses about the parameter β, as is typically done in empirical practice.

As is now well understood, the standard asymptotic approximations to the distribution of estimators and test statistics in linear IV may be quite poor if the first stage parameter π is small relative to the sample size. To derive better approximations for this weakly identified case, Staiger and Stock (1997) model the first-stage parameter π as changing with the sample size, in particular taking π_T, the first-stage coefficient for sample size T, to be $\pi_T = \frac{c}{\sqrt{T}}$ for a fixed vector $c \in \mathbb{R}^k$. Using this device, Staiger and Stock derive asymptotic approximations to the distribution of estimators and test statistics in the model (3) which better reflect their finite-sample behavior for small values of π. In particular, they show that the (1949) Anderson-Rubin test for $H_0 : \beta = \beta_0$ controls size when the data (Y_t, X_t, Z_t') are independent and the errors $(V_{1,t}, V_{2,t})$ are homoskedastic, and a large subsequent literature including Stock and Wright (2000), Kleibergen (2002), Moreira (2003), Kleibergen (2005), Olea (2012), and Andrews and Cheng (2012) has extended these results to more general models and alternative identification-robust tests.

To derive the limit problem (2) for the weak IV model, define $f_T(\beta) = \frac{1}{T}\sum f_t(\beta)$ and let Ω be the asymptotic variance matrix of $\sqrt{T}\left(f_T(\beta_0)', -\frac{\partial}{\partial\beta} f_T(\beta_0)'\right)'$, i.e.

$$\Omega = \begin{pmatrix} \Omega_{ff} & \Omega_{f\beta} \\ \Omega_{\beta f} & \Omega_{\beta\beta} \end{pmatrix} = \lim_{T\to\infty} Var\left(\sqrt{T}\begin{pmatrix} f_T(\beta_0) \\ -\frac{\partial}{\partial\beta} f_T(\beta_0) \end{pmatrix}\right). \tag{5}$$

We assume that Ω_{ff} is full-rank. For $\hat{\Omega}$ a consistent estimator of Ω, define

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

$$\Delta g_T(\beta) = -\sqrt{T}\hat{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial \beta} f_T(\beta),$$

and $\hat{\gamma} = vec\left(\hat{\Omega}\right)$. For $\theta = \beta$, $\Theta = \mathbb{R}$, $\gamma = vec(\Omega)$, and Γ the set of of values γ such that $\Omega(\gamma)$ is symmetric and positive definite, we can see that for all $(\theta, \gamma) \in \Theta \times \Gamma$ the Continuous Mapping Theorem implies

$$\begin{pmatrix} g_T(\beta_0) \\ \Delta g_T(\beta_0) \end{pmatrix} \to_d \begin{pmatrix} g \\ \Delta g \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ \mu \end{pmatrix}, \begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix}\right) \tag{6}$$

so (1) and (2) hold here with $m = \Omega_{ff}^{-\frac{1}{2}} Q_Z c(\beta - \beta_0)$, $\mu = \Omega_{ff}^{-\frac{1}{2}} Q_Z c \in \mathbb{M} = \mathbb{R}^k$, $\Sigma_{g\theta} = \Omega_{ff}^{-\frac{1}{2}} \Omega_{f\beta} \Omega_{ff}^{-\frac{1}{2}}$, and $\Sigma_{\theta\theta} = \Omega_{ff}^{-\frac{1}{2}} \Omega_{\beta\beta} \Omega_{ff}^{-\frac{1}{2}}$. Note that for any μ we have that $m \in \mathcal{M}(\mu) = \{b \cdot \mu : b \in \mathbb{R}\}$ and that $m = 0$ when $\beta = \beta_0$. To derive this limit problem we have imposed very little structure on the underlying data generating process, and so can easily accommodate heteroskedastic, clustered, or serially correlated errors V_1 and V_2 and other features commonly encountered in applied work.□

Example II: Minimum Distance A common approach to estimating econometric models is to choose structural parameters to match some vector of sample moments or reduced-form parameter estimates. Canova and Sala (2010), for example, discuss estimation of Dynamic Stochastic General Equilibrium (DSGE) models by matching impulse responses. Other papers that apply a minimum distance approach in the DSGE context include Christiano and Eichenbaum (1992), Rotemberg and Woodford (1997), and Ruge-Murcia (2010). Minimum distance and moment matching approaches are also common in a wide range of other applications, and encompass both indirect inference as discussed in Gourieroux, Montfort and Renault (1993) and simulated method of moments as in McFadden (1989) and Duffie and Singleton (1993) and much of the subsequent literature.

In minimum distance or moment-matching models, for θ a $p \times 1$ vector of structural parameters and η a $k \times 1$ vector of reduced-form parameters or moments, the model implies that $\eta = f(\theta)$ for some function f. We assume that $f(\theta)$ is continuously differentiable and that $f(\theta)$ and its Jacobian can be calculated either directly or by simulation. Suppose we have an estimator $\hat{\eta}$ for

the reduced-form parameter η that, together with an estimator $\hat{\Omega}_\eta$ for the variance of $\hat{\eta}$, satisfies

$$\hat{\Omega}_\eta^{-\frac{1}{2}} (\hat{\eta} - \eta) \to_d N(0, I).$$

Under strong identification asymptotics, we have that $\hat{\eta} - \eta = O_p\left(\frac{1}{\sqrt{T}}\right)$ and can use $\Delta-$method arguments to derive the usual asymptotic approximations to the distribution of the structural parameter estimates

$$\hat{\theta} = \arg\min_\theta (\hat{\eta} - f(\theta))' \hat{\Omega}_\eta^{-1} (\hat{\eta} - f(\theta))$$

and the standard test statistics (see e.g. Newey and McFadden (1994)). As Canova and Sala (2010) highlight, however, if there is limited information about the structural parameters θ these approximations may be quite poor. One way to model such weak identification is to take the variance of the reduced-form parameter estimates to be constant, with $\hat{\Omega}_\eta \to_p \Omega_\eta$ for Ω_η non-degenerate, which implies that $\hat{\eta}$ is not consistent for η. Such sequences can often be justified by modeling the variance of the data generating process as growing with the sample size. Let $g_T(\theta) = \hat{\Omega}_\eta^{-\frac{1}{2}} (\hat{\eta} - f(\theta))$, $\Delta g_T(\theta) = \frac{\partial}{\partial \theta'} g_T(\theta) = \hat{\Omega}_\eta^{-\frac{1}{2}} \frac{\partial}{\partial \theta'} f(\theta)$, and $\hat{\gamma} = vec\left(\hat{\Omega}_\eta\right)$. For $\gamma = vec(\Omega_\eta)$ and Γ again the set of γ values corresponding to symmetric positive definite matrices, we have that under $(\theta, \gamma) \in \Theta \times \Gamma$, $\hat{\gamma} \to_p \gamma$ and

$$\begin{pmatrix} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{pmatrix} \to_d \begin{pmatrix} g \\ vec(\Delta g) \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ vec(\mu) \end{pmatrix}, \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\right) \tag{7}$$

where $m \in \mathcal{M} = \left\{\Omega_\eta^{-\frac{1}{2}} (f(\theta) - f(\theta_0)) : \theta \in \Theta\right\}$ and $\mu = \Omega_\eta^{-\frac{1}{2}} \frac{\partial}{\partial \theta'} f(\theta_0)$ (see Appendix A for details).□

As these examples highlight, limit problems of the form (2) arise in a wide variety of econometric models with weak identification. Appendix B shows that general GMM models that are weakly identified in the sense of Stock and Wright (2000) generate limit problems of the form (2), and Example I could be viewed as a special case of this result. As Examples II illustrates, however, the limit problem (2) is more general. Another non-GMM example is provided in Section 9, which discusses a weakly identified generalized minimum distance model studied by Magnusson and Mavroeidis (2010).[1]

[1]Other examples may be found in Olea (2012), who shows that for appropriately defined $g_T(\theta_0)$ and $\Delta g_T(\theta)$ convergence of the form (2) holds in several weakly identified extremum estimation examples, including a probit model with endogenous regressors and a nonlinear regression model. Guggenberger and Smith (2005, proofs of

Since the limit problem (2) appears in a wide range of weakly identified contexts, for the next several sections we focus on tests in this limit problem. In particular, similar to Mueller (2011) we consider the problem of testing $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m \in \mathcal{M}(\mu)\backslash\{0\}, \mu \in \mathbb{M}$ with the limiting random variables (g,Δg,γ) observed and seek to derive tests with good properties. In Section 8 we return to the original problem, and argue that under mild assumptions the results we derive in the limit problem (2) can be viewed as asymptotic results along sequences of models satisfying (1).

3 Pivotal Statistics Under Weak Identification

As noted in the introduction, under weak identification many commonly used test statistics are no longer asymptotically pivotal under the null. To address this issue, much of the literature on identification-robust testing has focused on deriving statistics that are asymptotically pivotal or conditionally pivotal under the null even when the model is weakly identified. Many of the statistics proposed in this literature can be written as functions of the S statistic of Stock and Wright (2000) and the K and D statistics of Kleibergen (2005), or their appropriately-defined analogs in non-GMM settings. In this section we define these statistics, which will play a central role in the remainder of the paper, and develop some results concerning their properties.

3.1 The S Statistic

When testing $H_0 : m = 0, \mu \in \mathbb{M}$ in the limit problem (2), a natural statistic to consider is

$$S = g'g \sim \chi^2_k \left(m'm\right). \tag{8}$$

Under the null S will be χ^2 distributed with k degrees of freedom, while under the alternative it will be non-central χ^2 distributed with non-centrality parameter $m'm = \|m\|^2$. Statistics that are asymptotically equivalent to (8) for appropriately defined g_T and Δg_T have been suggested in a wide range of contexts, including the AR statistic for instrumental variables models proposed by Anderson and Rubin (1949) and discussed in Staiger and Stock (1997), the S statistic of Stock and Wright (2000) for GMM models, the MD-AR statistic of Magnusson and Mavroeidis (2010)

theorems 4 and 6) show that for appropriately defined $g_T(\theta_0)$ and $\Delta g_T(\theta)$ such convergence also holds in weakly identified Generalized Empirical Likelihood (GEL) models with independent data, both with and without strongly identified nuisance parameters. Guggenberger, Ramalho, and Smith (2012, proofs of theorems 3.2 and 4.2) extend these results to time series GEL applications, further highlighting the relevance of the limit problem (2).

for minimum distance models, and a number of statistics for GEL models discussed in Ramalho and Smith (2004), Guggenberger and Smith (2005), Otsu (2006), Guggenberger and Smith (2008), and Guggenberger, Ramalho, and Smith (2012) that are asymptotically equivalent to Generalized Empirical Likelihood Ratio statistics: see Guggenberger, Ramalho, and Smith (2012) for discussion. In each of these cases, the S statistic is equivalent, at least asymptotically, to an appropriately scaled version of some estimation objective function, for example a continuously updating GMM objective function in Stock and Wright (2000) and a continuously updating minimum distance objective function in Magnusson and Mavroeidis (2010).

3.2 The K Statistic

While S is a natural statistic for testing $H_0 : m = 0, \mu \in \mathbb{M}$, in some ways it is not ideal. In particular, under strong identification the limit problem (2) generally arises when we consider local alternatives to the null hypothesis, in which case Taylor expansion arguments yield that Δg is non-random (so $\Sigma_{\theta\theta} = \Sigma_{g\theta} = 0$) and $m = \Delta g(\theta - \theta_0)$.[2] In such cases, rather than considering the S statistic we might instead considered the statistic

$$LM = g'\Delta g \left(\Delta g'\Delta g\right)^{-1} \Delta g' g = g' P_{\Delta g} g$$

which in GMM corresponds to the score statistics discussed in Newey and West (1987). If $p = k$ this is exactly the same as the S statistic (assuming Δg is full rank). If $p < k$, however, LM and S differ and we can see that under strong identification the LM statistic statistic has non-centrality $m'm = (\theta - \theta_0)'\Delta g'\Delta g(\theta - \theta_0)$ under alternative θ, which is the same as the non-centrality of the S statistic. Under the null, however the LM statistic is χ^2 distributed with $p < k$ degrees of freedom so tests based on this statistic use smaller critical values than those based on the S statistic. Thus, under strong identification the level α test that rejects for large values of LM dominates the test that rejects for large values of S, in the sense that it has the same size and higher power against all alternatives. The source of this difference is that under strong identification tests based on the S statistic have power against all violations of $H_0 : m = 0$, and so test the parametric restriction $\theta = \theta_0$ together with the over-identifying restriction $m'\left(I - P_{\Delta g}\right) m = 0$, while tests based on the LM statistic test only $\theta = \theta_0$. In part as a result of this difference, under strong identification tests of based on LM are typically asymptotically equivalent to tests based on Wald or Quasi-Likelihood

[2]For discussion of this point in a GMM context see Appendix B. For a more extensive analysis see Newey and McFadden (1994), Section 9.

Ratio statistics, while those based on S are not.

Unfortunately, in weakly identified models where $\Sigma_{\theta g} \neq 0$ tests based on the LM statistic generally do not control size. As Kleibergen (2005) highlights in a general GMM context, the issue is that in such models Δg may be random and correlated with g. When this occurs the limiting distribution of LM under the null will depend on $\Sigma_{\theta g}$, $\Sigma_{\theta\theta}$, and μ. Since μ cannot be consistently estimated, this dependence makes drawing reliable inferences on the basis of LM challenging at best. To address this problem in weak IV Moreira (2001) and Kleibergen (2002) propose a modified score statistic that remains pivotal under weak identification, and Kleibergen (2005) generalizes this statistic to GMM. The key insight is that by replacing Δg by a matrix D that is independent of g we can obtain a statistic that remains pivotal under weak identification. In particular, analogous to Kleibergen (2005) let us define D as the $k \times p$ matrix such that

$$vec(D) = vec(\Delta g) - \Sigma_{\theta g} g$$

and note that

$$\begin{pmatrix} g \\ vec(D) \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ vec(\mu_D) \end{pmatrix}, \begin{pmatrix} I & 0 \\ 0 & \Sigma_D \end{pmatrix} \right)$$

where $vec\left(\mu_D\right) = vec\left(\mu\right) - \Sigma_{\theta g} m$, $\mu_D \in \mathbb{M}_D$, $\Sigma_D = \Sigma_{\theta\theta} - \Sigma_{\theta g}\Sigma_{g\theta}$, and $m \in \mathcal{M}_D\left(\mu_D\right)$ for

$$\mathcal{M}_D\left(\mu_D\right) = \left\{ m : m \in \mathcal{M}\left(\mu\right) \text{ for } vec\left(\mu\right) = vec\left(\mu_D\right) + \Sigma_{\theta g} m \right\}.$$

$\mathcal{M}_D$ plays a role similar to $\mathcal{M}$, defining the set of values m consistent with a given mean μ_D for D. The matrix D can be interpreted as the part of Δg that is uncorrelated with g which, since D and g are jointly normal, implies that D and g are independent. In many models D contains information about the strength of identification: in linear IV (Example I) for instance, D is a transformation of a particular first-stage parameter estimate.

Kleibergen then defines the K statistic

$$K = g' D \left(D'D\right)^{-1} D' g = g' P_D g. \tag{9}$$

Since D and g are independent we can see that the distribution of g conditional on $D = d$ is the same as the unconditional distribution, $g|D = d \sim N(m, I)$. Hence, if D is full rank (which we assume holds with probability one for the remainder of the paper) we can see that under the null

the conditional distribution of K is $K|D = d \sim \chi^2_p$. Thus, under the null K is independent of D and the unconditional distribution of K is χ^2_p as well. Note, further, that in the strongly identified limit problem g and Δg are uncorrelated, $\Sigma_{\theta g} = 0$, so $D = \Delta g$, $K = LM$, and the K statistic retains all the desirable properties of the LM statistic for this case.

Kleibergen (2005) shows that in GMM his K test is a score test based on the continuous updating GMM objective function, and subsequent work has proposed related statistics in a number of other settings, all of which yield the K statistic (9) in the appropriately defined limit problem. In particular, Magnusson and Mavroeidis (2010) propose what they term an MD-K statistic for weakly identified generalized minimum distance models, based on a quadratic form in the score of the continuous updating minimum distance objective function, while Ramalho and Smith (2004), Guggenberger and Smith (2005), Guggenberger and Smith (2008), and Guggenberger, Ramalho, and Smith (2012) discuss identification-robust statistics that are (asymptotically) equivalent to quadratic forms in the score of the GEL objective function: again, see Guggenberger, Ramalho, and Smith (2012) for discussion.

For the remainder of the paper we will focus on the class of tests that can be written as functions of the S, K, and D statistics. While, as the discussion above suggests, this class includes most of the identification-robust procedures proposed in the literature to date, it does rule out some robust tests. In particular, Andrews and Cheng (2012) and (2013) derive identification-robust Wald and Quasi-LR tests that cannot in general be written as functions of (S, K, D) and so fall outside the class studied in this paper.

3.3 Other Asymptotically Pivotal Statistics

A number of other identification-robust test statistics have been created using S, K, and D. Since some of these statistics will play an important role later in our analysis we briefly introduce them here. Kleibergen (2005) defines the J statistic as the difference between the S and K statistics

$$J = S - K = g' \left(I - D \left(D'D \right)^{-1} D \right) g = g' \left(I - P_D \right) g$$

and notes that under the null J is χ^2_{k-p} distributed and is independent of (K, D) regardless of the strength of identification. In a GMM context, one can show that under strong identification this statistic is asymptotically equivalent to Hansen's (1982) J statistic for testing over-identifying restrictions under the null and local alternatives. In cases where one may be concerned about

spurious declines in the power of the K test, Kleibergen (2005) suggests using JK tests that reject if either the K or J statistic is too large,

$$\phi_{JK} = \max\left\{1\left\{K > \chi^2_{p,1-\alpha_J}\right\}, 1\left\{J > \chi^2_{k-p,1-\alpha_K}\right\}\right\} \tag{10}$$

where the size of the resulting test is $E_{m=0}\left[\phi_{JK}\right] = \alpha_J + \alpha_K - \alpha_J\alpha_K$.[3] Magnusson and Mavroeidis (2010) and Guggenberger, Ramalho, and Smith (2012) derive analogs of Kleibergen's J statistic in the generalized minimum distance and GEL contexts, respectively, and Magnusson and Mavroeidis (2010) consider JK tests in their empirical analysis.

Moreira (2003) considers the problem of testing hypotheses on the parameter β in weak IV (Example I) when the instruments Z are fixed and the errors V are normal and homoskedastic with a known variance. Moreira derives a conditional likelihood ratio statistic which, for $p = 1$ and (J, K, D) defined appropriately, is

$$CLR = \frac{1}{2}\left(K + J - D'\Sigma_D^{-1}D + \sqrt{\left(K + J + D'\Sigma_D^{-1}D\right)^2 - 4J \cdot D'\Sigma_D^{-1}D}\right)$$

and which, under the null, is conditionally pivotal:

$$CLR|D = d \sim \frac{1}{2}\left(\chi^2_1 + \chi^2_{k-1} - d'\Sigma_D^{-1}d + \sqrt{\left(\chi^2_1 + \chi^2_{k-1} + d'\Sigma_D^{-1}d\right)^2 - 4\chi^2_{k-1} \cdot d'\Sigma_D^{-1}d}\right)$$

where χ^2_1 and χ^2_{k-1} are independent χ^2 random variables with 1 and $k-1$ degrees of freedom, respectively.

We can define analogs of CLR for more general contexts. In particular, for any function $r : D \to \mathbb{R}_+ \cup \infty$, define the quasi-CLR statistic

$$QCLR_r = \frac{1}{2}\left(K + J - r(D) + \sqrt{\left(K + J + r(D)\right)^2 - 4J \cdot r(D)}\right) \tag{11}$$

where $QCLR_r = K$ when $r(D) = \infty$. Exactly as above, we can see that conditional on $D = d$ the $QCLR_r$ statistic has distribution

$$QCLR_r|D = d \sim \frac{1}{2}\left(\chi^2_p + \chi^2_{k-p} - r(d) + \sqrt{\left(\chi^2_p + \chi^2_{k-p} + r(d)\right)^2 - 4\chi^2_{k-p} \cdot r(d)}\right) \tag{12}$$

under the null and can consider quasi-CLR (QCLR) tests that reject when $QCLR_r > q_\alpha(r\left(D\right))$

[3]For ϕ a (non-randomized) test, we take $\phi = 1$ to denote rejection and $\phi = 0$ failure to reject.

for $q_\alpha(r(d))$ the $1-\alpha$ quantile of (12). This class of QCLR tests nests the quasi-CLR tests of Kleibergen (2005), Smith (2007), and Guggenberger, Ramalho, and Smith (2012), all of which take $r(D)$ to be a test statistic for the hypothesis that μ_D is reduced rank.

3.4 Distribution of the J and K Statistics Under Weak Identification

Since the J and K statistics will play a central role in the remainder of the analysis we discuss their respective properties under weak identification. In particular, note that conditional on $D = d$ for d full rank, the K and J statistics are independent with distribution $K|D = d \sim \chi^2_p(\tau_K(d))$ and $J|D = d \sim \chi^2_{k-p}(\tau_J(d))$ where

$$\tau_K(D) = m'P_D m,\ \tau_J(D) = m'(I - P_D)\, m \tag{13}$$

are the squared length of the projection of m onto D and the residual from that projection, respectively.

The K statistic picks out a particular (random) direction corresponding to the span of D and restricts attention to deviations from $m = 0$ along this direction. Under strong identification, this direction corresponds to parametric alternatives, with the consequence that the K test

$$\phi_K = 1\left\{K > \chi^2_{p,1-\alpha}\right\} \tag{14}$$

is optimal in this case. Under weak identification, however, whether or not it makes sense to focus on the direction picked out by the K statistic will depend on the distribution of D and the set $\mathcal{M}_D(\mu_D)$ of possible values of m. In contrast to the K statistic, the S statistic treats all deviations from $m = 0$ equally and its power depends only on $\|m\|$, which may be quite appealing in cases where $\mathcal{M}_D(\mu_D)$ imposes few restrictions on the possible values of m. To give a more concrete sense of the properties of the K statistic, we return to Examples I and II introduced above.

Example II: Minimum Distance (Continued) We established in (7) that Δg is non-random, so $D = \Delta g = \mu = \mu_D$. To simplify the exposition, let us assume for this section that $\Omega_\eta = I$.[4] Since $\mathcal{M} = \{(f(\theta) - f(\theta_0)) : \theta \in \Theta\}$ we have that under alternative θ the non-centrality parameters

[4]Without this assumption the same intuition applies under the norm $||x||_{\Omega^{-1}} = \sqrt{x'\Omega_\eta^{-1}x}$.

in the J and K statistics are

$$(\tau_J(\theta), \tau_K(\theta)) = \left((f(\theta) - f(\theta_0))' (I - P_\mu) (f(\theta) - f(\theta_0)), (f(\theta) - f(\theta_0))' P_\mu (f(\theta) - f(\theta_0)) \right).$$

Since $\mu = \frac{\partial}{\partial \theta'} f(\theta_0)$, this means that under alternative θ the non-centrality parameter τ_K is the squared length of $f(\theta) - f(\theta_0)$ projected onto the model's tangent space at the null parameter value, while τ_J is the squared length of the residual from this projection. Hence, if $f(\theta)$ is linear so $f(\theta) = \frac{\partial}{\partial \theta'} f(\theta_0)(\theta - \theta_0)$ and $\mathcal{M} = \left\{ \frac{\partial}{\partial \theta'} f(\theta_0) \cdot b : b \in \mathbb{R}^p \right\}$ then $\tau_J \equiv 0$ and the K test ϕ_K will be optimal. As argued in Andrews and Mikusheva (2012), under strong identification minimum distance models are approximately linear, confirming the desirable properties of the K statistic under strong identification. Under weak identification, however, non-linearity of $f(\theta)$ may remain important even asymptotically. To take an extreme case, if there is any value $\theta \in \Theta$ such that $||f(\theta) - f(\theta_0)|| > 0$ and $\frac{\partial}{\partial \theta} f(\theta_0)' (f(\theta) - f(\theta_0)) = 0$, the K statistic will not help in detecting such an alternative and the optimal test against θ, $\phi_J = 1\left\{ J > \chi^2_{k-p,1-\alpha} \right\}$, depends on J alone.□

Example I: Weak IV (Continued) In the limit problem (6) Δg is random and may be correlated with g, so $D \neq \Delta g$ and $\mu_D = \mu - \Sigma_{\theta g} m$. Since $m = \mu(\beta - \beta_0)$ this means that

$$\mu_D = \mu - \Sigma_{\theta g} \mu (\beta - \beta_0) = (I - \Sigma_{\theta g} (\beta - \beta_0)) \mu.$$

Note that if μ is proportional to an eigenvector of $\Sigma_{\theta g}$ corresponding to a non-zero eigenvalue λ, then for $(\beta - \beta_0) = \lambda^{-1}$ we have that $\mu_D = 0$. Hence for some $(\Sigma_{\theta g}, \mu)$ combinations, while μ may be quite large relative to both $\Sigma_{\theta g}$ and $\Sigma_{\theta\theta}$ there will be some alternatives β under which $\mu_D = 0$. When this occurs the direction of the vector D bears no relation to the direction of m or μ and the K statistic picks a direction entirely at random and so loses much of its appeal. The well-known non-monotonicity of the power function for tests based on K alone is a consequence of this fact. A special case of this phenomenon appears when, as in the homoskedastic model considered by AMS, Ω as defined in (5) has Kronecker product structure, so $\Omega = A \otimes B$ for a 2×2 matrix A and a $k \times k$ matrix B. In this case $\Sigma_{g\theta} = \lambda \cdot I$ so when $(\beta - \beta_0) = \lambda^{-1}$, $\mu_D = 0$ regardless of the true value μ. This is precisely what occurs at the point β_{AR} discussed by AMS, where they show that the test

$$\phi_S = 1\left\{ S > \chi^2_{k,1-\alpha} \right\} \tag{15}$$

is optimal. This is entirely intuitive, since at this point D bears no relation to m and the best thing we can do is to ignore it entirely and focus on the S statistic.

The case where Ω has Kronecker product structure is extreme in that $\mu_D = 0$ at alternative β_{AR} regardless of the true value μ. However, tests based on the K statistic face other challenges in the non-Kronecker case. In particular, in the Kronecker product case $\mu_D \propto \mu$ and so as long as $\mu_D \neq 0$ the mean of D has the correct direction, while in contrast $\mu_D \not\propto \mu$ in the general (non-Kronecker) case. An extreme version of this issue arises if there is some value β^* such that $(I - \Sigma_{\theta g}(\beta^* - \beta_0))\,\mu \neq 0$ but $\mu'\,(I - \Sigma_{\theta g}(\beta^* - \beta_0))\,\mu = 0$. For this value of β^* we have that $\mu_D \neq 0$ but $\mu_D' m = 0$, and hence the K statistic tends to focus on a directions that yield low power against alternative β^*.$\square$

As these examples show, while tests rejecting for large values of the K statistic are efficient under strong identification, they can sometimes have low power when identification is weak. In contrast, the S test ϕ_S is inefficient under strong identification but has the appealing property that its power depends only on $||m||$, which measures the violation of the null hypothesis $m = 0$, and thus does not suffer from the spurious loss of power that can affect tests based on K. The question in constructing tests based on (S, K, D) (or equivalently (J, K, D)) is thus how best to use the information contained in D to combine the S and K statistics to retain the advantages of each while ameliorating their deficiencies.

4 Conditional Linear Combination Tests

To flexibly combine the S, K and D statistics we introduce the class of conditional linear combination tests. For a weight function $a : D \to [0, 1]$ the corresponding conditional linear combination test, $\phi_{a(D)}$, rejects when a convex combination of the S and K statistics weighted by $a\,(D)$ exceeds a conditional critical value:

$$\phi_{a(D)} = 1\left\{(1 - a\,(D)) \cdot K + a\,(D) \cdot S > c_\alpha(a(D))\right\} = 1\left\{K + a(D) \cdot J > c_\alpha(a(D))\right\}. \tag{16}$$

We take the conditional critical value $c_\alpha(a)$ to be $1-\alpha$ quantile of a $\chi^2_p + a \cdot \chi^2_{k-p}$ distribution. This choice ensures that $\phi_{a(D)}$ will be conditionally similar, and thus similar, for any choice of $a(D)$. Stated formally:

Theorem 1 *For any weight function $a : D \to [0, 1]$ the test $\phi_{a(D)}$ defined in (16) is conditionally*

similar with $E_{m=0,\mu_D}\left[\phi_{a(D)}|D\right] = \alpha$ *almost surely for all* $\mu_D \in \mathbb{M}_D$. *Hence,* $E_{m,\mu_D}\left[\phi_{a(D)}\right] = \alpha$ *for all* $(m, \mu_D) \in H_0$ *and* $\phi_{a(D)}$ *is a similar test.*

While we could construct a family of CLC tests based on some conditional critical value function other than $c_\alpha(a)$ that does not impose conditional similarity, restricting attention to conditionally similar tests is a simple way to ensure correct size regardless of our choice of $a(D)$.

Interestingly, the class of QCLR tests is precisely the same as the class of CLC tests. Formally, for any function $r : D \to \mathbb{R}_+ \cup \{\infty\}$ define the quasi-CLR statistic $QCLR_r$ as in (11) and let $q_\alpha(r(d))$ be the $1-\alpha$ quantile of the random variable (12). We obtain the following result:

Theorem 2 *For any function* $r : D \to \mathbb{R}_+ \cup \{\infty\}$ *if we take* $\phi_{QCLR_r} = 1\{QCLR_r > q_\alpha(r(D))\}$ *then for* $\tilde{a}(D) = \frac{q_\alpha(r(D))}{q_\alpha(r(D))+r(D)}$ *we have that* $\phi_{QCLR_r} \equiv \phi_{\tilde{a}(D)}$. *Conversely, for any* $a : D \to [0,1]$ *there exists an* $\tilde{r} : D \to \mathbb{R}_+ \cup \{\infty\}$ *such that* $\phi_{a(D)} \equiv \phi_{QCLR_{\tilde{r}}}$. *Hence, the class of CLC tests for* $a : D \to [0,1]$ *is precisely the same as the class of Quasi-CLR tests for* $r : D \to \mathbb{R}_+ \cup \{\infty\}$.

Theorem 2 shows that the QCLR test ϕ_{QCLR_r} is a linear combination test with weight function $a(D) = \frac{q_\alpha(r(D))}{q_\alpha(r(D))+r(D)}$. In particular, this result establishes that the CLR test of Moreira (2003) for linear IV with a single endogenous regressor is a CLC test. In the remainder of the paper our exposition focuses on CLC tests but by Theorem 2 all of our results apply to QCLR tests as well.

5 Optimality of CLC Tests in a Conditional Problem

The CLC tests $\phi_{a(D)}$ represent only one of many ways to combine the S, K, and D statistics. Nonetheless this class has a number of optimal power properties in the problem obtained by conditioning on D. In this section we show that CLC tests are admissible in this conditional problem as well as locally most powerful against particular sequences of alternatives and weighted average power maximizing for a continuum of weight functions. As a side result, we give a simple characterization of an essentially complete class of tests for $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m \in \mathcal{M}(\mu), \mu \in \mathbb{M}$ based on (S, K, D).

Conditional on the event $D = d$ (for d full rank), J and K are independent and distributed $\chi^2_{k-p}(\tau_J(d,m))$ and $\chi^2_p(\tau_K(d,m))$, respectively, for τ_J and τ_K as defined in (13). Once we condition on $D = d$ the non-centrality parameters τ_J and τ_K are fixed, though unknown, values and our null hypothesis $H_0 : m = 0, \mu \in \mathbb{M}$ can be re-written as $H_0 : \tau_J = \tau_K = 0$. Our first task is to characterize the set of possible values for the non-centrality parameters (τ_J, τ_K) under the

alternative H_1. Let $\mathbb{M}_D(d)$ denote the set of values $\mu_D \in \mathbb{M}_D$ such that d is in the support of D.[5] Letting $\widetilde{\mathcal{M}}(d) = \cup_{\mu_D \in \mathbb{M}_D(d)} \mathcal{M}(\mu_D)$, we see that conditional on $D = d$, m may take any value in $\widetilde{\mathcal{M}}(d)$ and still be consistent with both $m \in \mathcal{M}(\mu_D)$ and d lying in the support of D. Hence, the non-centrality parameters (τ_J, τ_K) may take any value in the set

$$\mathcal{T}(d) = \cup_{m \in \widetilde{\mathcal{M}}(d)} \left(\tau_J(d,m), \tau_K(d,m)\right).$$

This set $\mathcal{T}(d)$ plays a similar role to $\mathcal{M}(\mu)$ in the original problem (2). In particular, conditional on $D = d$ our problem becomes one of testing $H_0 : \tau_J = \tau_K = 0$ against the alternative $H_1 : (\tau_J, \tau_K) \in \mathcal{T}(d) \backslash \{0\}$ based on observing $(J, K) \sim \left(\chi^2_{k-p}(\tau_J), \chi^2_p(\tau_K)\right)$.

5.1 CLC Tests are Admissible in the Conditional Problem

We say that a test ϕ is admissible if there is no other test $\tilde{\phi}$ with size less than or equal to ϕ and power greater than or equal to ϕ at all points and strictly higher power (or smaller size) at some point (that is, if there is no test $\tilde{\phi}$ that dominates ϕ). A result from Marden (1982) establishes that the class of admissible tests in the conditional problem has a simple form when $\mathcal{T}(d) = \mathbb{R}^2_+$.

Theorem 3 *(Marden 1982) Conditional on $D = d$, let $J \sim \chi^2_{k-p}(\tau_J)$ and $K \sim \chi^2_p(\tau_K)$ be independent and let ϕ be a test of $H_0 : \tau_J = \tau_K = 0$ against $H_1 : (\tau_J, \tau_K) \in \mathbb{R}^2_+ \backslash \{0\}$. ϕ is admissible in the conditional problem given $D = d$ if and only if it is almost surely equal to $1\left\{\left(\sqrt{J}, \sqrt{K}\right) \notin C_d\right\}$ for some set C_d satisfying*

1. *C_d is closed and convex*
2. *C_d is monotone decreasing: i.e. $x \in C_d$ and $y_i \leq x_i \ \forall i$ implies $y \in C_d$*

Thus, a test ϕ is admissible in the conditional problem if and only if its acceptance region in $\left(\sqrt{J}, \sqrt{K}\right)$ space is almost-everywhere equal to a closed, convex, monotone decreasing set. Using this result, it is straightforward to show that CLC tests are admissible in the conditional problem for all $a : D \to [0,1]$ and all d.

Corollary 1 *For all weight functions $a : D \to [0,1]$ the CLC test $\phi_{a(D)}$ is admissible in the problem conditional on $D = d$ for all d.*

The test ϕ_{JK} defined in (10) is likewise admissible for all d and all α_J and α_K.

[5] If Σ_D is full rank then $\mathbb{M}_D(d) = \mathbb{M}_D$, since the support of D is the same for all $\mu_D \in \mathbb{M}_D$, but if Σ_D is reduced rank (e.g. $\Sigma_D = 0$) then we may have $\mathbb{M}_D(d) \subset \mathbb{M}_D$.

5.1.1 An Essentially Complete Class of Tests

It is important to note that Theorem 3 concerns only admissibility in the problem where we have conditioned on $D = d$. Admissibility in this conditional problem for all values d is not sufficient for admissibility as a test of $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m = \mathcal{M}(\mu) \backslash \{0\}, \mu \in \mathbb{M}$ in the original problem. However, the set of tests which are admissible in the conditional problem for almost all d form an essentially complete class: that is, for any test ϕ we can find a test $\tilde{\phi}$ which is weakly better than ϕ and is admissible in the conditional problem for almost every d (with respect to the distribution of D, F_D).[6]

Corollary 2 *For any test ϕ, there exists a test $\tilde{\phi}$ which is admissible in the conditional problem F_D-almost-everywhere such that for all $\mu \in \mathbb{M}$, $E_{m=0,\mu}\left[\tilde{\phi}\right] \leq E_{m=0,\mu}\left[\phi\right]$ and for all $m \neq 0$ $E_{m,\mu}\left[\tilde{\phi}\right] \geq E_{m,\mu}\left[\phi\right]$. Thus, the set of tests admissible conditional on $D = d$ for almost every d forms an essentially complete class.*

Since CLC tests are admissible in the conditional problem for all d they belong to this essentially complete class.

5.2 Local and Weighted Average Power Optimality of CLC Tests

While the admissibility of CLC tests in the conditional problem is certainly a desirable property, the essentially complete class of tests defined by Corollary 2 is quite large and contains many procedures. Here we show that CLC tests have additional optimality properties in the conditional problem not shared by these other tests. Specifically, we show that CLC tests are locally most powerful against sequences of alternatives approaching $(\tau_J, \tau_K) = 0$ linearly and weighted average power maximizing for a continuum of weight functions in the conditional problem. Weighted average power is a commonly-used optimality criterion in statistics and econometrics: see Lehman and Romano (2005) for discussion of this approach, and Mueller (2011), Elliott, Mueller, and Watson (2012), and Olea (2012) for recent applications in econometrics and extensive references.

Theorem 4 *Fix a conditional linear combination test $\phi_{a(D)}$ and a value d. Let $\Phi_\alpha(d)$ denote the class of tests which have size α conditional on $D = d$, $E_{m=0,\mu}\left[\phi | D = d\right] = \alpha \forall \phi \in \Phi_\alpha(d)$.*

[6] We could alternatively state this result requiring admissibility for all d, but do not wish to rule out tests which have poor behavior on sets of measure zero, e.g. when d is singular.

1. *Let* $(\tau_J, \tau_K) = \lambda \cdot \left(a(d)\frac{k-p}{p}, 1\right)$. *For any test* $\phi \in \Phi_\alpha(d)$ *there exists* $\bar{\lambda} > 0$ *such that if* $0 < \lambda < \bar{\lambda}$,

$$E_{(\tau_J,\tau_K)}[\phi | D = d] \leq E_{(\tau_J,\tau_K)}\left[\phi_{a(D)} | D = d\right].$$

2. *Let* $F_{t_J,t_K}(\tau_J, \tau_K)$ *be the distribution function corresponding to* $(\tau_J, \tau_K) \sim \left(t_J \cdot \chi^2_{k-p}, t_K \cdot \chi^2_p\right)$. *For any* (t_J, t_K) *with* $\frac{t_J}{t_K}\frac{t_K+1}{t_J+1} = a(d)$ *the conditional linear combination test* $\phi_{a(D)}$ *solves the conditional weighted average power maximization problem*

$$\phi_{a(D)} \in \arg\max_{\phi \in \Phi_a(d)} \int E_{(\tau_J,\tau_K)}\left[\phi_{a(D)} | D = d\right] dF_{t_J,t_K}(\tau_J, \tau_K).$$

Theorem 4 follows immediately from results in Monti and Sen (1976) and Koziol and Perlman (1978) on the optimal combination of independent non-central χ^2 random variables. Theorem 4(1) shows that conditional on $D = d$ the CLC test $\phi_{a(D)}$ is locally most powerful against sequences of alternatives with $\tau_J/\tau_K = a(d)\frac{k-p}{p}$, in the sense that it has power at least as good as any other test ϕ once we come sufficiently close to the null along this direction. Theorem 4(2) establishes that $\phi_{a(D)}$ maximizes weighted average power in the conditional problem for a continuum of different weight functions corresponding to scaled χ^2 distributions.

Together, the results of this section establish that conditional linear combination tests $\phi_{a(D)}$ have a number of desirable power properties conditional on $D = d$. We can see, however, that the weight function $a(D)$ plays a central role in determining the power of $\phi_{a(D)}$ in the conditional problem, for example determining against which directions CLC tests are locally most powerful and against which weight functions they maximize weighted average power. Thus, the central question in choosing from the class of CLC tests is what weight function to use. We turn to this question in the next section.

6 Optimal CLC Tests

For any weight function $a : D \to [0, 1]$ we can define a CLC test $\phi_{a(D)}$ for H_0 against H_1 using (16). While any such test controls size by Theorem 1, the class of such CLC tests is large and we would like a systematic way to pick weight functions a yielding tests with good power properties.

A natural optimality criterion, after restricting attention to the class of CLC tests, is minimax regret: see Stoye (2009) for an introduction to this approach and extensive discussion of its recent application in economic theory and econometrics. To define a minimax regret CLC test, for any

$(m, \mu_D) \in H_1$, define $\beta^*_{m,\mu_D} = \sup_{a \in \mathcal{A}} E_{m,\mu_D}\left[\phi_{a(D)}\right]$ for $\mathcal{A}$ the class of Borel-measurable functions $a : D \to [0,1]$. β^*_{m,μ_D} gives the highest attainable power against alternative (m, μ_D) in the class of CLC tests and, as we vary (m, μ_D), defines the power envelope for this class. For a given $a \in \mathcal{A}$ we can then define the regret associated with $\phi_{a(D)}$ against alternative (m, μ_D) as $\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]$, which is the amount by which the power of the test $\phi_{a(D)}$ falls short of the highest power we might have attained against this alternative by choosing some other CLC test. We can then define the maximum regret for a test $\phi_{a(D)}$ as

$$\sup_{(m,\mu_D) \in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right),$$

which is the largest amount by which the power function of $\phi_{a(D)}$ falls short of the power envelope for the class of CLC tests or, equivalently, the sup-norm distance between the power function of $\phi_{a(D)}$ and the power envelope. A minimax regret choice of $a \in \mathcal{A}$ is

$$a_{MMR} \in \arg\min_{a \in \mathcal{A}} \sup_{(m,\mu_D) \in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right).$$

A minimax regret test $\phi_{MMR} = \phi_{a_{MMR}(D)}$ is one whose power function is as close as possible to the power envelope for the class of conditional linear combination tests in the sup norm. As an optimality criterion this is an intuitive choice: having already restricted attention to the class of CLC tests, focusing on MMR tests minimizes the maximal extent to which the test we choose could under-perform relative to other CLC tests we might have picked.

Given the way that MMR tests are defined, it is not obvious that an MMR test exists, i.e. that

$$\inf_{a \in \mathcal{A}} \sup_{(m,\mu_D) \in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right)$$

is achieved by any $a \in \mathcal{A}$. Theorem 5 establishes that an MMR test ϕ_{MMR} always exists.

Theorem 5 *For any non-empty set of alternatives $H_1 : m = \mathcal{M}_D(\mu_D) \backslash \{0\}, \mu_D \in \mathbb{M}_D$ in the limit problem (2) there exists a weight function $a^* \in \mathcal{A}$ such that*

$$\sup_{(m,\mu_D) \in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a^*(D)}\right]\right) = \inf_{a \in \mathcal{A}} \sup_{(m,\mu_D) \in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right).$$

Hence, an MMR test $\phi_{MMR} = \phi_{a^*(D)}$ exists.

Example II: Minimum Distance (Continued) Calculating the MMR test in Example II is straightforward. In particular D is non-random and $a(D) \equiv a(\mu)$, so rather than picking a function from D to $[0,1]$ we are simply picking a number a in $[0,1]$. Hence, we have that $\beta^*_{m,\mu_D} = \sup_{a\in[0,1]} E_{m,\mu_D}[\phi_a]$. Moreover, we know that in this example $m = m(\theta) = \Omega_\eta^{-\frac{1}{2}}(f(\theta) - f(\theta_0))$ and $\mu_D = \mu = \Omega_\eta^{-\frac{1}{2}} \frac{\partial}{\partial\theta} f(\theta_0)$, so the maximum attainable power against alternative θ is simply $\beta^*_\theta = \sup_{a\in[0,1]} E_{m(\theta),\mu}[\phi_a]$ which we can calculate for any value θ. To solve for the MMR test ϕ_{MMR}, we need only calculate

$$a_{MMR} = \arg\min_{a\in[0,1]} \sup_{\theta\in\Theta} \left(\beta^*_\theta - E_{m(\theta),\mu}[\phi_a]\right). \square$$

6.1 Plug-in Minimax Regret Tests

While finding the MMR test is straightforward in Example II, Example I is less tractable in this respect. In this example D is random and as a result rather than merely optimizing over numbers in $[0,1]$ solving for ϕ_{MMR} requires that we optimize over the set $\mathcal{A}$ of functions. In most cases finding even an approximate solution to this optimization problem is extremely computationally costly, rendering ϕ_{MMR} unattractive in many applications. To overcome this difficulty we suggest a computationally tractable class of plug-in tests.

There are two aspects of Example II that make calculating ϕ_{MMR} straightforward. First, rather than optimizing over the space of measurable functions $\mathcal{A}$ we need only optimize over numbers in $[0,1]$. Second, $\mu = \mu_D$ is known so in solving the minimax problem we need only search over $\theta \in \Theta$ rather than over some potentially higher dimensional space of values for $(m, \mu_D) \in H_1$.

To construct a test for the general case with similarly modest computational requirements, suppose first that μ_D is known. Let us restrict attention to unconditional linear combination tests with $a(D) \equiv a(\mu_D) \in [0,1]$, where we write a as a function of μ_D to emphasize its dependence on this parameter. The power envelope for this class of *unconditional* linear combination tests is $\beta^u_{m,\mu_D} = \sup_{a\in[0,1]} E_{m,\mu_D}[\phi_a]$. The minimax regret unconditional (MMRU) test $\phi_{MMRU} = \phi_{a_{MMRU}(\mu_D)}$ can then be derived by finding

$$a_{MMRU}(\mu_D) \in \arg\min_{a\in[0,1]} \sup_{m\in\mathcal{M}_D(\mu_D)} \left(\beta^u_{m,\mu_D} - E_{m,\mu_D}[\phi_a]\right).$$

Just as when we derived ϕ_{MMR} for Example II above, here we need only optimize over $a \in [0,1]$ and $m \in \mathcal{M}_D(\mu_D)$, rather than over $a \in \mathcal{A}$ and $(m, \mu_D) \in H_1$.

In defining ϕ_{MMRU} we assumed that μ_D was known, which is unlikely to hold in contexts like Example I where D is random. Note, however, that for any estimator $\hat{\mu}_D$ which depends only on D, $a_{MMRU}(\hat{\mu}_D)$ can be viewed as a particular weight function $a(D)$ and the plug-in minimax regret (PI) test

$$\phi_{PI} = \phi_{a_{PI}(D)} = 1\{K + a_{MMRU}(\hat{\mu}_D) \cdot J > c_\alpha(a_{MMRU}(\hat{\mu}_D))\}$$

is a CLC test and so controls size by Theorem 1. Moreover, to calculate this test all we need to do is solve for a_{MMRU} taking the estimate $\hat{\mu}_D$ to be the true value, so this test remains quite computationally tractable.

It is important to note that ϕ_{PI} is not in general a true MMR test. First, ϕ_{PI} treats the estimated value $\hat{\mu}_D$ as the true value, and hence does not account for any uncertainly in the estimation of μ_D. Second, even taking the value μ_D as given ϕ_{PI} restricts attention to unconditional linear combination tests, which represent a strict subset of the possible functions $a \in \mathcal{A}$. Despite these potential shortcomings, we find in Sections 7 and 9 below that PI tests perform quite well in weakly identified simulation examples, and show in Section 8 that PI tests will be asymptotically optimal under strong identification in our examples.

To use PI tests in a given context we need only choose the estimator $\hat{\mu}_D$. While the MLE for μ_D based on D, $\hat{\mu}_D = D$, is a natural choice we may be able to do better in many contexts. In particular, in weak IV (Example I) with homoskedastic errors estimation of $\hat{\mu}_D$ is related to a problem of non-centrality parameter estimation, allowing us to use results from that literature.

Example I: Weak IV (Continued) Consider again the case studied by AMS where $\Omega = A \otimes B$ has Kronecker product structure. Results in AMS show that $(J, K, D'\Sigma_D^{-1}D)$ is a maximal invariant under rotations of the instruments, where $D'\Sigma_D^{-1}D \sim \chi^2_k\left(\mu_D'\Sigma_D^{-1}\mu_D\right)$.[7] AMS show that the distribution of $(J, K, D'\Sigma_D^{-1}D)$ depends on $c = \sqrt{T}\pi_T$ only through the non-centrality parameter $r = \mu_D'\Sigma_D^{-1}\mu_D$, which controls the strength of identification.

Note that the MLE $\hat{\mu}_D = D$ for μ_D based on D implies a severely biased estimator for r, $\hat{r} = D'\Sigma_D^{-1}D$, since

$$E[\hat{r}] = E\left[D'\Sigma_D^{-1}D\right] = r + k.$$

The problem of estimating r relates to the well-studied problem of estimating the non-centrality parameter of a non-central χ^2 distribution and a number of different estimators have been proposed

[7] It suffices to note that $(J, K, D'\Sigma_D^{-1}D)$ is a one-to-one transformation of Q as defined in AMS.

for this purpose, including $\hat{r}_{MLE}$, the MLE for r based on $\hat{r}$ (which is not available in closed form) and $\hat{r}_{PP} = \max\{\hat{r} - k, 0\}$ which is the positive part of the bias corrected estimator $\hat{r} - k$.[8] Both $\hat{r}_{MLE}$ and $\hat{r}_{PP}$ are zero for a range of values $\hat{r} > 0$ so we also consider an estimator proposed by Kubokawa, Robert and Saleh (1993),

$$\hat{r}_{KRS} = \hat{r} - k + e^{-\frac{\hat{r}}{2}} \left(\sum_{j=0}^{\infty} \left(-\frac{\hat{r}}{2} \right)^j \frac{1}{j!\,(k+2j)} \right)^{-1}$$

which is smooth in $\hat{r}$ and greater than zero whenever $\hat{r} > 0$. We show in Section 7 below that estimators $\hat{\mu}_D$ corresponding to all three non-centrality estimators $\hat{r}_{MLE}$, $\hat{r}_{PP}$, and $\hat{r}_{KRS}$ yield PI tests ϕ_{PI} with good power properties.□

6.2 Implementing MMRU and PI Tests

The weight functions $a_{MMRU}(\mu_D)$ and $a_{PI}(D)$ are not typically available in closed form, so to implement MMRU and PI tests we need to approximate these weight functions numerically. Since PI tests are simply MMRU tests that plug in an estimate for μ_D, we focus on the problem of evaluating MMRU tests. For details on our application of this approach to the heteroskedastic linear IV model discussed in the next section, see Appendix D.

Calculating a good approximation to the weight function a_{MMRU} is straightforward when $\mathcal{M}_D(\mu_D)$ is compact. In particular, since μ_D and Σ_D are both known in the MMRU problem, the only unknown parameter that affects the power of linear combination tests ϕ_a is m. Moreover, it is clear that $E_{m,\mu_D}[\phi_a]$ is continuous in (m, a). Thus, for any compact set of values $\mathcal{M}_D(\mu_D)$ we can see that if we take sufficiently fine grids of values $M \subset \mathcal{M}_D(\mu_D)$ and $A \subset [0,1]$, the approximate value for $a_{MMRU}(\mu_D)$

$$a^*_{MMRU}(\mu_D) = \arg\min_{a \in A} \sup_{m \in M} \left(\sup_{\tilde{a} \in A} E_{m,\mu_D}[\phi_{\tilde{a}}] - E_{m,\mu_D}[\phi_a] \right)$$

will have maximum regret arbitrarily close to that of the true MMRU choice

$$\sup_{m \in \mathcal{M}_D(\mu_D)} \left(\beta^u_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a^*_{MMRU}(\mu_D)} \right] \right) \approx \min_{a \in [0,1]} \sup_{m \in \mathcal{M}_D(\mu_D)} \left(\beta^u_{m,\mu_D} - E_{m,\mu_D}[\phi_a] \right).$$

To calculate $a^*_{MMRU}(\mu_D)$, however, it suffices to evaluate $E_{m,\mu_D}[\phi_a]$ for all $(a, m) \in A \times M$,

[8] $\hat{r}_{PP}$ has been shown to dominate $\hat{r}_{MLE}$ in terms of mean squared error but is itself inadmissible (Saxena and and Alam, 1982).

which can be done by simulation.[9] Once we have these values, they immediately imply a value for $a^*_{MMRU}(\mu_D)$ which we can use to evaluate the MMRU test.

Even if the initial set of alternatives $\mathcal{M}_D(\mu_D)$ is non-compact, it is typically reasonable to restrict attention to some bounded neighborhood of the null, for example the set of alternatives against which the S test ϕ_S attains power less than some pre-specified level β. This amounts, however, to considering $\widetilde{\mathcal{M}}_D(\mu_D) = \mathcal{M}_D(\mu_D) \cap B_{\tilde{C}}$ where $B_C = \left\{m \in \mathbb{R}^k : ||m||^2 < C\right\}$ and $\tilde{C}$ solves $Pr\left\{\chi^2_k\left(\tilde{C}\right) > \chi^2_{k,1-\alpha}\right\} = \beta$. Once we restrict attention to $\widetilde{\mathcal{M}}_D(\mu_D)$ we can approximate $a_{MMRU}(\mu_D)$ as discussed above.

7 Performance of PI Tests in Weak IV

In this section, we examine the performance of PI tests in linear IV with weak instruments (Example I). We begin by considering the model studied by AMS, whose assumptions imply Kronecker product structure for the covariance matrix Ω. Since data encountered in empirical practice commonly violate this assumption, we then consider the performance of PI tests in a model calibrated to match the heteroskedastic time-series data used by Yogo in his (2004) study on the effect of weak instruments on estimation of the elasticity of inter-temporal substitution.

7.1 Homoskedastic Linear IV

AMS consider the linear IV model in Example I under the additional restriction that the errors V are independent of Z, normal, and independent across t with $corr(V_{1,t}, V_{2,t}) = \rho$. AMS show that under these restrictions the CLR test of Moreira (2003) is nearly uniformly most powerful in a class of two-sided tests invariant to rotations of the instruments, in the sense that the power function of the CLR test is uniformly close to the power envelope for this class. Mueller (2011) notes that using his Theorems 1 and 2 one can extend this result to show that the CLR test is nearly asymptotically uniformly most powerful in the class of all invariant two-sided tests that have correct size under the weak convergence assumption (6) with the additional restriction that $\Omega = A \otimes B$ for A and B symmetric positive-definite matrices of dimension 2×2 and $k \times k$, respectively. As Mueller notes, matrices Ω of this form arise naturally only for serially uncorrelated homoskedastic IV models, limiting the applicability of this result. Nonetheless, the asymptotic optimality of CLR under the assumption that $\Omega = A \otimes B$ provides a useful benchmark against which to evaluate the performance

[9] For this step it is helpful to note that $E_{m,\mu_D}[\phi_a] = \int E_{\tau_J(D),\tau_K(D)}[\phi_a]\, dF_D$, so we can tabulate $E_{\tau_J,\tau_K}[\phi_a]$ in advance and then calculate this integral by simulation. See Appendix D for discussion of this approach.

of ϕ_{PI}. In particular, by Theorem 2 the CLR test is a CLC test. Hence, if our plug-in minimax regret approach for selecting CLC tests is to work well in this benchmark case, it should match the near-optimal performance of the CLR test.

We begin by directly comparing the CLR test's weight function $a_{CLR}(D)$ to the weight functions implied by the plug-in MMR approach for the different estimates for $r = \mu_D' \Sigma_D^{-1} \mu_D$ described in Section 6.1. Next, we simulate the power of the various PI tests considered and show that tests based on reasonable estimators for r match the near-optimal performance of the CLR test.

7.1.1 Weight Function Comparison

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

Lemma 1 *The function $a_{MMRU}(\mu_D)$ in the limit problem (6) with*

$$\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix} \otimes I$$

depends only on $r = \mu_D' \Sigma_D^{-1} \mu_D$.

Note that the assumption

$$\Omega = A \otimes B = \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \otimes B$$

implies that Σ is of the form discussed in Lemma 1, so we can write our MMRU weight function $a_{MMRU}(\mu_D)$ as a function of r alone, $a_{MMRU}(r)$. Hence, since the weights of both the CLR test and the plug-in approaches discussed in Section 6.1 depend on $\hat{r}$ alone, we can plot the values of $a_{CLR}(\hat{r})$, $a_{MMRU}(\hat{r})$, $a_{MMRU}(\hat{r}_{MLE})$, $a_{MMRU}(\hat{r}_{PP})$, and $a_{MMRU}(\hat{r}_{KRS})$ as functions of $\hat{r}$. Since the weight functions also depend on the number of instruments, we plot these functions of $\hat{r}$ for $k = 2$, $k = 5$, and $k = 10$ in Figures 1-3. As these plots make clear all the weight functions show similar qualitative behavior, placing large weight on S for small values of $\hat{r}$ and increasing the weight on K as $\hat{r}$ grows, but there are some notable differences. Perhaps most pronounced, $a_{MMRU}(\hat{r})$ is typically lower than any of the other functions, and the difference seems to be increasing in k. This is intuitively reasonable given that $\hat{r}$ tends to overestimate r and that this bias is increasing

in k. As previously noted both $\hat{r}_{MLE}$ and $\hat{r}_{PP}$ are zero for a range of strictly positive values $\hat{r}$, and this range is clearly visible in the plots. Also notable, for $k = 5$ and $k = 10$ we have that $a_{MMRU}(0) < 1$: this reflects the fact that for large values of k the MMRU test down-weights S even when $r = 0$ due to its high degrees of freedom.

7.1.2 Power Simulation

To compare the power of the PI tests to that of the CLR test, we follow the simulation design of AMS and consider a homoskedastic normal model with a known reduced-form covariance matrix. Following AMS we consider models with five instruments, reduced-form error correlation ρ equal to 0.5 or 0.95, and concentration (identification strength) parameter $\lambda = \mu'\mu$ equal to 5 and 20. To examine the effect of changing the number of instruments, like AMS we also consider models with two and ten instruments, in each case fixing ρ equal to 0.5 and letting λ equal 5 and 20. The power functions of the resulting tests (based on 10,000 simulations) are given in Figures 4 and 5.

As these figures illustrate, with the exception of the test using the badly biased estimator $\hat{r}$ for r the PI tests match the near-optimal performance of the CLR test in all cases. Since the power differences between the CLR and PI tests are for the most part quite small, and thus difficult to see in the figures, in Table 1 we report the maximal distance from the power functions of the CLR, PI, AR, and K tests to the power envelope for the class consisting of these tests alone. In particular, for each k in $\{2, 5, 10\}$ we calculate the point-wise maximal power of the tests studied at each of the alternatives used to generate Figures 4 and 5. For each test, we then find the largest margin by which the power of that test falls short of point-wise maximal power. Hence, the values reported in Table 1 can be viewed as a measure of the maximum regret, where we restrict the set of tests and parameter values to those used to generate Figures 4 and 5. As Table 1 makes clear, most of the PI tests considered match the near-optimal performance of the CLR test. We again see that the one exception is the PI test using the highly biased estimator $\hat{r}$ for r, which tends to overweight the K statistic and consequently under-performs relative to the other tests considered. As these results highlight, in one of the only weakly identified contexts where a near-UMP test is known, reasonable implementations of the plug-in testing approach suggested in this paper are near-optimal as well.

7.2 Linear IV with Unrestricted Covariance Matrix

The near-optimal performance of PI tests in linear IV models where Ω has Kronecker product structure is promising, but is of limited relevance for empirical work. Economic data frequently

exhibit of heteroskedasticity, serial dependence, clustering, and other features that render a Kronecker structure assumption for Ω implausible. It is natural to ask whether PI tests continue to have good power properties in this more general case. For comparison, we consider Kleibergen (2005)'s quasi-CLR test, which takes $r(D) = D'\Sigma_D^{-1}D$ and can be viewed as a heteroskedasticity and autocorrelation-robust version of the CLR test, as well as the K and Anderson Rubin (S) statistics.[10]

There are a multitude of ways in which a Kronecker structure assumption on Ω might be violated. As noted in Section 3.4, for some values of μ and Σ the K statistic may focus on directions irrelevant to the parametric hypothesis of interest, leading K tests to have low power. Moreover, when Σ is unrestricted these issues can arise in models with μ_D large and Σ_D small, where $r(D) = D'\Sigma_D^{-1}D$ will tend to be large. Since the $QCLR$ statistic behaves much like K for $r(D)$ large it seems that, unlike in the Kronecker case, in models with Σ unrestricted the QCLR test may have low power against some alternatives. Less clear, however, are whether such power problems for the QCLR test arise at empirically plausible parameter values and whether PI tests are successful in correcting these issues. To asses the relative performance of these tests at parameter values relevant for empirical practice we calibrate a simulation based on data from Yogo (2004). Yogo considers estimation of the elasticity of inter-temporal substitution in eleven developed countries using linear IV and argues that estimation of this parameter appears to suffer from a weak instruments problem.[11] Yogo notes that both the strength of identification and the degree of heteroskedasticity appears to vary across countries, making his data-set especially interesting for our purposes since it allows us to explore the behavior of the tests considered for a range of empirically relevant parameter values.

Unfortunately, unlike in the homoskedastic case above $a_{PI}(D)$ may depend on all of D, rather than only on the scalar $\hat{r}$. Consequently visual comparison of $a_{PI}(D)$ to the weight function $a_{QCLR}(D)$ implied by the QCLR test is impractical and we move directly to our simulation results.

[10] Moreira and Moreira (2013) also consider hypothesis testing in linear IV models with non-Kroneker Ω as an example, and discuss numerical approximation of optimal tests over a class (their SU tests) which includes the CLC tests. Comparison of their tests to the PI tests suggested here is of considerable interest for future work.

[11] The countries considered are Australia, Canada, France, Germany, Italy, Japan, the Netherlands, Sweden, Switzerland, the United Kingdom, and the United States. For comparability we use Yogo's quarterly data for all countries, which in each case covers a period beginning in the 1970's and ending in the late 1990's.

7.2.1 Power Simulation

We focus on simulating the behavior of tests in the weak IV limit problem (6), and so require estimates for μ and Ω. To obtain these estimates, for each of the 11 countries in Yogo's data we calculate $\hat{\mu}$ and $\hat{\Omega}$ based on two-stage least squares estimates for the elasticity of inter-temporal substitution, where $\hat{\Omega}$ is a Newey-West covariance matrix estimator using three lags.[12] A detailed description of this estimation procedure, together with the implementation of all tests considered, is given in Appendix D. In particular, for the PI test we consider an estimator $\hat{\mu}_D$ which corresponds to the positive-part non-centrality estimator $\hat{r}_{PP}$ in the homoskedastic case discussed above. The resulting power curves (based on 10,000 simulations) are plotted in Figures 6-8. Since for many countries the power curves are difficult to distinguish visually, in Table 2 we list the maximum regret for each test relative to the other tests over the range of parameter values considered for each country, repeating the same exercise described above for the homoskedastic case.

Both the figures and the table highlight that while for many of the countries the K, QCLR, and PI tests all perform well, as in the homoskedastic case there are some parameter values where the K test suffers from substantial declines in power relative to the other tests. In contrast to the homoskedastic case, however, the QCLR test does not fully resolve these issues. Instead, for parameter values where the K test exhibits especially large power declines, as in the simulations calibrated to match data from Japan and the United Kingdom, the QCLR test suffers from power loss as well. While the QCLR test reduces power loss relative to the K test, the PI test does substantially better, and has the smallest maximal power shortfall of any of the tests considered. While the power of the AR test is stable, for all countries its maximal power shortfall exceeds 10%.

The relatively poor performance of the QCLR test is driven by the fact, discussed above, that in the non-homoskedastic case the K statistic may focus on directions yielding low power. Since Kleibergen's QCLR test uses the CLR weight function, which is optimal in the homoskedastic case, it does not account for the fact that K may have even worse performance when Σ lacks Kronecker product structure. In contrast, the PI test takes both the structure of Σ and the estimated value $\hat{\mu}_D$ into account when calculating $a_{PI}(D)$, and so performs well in both the homoskedastic and non-homoskedastic cases. Unlike in the homoskedastic case we can see that that none of the tests considered is even approximately uniformly most powerful, but the PI test delivers powerful, stable performance.

[12] While the model assumptions imply that the GMM residuals $f_t(\beta)$ are serially uncorrelated at the true parameter value, the derivatives of the moment conditions $\frac{\partial}{\partial \beta} f_t(\beta)$ may be serially dependent.

8 Asymptotic Properties of CLC Tests

The results of Sections 3-7 treat the limiting random variables $(g, \Delta g, \gamma)$ as observed and consider the problem of testing $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m \in \mathcal{M}(\mu)\backslash\{0\}, \mu \in \mathbb{M}$. In this section, we show that under mild assumptions our results for the limit problem (2) imply asymptotic results along sequences of models satisfying (1). We first introduce a useful invariance condition for the weight function a and then prove results concerning the asymptotic size and power of CLC tests.

8.1 Postmultiplication Invariant Weight Functions

Let us introduce finite-sample analogs to the limiting random variables S, D, K, and J

$$S_T = g_T' g_T$$

$$vec\left(D_T\right) = vec\left(\Delta g_T\right) - \hat{\Sigma}_{\theta_i g} g_T$$

$$K_T = g_T' D_T \left(D_T' D_T\right)^{-1} D_T' g_T$$

$$J_T = S_T - K_T.$$

We previously wrote the weight functions a of CLC tests as functions of D alone, since in the limit problem the parameter γ is fixed and known. In practice, however, we will plug in the estimator $\hat{\gamma}$ for γ so in this section it is helpful to instead write $a(D, \gamma)$. Likewise, since the estimator $\hat{\mu}_D$ used in plug-in tests may depend on γ in this section we will write it as $\hat{\mu}_D\left(D, \gamma\right)$.

Our weak convergence assumption (1), together with the continuous mapping theorem, implies that $D_T \to_d D$ for D normally distributed, where we assume that D is full rank almost surely for all $(\theta, \gamma) \in \Theta \times \Gamma$. In many applications such convergence will only hold if we choose an appropriate normalization when defining Δg_T, which may seem like an obstacle to applying our approach. In the linear IV model for instance, the appropriate definition for Δg_T will depend on the strength of identification.

Example I: Weak IV (Continued) In Section 2 we assumed that the instruments were weak, with $\pi_T = \frac{c}{\sqrt{T}}$, and showed that $\Delta g_T = \sqrt{T}\hat{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial \beta} f_T(\beta_0)$ converged in distribution. If on the other hand the instruments are strong, $\pi_T = \pi_1$ and $||\pi_1|| > 0$, then $\frac{\partial}{\partial \beta} f_T(\beta) \to_p E\left[X_t Z_t\right] \neq 0$ so $\sqrt{T}\hat{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial \beta} f_T(\beta_0)$ diverges and we should instead take $\Delta g_T = \hat{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial \beta} f_T(\beta_0)$.□

This apparent dependence on normalization is not a problem, however, since CLC tests are typically invariant to renormalization of $(g_T, \Delta g_T, \hat{\gamma})$. In particular, for A any full rank $p \times p$ matrix consider the transformations

$$h_{\Delta g}(\Delta g_T; A) = \Delta g_T A$$

$$h_{\Sigma}(\Sigma; A) = \left(\begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} \otimes I_k \right)' \Sigma \left(\begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} \otimes I_k \right)$$

and let $h_{\gamma}(\gamma; A)$ be the transformation of γ such that $\Sigma(h_{\gamma}(\gamma; A)) = h_{\Sigma}(\Sigma(\gamma); A)$. Let

$$h(g_T, \Delta g_T, \hat{\gamma}; A) = (g_T, h_{\Delta g}(\Delta g_T; A), h_{\gamma}(\hat{\gamma}; A)) \tag{17}$$

and note that the statistics J_T and K_T are invariant to this transformation for all full rank matrices A, in the sense that their values based on $(g_T, \Delta g_T, \hat{\gamma})$ are the same as those based on $h(g_T, \Delta g_T, \hat{\gamma}; A)$. Thus if we choose a weight function $a(D, \gamma)$ that is also invariant, the CLC test $\phi_{a(D_T, \hat{\gamma})}$ will be invariant to transformations of the form (17). Formally, we say that the weight function $a(D, \gamma)$ is invariant to postmultiplication if for all full-rank $p \times p$ matrices A we have

$$a(D, \gamma) = a(h_{\Delta g}(D; A), h_{\gamma}(\gamma; A)),$$

where we have used the fact that D calculated using $h(g, \Delta g, \gamma; A)$ is equal to $h_{\Delta g}(D; A)$.

Invariance to postmultiplication is a useful property, since to obtain results for invariant tests based on $(\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\gamma})$ it will suffice that there exist some sequence of matrices A_T such that

$$(g_T, \Delta g_T, \hat{\gamma}) = h(\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\gamma}; A_T)$$

satisfies the weak convergence assumption (1), without any need to know the correct sequence A_T for a given application. In the linear IV example discussed above, for instance, we can take Δg_T as originally defined and consider $A_T = 1$ under weak identification and $A_T = \frac{1}{\sqrt{T}}$ under strong identification, and thus can make use of results derived under the convergence assumption (6) without knowing the identification strength in a given context.

The class of postmultiplication-invariant weight functions a is quite large, and includes all the weight functions discussed above for tests based on $(g_T, \Delta g_T, \hat{\gamma})$. In particular, since the

distribution of the J and K statistics is unchanged by such transformations and $h_{\Delta g}(D;A)$ is a one-to-one function of D, it is easy to see that we can choose the minimax regret weight function a_{MMR} to be invariant to postmultiplication. Likewise, provided we take the estimator $\hat{\mu}_D(D,\gamma)$ to be equivariant under transformation by h, so that $h_{\Delta g}(\hat{\mu}_D(D,\gamma);A) = \hat{\mu}_D\left(h_{\Delta g}(D;A), h_\gamma(\gamma;A)\right)$, we can see that the plug-in weight function a_{PI} will be invariant as well.

8.2 Asymptotic Size and Power of CLC Tests

Let $F(g,\Delta g,\gamma)$ denote the distribution of $(g,\Delta g,\gamma)$ in the limit problem, noting that since we assume that $\hat{\gamma}$ is consistent the marginal distribution for γ in the limit problem is a point mass. Since we have assumed that D is full rank almost surely we have that J and K are F-almost-everywhere continuous functions of $(g,\Delta g,\gamma)$, so the continuous mapping theorem implies

$$(J_T, K_T, D_T) \to_d (J, K, D).$$

To obtain asymptotic size control for the CLC test

$$\phi_{a(D_T,\hat{\gamma})} = 1\left\{(1 - a(D_T,\hat{\gamma}))\cdot K_T + a(D_T,\hat{\gamma})\cdot S_T > c_\alpha\left(a(D_T,\hat{\gamma})\right)\right\}$$

all we require is that a be almost-everywhere continuous with respect to F. Indeed, this test is asymptotically conditionally similar in the sense discussed by Jansson and Moreira (2006).

Proposition 1 *Assume $(g_T,\Delta g_T,\hat{\gamma})$ satisfies the weak convergence assumption (1) and let $a(D,\gamma)$ be almost-everywhere continuous with respect to the limit distribution $F(g,\Delta g,\gamma)$ for $(\theta_0,\gamma) \in \{\theta_0\}\times\Gamma$. Then under (θ_0,γ) we have that*

$$\lim_{T\to\infty} E_{T,(\theta_0,\gamma)}\left[\phi_{a(D_T,\hat{\gamma})}\right] = \alpha. \tag{18}$$

Moreover, for $\mathcal{F}$ the set of bounded functions $f(D)$ which are almost-everywhere continuous with respect to $F(g,\Delta g,\gamma)$ under (θ_0,γ),

$$\lim_{T\to\infty} E_{T,(\theta_0,\gamma)}\left[\left(\phi_{a(D_T,\hat{\gamma})} - \alpha\right) f(D_T)\right] = 0 \ \forall f \in \mathcal{F}. \tag{19}$$

It is important to note that Proposition 1 only establishes sequential size control, and depending on the underlying model establishing uniform size control may require substantial further restric-

tions. In Example I, however, we can use results from Andrews, Cheng, and Guggenberger (2011) (henceforth ACG) to prove that a large class of CLC tests based on postmultiplication-invariant weight functions control size uniformly in heteroskedastic linear IV.

Example I: Weak IV (Continued) Define $\hat{\Omega}$ and $\hat{\Sigma}$ in the usual way (detailed in the proof of Proposition 2 in the Appendix). Define a parameter space Λ of null distributions as in ACG Section 3, noting that γ consists of the elements of $(\Omega_F, \Gamma_F, \Sigma_F)$ in the notation of ACG. Building on results in ACG it is straightforward to prove the following proposition:

Proposition 2 *Consider the CLC test $\phi_{a(D_T,\hat{\gamma})}$ based on a postmultiplication-invariant weight function $a(D,\gamma)$ which is continuous in D and γ at all points with $||D|| > 0$ and satisfies*

$$\lim_{\delta \to 0} \left(\sup_{(D,\gamma):||D||>\varepsilon, maxeig(\Sigma_D)\leq\delta} a(D,\gamma) \right) = \lim_{\delta \to 0} \left(\inf_{(D,\gamma):||D||>\varepsilon, maxeig(\Sigma_D)\leq\delta} a(D,\gamma) \right) = a_0 \qquad (20)$$

for some constant $a_0 \in [0,1]$, $maxeig(A)$ the maximal eigenvalue of A, and all $\varepsilon > 0$. The test $\phi_{a(D_T,\hat{\gamma})}$ is uniformly asymptotically similar on Λ:

$$\lim_{T\to\infty} \inf_{\lambda\in\Lambda} E_{T,\lambda}\left[\phi_{a(D_T,\hat{\gamma})}\right] = \lim_{T\to\infty} \sup_{\lambda\in\Lambda} E_{T,\lambda}\left[\phi_{a(D_T,\hat{\gamma})}\right] = \alpha.$$

The assumption (20), together with the assumed postmultiplication invariance of $a(D,\gamma)$ and the restrictions on the parameter space Λ, ensures that under sequences with $\sqrt{T}||\pi_T|| \to \infty$ we have that $a(D_T,\hat{\gamma}) \to_p a_0$ asymptotically, and hence that under all strongly identified sequences the test converges to the linear combination test ϕ_{a_0}. We show in the next section that for $a_0 = 0$ this condition plays an important role in establishing asymptotic efficiency of CLC tests in linear IV under strong identification, and will verify that this condition holds for PI tests ϕ_{PI} in linear IV. The conditions needed to ensure that a_{PI} satisfies the continuity conditions in Proposition 2 are much less clear, but we can always create a sufficiently continuous weight function $\tilde{a}$ which approximates a_{PI} arbitrarily well by calculating a_{PI} on a grid of values for (D,γ) and taking $\tilde{a}$ to continuously interpolate between these values.[13]□

Power results in the limit problem (2) also imply asymptotic power results under (1). In particular, for $a(D,\gamma)$ almost everywhere continuous with respect to $F(g,\Delta g,\gamma)$, the asymptotic power of $\phi_{a(D_T,\hat{\gamma})}$ is simply the power of $\phi_{a(D,\gamma)}$ in the limit problem.

[13]To ensure that $\tilde{a}$ is invariant to postmultiplication we can fix $||D|| = 1$ in the grid used to calculate $\tilde{a}$ and evaluate $\tilde{a}$ for other values by rescaling the problem to $||D|| = 1$ using the transformation (17).

Proposition 3 *Assume $(g_T, \Delta g_T, \hat{\gamma})$ satisfies the weak convergence assumption (1) and let $a(D, \gamma)$ be almost-everywhere continuous with respect to the limit distribution $F(g, \Delta g, \gamma)$ for some $(\theta, \gamma) \in \Theta \times \Gamma$. Then under (θ, γ)*

$$\lim_{T\to\infty} E_{T,(\theta,\gamma)}\left[\phi_{a(D_T,\hat{\gamma})}\right] = E_{m,\mu_D,\gamma}\left[\phi_{a(D,\gamma)}\right]$$

where $m = m(\theta, \theta_0, \gamma)$ and μ_D are the parameters in the limit problem.

Thus, under mild continuity conditions on $a(D, \gamma)$, the asymptotic size and power of tests under (1) are just their size and power in the limit problem. Moreover, sufficiently continuous postmultiplication invariant weight functions $a(D, \gamma)$ which select a fixed weight a_0 under strong identification yield uniformly asymptotically similar tests in heteroskedastic linear IV.

8.3 Asymptotic Efficiency Under Strong Identification

The power results above concern the asymptotic properties of CLC tests under general conditions that allow for weak identification, but since the commonly-used non-robust tests are efficient under strong identification we may particularly want to ensure that our CLC tests share this property.

As noted in Section 3.2, under strong identification we typically have that $\Sigma_{\theta\theta} = \Sigma_{\theta g} = 0$, that μ is full rank, and that $\mathcal{M}(\mu) = \{\mu \cdot c : c \in \mathbb{R}^p\}$. We say that $(g_T, \Delta g_T, \hat{\gamma})$ converges to a Gaussian shift model under (θ, γ) if $(g_T, \Delta g_T, \hat{\gamma}) \to_d (g, \Delta g, \gamma)$ for

$$\begin{pmatrix} g \\ vec(\Delta g) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \cdot b \\ vec(\mu) \end{pmatrix}, \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}\right) \tag{21}$$

for μ full rank and $b \in \mathbb{R}^p$. We show in Appendix B that under strong identification general GMM models parametrized in terms of local alternatives converge to Gaussian shift models. In many cases strong identification is not necessary to obtain convergence to (21), however, and sequences of models between the polar cases of weak and strong identification, like the "semi-strong" case discussed in Andrews and Cheng (2012), often yield Gaussian shift limit problems under appropriately defined sequences of local alternatives.

Example I: Weak IV (Continued) Suppose that $\pi_T = r_T c$ for $c \in \mathbb{R}^p$ with $||c|| > 0$ for any sequence $\{r_T\}_{T=1}^{\infty}$ such that $r_T \to r$ as $T \to \infty$ and $\sqrt{T} r_T \to \infty$. For $0 < r < \infty$ this is the usual, strongly identified case, while for $r = 0$ this is falls into the "semi-strong"

category of Andrews and Cheng: the first stage converges to zero, but at a sufficiently slow rate that many standard asymptotic results are preserved. Let $\tilde{\Omega}$ be a consistent estimator for $\lim_{T\to\infty} Var\left(\left(\sqrt{T}f_T(\beta_0)', r_T^{-1}f_T(\beta_0)'\right)'\right)$ and define $\tilde{g}_T(\beta) = \sqrt{T}\tilde{\Omega}_{ff}^{-\frac{1}{2}} f_T(\beta)$ and $\tilde{\gamma} = vec\left(\tilde{\Omega}\right)$ as before. Consider sequences of local alternatives with $\beta_T = \beta_0 + \frac{b^*}{r_T\sqrt{T}}$ and let $\Delta\tilde{g}_T = r_T^{-1}\tilde{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial\beta} f_T(\beta)$. As $T \to \infty$, $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$ converges to the Gaussian shift limit problem (21) with $\mu = E[Z_t Z_t']\, c$ and $b = b^*$.□

In the Gaussian shift limit problem (21), application of the Neyman Pearson Lemma yields that the uniformly most powerful level α test of $H_0 : m = 0, \mu \in \mathbb{M}$ against $H_1 : m \in \mathcal{M}(\mu)\backslash\{0\}, \mu \in \mathbb{M}$ based on (J, K, D) is ϕ_K as defined in (14). Correspondingly, it is straightforward to show that a CLC test based on the weight function $a(D, \gamma)$ will be asymptotically efficient under sequences converging to (21) if and only if $a(D_T, \hat{\gamma}) \to_p 0$ under such sequences.

Proposition 4 *Denote by $\mathcal{A}_c$ the class of weight functions functions $a(D, \gamma) \to [0, 1]$ that are continuous in both D and γ for all full-rank D. Fix $(\theta, \gamma) \in \Theta \times \Gamma$ with $\theta \neq \theta_0$ and suppose that $(g_T, \Delta g_T, \hat{\gamma})$ converges weakly to the Gaussian shift limit problem (21) with $b \neq 0$. For $a(D, \gamma)$ almost-everywhere continuous with respect to the limiting measure $F(g, \Delta g, \gamma)$ under (θ, γ),*

$$\lim_{T\to\infty} E_{T,(\theta,\gamma)}\left[\phi_{a(D_T,\hat{\gamma})}\right] = \sup_{\tilde{a}\in\mathcal{A}_c} \lim_{T\to\infty} E_{T,(\theta,\gamma)}\left[\phi_{\tilde{a}(D_T,\hat{\gamma})}\right]$$

if and only if $a(D, \gamma) = 0$ almost surely with respect to $F(g, \Delta g, \gamma)$. Thus, for $\phi_{K_T} = 1\left\{K_T > \chi^2_{p,1-\alpha}\right\}$,

$$\lim_{T\to\infty} E_{T,(\theta,\gamma)}[\phi_{K_T}] = \sup_{\tilde{a}\in\mathcal{A}_c} \lim_{T\to\infty} E_{T,(\theta,\gamma)}\left[\phi_{\tilde{a}(D_T,\hat{\gamma})}\right].$$

Using this proposition, it is easy to see that the condition (20) that we used to ensure uniformly correct size for CLC tests in linear IV Example I will also ensure asymptotic efficiency under strong and semi-strong identification provided $a_0 = 0$.

It is straightforward to give conditions under which MMRU tests select $a(\mu_D, \gamma) = 0$ asymptotically in sequences of models converging to Gaussian shift experiments:

Theorem 6 *Suppose that for some pair $(\mu_D, \gamma) \in \mathbb{M}_D \times \Gamma$ with μ_D full-rank and $\Sigma_{\theta g}(\gamma) = \Sigma_{\theta\theta}(\gamma) = 0$, for all $C > 0$ and all sequences $(\mu_{D,n}, \gamma_n) \in \mathbb{M}_D \times \Gamma$ such that $(\mu_{D,n}, \gamma_n) \to (\mu_D, \gamma)$ we have*

$$d_H\left(\mathcal{M}_D(\mu_{D,n}, \gamma_n) \cap B_C, \{\mu_D \cdot b : b \in \mathbb{R}^p\} \cap B_C\right) \to 0$$

where $B_C = \{m : ||m|| \leq C\}$ *and* $d_H(A_1, A_2)$ *is the Hausdorff distance between the sets* A_1 *and* A_2*,*

$$d_H(A_1, A_2) = \max\left\{\sup_{x_1 \in A_1} \inf_{x_2 \in A_2} \|x_1 - x_2\|, \sup_{x_2 \in A_2} \inf_{x_1 \in A_1} \|x_1 - x_2\|\right\}.$$

Then for $\beta^u_{m,\mu_{D,n},\gamma_n} = \sup_{a \in [0,1]} E_{m,\mu_{D,n},\gamma_n}[\phi_a]$ *and all* $(\mu_{D,n}, \gamma_n) \to (\mu_D, \gamma)$ *the MMRU weight*

$$a_{MMRU}(\mu_{D,n}, \gamma_n) = \arg\min_{a \in [0,1]} \sup_{m \in \mathcal{M}_D(\mu_{D,n}, \gamma_n)} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}[\phi_a]\right)$$

satisfies $a_{MMRU}(\mu_{D,n}, \gamma_n) \to 0$*.*

Using Theorem 6, it is straightforward to establish that PI tests will be efficient under strong and semi-strong identification in Example I, while MMR tests will be efficient under strong and semi-strong identification in Example II, where the MMR and MMRU tests coincide.

Example I: Weak IV (Continued) Define $(g_T, \Delta g_T, \hat{\gamma})$ as in Section 1, and as above let $\pi_T = r_T c$ for $c \in \mathbb{R}^p$ with $||c|| > 0$. For simplicity we take $\hat{\mu}_D = D_T$ but the extension to other estimators is straightforward. We have the following result:

Corollary 3 *Provided* $\sqrt{T} r_T \to \infty$*, we have that in the linear IV model* $a_{PI}(\hat{\mu}_D, \hat{\gamma}) \to_p 0$ *and thus that the PI test based on* $(g_T, \Delta g_T, \hat{\gamma})$ *is efficient under strong and semi-strong identification.*□

Example II: Minimum Distance (Continued) We can model semi-strong identification in this example by taking $\Omega_\eta = r_T \Omega_{\eta,0}$ where $r_T \to 0$ and $r_T^{-1} \hat{\Omega}_\eta \to_p \Omega_{\eta,0}$, noting that $r_T = \frac{1}{T}$ is the typical strongly identified case. Again define $\hat{\gamma} = vec\left(\hat{\Omega}_\eta\right)$ and note that $\mathcal{M}(\hat{\gamma}) = \left\{\hat{\Omega}_\eta^{-\frac{1}{2}}(f(\theta) - f(\theta_0))\right\}$. Defining $g_T(\theta) = \hat{\Omega}_\eta^{-\frac{1}{2}}(\hat{\eta} - f(\theta))$, and $\Delta g_T(\theta) = \frac{\partial}{\partial \theta} g_T(\theta) = \hat{\Omega}_\eta^{-\frac{1}{2}} \frac{\partial}{\partial \theta} f(\theta)$ as before, a global identification assumption yields that PI tests are asymptotically efficient.

Corollary 4 *For* $r_T^{-1} \to \infty$*, assume that* θ *is in the interior of* Θ *and that for all* $\delta > 0$ *there exists* $\varepsilon(\delta) > 0$ *such that* $||f(\tilde{\theta}) - f(\theta)|| < \varepsilon(\delta)$ *implies that* $||\tilde{\theta} - \theta|| < \delta$*. Then the MMR weight function* a_{MMR} *satisfies* $a_{MMR}(\hat{\gamma}) \to_p 0$ *and the MMR test is efficient under strong and semi-strong identification.*□

Hence, in our examples the plug-in test ϕ_{PI} is efficient under asymptotically strong and semi-strong identification.

9 Simulation: Inference on the New Keynesian Phillips Curve

To illustrate the application of PI tests to a nonlinear example, we study the performance of robust minimum distance inference on new Keynesian Phillips curve (NKPC) parameters. There is considerable evidence that some NKPC parameters are weakly identified: Mavroeidis, Plagborg-Moller, and Stock (2013) review the empirical literature on the role of expectations in the NKPC and find that parameter estimates are extremely sensitive to model specification and, conditional on correct specification, suffer from weak identification. To address these weak identification issues Magnusson and Mavroeidis (2010) (henceforth MM) propose identification-robust S and K statistics for testing hypotheses on NKPC parameters using a minimum distance approach. These statistics will form the basis for our analysis.

MM study a simple New Keynesian Phillips Curve model

$$\pi_t = \frac{(1-\nu)^2}{\nu(1+\rho)} x_t + \frac{1}{1+\rho} E\left[\pi_{t+1} | \mathcal{I}_t\right] + \frac{\rho}{1+\rho}\pi_{t-1} + \varepsilon_t \tag{22}$$

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

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

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

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

MM show that the NKPC model (22) implies that the true parameter values φ and θ satisfy $f(\varphi, \theta) = 0$, and propose testing $H_0 : \theta = \theta_0$ using an identification-robust minimum distance approach.

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

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

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

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

To model weak identification in this context MM assume that $\frac{\partial}{\partial \theta'} f(\varphi_T, \theta) = \frac{1}{\sqrt{T}} C$ for a fixed matrix C, with the result that $\sqrt{T} \frac{\partial}{\partial \theta'} f(\varphi_T, \theta)$ is constant across T. This leads to the usual issues associated with weak identification, including nonstandard limiting distributions for non-robust test statistics. Here, we will take a more flexible approach and assume only that $\frac{\partial}{\partial \theta'} f(\varphi_T, \theta)$ drifts towards some, potentially reduced-rank, matrix as the sample size grows.

To apply our robust testing approach in this context, define $\hat{\Omega}_{ff} = \frac{\partial}{\partial \varphi'} f(\hat{\varphi}, \theta_0) \hat{\Sigma}_{\varphi\varphi} \frac{\partial}{\partial \varphi'} f(\hat{\varphi}, \theta_0)'$ which is a consistent, Δ-method-based estimator for $\Omega_{ff} = \lim_{T\to\infty} T \cdot Var\left(f(\hat{\varphi}, \theta_0)'\right)$. We can then define

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

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

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

$$\begin{pmatrix} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{pmatrix} \to_d \begin{pmatrix} g \\ \Delta g \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ \mu \end{pmatrix}, \begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix}\right). \tag{23}$$

where Δg is full rank almost surely, $m \in \mathcal{M}(\mu, \gamma)$ for $\mathcal{M}(\mu, \gamma)$ appropriately defined, Σ is consistently estimable, and details on all terms may be found in Appendix E. Hence, this model falls into the class considered in this paper. While Δg_T depends on the (generally unknown) sequence of

normalizing matrices A_T, provided we restrict attention to postmultiplication-invariant CLC tests we can instead conduct tests based on the feasible statistics $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma}) = h\left(g_T, \Delta g_T, \hat{\gamma}; A_T^{-1}\right)$. For $\hat{\gamma}$ as defined in Appendix E the statistics S_T and K_T based on $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$ are equivalent to the $MD - AR$ and $MD - K$ statistics discussed in MM.

9.1 Coverage Simulations

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

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

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

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

where as before

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

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

To compare the performance of the PI test to the tests discussed by MM, we calibrate a simulation example based on the empirical application of MM. In particular, we estimate structural and reduced-form parameters using the data studied by MM and generate samples of 100 observations based on these estimates together with the assumption of Gaussian errors ϵ_t (see Appendix E for details).[14] We calculate the true size of nominal 5% tests, based on 10,000 simulations, and report the results in Table 3. We find that all the tests over-reject, which is unsurprising given the non-linearity of the model together with the small sample size, but that only the JK and S tests has true size exceeding 10%.

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

Next, we simulate false coverage probabilities for confidence sets formed by inverting these tests. In particular we calculate the rejection rates for PI, JK, S, and K tests of hypotheses $H_0 : \theta = \theta_0$ for θ_0 not equal to the true parameter value.[15] Table 4 reports the maximal difference in point-wise false coverage probability across tests, based on 500 simulations. For each test we report the largest margin by which the rejection probability of that test falls short relative to that of the other tests considered over $\theta_0 \in (0,1)^2$, which is the parameter space for the model.[16] For example, the second entry of the first row of Table 4 reports

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

where ϕ_{PI_T,θ_0} and ϕ_{JK_T,θ_0} denote the PI and JK tests of $H_0 : \theta = \theta_0$, respectively, and $\tilde{\theta} = (\tilde{\nu}, \tilde{\rho}) = (0.96, 0.48)$ is the true parameter value in the simulations. As these results make clear, the PI test outperforms the other tests studied and has the smallest maximal rejection rate shortfall. The JK test also performs reasonably well, with a much smaller maximal rejection rate shortfall than the S and K tests. Interpreting these results is complicated by the fact that, while all the tests considered have correct asymptotic size under weak identification, their finite sample size differs substantially. To account for such size differences, Table 5 reports results analogous to those of Table 4 based on (infeasible) size-corrected versions of all four tests. As in Table 4, we can see that the PI test offers the best performance, followed by the JK test.[17]

After simulating false coverage probabilities, it is easy to calculate the expected area of confidence sets obtained by inverting the PI, JK, S, and K tests. The expected area for confidence sets formed by inverting both the feasible and size-corrected tests is reported in Table 6. As we can see, using size-corrected tests increases the area of all confidence sets. In each case the PI test

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

[16] For computational reasons, our simulations use a discretized version of this parameter space- see Appendix E.

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

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

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

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

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

produces confidence sets with the smallest expected area, while the S test yields confidence sets with the largest expected area. The feasible JK test yields smaller confidence sets than the feasible K test, but size correction reveals that this is due in part to finite-sample size distortions for the JK test: when we invert size-corrected tests, we find that JK confidence sets have higher expected area than K confidence sets. A further advantage of the PI-test-based confidence sets is that, like K-test-based confidence sets, they are non-empty in all 500 simulations, whereas confidence sets formed by inverting the JK and S tests are empty in 3.2% and 4.8% of simulations, respectively.[18] These results confirm that the PI test outperforms the other tests considered.

10 Conclusion

This paper considers the problem of constructing powerful identification-robust tests for a broad class of weakly identified models. We define the class of conditional linear combination (CLC) tests and show that this class is equivalent to an appropriately defined class of quasi-conditional likelihood ratio (quasi-CLR) tests. We show that CLC tests are admissible, locally most powerful, and weighted average power maximizing conditional on D. To pick from the class of CLC tests we suggest using minimax regret (MMR) tests when feasible and plug-in (PI) tests when MMR tests are too difficult to compute. We show that PI tests match the near-optimal performance of the CLR test of Moreira (2003) in homoskedastic linear IV, outperform alternative approaches in simulations calibrated to match an IV model with heteroskedastic time-series data and a new Keynesian Phillips curve model, and are efficient in linear IV models with strong instruments.

Our results suggest interesting directions for further research. In particular, for models where the true MMR test is difficult to compute, the PI tests discussed in this paper represent only one of many possible approaches to selecting CLC tests. Exploring alternative approaches to constructing CLC tests and comparing their performance to that of PI tests may yield both additional powerful tests and further insight into the problem of testing under weak identification. Relatedly, the identification-robust statistics proposed by Chaudhuri and Zivot (2011) for testing composite hypotheses with weakly identified nuisance parameters are similar in structure to the J and K statistics considered here, and application of the approach of this paper to those statistics may yield powerful tests for hypotheses with weakly identified nuisance parameters. Finally, the characterization of CLC tests as weighted average power maximizing tests in a conditional problem, discussed

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

in Theorem 4, suggests a potentially fruitful route to studying the admissibility of particular CLC tests in various contexts, including the CLR test in linear IV.

Appendix A: Derivation of Limit Problems for Examples

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

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

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

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

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

we can see that under the assumptions already made the weak-instruments sequence $\pi_T = \frac{c}{\sqrt{T}}$ implies that under true parameter value β,

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

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

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

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

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

Appendix B: Limit Problem for Weak GMM Models

In this appendix, we prove some additional results for GMM models that are weakly identified in the sense of Stock and Wright (2000). Suppose we begin with a moment function $f_t(\psi)$ which is differentiable in the parameter ψ and satisfies the usual GMM identifying assumption that $E_\psi\left[f_t(\psi)\right] = 0$,

and are interested in testing $H_0 : \psi = \psi_0$. Suppose that, much like in Stock and Wright (2000), our parameter vector $\psi = (\psi_1, \psi_2)$ is such that ψ_1 is weakly identified while ψ_2 is strongly identified, and that the expectation of $f_t(\psi_0)$ under alternative ψ is

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

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

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

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

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

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

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

and

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

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

$$h(\theta) = \frac{1}{\sqrt{T}} \frac{\partial}{\partial \psi_1'} \tilde{h}_1 \left(\psi_{1,0} \right) \theta_1 + \frac{1}{\sqrt{T}} \tilde{h}_2 \left(\psi_{1,0}, \theta_2 \right). \tag{24}$$

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

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

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

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

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

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

and

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

note that

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

where $\mu = \lim_{T\to\infty} E_\theta\left[\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \theta'} g_t(\theta_0)\right]$ provided this limit exists and $m = h(\theta) \in \mathcal{M}(\mu, \gamma)$, where $\mathcal{M}(\mu, \gamma)$ will depend on the structure of the problem at hand: in some cases it may be that without additional structure we cannot restrict the set of possible values m and have $\mathcal{M}(\mu) = \mathbb{R}^k$ while in others, like Example I, we may be able to obtain further restrictions. Note further, that while we framed the analysis here using reparameterization in terms of local alternatives for strongly identified parameters, we could equivalently have formulated Δg_T using the Jacobian of the original moment function, $\frac{\partial}{\partial \psi'} f_t\left(\psi_0\right)$, post-multiplied by an appropriate sequence of normalizing matrices A_T, as in Section 8.1.

We can say a bit more regarding the strongly identified parameters θ_1. Note that by the definition of θ, $\frac{\partial}{\partial \theta_1'} g_t(\theta_0) = \frac{1}{\sqrt{T}} \frac{\partial}{\partial \psi_1'} f_t(\psi_0)$. Hence, $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \theta_1'} g_t(\theta_0) = \frac{1}{T} \sum \frac{\partial}{\partial \psi_1'} f_t(\psi_0)$ and we can re-write μ as $\lim_{T\to\infty} E_\theta \left[\begin{array}{cc} \frac{1}{T} \sum \frac{\partial}{\partial \psi_1'} f_t(\psi_0) & \frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \psi_2'} f_t(\psi_0) \end{array} \right]$. Further, the central limit theorem we have assumed for $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \psi_1'} f_t(\psi_0)$ implies that $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \theta_1'} g_t(\theta_0) \to_p \mu_1 = \lim_{T\to\infty} E_\theta \left[\frac{1}{T} \sum \frac{\partial}{\partial \psi_1'} f_t(\psi_0) \right]$. Together with (24) this implies that under standard regularity conditions (see e.g. Newey and McFadden, 1994) $h(\theta_1, \theta_{2,0}) = \mu_1 \cdot \theta_1$ and hence that in the special case where all parameters are strongly identified

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

Appendix C: Proofs

Proof of Theorem 1

By the independence of J, K, and D under the null, conditional on the event $D = d$

$$K + a(D) \cdot J | D = d \sim \chi^2_p + a(d) \cdot \chi^2_{k-p}.$$

Hence

$$Pr\left\{ K + a(D) \cdot J > c_\alpha \left(a\left(D \right) \right) \middle| \, D = d \right\} = \alpha$$

so $E_{m=0,\mu_D} \left[\phi_{a(D)} \middle| D = d \right] = \alpha$ for all d in the support of D and all values μ_D. $E_{m=0,\mu_D} \left[\phi_{a(D)} \right]$ can then be written as

$$\int E_{m=0,\mu_D} \left[\phi_{a(D)} \middle| D = d \right] dF_D = \int \alpha dF_D = \alpha$$

for F_D the distribution of D, proving the theorem.

Proof of Theorem 2

We first argue that conditional on $D = d$ the test ϕ_{QCLR_r} is exactly equivalent to the level α test that rejects for large values of the statistic $K + \frac{q_\alpha(r(d))}{q_\alpha(r(d)) + r(d)} \cdot J$. This result is trivial for $r\left(d\right) = \infty$. For $r\left(d\right) < \infty$ and $K > 0$, note first that for fixed d the QCLR statistic is strictly increasing in (J, K). Further, for any $L > 0$, the L level set of the $QCLR_r$ statistic is of the form $L = K + \frac{L}{L + r(d)} \cdot J$ so that fixing $D = d$,

$$\left\{(J,K) \in \mathbb{R}_+^2 : QCLR_r = L\right\} = \left\{(J,K) \in \mathbb{R}_+^2 : L = K + \frac{L}{L + r(d)} \cdot J\right\}.$$

To verify that this is the case, note that if we plug $K = L - \frac{L}{L+r(d)} \cdot J$ into the $QCLR_r$ statistic and collect terms we have

$$QCLR_r = \frac{1}{2}\left(L + \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J - r(d) + \sqrt{\left(L + r\left(d\right) + \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J)\right)^2 - 4J \cdot r(d)}\right).$$

However,

$$\left(L + r(d) + \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J\right)^2 - 4J \cdot r(d) = \left(L + r\left(d\right) - \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J\right)^2$$

and thus for $K = L - \frac{L}{L+r(d)} \cdot J$,

$$QCLR = \frac{1}{2}\left(L + \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J - r(d) + \sqrt{\left(L + r\left(d\right) - \frac{r\left(d\right)}{L + r\left(d\right)} \cdot J\right)^2}\right).$$

Since we've taken $K = L - \frac{L}{L+r(D)} \cdot J$ and we know $K \geq 0$, we have that $J \leq L + r\left(d\right)$. Thus $L + r\left(d\right) - \frac{r(d)}{L+r(d)} \cdot J \geq 0$ and we can open the square root and collect terms to obtain $QCLR_r = L$ on the set $\left\{(J,K) \in \mathbb{R}_+^2 : L = K + \frac{L}{L+r(d)} \cdot J\right\}$, as we claimed.

This implies that, conditional on $D = d$ the rejection region of ϕ_{QCLR_r} is

$$\left\{(J,K) \in \mathbb{R}_+^2 : q_\alpha\left(r(d)\right) < K + \frac{q_\alpha\left(r(d)\right)}{q_\alpha\left(r(d)\right) + r(d)} \cdot J\right\}.$$

Since J and K are pivotal under the null, conditional on $D = d$

$$Pr_{m=0,\mu_D}\left\{q_\alpha\left(r(d)\right) < K + \frac{q_\alpha\left(r(d)\right)}{q_\alpha\left(r(d)\right) + r(d)} \cdot J \middle| D = d\right\} = \alpha,$$

so since $K + \frac{q_\alpha(r(d))}{q_\alpha(r(d))+r(d)} \cdot J$ is continuously distributed with support equal $\mathbb{R}_+$, $q_\alpha\left(r(d)\right)$ must be the $1-\alpha$ quantile of this random variable. Hence, if we define the test $\phi_{\tilde{a}(D)}$ as in (16) with $\tilde{a}(D) = \frac{q_\alpha(r(D))}{q_\alpha(r(D))+r(D)}$, we can see that $c_\alpha\left(\tilde{a}\left(d\right)\right) = q_\alpha\left(d\right)$ and thus that $\phi_{QCLR_r} = \phi_{\tilde{a}(d)}$ conditional on $D = d$. Since this hold for all d, $\phi_{QCLR_r} \equiv \phi_{\tilde{a}(D)}$. Thus, for any function $r : D \to \mathbb{R}_+ \cup \{\infty\}$ there is a function $\tilde{a} : D \to [0,1]$ such that $\phi_{QCLR_r} \equiv \phi_{\tilde{a}(D)}$.

To prove the converse, that for any CLC test $\phi_{a(D)}$ for $a : D \to [0,1]$ we can find a function r :

$D \to \mathbb{R}_+ \cup \{\infty\}$ yielding the same test, fix the function $a(D)$ and note that $q_\alpha(r(D))$ is a continuous function of $r(D)$ which is deceasing in $r(D)$ and is bounded below by $\chi^2_{p,1-\alpha}$ and above by $\chi^2_{k,1-\alpha}$ (see Moreira (2003)). Hence for any value d, as $r(d)$ goes from zero to infinity $\frac{q_\alpha(r(d))}{q_\alpha(r(d))+r(d)}$ varies continuously between zero and one, with $\lim_{r(d)\to 0} \frac{q_\alpha(r(d))}{q_\alpha(r(d))+r(d)} = 1$ and $\lim_{r(d)\to\infty} \frac{q_\alpha(r(d))}{q_\alpha(r(d))+r(d)} = \frac{q_\alpha(\infty)}{q_\alpha(\infty)+\infty} = 0$. If $a(d) = 0$ define $\tilde{r}(d) = \infty$. If $a(d) > 0$, note that there exists a value $r^* < \infty$ such that $a(d) > \frac{q_\alpha(r^*)}{q_\alpha(r^*)+r^*}$, so by the intermediate value theorem we can pick $\tilde{r}(d) \in [0, r^*]$ such that $a(d) = \frac{q_\alpha(\tilde{r}(d))}{q_\alpha(\tilde{r}(d))+\tilde{r}(d)}$. Repeating this exercise for all values d we can construct a function $\tilde{r} : D \to \mathbb{R}_+ \cup \{\infty\}$ such that $\phi_{a(D)} \equiv \phi_{\tilde{r}(D)}$, completing the proof.

Proof of Corollary 1

Conditional on $D = d$ the CLC test $\phi_{a(D)}$ fails to reject if and only if $K + a(d) \cdot J \leq c_\alpha(a(d))$, where $a(d) \in [0, 1]$. Thus, for this test we can define

$$C_d = \left\{ \left(\sqrt{J}, \sqrt{K} \right) : K + a(d) \cdot J \leq c_\alpha(a(d)) \right\}.$$

This set trivially satisfies requirement (2) of Theorem 3. To verify that it satisfies (1), note that C_d is closed and if we have two pairs $\left(\sqrt{J_1}, \sqrt{K_1}\right), \left(\sqrt{J_2}, \sqrt{K_2}\right) \in C_d$ then for all $\lambda \in [0, 1]$

$$\lambda \left(\sqrt{J_1}, \sqrt{K_1} \right) + (1 - \lambda) \left(\sqrt{J_2}, \sqrt{K_2} \right) \in C_d$$

since by the convexity of the function $f(x) = x^2$

$$\left(\lambda \sqrt{J_1} + (1 - \lambda) \sqrt{J_2} \right)^2 \leq \lambda J_1 + (1 - \lambda) J_2$$

$$\left(\lambda \sqrt{K_1} + (1 - \lambda) \sqrt{K_2} \right)^2 \leq \lambda K_1 + (1 - \lambda) K_2.$$

Proof of Corollary 2

Given a test ϕ, note that if ϕ is admissible in the conditional problem for almost every d, we can take $\tilde{\phi} = \phi$ and the result is immediate.

Suppose instead that ϕ is inadmissible conditional on $D = d$ if and only if $d \in \mathcal{D}$ where $\mathcal{D}$ has positive measure under F_D. For each $d \in \mathcal{D}$, by Theorem 3 we know that there is a test $\tilde{\phi}_d$ which

dominates ϕ conditional on $D = d$. Repeating this exercise for each $d \in \mathcal{D}$ we can construct $\tilde{\phi}$

$$\tilde{\phi} = 1\{D \notin \mathcal{D}\}\phi + 1\{D \in \mathcal{D}\}\tilde{\phi}_D$$

where for all $d \in \mathcal{D}$, $\tilde{\phi}_d$ is admissible in the conditional problem, and we choose $\tilde{\phi}_D$ to ensure that $\tilde{\phi}$ is measurable with respect to (J, K, D) (and thus is a test). Note that for all $d \in \mathcal{D}$ and all $\mu \in \mathbb{M}$, $m \in \mathbb{R}^k$

$$E_{m=0,\mu}\left[\tilde{\phi}|D = d\right] \leq E_{m=0,\mu}\left[\phi|D = d\right]$$

and

$$E_{m\neq 0,\mu}\left[\tilde{\phi}|D = d\right] \geq E_{m\neq 0,\mu}\left[\phi|D = d\right]$$

since $\tilde{\phi}_d$ was chosen to dominate ϕ. On the other hand, conditional on $D \notin \mathcal{D}$ we have that $\tilde{\phi} = \phi$. Hence, since we choose $\tilde{\phi}_d$ such that $E_{m=0,\mu}\left[\tilde{\phi}|D = d\right]$ is measurable as a function of d we can see that

$$E_{m=0,\mu}\left[\tilde{\phi}\right] = \int E_{m=0,\mu}\left[\tilde{\phi}|D = d\right] dF_D(d) \leq \int E_{m=0,\mu}\left[\phi|D = d\right] dF_D(d) = E_{m=0,\mu}\left[\phi\right]$$

and

$$E_{m\neq 0,\mu}\left[\tilde{\phi}\right] = \int E_{m\neq 0,\mu}\left[\tilde{\phi}|D = d\right] dF_D(d) \geq \int E_{m\neq 0,\mu}\left[\phi|D = d\right] dF_D(d) = E_{m\neq 0}\left[\phi\right].$$

Proof of Theorem 4

Both (1) and (2) follow from results in Monti and Sen (1976) and Koziol and Perlman (1978). Specifically, both papers note that if $(A, B) \sim \left(\chi^2_{k-p}(\tau_A), \chi^2_p(\tau_B)\right)$ and $(\tau_A, \tau_B) = \lambda \cdot (t_A, t_B)$ for t_A, $t_B \geq 0$ then for ϕ any size α test for $H_0 : \tau_A = \tau_B = 0$ based on (A, B) there exists some $\bar{\lambda} > 0$ such that for $0 < \lambda < \bar{\lambda}$,

$$E_{(\tau_A,\tau_B)}\left[\phi\right] \leq E_{(\tau_A,\tau_B)}\left[1\left\{\frac{t_A}{k-p}A + \frac{t_B}{p}B > c\right\}\right]$$

for c the $1-\alpha$ quantile of a $\frac{t_A}{k-p}\chi^2_{k-p} + \frac{t_B}{p}\chi^2_p$ distribution. Statement (1) then follows immediately from the fact that $(J, K)|D = d \sim \left(\chi^2_{k-p}(\tau_J), \chi^2_p(\tau_K)\right)$.

Establishing statement (2) is similarly straightforward. In particular for F_{t_K,t_J} as described in Theorem 4, Koziol and Perlman note that we can use the Neyman Pearson Lemma to establish

that the weighted average power maximizing level α test based on $(A,B) \sim \left(\chi^2_{k-p}(\tau_A), \chi^2_p(\tau_B)\right)$ is $\phi^*_F = 1\left\{\frac{t_K}{t_K+1}A + \frac{t_J}{t_J+1}B > c\right\}$, where c is the $1-\alpha$ quantile of a $\frac{t_K}{t_K+1}\chi^2_p + \frac{t_J}{t_J+1}\chi^2_{k-p}$ distribution. In particular, for Φ_α the class of level α tests based on (A,B),

$$\phi^*_F \in \arg\max_{\phi\in\Phi_\alpha} \int_{\mathcal{T}(d)} E_{\tau_A,\tau_B}[\phi]\, dF(\tau_A,\tau_B).$$

Statement (2) again follows from the fact that $(J,K)\,|D=d \sim \left(\chi^2_{k-p}(\tau_J), \chi^2_p(\tau_K)\right)$.

Proof of Theorem 5

For this proof we assume that D has a density with respect to Lebesgue measure. The proof for D degenerate follows along the lines. By the definition of the infimum we know that there exists a sequence of functions $\{a_n\}_{n=1}^\infty \subset \mathcal{A}$ such that

$$\lim_{n\to\infty} \sup_{(m,\mu_D)\in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a_n(D)}\right]\right) = \inf_{a\in\mathcal{A}} \sup_{(m,\mu_D)\in H_1} \left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right). \tag{25}$$

Since we have defined $\mathcal{A}$ to consist of Borel-measurable functions, by Theorem A.5.1 in Lehman and Romano (2005) there exists a sub-sequence $\{a_{n_k}\}_{k=1}^\infty \subset \{a_n\}_{n=1}^\infty$ and a function $a^* \in \mathcal{A}$ such that for ν Lebesgue measure on $\mathbb{R}^{kp+2}$

$$\int a_{n_k}(D)h(J,K,D)d\nu \to \int a^*(D)h(J,K,D)d\nu \tag{26}$$

for all ν-integrable functions $h(J,K,D)$.

For any value $(m,\mu) \in H_1$ denote the joint distribution of (J,K,D) under this value by $F_{JKD}(m,\mu)$ with density $f_{JKD}(m,\mu_D)$ with respect to Lebesgue measure ν. Note that for any bounded, continuous function n of (J,K,D), $n(J,K,D)\, f_{JKD}(m,\mu_D)$ is an integrable function with respect to Lebesgue measure ν. Hence, (26) implies that

$$\int a_{n_k}(D)n(J,K,D)dF_{JKD} \to \int a^*(D)n(J,K,D)dF_{JKD}$$

for all bounded continuous functions n. By the Portmanteau Lemma (see Lemma 2.2 in Van der Vaart 2000) this implies that, viewed as a random variable, $(J,K,a_{n_k}(D)) \to_d J(,K,a^*(D))$. Since

$c_\alpha(a)$ is a continuous function of a, the Continuous Mapping Theorem implies that

$$K + a_{n_k}(D) \cdot J - c_\alpha(a_{n_k}(D)) \to_d K + a^*(D) \cdot J - c_\alpha(a^*(D)).$$

Since zero is a continuity point of distribution of the random variable on the right hand side this implies that $E_{m,\mu}\left[\phi_{a_{n_k}(D)}\right] \to E_{m,\mu}\left[\phi_{a^*(D)}\right]$. This suffices to establish

$$\sup_{(m,\mu_D)\in H_1}\left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a^*(D)}\right]\right) = \inf_{a\in\mathcal{A}}\sup_{(m,\mu_D)\in H_1}\left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right).$$

To see that this is the case, note that the right hand side is weakly smaller than the left hand side by construction. If the right hand side is strictly smaller then there exists some value $(m^*, \mu^*_D) \in H_1$ such that

$$\beta^*_{m^*,\mu^*_D} - E_{m^*,\mu^*_D}\left[\phi_{a^*(D)}\right] > \inf_{a\in\mathcal{A}}\sup_{(m,\mu_D)\in H_1}\left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right) + \varepsilon$$

for some $\varepsilon > 0$, which since $E_{m^*,\mu^*_D}\left[\phi_{a^*(D)}\right] = \lim_{n\to\infty} E_{m^*,\mu^*_D}\left[\phi_{a_n(D)}\right]$ implies that

$$\lim_{n\to\infty}\sup_{(m,\mu_D)\in H_1}\left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a_n(D)}\right]\right) \geq \inf_{a\in\mathcal{A}}\sup_{(m,\mu_D)\in H_1}\left(\beta^*_{m,\mu_D} - E_{m,\mu_D}\left[\phi_{a(D)}\right]\right) + \varepsilon$$

which contradicts (25).

Proof of Lemma 1

To prove this result, it is easier to work with the formulation of the problem discussed in AMS. In particular, consider $k \times 1$ random vectors $\widetilde{S}$ and $\widetilde{T}$ (denoted by S and T in AMS) with

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

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

limit problem (6) with $\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix}$, the maximal invariant under rotations of the instruments $\left(\widetilde{S}'\widetilde{S}, \widetilde{S}'\widetilde{T}, \widetilde{T}'\widetilde{T}\right)$ is a one-to-one transformation of $(J, K, D'\Sigma_D D)$.

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

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

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

Proof of Proposition 1

The discussion preceding Proposition 1 establishes that under (θ_0, γ) we have $(J_T, K_T, D_T) \to_d (J, K, D)$ and $\hat{\gamma} \to_p \gamma$. Since we assume that $a(D, \gamma)$ is almost everywhere continuous with respect to the limiting distribution F and $c_\alpha(a)$ is a continuous function of a, the Continuous Mapping Theorem establishes that

$$K_T + a(D_T, \hat{\gamma}) J_T - c_\alpha(a(D_T, \hat{\gamma})) \to_d K + a(D, \gamma) J - c_\alpha(a(D, \gamma)).$$

Since zero is a point of continuity of the distribution of the right hand side this establishes that

$$Pr_{T,(\theta_0,\gamma)}\{K_T + a(D_T, \hat{\gamma}) J_T > c_\alpha(a(D_T, \hat{\gamma}))\} \to Pr_{m=0,\mu_D}\{K + a(D, \gamma) J > c_\alpha(a(D, \gamma))\} = \alpha$$

which proves (18). To prove (19) note that the results above establish that $\phi_{a(D,\gamma)}$ is almost-everywhere continuous with respect to F, and hence for $f \in \mathcal{F}$

$$\left(\phi_{a(D_T,\hat{\gamma})} - \alpha\right) f(D_T) \to_d \left(\phi_{a(D,\gamma)} - \alpha\right) f(D).$$

Since the left hand side is bounded, convergence in distribution implies convergence in expectation, proving (19).

Proof of Proposition 2

Let us take the estimator $\hat{\Omega}$ to be

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{ff} & \hat{\Omega}_{f\beta} \\ \hat{\Omega}_{\beta f} & \hat{\Omega}_{\beta\beta} \end{pmatrix} = \frac{1}{T}\sum_t \begin{pmatrix} f_t(\beta_0) - f_T(\beta_0) \\ \frac{\partial}{\partial\beta} f_t(\beta_0) - \frac{\partial}{\partial\beta} f_T(\beta_0) \end{pmatrix} \begin{pmatrix} f_t(\beta_0)' - f_T(\beta_0)' & \frac{\partial}{\partial\beta} f_t(\beta_0)' - \frac{\partial}{\partial\beta} f_T(\beta_0)' \end{pmatrix}$$

and

$$\hat{\Sigma} = \begin{pmatrix} I_k & \hat{\Omega}_{ff}^{-\frac{1}{2}} \hat{\Omega}_{f\beta} \hat{\Omega}_{ff}^{-\frac{1}{2}} \\ \hat{\Omega}_{ff}^{-\frac{1}{2}} \hat{\Omega}_{\beta f} \hat{\Omega}_{ff}^{-\frac{1}{2}} & \hat{\Omega}_{ff}^{-\frac{1}{2}} \hat{\Omega}_{\beta\beta} \hat{\Omega}_{ff}^{-\frac{1}{2}} \end{pmatrix}.$$

These choices imply that our S_T and K_T coincide exactly with AR and LM in ACG, and that our D_T is $\sqrt{T}\hat{\Omega}_{ff}^{-\frac{1}{2}}\hat{D}$ for $\hat{D}$ as in ACG. To prove the result, we will rely heavily on their results. ACG consider two cases: sequences λ_T for which $\sqrt{T}||\pi_T||$ converges to a constant and those for which it diverges to infinity.

Let us begin by considering the case where $\sqrt{T}||\pi_T||$ converges. ACG establish that for this case their $(LM, AR, \hat{D})$ converges in distribution to $\left(\chi^2_1, \chi^2_{k-1}, \tilde{D}\right)$ where all three random variables are independent and $\tilde{D}$ has a non-degenerate Gaussian distribution. Since $\hat{\Omega}_{ff} \to_p \Omega_{ff}$ which is full-rank by assumption, this proves that $(K_T, S_T, D_T) \to_d \left(\chi^2_1, \chi^2_{k-1}, D\right)$ where again all the variables on the RHS are mutually independent and D has a non-degenerate Gaussian distribution. Thus, by the Continuous Mapping Theorem and consistency of $\hat{\Sigma}_{\theta g}$ and $\hat{\Sigma}_{\theta\theta}$, which under the null follows from (6.7) and (6.9) in ACG, we have that

$$(1 - a(D_T, \hat{\gamma}))\, K_T + a(D_T, \hat{\gamma}) S_T - c_\alpha\left(a(D_T, \hat{\gamma})\right) \to_d (1 - a(D, \gamma))\, K + a(D, \gamma) S - c_\alpha\left(a(D, \gamma)\right)$$

which establishes correct asymptotic size under sequences with $\sqrt{T}||\pi_T||$ converging.

Next, consider the case where $\sqrt{T}||\pi_T||$ diverges. Let $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma}) = h\left(g_T, \Delta g_T, \hat{\gamma}; ||\pi_T||^{-1}\right)$, and define the random variables $\tilde{D}_T$, $\tilde{\Sigma}$, and $\tilde{\Sigma}_D$ accordingly. ACG equation (6.22) establishes that for this parametrization $\tilde{D}_T \to_p D^*$ for $||D^*|| > 0$, and equations (6.7) and (6.21) together establish that $\tilde{\Sigma}_D \to_p 0$. Our assumption on $a(\tilde{D}_T, \tilde{\gamma})$ thus implies that $a(\tilde{D}_T, \tilde{\gamma}) \to_p a_0$. Since ACG establish the convergence in distribution of (LM, AR) under sequences of this type, we have

that

$$\left(1 - a(\tilde{D}_T, \tilde{\gamma})\right) K_T + a(D_T, \xi_T) S_T - c_\alpha \left(a(D_T, \xi_T)\right) \to_d (1 - a_0) K + a_0 S - c_\alpha (a_0)$$

and thus that the CLC test $\phi_{a\left(\tilde{D}_T, \tilde{\gamma}\right)}$ has asymptotic rejection probability equal to α under these sequences. By the assumed invariance the postmultiplication, however, this implies that $\phi_{a(D_T,\gamma)}$ has asymptotic rejection probability α as well.

To complete the proof, following ACG we can note that the above argument verifies their Assumption B^* and that we can thus invoke ACG Corollary 2.1 to establish the result.

Proof of Proposition 3

Follows by the same argument as the first part of Proposition 1.

Proof of Proposition 4

As discussed in the text, ϕ_K is efficient in the limit problem (21) by the Neyman-Pearson Lemma, and $\phi_{K_T} = \phi_{\tilde{a}(D_T,\hat{\gamma})}$ for $\tilde{a}(D, \gamma) \equiv 0$, $\tilde{a} \in \mathcal{A}_c$, so

$$\lim_{T\to\infty} E_{T,(\theta,\gamma)} [\phi_{K_T}] = \sup_{\tilde{a}\in\mathcal{A}_c} \lim_{T\to\infty} E_{T,(\theta,\gamma)} \left[\phi_{\tilde{a}(D_T,\hat{\gamma})}\right]$$

follows from Proposition 3.

If $a(D, \gamma) = 0$ almost surely, then we have that $\lim_{T\to\infty} E_{T,(\theta,\gamma)} \left[\phi_{a(D_T,\hat{\gamma})}\right] = \lim_{T\to\infty} E_{T,(\theta,\gamma)} [\phi_{K_T}]$ by Proposition 3. If, on the other hand, $Pr\{a(D, \gamma) \neq 0\} = \delta > 0$, note that $D = \mu$ is non-random in the limit problem, so this implies that $a(\mu, \gamma) = a^* \neq 0$. Note, however, that the test ϕ_{a^*} does not satisfy the necessary condition for a most powerful test given in Theorem 3.2.1 in Lehmann and Romano and thus has strictly lower power than the test ϕ_K in the limit problem, which together with Proposition 3 implies that $\lim_{T\to\infty} E_{T,(\theta,\gamma)} \left[\phi_{a(D_T,\hat{\gamma})}\right] < \lim_{T\to\infty} E_{T,(\theta,\gamma)} [\phi_{K_T}]$.

Proof of Theorem 6

Define $\mathcal{M}_L = \{\mu_D \cdot b : b \in \mathbb{R}^p\}$. Note that for any $\zeta > 0$, there exists $C_\zeta > 0$ such that

$$\inf_{a\in[0,1]} \inf_{m\in\mathcal{M}_L : ||m|| > C_\zeta} E_{m,\mu_D,\gamma} [\phi_a] > 1 - \zeta.$$

Note further that $C_\zeta \to \infty$ as $\zeta \to 0$. Since the test ϕ_K is UMP against $m \in \mathcal{M}_L$ for $\Sigma_D = 0$, we can see that for $\beta^u_{m,\mu_D,\gamma} = \sup_{a\in[0,1]} E_{m,\mu_D,\gamma}[\phi_a]$ we have $\beta^u_{m,\mu_D,\gamma} = E_{m,\mu_D,\gamma}[\phi_K]\ \forall m \in \mathcal{M}_L$. Thus,

$$\sup_{m\in\mathcal{M}_L}\left(\beta^u_{m,\mu_D,\gamma} - E_{m,\mu_D,\gamma}[\phi_a]\right) = \sup_{m\in\mathcal{M}_L}\left(E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]\right).$$

Next, note that since as discussed in the proof of Proposition 4 none of the tests $\phi_a : a \in (0,1]$ satisfy the necessary condition for an optimal test against $m \in \mathcal{M}_L$ for $\Sigma_D = 0$ given in Lehman and Romano Theorem 3.2.1. Thus if we define

$$\varepsilon(a) = \sup_{m\in\mathcal{M}_L}\left(E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]\right)$$

we have that $\varepsilon(a) > 0\ \forall a \in (0,1]$. Moreover for all a there is some $m^* \in \mathcal{M}_L$ such that

$$\varepsilon(a) = E_{m^*,\mu_D,\gamma}[\phi_K] - E_{m^*,\mu_D,\gamma}[\phi_a],$$

which can be seen by noting that for $\zeta = \frac{\varepsilon(a)}{2}$, $B_C = \{m : ||m|| \le C\}$, and $A = \mathcal{M}_L \cap B^C_{C_\zeta}$ (for $B^C_{C_\zeta}$ the complement of B_{C_ζ})

$$\sup_{m\in A}\left(E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]\right) \le 1 - \frac{\varepsilon(a)}{2}$$

by the definition of C_ζ. Thus, for $\tilde{A} = \mathcal{M}_L \cap B_{C_\zeta}$,

$$\varepsilon(a) = \sup_{m\in\tilde{A}}\left(E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]\right).$$

Since $\tilde{A}$ is compact and $E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]$ is continuous in m, the sup must be attained by some $m^* \in \tilde{A}$.

Since $E_{m,\mu_D,\gamma}[\phi_a]$ is continuous in a for all m, the fact that

$$\varepsilon(a) = \sup_{m\in\mathcal{M}_L}\left(E_{m,\mu_D,\gamma}[\phi_K] - E_{m,\mu_D,\gamma}[\phi_a]\right)$$

is achieved implies that $\varepsilon(a)$ is continuous in a. We know that $\varepsilon(0) = 0$ by definition, so 0 is the unique minimizer of $\varepsilon(a)$ over $[0,1]$. By the compactness of $[0,1]$, this implies that for any $\delta > 0$ there exists $\bar{\varepsilon}(\delta) > 0$ such that $\varepsilon(a) < \bar{\varepsilon}(\delta)$ only if $a < \delta$. Further, by the intermediate value theorem there exists $a(\delta) > 0$ such that $\varepsilon(a(\delta)) = \frac{\bar{\varepsilon}(\delta)}{2}$.

To prove Theorem 6 we want to show that under the assumptions of the theorem, for all $\nu > 0$ there exists N such that $n > N$ implies

$$\arg\min_{a\in[0,1]} \sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_{a^*}\right]\right) < \nu.$$

Fixing ν, let $\bar{\varepsilon}^* = \bar{\varepsilon}\left(\nu\right)$, $a^* = a(\nu)$, for $\bar{\varepsilon}\left(\cdot\right)$ and $a\left(\cdot\right)$ as defined above. Let $\zeta^* = \frac{\bar{\varepsilon}^*}{4}$, and take C^* to be such that

$$\inf_{m\in\mathbb{R}^k:||m||>C^*} E_{m,\mu_D,\gamma}\left[\phi_{a^*}\right] > 1 - \zeta^*.$$

Under our assumptions and the continuity of $E_{m,\mu_D,\gamma}\left[\phi_a\right]$ in (m, μ_D, γ, a), there exists some N such that for $n > N$,

$$\inf_{a\in[\nu,1]} \sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)\cap B_{C^*}} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_a\right]\right) > 3\zeta^*$$

while

$$\sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)\cap B_{C^*}} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_{a^*}\right]\right) < 3\zeta^*$$

and

$$\sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)\cap B^C_{C^*}} \left(\beta^u_{m,n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_{a^*}\right]\right) < 2\zeta^*.$$

Thus, for $n > N$ we have

$$\sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_{a^*}\right]\right) < \inf_{a\in[\nu,1]} \sup_{m\in\mathcal{M}_D\left(\mu_{D,n},\gamma_n\right)\cap B_{C^*}} \left(\beta^u_{m,\mu_{D,n},\gamma_n} - E_{m,\mu_{D,n},\gamma_n}\left[\phi_a\right]\right)$$

and thus that $a\left(\mu_{D,n},\gamma_n\right) < \nu$ since $a^* < \nu$. Since we can repeat this argument for all $\nu > 0$ we obtain that $a\left(\mu_{D,n},\gamma_n\right) \to 0$ as desired.

Proof of Corollary 3

Let $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma}) = h\left(g_T, \Delta g_T, \hat{\gamma}; r_T^{-1}/\sqrt{T}\right)$ for h as defined in (17), and note that this coincides with the definition of $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$ given near the beginning of Section 8.3. By the postmultiplication-invariance of plug-in tests with equivariant $\hat{\mu}_D$, tests based on $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$ with plug-in estimate $\tilde{\mu}_D = \tilde{D}_T$ will be the same as those based on $(g_T, \Delta g_T, \hat{\gamma})$ with estimate $\hat{\mu}_D = D_T$. To prove the result we will focus on tests based on $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$.

As established in the main text, $(\tilde{g}_T, \Delta\tilde{g}_T, \tilde{\gamma})$ converges in distribution to $(g, \Delta g, \gamma)$ in a Gaussian shift model with $\mu = E\left[Z_t Z_t\right]' c$ and $b = b^*$. Note that in linear IV we have

$$\mathcal{M}_D\left(\mu_D, \gamma\right) = \left\{\left(I - \Sigma_{\beta g} \cdot b\right)^{-1} \mu_D \cdot b : b \in \mathbb{R}\right\}.$$

Hence, for any sequence $(\mu_{D,n}, \gamma_n)$ with $\mu_{D,n} \to \mu$, $||\mu|| > 0$, and $\Sigma_{\beta g}\left(\gamma_n\right) \to 0$ we can see that for any $C > 0$

$$d_H\left(\mathcal{M}_D\left(\mu_{D,n}, \gamma_n\right) \cap B_C, \left\{\mu_D \cdot b : b \in \mathbb{R}^p\right\} \cap B_C\right) \to 0,$$

so by Theorem 6 we have that $a_{PI}\left(\mu_{D,n}, \gamma_n\right) \to 0$. Note, however, that under our assumptions $(\hat{\mu}_D, \hat{\gamma}) \to_p (\mu, \gamma)$ with $||\mu|| > 0$ and $\Sigma_{\beta g}(\gamma) = \Sigma_D(\gamma) = 0$. Thus, the Continuous Mapping Theorem yields that $a_{PI}\left(\hat{\mu}_D, \hat{\gamma}\right) \to_p 0$.

Proof of Corollary 4

Note that

$$\mathcal{M}\left(\hat{\gamma}\right) = \left\{\hat{\Omega}_\eta^{-\frac{1}{2}}\left(f\left(\theta\right) - f\left(\theta_0\right)\right) : \theta \in \Theta\right\} = r_T^{-\frac{1}{2}}\left\{\left(r_T^{-1}\hat{\Omega}_\eta\right)^{-\frac{1}{2}}\left(f\left(\theta\right) - f\left(\theta_0\right)\right) : \theta \in \Theta\right\}.$$

For any sequence $r_T^{-1}\Omega_{\eta,T} \to \Omega_{\eta,0}$ and $B_C = \{m \in \mathbb{R}^p : ||m|| \leq C\}$ for $C > 0$ we have that

$$\lim_{T\to\infty} d_H\left(\left\{r_T^{-\frac{1}{2}}\left(r_T^{-1}\Omega_{\eta,T}\right)^{-\frac{1}{2}}\left(f\left(\theta\right) - f\left(\theta_0\right)\right) : \theta \in \Theta\right\} \cap B_C, \left\{r_T^{-\frac{1}{2}}\Omega_{\eta,0}^{-\frac{1}{2}}\left(f\left(\theta\right) - f\left(\theta_0\right)\right) : \theta \in \Theta\right\} \cap B_C\right) = 0.$$

From the definition of differentiability, we know that

$$\lim_{\theta\to\theta_0} \frac{f\left(\theta\right) - f\left(\theta_0\right) - \frac{\partial}{\partial\theta'} f\left(\theta_0\right)\left(\theta - \theta_0\right)}{||\theta - \theta_0||} = 0.$$

Thus, for any sequence $\delta_T \to 0$ we have that

$$\lim_{\delta_T\to 0} \sup_{||\theta-\theta_0||\leq\delta_T} \frac{1}{\delta_T}\left(f\left(\theta\right) - f\left(\theta_0\right) - \frac{\partial}{\partial\theta'} f\left(\theta_0\right)\left(\theta - \theta_0\right)\right) = 0.$$

Moreover, by our identifiability assumption on θ_0 we know that for any constant $K > 0$,

$$\lim_{T\to\infty} \sup_{\theta : r_T^{-\frac{1}{2}}||\Omega_{\eta,0}^{-\frac{1}{2}} f(\theta) - \Omega_{\eta,0}^{-\frac{1}{2}} f(\theta_0)|| \leq K} ||\theta - \theta_0|| = 0.$$

Combined with the previous equation, this implies that

$$\lim_{T\to\infty} \sup_{\theta: r_T^{-\frac{1}{2}} ||\Omega_{\eta,0}^{-\frac{1}{2}} f(\theta) - \Omega_{\eta,0}^{-\frac{1}{2}} f(\theta_0)|| \leq K} r_T^{-\frac{1}{2}} \left\| \Omega_{\eta,0}^{-\frac{1}{2}} \left(f(\theta) - f(\theta_0) \right) - \Omega_{\eta,0}^{-\frac{1}{2}} \frac{\partial}{\partial \theta'} f(\theta_0) (\theta - \theta_0) \right\| = 0$$

which in turn shows that for any $C > 0$, provided θ belongs to the interior of Θ

$$d_H \left(r_T^{-\frac{1}{2}} \left\{ \Omega_{\eta,0}^{-\frac{1}{2}} \left(f(\theta) - f(\theta_0) \right) : \theta \in \Theta \right\} \cap B_C, r_T^{-\frac{1}{2}} \left\{ \Omega_{\eta,0}^{-\frac{1}{2}} \frac{\partial}{\partial \theta'} f(\theta_0) \cdot b : b \in \mathbb{R}^p \right\} \cap B_C \right) \to 0.$$

Thus, we see that for any $r_T^{-1} \Omega_{\eta,T} \to \Omega_{\eta,0}$ the convergence required by Theorem 6 holds, so for the corresponding sequence $\{\gamma_T\}_{T=1}^{\infty}$ we have that $a_{MMR}(\gamma_T) \to 0$. Hence, by the Continuous Mapping Theorem we have that under our assumptions $a_{MMR}(\hat{\gamma}) \to_p 0$.

One can show that sequences of local alternatives of the form $\theta_T = \theta_0 + r_T^{\frac{1}{2}} b^*$ yield Gaussian Shift limit problems in this model. The fact that $a_{MMR}(\hat{\gamma}) \to_p 0$ implies, by Proposition 4, that the MMR test is asymptotically efficient against such sequences, and hence that the MMR test is asymptotically efficient under strong and semi-strong identification, as we wanted to prove.

Appendix D: Details of Weak IV Simulation

In Section 7.2 we discuss simulation results in weak IV limit problems calibrated to match parameters estimated using data from Yogo (2004). This section details the estimates, simulation design, and implementation of CLR and PI tests underlying these results.

Appendix D.1: Estimation of Parameters for the Limit Problem

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

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

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

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

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

Appendix D.2: Simulation Design

For each country i we consider the problem of testing $H_0 : \beta = \beta_0$ in the limit problem. For true parameter value β, in simulation runs $b = 1, ..., B$ we draw

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

where

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

Note that this is the limiting distribution (6) of the normalized moment condition and Jacobian $(g_T, \Delta g_T)$ in a weak IV problem with true parameters β, $\Omega = \hat{\Omega}_i$, and $\mu = \hat{\mu}_i$. We then calculate the S and K tests $\phi_{S,b}$, $\phi_{K,b}$ as in (15) and (14). We define the QCLR test as in Theorem 2 and, following Kleibergen (2005), take $r = D_b' \hat{\Sigma}_{D,i}^{-1} D_b$ for $\hat{\Sigma}_{D,i} = \hat{\Sigma}_{\theta\theta,i} - \hat{\Sigma}_{\theta g,i} \hat{\Sigma}_{g\theta,i}$. Finally, to calculate the PI test we take

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

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

For each value β in a grid we estimate the power of each test against this alternative by averaging

over $B = 10,000$ simulations, e.g. estimating the power of ϕ_K by $\frac{1}{B}\sum \phi_{K,b}$, and repeat this exercise for each of the eleven countries considered.

Appendix D.3: Implementation of PI Test

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

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

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

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

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

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

We next consider a grid B of 41 values for the alternative β_h. For each value β_h we solve for

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix E: Details of NKPC Example

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

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

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

and note that given our assumptions this will be consistent for

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

To derive the weak convergence (23), as well as the form of the matrices A_T, note that since we have assumed $\varphi_T \to \varphi$ and $\sqrt{T}\left(\hat{\varphi} - \varphi_T\right) \to_d N\left(0, \Sigma_{\varphi\varphi}\right)$, the Δ-method (Theorem 3.1 in Van der

Vaart (2000)) yields that

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

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

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

E.1 Possible Sources of Weak Identification

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

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

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

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

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

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

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

so if

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

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

E.2 Derivation of the Limit Problem

To derive the form of the limit problem (23) we need to understand the behavior of g_T and Δg_T under alternatives. Note that for alternative θ_T and true reduced-form parameter value φ_T, we have that since $f(\varphi_T,\theta_T)=0$,

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

Define

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

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

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

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

and hence

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

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

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

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

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

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

$$m(\varphi_T,\theta_T) = \left(\frac{1}{1+\rho_T} - \frac{1}{1+\rho_0}\right)\left(h_2(\theta_0)\left[\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right]_2 - h_3(\theta_0)h_1(\theta_0)\left[\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right]_1 + e_\pi\right)$$

$$+\left(\frac{(1-\nu_T)^2}{\nu_T(1+\rho_T)} - \frac{(1-\nu_0)^2}{\nu_0(1+\rho_0)}\right)h_1(\theta_0)\left[\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right]_1 + \left(\frac{\rho_T}{1+\rho_T} - \frac{\rho_0}{1+\rho_0}\right)e_\pi$$

$$= h_4(\theta_0,\theta_T)\left[\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right]_1 + h_5(\theta_0,\theta_T)\left[\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)\right]_2$$

for

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

and

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

Thus, knowledge of $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)$ suffices to let us calculate $m(\varphi_T,\theta)$ for any alternative θ in the sample of size T. Consequently, in each sample size T, an estimate of $\mu_T = \frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)A_T$ implies a corresponding set $\mathcal{M}_T(\mu_T) = \left\{\sqrt{T}m(\varphi_T,\theta) : \theta\in\Theta\right\}$. For a given convergent sequence φ_T, we can then define $\mathcal{M}$ in the limit problem as $\mathcal{M}(\mu) = \lim_T(\mathcal{M}_T(\mu)\cap C)$ for any compact set C: the restriction to the set C ensures convergence, and has the effect of restricting attention to a particular neighborhood of fixed alternatives for weakly identified parameters and local alternatives for strongly identified parameters. Note that in any given sample size we need not know A_T to calculate $\mathcal{M}_T(\mu_T)$ once given an estimate of $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)$, so if we just treat $\frac{\partial}{\partial\theta'}f(\hat{\varphi},\theta_0)$ as a Gaussian random matrix with mean $\frac{\partial}{\partial\theta'}f(\varphi_T,\theta_0)$ and proceed accordingly, this will (asymptotically) correspond to using the correct $\mathcal{M}(\mu)$ under all sequences yielding limit problems in this class. Indeed, this is the approach we adopt to calculate plug-in tests in our simulations.

E.3 NKPC Simulation Details

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

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

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

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

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

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

References

Anderson T.W. and H. Rubin (1949): "Estimators for the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570-582.

Andrews D.W.K, M.J. Moreira and J. Stock (2006): "Optimal Two-Sided Invariant Similar Tests of Instrumental Variables Regression," *Econometrica*, 74, 715-752.

Andrews D.W.K, and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong, and Strong Identification," *Econometrica*, 80, 2153-2211.

Andrews D.W.K, and X. Cheng (2013): "GMM Estimation and Uniform Subvector Inference with Possible Identification Failure," *Econometric Theory*, forthcoming.

Andrews D.W.K., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing

the Asymptotic Size of Confidence Sets and Tests," Cowles Foundation working paper.

Andrews I. and A. Mikusheva (2012): "A Geometric Approach to Weakly Identified Econometric Models," Unpublished Manuscript.

Canova and Sala (2010): "Back to Square One: Identification Issues in DSGE Models," *Journal of Monetary Economics*, 56, 431-449.

Chaudhuri S. and E. Zivot (2011): "A New Method of Projection-Based Inference in GMM with Weakly Identified Nuisance Parameters", *Journal of Econometrics*, 164, 239-251.

Christiano L.J. and M. Eichenbaum (1992): "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations", *American Economic Review*, 82, 430-450.

Duffie D. and K. J. Singleton (1993): "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929-952.

Elliott G., U.K. Mueller, and M.W. Watson (2012): "Nearly Optimal Tests When a Nuisance Parameter is Present Under The Null Hypothesis," Unpublished Manuscript.

Gourieroux C., A. Montfort and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85-S118.

Guggenberger P. and R.J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory*, 21, 667-709.

Guggenberger P. and R.J. Smith (2008): "Generalized Empirical Likelihood Ratio Tests in Time Series Models with Potential Identification Failure," *Journal of Econometrics*, 142, 134-161.

Guggenberger P., J.J.S. Ramalho, R.J. Smith (2012): "GEL Statistics Under Weak Identification," *Journal of Econometrics*, 170, 331-349.

Hansen L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Jansson and Moreira (2006): "Optimal Inference in a Model with Nearly Integrated Regressors," *Econometrica*, 74, 681-714.

Kleibergen F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781-1803.

Kleibergen F. (2005): "Testing Parameters in GMM Without Assuming They Are Identified," *Econometrica*, 73, 1103-1123.

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

Koziol J.A. and M.D. Perlman (1978): "Combining Independent Chi Squared Tests," *Journal of the American Statistical Association*, 73, 753-763.

Kubokawa T., C. P. Roberts, and A. K. Md. E. Saleh (1993): "Estimation of Noncentrality Parameters," *The Canadian Journal of Statistics*, 21, 45-57.

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

Magnusson L.M. and S. Mavroeidis (2010): "Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve," *Journal of Money, Banking, and Credit*, 42, 465-481.

Marden J.I. (1982): "Combining Independent Noncentral Chi Squared or F Tests," *The Annals of Statistics*, 10, 266-277.

Mavroeidis S., M. Plagborg-Moller, J. Stock (2013): "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve," Unpublished Manuscript.

McFadden D.L. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995-1026.

Monti K.L. and P.K. Sen (1976): "The Locally Optimal Combination of Independent Test Statistics," *Journal of the American Statistical Association*, 71, 903-911.

Moreira H. and Moreira M. J. (2013): "Contributions to the Theory of Optimal Tests," Unpublished Manuscript.

Moreira M.J. (2001): "Tests with Correct Size when Instruments Can Be Arbitrarily Weak," Unpublished Manuscript, University of California, Berkeley.

Moreira M.J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027-1048.

Mueller U.K. (2011): "Efficient Tests Under a Weak Convergence Assumption," *Econometrica*, 79, 395-435.

Newey W. and D.L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," Chapter 36 of D.L. McFadden and R.F. Engle Eds, *Handbook of Econometrics, Volume 4*, Elsever.

Newey W.K. and K.D. West (1987): "Hypothesis Testing With Efficient Method of Moments Estimation," *International Economic Review*, 28, 777-787.

Olea J. L. M. (2012): "Efficient Conditionally Similar Tests: Finite-Sample Theory and Large-Sample Applications," working paper.

Otsu T. (2006): "Generalized Empirical Likelihood Inference for Nonlinear and Time Series Models Under Weak Identification," *Econometric Theory*, 22, 513-527.

Ramalho J.J.S. and R.J. Smith (2004): "Goodness of Fit Tests for Moment Condition Models," Unpublished Manuscript, University of Cambridge.

Rotemberg J., and M. Woodford (1997): "An Optimization Based Econometric Framework for the Evaluation of Monetary Policy," *NBER Macroeconomic Annual*, 12, 297-346.

Ruge-Murcia F. (2010): "Estimating Nonlinear DSGE Models by the Simulated Method of Moments: with an Application to Business Cycles," *Journal of Economic Dynamics and Control*, 36, 914-938.

Saxena K.M.L and K. Alam (1982): "Estimation of the Non-centrality Parameter of a Chi Squared Distribution," *The Annals of Statistics*, 10, 1012-1016.

Sbordone, A. M. (2005), "Do expected future marginal costs drive inflation dynamics?" *Journal of Monetary Economics*, 52, 1183-1197.

Smith, R.J., 2007. Weak instruments and empirical likelihood: a discussion of the papers by D.W.K. Andrews and J.H. Stock and Y. Kitamura. In: Blundell, R.W., Newey, W.K., Persson, T. (Eds.), Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, vol. 3. In: Econometric Society Monograph Series, ESM, vol. 43. Cambridge University Press, Cambridge, pp. 238–260 (Chapter 8).

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

Stock J. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stoye, Jörg. "minimax regret." The New Palgrave Dictionary of Economics. Online Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2009.

Van der Vaart A.W. (2000): Asymptotic Statistics. New York, Cambridge University Press.

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

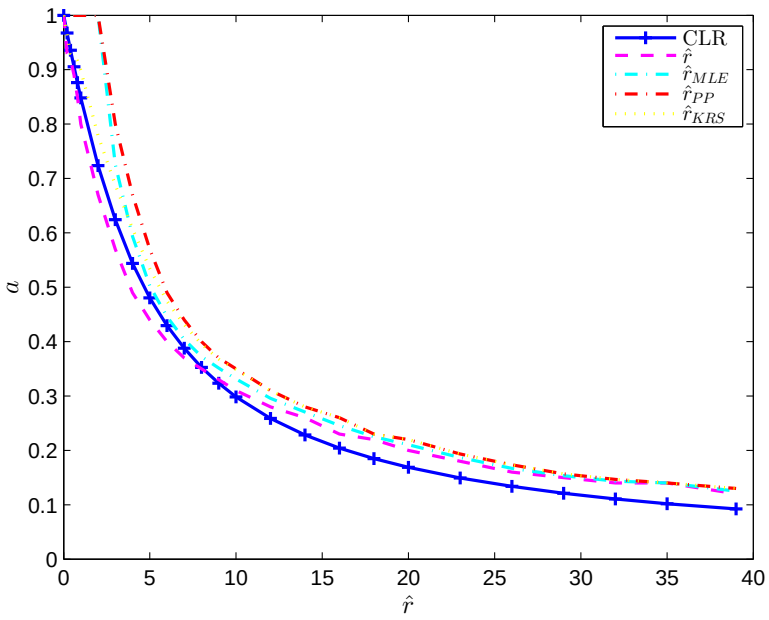

Figure 1: Weight functions $a_{CLR}(\hat{r})$ for CLR and $a_{MMRU}\left(\tilde{r}\left(\hat{r}\right)\right)$ for PI tests with different estimators $\tilde{r}$ of r discussed in Section 6.1 for linear IV with two instruments and homoskedastic errors.

	CLR	PI-$\hat{r}$	PI-$\hat{r}_{MLE}$	PI-$\hat{r}_{PP}$	PI-$\hat{r}_{KRS}$	AR	K
$k=2$	1.18%	1.44%	0.72%	0.72%	0.88%	9.40%	29.96%
$k=5$	2.14%	5.90%	1.37%	1.07%	2.04%	25.05%	53.71%
$k=10$	3.51%	13.21%	2.29%	2.18%	4.00%	30.76%	64.62%

Table 1: Maximal power shortfall relative to other tests considered, in linear IV model with homoskedastic errors. For each k (number of instruments) we calculate the point-wise maximal power of the tests studied at each of the alternatives used to generate Figures 4 and 5. For each test, we report the largest margin by which the power of that test falls short of point-wise maximal power. CLR denotes the CLR test of Moreira (2003) while PI-$\hat{r}$, PI-$\hat{r}_{MLE}$, PI-$\hat{r}_{PP}$, and PI-$\hat{r}_{KRS}$ denote the PI tests with weight functions $a_{MMRU}\left(\hat{r}\right)$, $a_{MMRU}\left(\hat{r}_{MLE}\right)$, $a_{MMRU}\left(\hat{r}_{PP}\right)$, and $a_{MMRU}\left(\hat{r}_{KRS}\right)$, respectively. AR is the Anderson Rubin test (equivalent to the S test) and K is Kleibergen's (2002) K test.

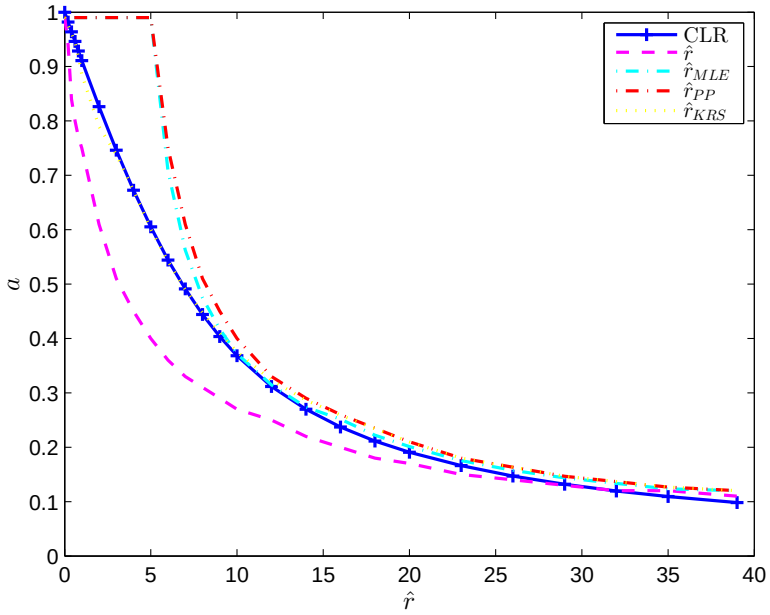

Figure 2: Weight functions $a_{CLR}(\hat{r})$ for CLR and $a_{MMRU}\left(\tilde{r}\left(\hat{r}\right)\right)$ for PI tests with different estimators $\tilde{r}$ of r discussed in Section 6.1 for linear IV with five instruments and homoskedastic errors.

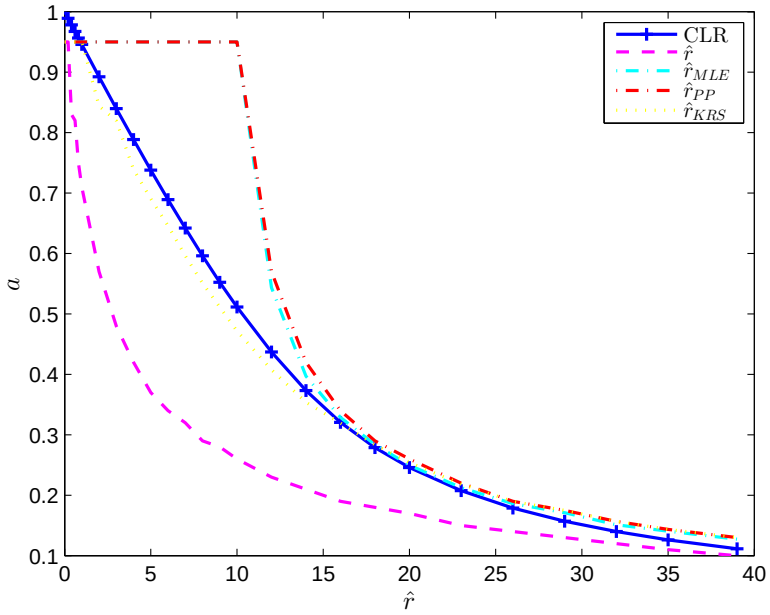

Figure 3: Weight functions $a_{CLR}(\hat{r})$ for CLR and $a_{MMRU}\left(\tilde{r}\left(\hat{r}\right)\right)$ for PI tests with different estimators $\tilde{r}$ of r discussed in Section 6.1 for linear IV with ten instruments and homoskedastic errors.

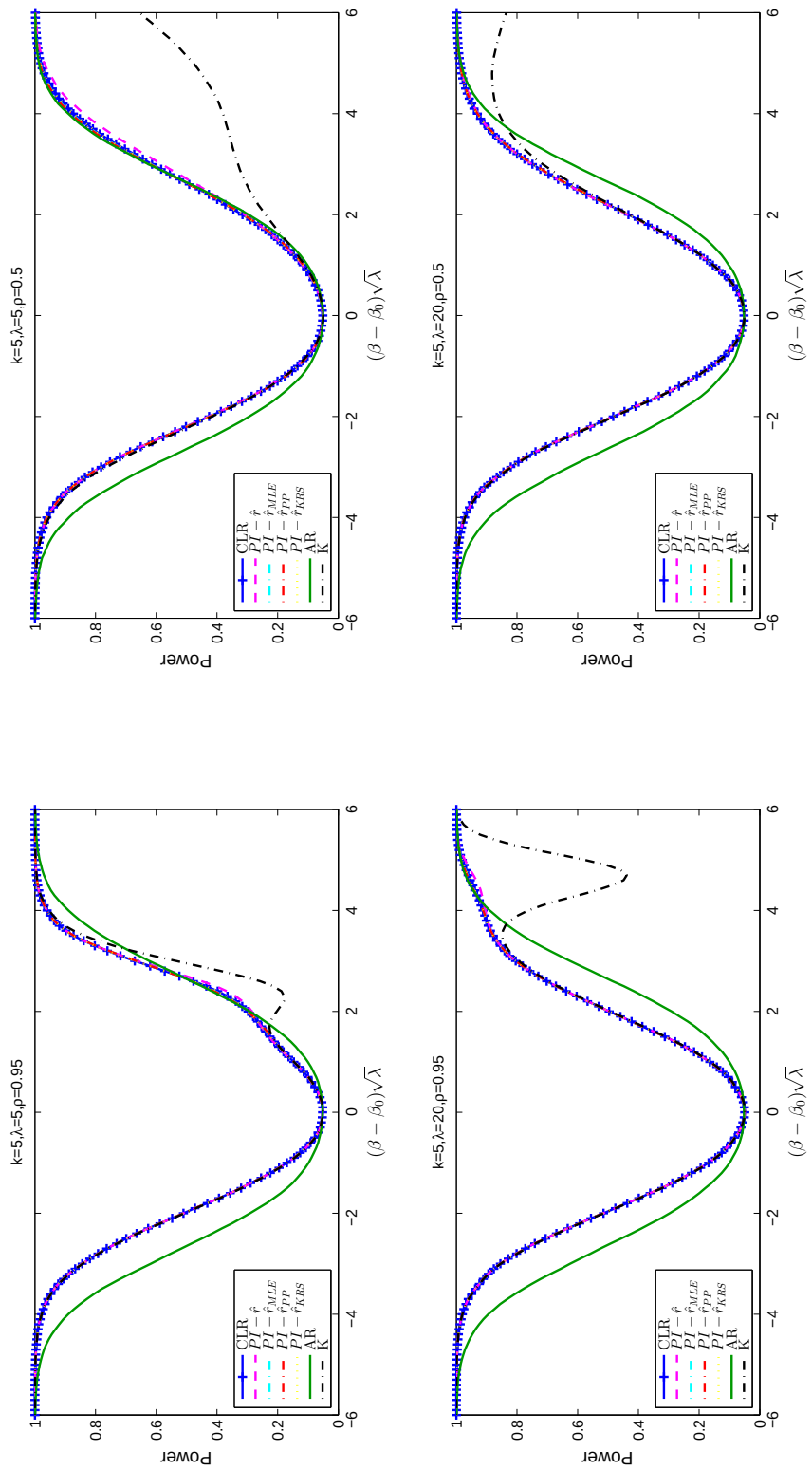

Figure 4: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with five instruments, discussed in Section 7.1.

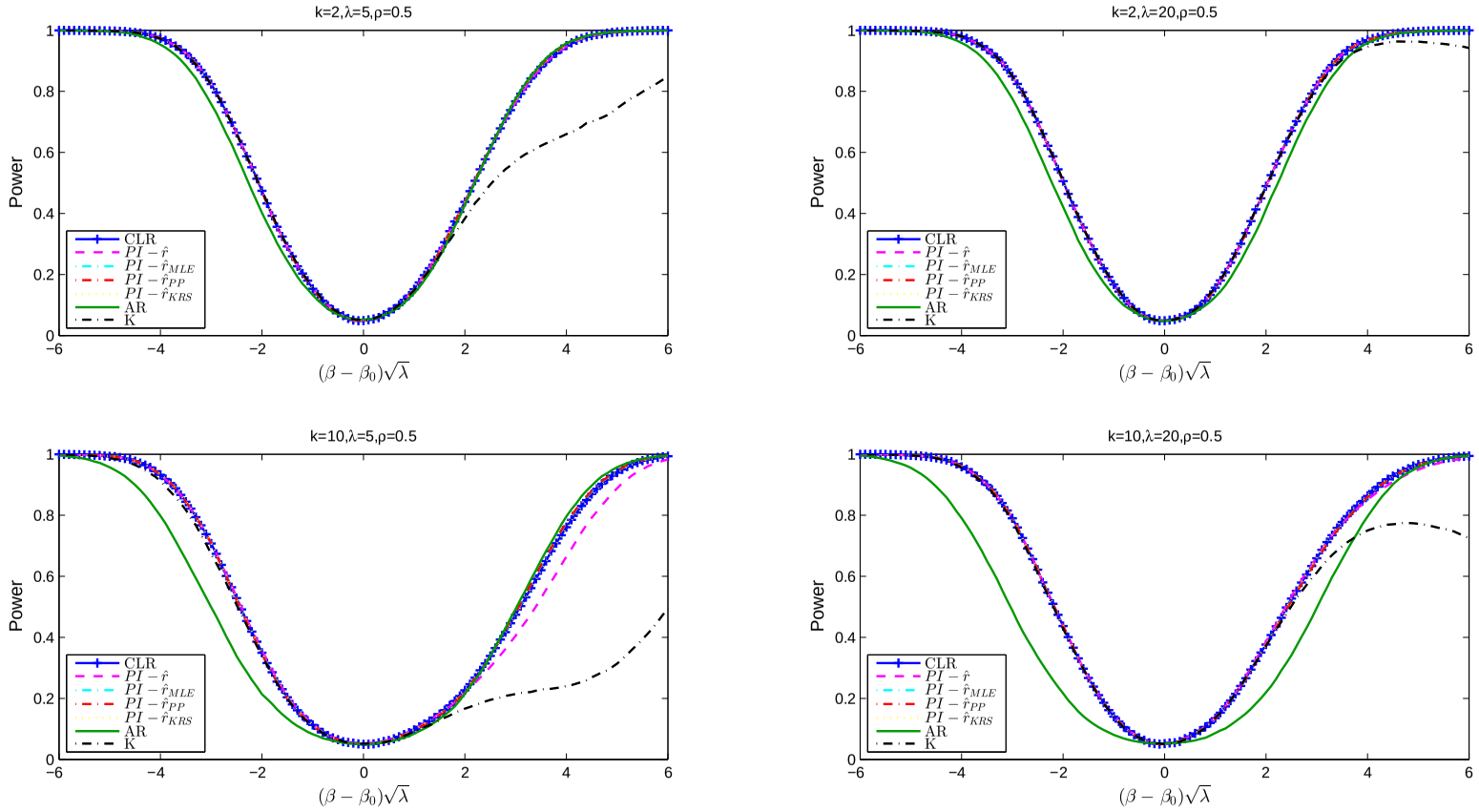

Figure 5: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with two instruments and ten instruments, discussed in Section 7.1.

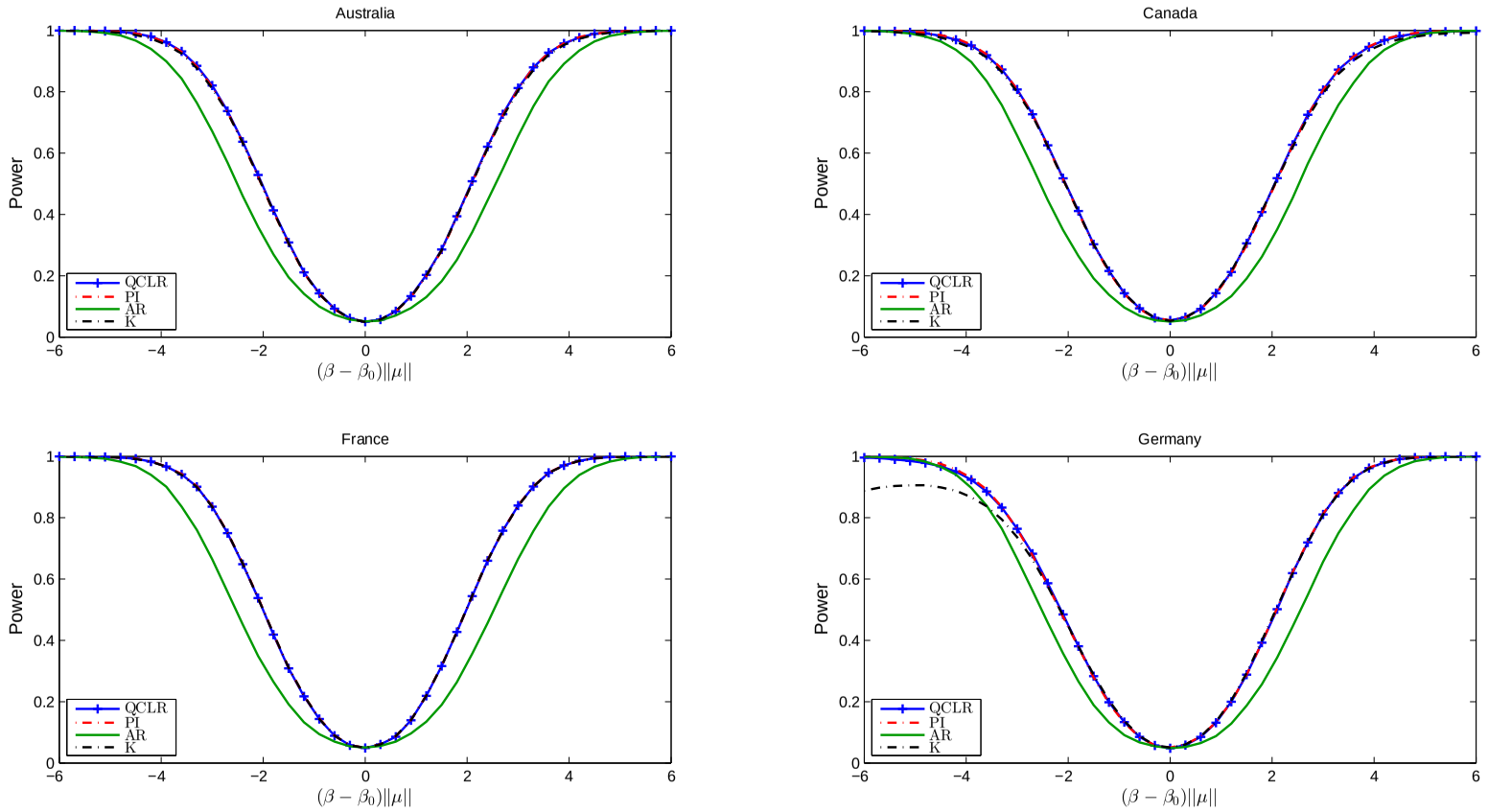

Figure 6: Power functions for QCLR, K, AR (or S), and PI tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 7.2 .

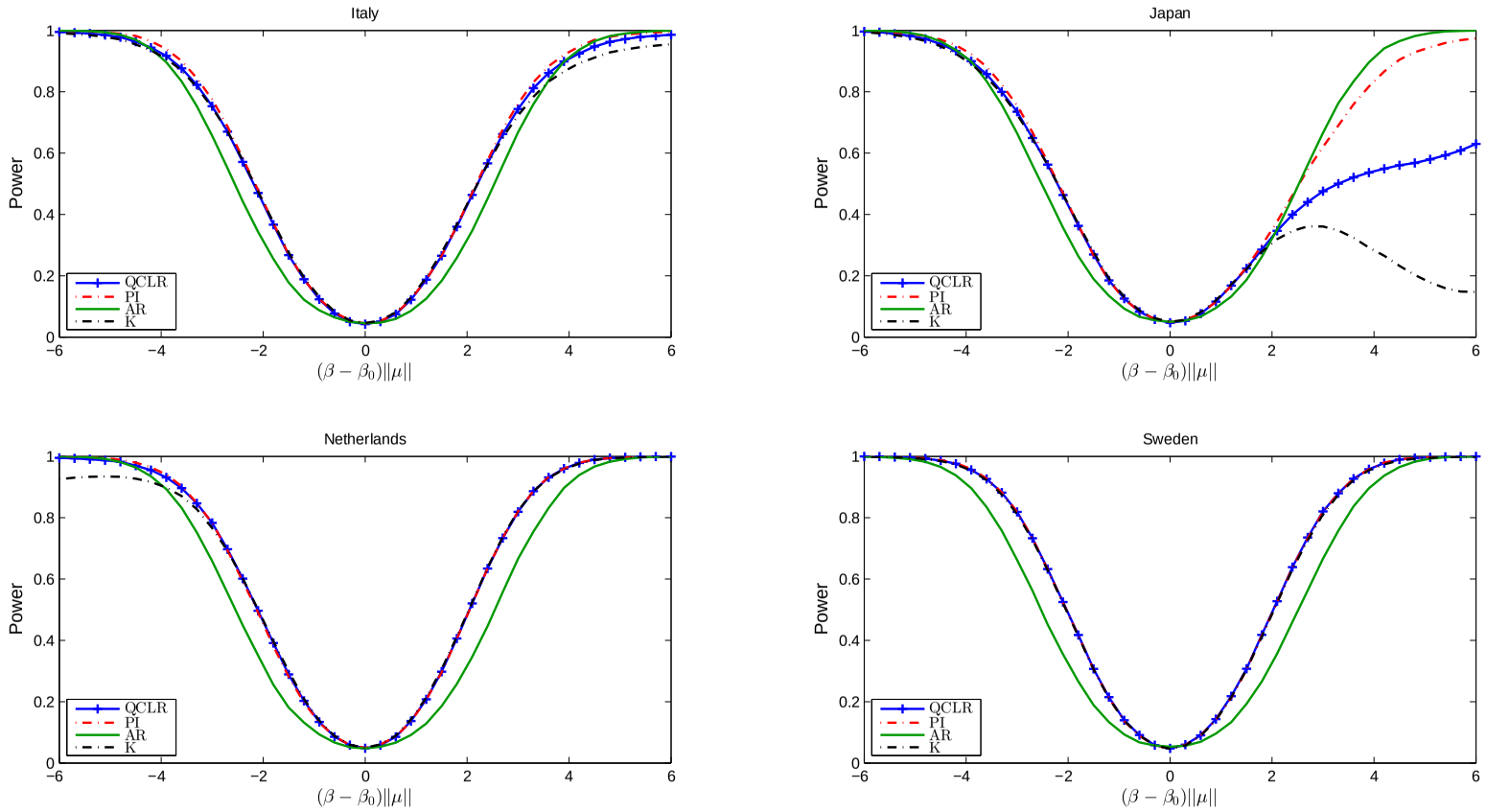

Figure 7: Power functions for QCLR, K, AR (or S), and PI tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 7.2.

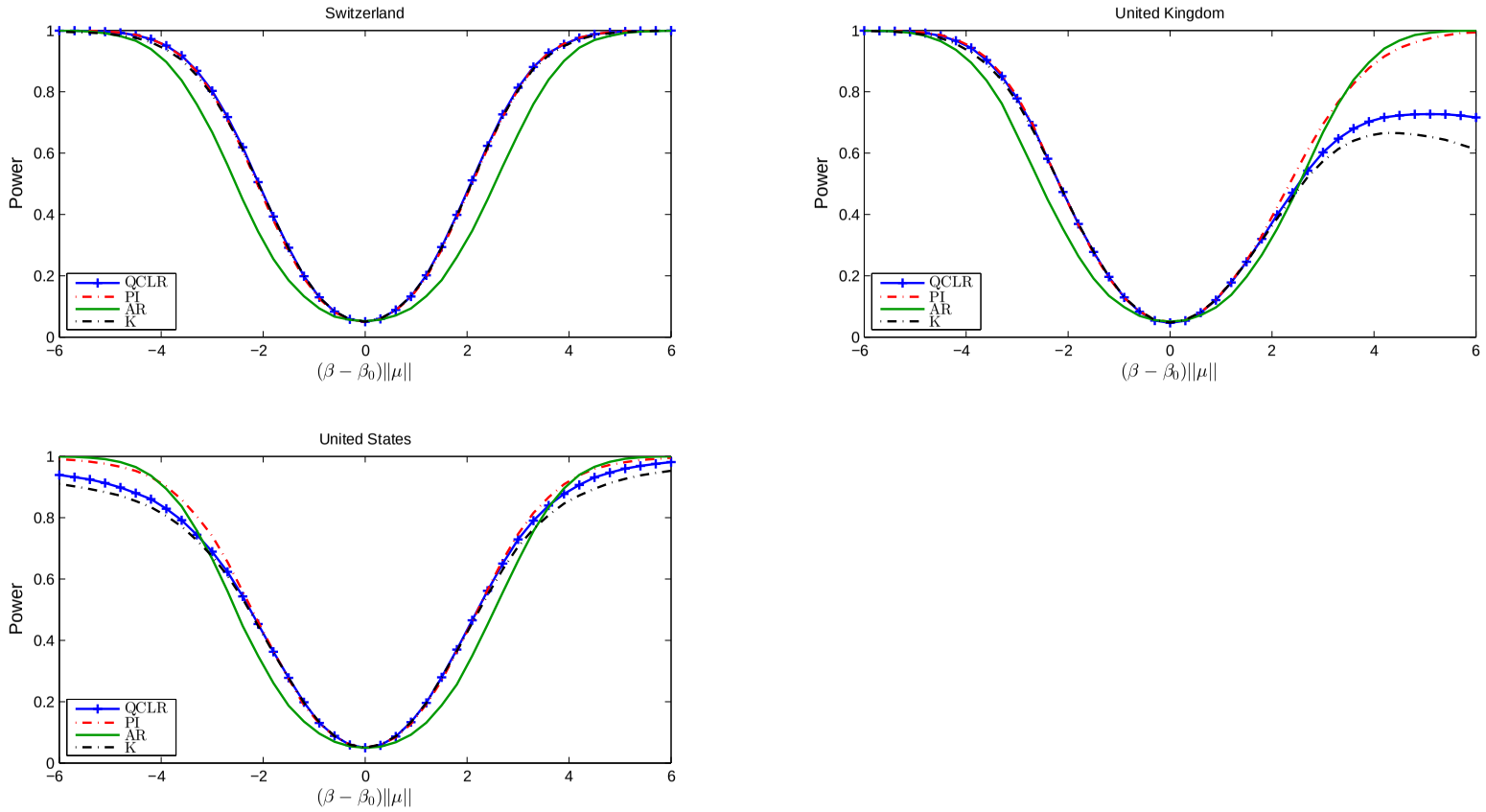

Figure 8: Power functions for QCLR, K, AR (or S), and PI tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 7.2.

	QCLR	PI	AR	K
Australia	0.16%	0.53%	17.85%	0.95%
Canada	0.40%	0.78%	17.85%	1.38%
France	0.33%	0.11%	20.49%	0.26%
Germany	0.81%	0.92%	17.76%	11.09%
Italy	2.40%	0.94%	14.05%	5.80%
Japan	41.39%	7.86%	11.23%	85.22%
Netherlands	1.05%	1.76%	19.07%	7.42%
Sweden	0.54%	0.53%	18.08%	0.94%
Switzerland	0.21%	1.33%	17.53%	1.53%
United Kingdom	28.30%	2.59%	14.09%	38.70%
United States	8.53%	1.57%	11.62%	11.09%

Table 2: Maximal point-wise power shortfall relative to other tests considered, for simulations calibrated to match data in Yogo (2004). In particular, we calculate the point-wise maximal power of the four tests studied at each of the alternatives used to generate Figures 6-8. For each test, we report the largest margin by which the power of that test falls short of point-wise maximal power. QCLR denotes the quasi-CLR test of Kleibergen (2005) while PI is the plug-in test discussed in Section 7.2. AR is the Anderson Rubin test (equivalent to the S test) and K is Kleibergen's (2005) K test.

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

Table 3: Size of nominal 5% tests in NKPC simulation example discussed in Section 9, based on 10,000 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

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

Table 4: Maximal point-wise differences in false coverage probability of nominal 5% tests in NKPC example discussed in Section 9. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test fall short relative to the JK test is 3%. Based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

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

Table 5: Maximal point-wise differences in false coverage probability of **size corrected** 5% tests in NKPC example discussed in Section 9. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test fall short relative to the JK test is 1.6%. Based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

	PI	JK	S	K
Expected Area: feasible confidence sets	0.084	0.873	0.110	0.094
Expected Area: corrected confidence sets	0.131	0.141	0.169	0.138

Table 6: Expected area of 95% confidence sets formed by inverting tests in NKPC example discussed in Section 9, based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.