Essays in Weak Identification
by
Isaiah Smith Andrews
B.A., Yale University (2009)
Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2014
© Isaiah Smith Andrews, MMXIV. All rights reserved.
The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Signature redacted
Department of Economics
May 15, 2014

Signature redacted
Castle-Krob Career Development Associate Professor of Economics
Thesis Supervisor

Signature redacted
Jane Berkowitz Carlton and Dennis William Carlton Professor of
Microeconomics
Thesis Supervisor

Signature redacted
Jerry Hausman
John and Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Signature redacted
Michael Greenstone
3M Professor of Environmental Economics
Chairman, Department Committee on Graduate Studies
Essays in Weak Identification

by

Isaiah Smith Andrews

Submitted to the Department of Economics
on May 15, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

Economic researchers and policymakers need reliable tools both to estimate economic relationships and to measure the uncertainty surrounding their estimates. Unfortunately, economic data sometimes contains limited information useful for estimating relationships of interest. In such cases, the statistical techniques commonly used in applied economics can break down and fail to accurately reflect the level of uncertainty present in the data. If they rely on such tools, researchers and policymakers may come away with serious misconceptions about the precision and reliability of their estimates. Econometricians refer to models where the lack of information in the data causes common statistical techniques to break down as weakly identified.

In this thesis, I examine several questions relating to weak identification. In the first chapter, I introduce the class of conditional linear combination tests. These tests control size under weak identification and have a number of optimality properties in a conditional problem. I 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.

In the second chapter, I consider the problem of detecting weak identification. When weak identification is a concern researchers frequently calculate confidence sets in two steps, first assessing the strength of identification and then deciding whether to use an identification-robust confidence set. Two-step procedures of this sort may generate highly misleading confidence sets, and I demonstrate that two-step confidence sets based on the first stage F-statistic can have extremely poor coverage in linear instrumental variables models with heteroskedastic errors. I introduce a simple approach to detecting weak identification and constructing two-step confidence sets which controls coverage distortions.

In the third chapter, joint with Anna Mikusheva, we consider minimum distance statistics and show that in a broad class of models the problem of testing under weak identification is closely related to the problem of testing a “curved null” in a finite-sample Gaussian model. Using the curvature of the model, we develop new finite-sample bounds on the distribution of minimum-distance statistics, which we show can be used to detect weak identification and to construct tests robust to weak
identification.

Thesis Supervisor: Anna Mikusheva
Title: Castle-Krob Career Development Associate Professor of Economics

Thesis Supervisor: Whitney Newey
Title: Jane Berkowitz Carlton and Dennis William Carlton Professor of Microeconomics

Thesis Supervisor: Jerry Hausman
Title: John and Jennie S. MacDonald Professor of Economics
Acknowledgments

I am immeasurably indebted to my advisors. Anna Mikusheva has been a truly fantastic mentor, and has taught me an incredible amount. Her patience, encouragement, and drive have helped me tremendously, and I truly lack the words to express how grateful I am. Whitney Newey’s enthusiasm for new ideas encouraged my growth as a researcher, while his guidance greatly helped in navigating both graduate school and the early stages of my life as a professional economist. Jerry Hausman’s advice and feedback has been invaluable, while his dedication to his students is an inspiration.

I have benefited from interaction with many other faculty members at MIT, and particularly owe thanks to Josh Angrist, Victor Chernozhukov, Amy Finklestein, Parag Pathak, Jim Poterba, and Ben Olken. I have also had the privilege of interacting with many remarkable classmates, and am especially grateful to Daniel Barron, Arun Chandrasekhar, Denis Chetverikov, Conrad Miller, Miikka Rokkenen, Brad Shapiro, Adam Sacarny, and Nils Wernerfelt. I am also indebted to the professors at Yale who set me on the path to graduate school in the first place including Nick Barberis, Truman Bewley, Gary Gorton, and especially Don Andrews.

I am grateful to my mother Cheryl and her husband Steve, as well as to my father Marcellus and his wife Kanwal, all of whose support has meant so much to me over the course of my PhD. Finally, I have been tremendously lucky to have Welton Blount in my life, whose support, encouragement, perspective, patience, and humor has meant more to me than I can say.

All the work contained in this thesis was conducted while I was on a National Science Foundation graduate research fellowship, and support under grant number 1122374 is gratefully acknowledged. I am also grateful for support from the Ford Foundation predoctoral fellowship earlier in my graduate career.
Chapter Acknowledgments

For the first chapter, I am 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, Bruce Hansen, Sally Hudson, Peter Hull, Conrad Miller, Scott Nelson, Jose Montiel Olea, Miikka Rokkanen, Adam Sacarny, Annalisa Scognamiglio, Brad Shapiro, Ashish Shenoy, Stefanie Stantcheva, the participants of the MIT Econometrics Lunch, and seminar participants at Penn State, Yale, Wisconsin, and Princeton for helpful comments. I thank Marcelo Moreira for help in finding and resolving a problem with the implementation of the Moreira and Moreira (2013) tests in an earlier draft, and for productive discussions about the Moreira and Moreira (2013) procedures.

For the second chapter, I am grateful to Anna Mikusheva, Whitney Newey, and Jerry Hausman for their guidance and support, and to Josh Angrist, Victor Chernozhukov, Jose Montiel Olea, Nils Wernerfelt, and the participants of the MIT Econometrics Lunch for helpful comments.

For the third chapter, Anna Mikusheva acknowledges support from the Castle-Krob Career Development Chair and Sloan Research Fellowship. We are extremely grateful to Professor of Differential Geometry Toby Colding for discussing geometric issues with us, and for checking our geometric proofs. We are grateful to Gary Chamberlain, Victor Chernozhukov, Jerry Hausman, Mattias Cattaneo, Whitney Newey and Jim Stock, as well as participants at numerous seminars, for helpful comments.
## Contents

1 Conditional Linear Combination Tests for Weakly Identified Models 11
   1.1 Weakly Identified Limit Problems .............................. 16
   1.2 Pivotal Statistics Under Weak Identification .................. 21
      1.2.1 Other Asymptotically Pivotal Statistics .................. 23
      1.2.2 Distribution of $J$ and $K$ Under Weak Identification ...... 24
   1.3 Conditional Linear Combination Tests .......................... 27
      1.3.1 Optimality of CLC Tests in a Conditional Problem .......... 28
   1.4 Optimal CLC Tests ............................................. 31
      1.4.1 Plug-in Minimax Regret Tests .............................. 32
      1.4.2 Implementing MMRU and PI Tests ......................... 35
   1.5 Performance of PI Tests in Weak IV ............................ 36
      1.5.1 Homoskedastic Linear IV ................................. 36
      1.5.2 Linear IV with Unrestricted Covariance Matrix .......... 39
   1.6 Asymptotic Properties of CLC Tests ........................... 46
      1.6.1 Postmultiplication Invariant Weight Functions ........... 46
      1.6.2 Asymptotic Size and Power of CLC Tests .......................... 48
      1.6.3 Asymptotic Efficiency Under Strong Identification ....... 50
   1.7 Conclusion ..................................................... 54
   1.8 Appendix ...................................................... 54

2 Two Step Confidence Sets, and the Trouble with the First Stage F-Statistic 95
   2.1 The Trouble with the First Stage F-Statistic .................. 103
2.1.1 The First Stage F-Statistic Under Homoskedasticity ........................................ 106
2.1.2 The First Stage F-Statistic Under Heteroskedasticity .................................... 106
2.2 A Simple Two-Step Approach: A User's Guide .................................................. 109
  2.2.1 GMM Model and Test Statistics ................................................................. 110
  2.2.2 A Generalized Weak Identification-Robust Test Statistic ............................. 111
  2.2.3 A Robust Two-Step Confidence Set ......................................................... 114
  2.2.4 Choosing the Value $a$ and Computing Critical Values ............................... 115
  2.2.5 Reporting Results ...................................................................................... 117
2.3 Two-Step Confidence Sets in Linear IV ............................................................ 118
  2.3.1 Two Step Confidence Sets Under Homoskedasticity .................................. 118
  2.3.2 Two Step Confidence Sets Heteroskedasticity ............................................. 119
2.4 Empirical Illustration ......................................................................................... 120
  2.4.1 Returns to Schooling .................................................................................... 120
  2.4.2 Income and Democracy ............................................................................... 124
2.5 A Simple Two-Step Confidence Set: Theory ....................................................... 128
2.6 Two-Step Confidence Sets for GMM .................................................................. 131
  2.6.1 Behavior Under Weak Identification ......................................................... 132
  2.6.2 Behavior Under Strong Identification ....................................................... 135
2.7 Conclusion ........................................................................................................... 138
2.8 Appendix ............................................................................................................. 139

3 A Geometric Approach to Weakly Identified Econometric Models 159

  3.1 Model setting ..................................................................................................... 163
    3.1.1 Strength of identification and linearity. ..................................................... 165
  3.2 Geometry .......................................................................................................... 169
    3.2.1 Manifolds, tangent spaces, curvature ....................................................... 169
    3.2.2 Geometric bounds ..................................................................................... 172
    3.2.3 Stochastic bound ....................................................................................... 178
  3.3 Statistical application of the stochastic bound and its modifications .............. 179
    3.3.1 Modification 1: curvature over a smaller set ............................................. 180
<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.2 Modification 2: Working with a subset of parameters</td>
<td>181</td>
</tr>
<tr>
<td>3.3.3 Modification 3: Pre-test for weak identification</td>
<td>182</td>
</tr>
<tr>
<td>3.4 Asymptotic procedure</td>
<td>184</td>
</tr>
<tr>
<td>3.4.1 Uniformity result</td>
<td>184</td>
</tr>
<tr>
<td>3.4.2 Curvature of strongly identified parameters</td>
<td>185</td>
</tr>
<tr>
<td>3.4.3 Comparison with other methods available for testing hypotheses with weak nuisance parameters</td>
<td>187</td>
</tr>
<tr>
<td>3.5 Empirical application: new Keynesian Phillips curve</td>
<td>189</td>
</tr>
<tr>
<td>3.5.1 Model setting</td>
<td>190</td>
</tr>
<tr>
<td>3.5.2 Empirical Results</td>
<td>191</td>
</tr>
<tr>
<td>3.5.3 Size Simulation</td>
<td>193</td>
</tr>
<tr>
<td>3.6 Simulation Example: DSGE Model</td>
<td>194</td>
</tr>
<tr>
<td>3.6.1 Simulation study: A Small-scale DSGE Model</td>
<td>196</td>
</tr>
<tr>
<td>3.7 Appendix</td>
<td>199</td>
</tr>
</tbody>
</table>
Chapter 1

Conditional Linear Combination

Tests for Weakly Identified Models

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 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 selects CLC tests with power functions 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 could 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 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.

We show that our plug-in tests perform well in linear IV. Specifically, in linear IV with homoskedastic Gaussian errors and one endogenous regressor we show that plug-in minimax regret tests using reasonable plug-in estimators match the near-optimal performance of the CLR test established by (Andrews et al., 2006, henceforth AMS). 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 (2005)'s quasi-CLR test for general GMM models. We further find that our approach offers power competitive with the weighted average power optimal MM1-SU and MM2-SU tests of (Moreira and Moreira, 2013, henceforth MM).

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. As we might hope given their minimax-regret motivation, the PI tests have smaller power maximal shortfalls, relative to the other tests considered in our simulations, than do the MM tests. On the other hand, by construction the MM tests have higher weighted average power with respect to the MM weights.

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 classes 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.

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 et al. (2012), Olea (2012), and MM. 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. Indeed, as MM show in their linear IV simulations, weighted average power optimal tests may sometimes have low power over empirically relevant regions of the parameter space. MM address this issue by restricting attention to classes of locally unbiased tests (their LU and SU tests).\(^1\) Here, we take a different approach and adopt a

\(^1\)A previous version of the paper, Moreira and Moreira (2010), discusses the issue of approximating weighted average power optimal similar tests of a given size, but don’t discuss the IV example or consider unbiased tests.
minimax regret perspective which attempts to pick tests that 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, yield easy-to-characterize behavior under strong-identification asymptotics. In contrast, weighted average power optimal tests need not be efficient under strong identification: of the tests discussed above, only Elliott et al. (2012) give results guaranteed to deliver efficient tests under strong identification in general contexts. Unfortunately, however, implementing their procedures is extremely computationally costly in many cases of econometric interest, including linear IV with non-homoskedastic errors and a moderate or large number of instruments.

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 1.2 we define several statistics including $S$, $K$, and $D$, and discuss tests which have been proposed based on these statistics. Section 1.3 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 1.3.1 then shows that CLC tests are admissible, locally most powerful, and weighted average power maximizing conditional on $D$. Section 1.4 defines minimax regret CLC tests, shows that such tests exist, defines plug-in tests, and discusses implementation of these procedures. Section 1.5 shows that suitably defined plug-in minimax regret tests match the near-optimal performance of the CLR test under homoskedasticity and compare favorably to existing alternatives in simulations calibrated to Yogo (2004)'s
data. Section 1.6 derives a number of asymptotic results, including uniform size control for CLC tests in heteroskedastic linear IV and efficiency of plug-in tests in our examples under strong asymptotics. Proofs for Theorems 1 and 2 are given in the main appendix, while the remaining proofs, as well as further details on our examples and additional simulation results, may be found in the supplementary appendix (available on the author’s website). To illustrate the application of our approach to a non-linear example, in the supplement we apply our results to a generalized minimum distance approach to inference on new Keynesian Phillips curve parameters studied in Magnusson and Mavroeidis (2010).

1.1 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 with two examples. We assume a sequence of models indexed by sample size $T$, where sample $T$ has distribution $F_T(\theta, \gamma)$ for $\theta \in \Theta$ a $p$-dimensional parameter of interest and $\gamma \in \Gamma$ 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 typically be an appropriately scaled moment vector or distance function, 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 $\theta$, and an estimate $\hat{\gamma}$ for $\gamma$. We assume that for all fixed $(\theta, \gamma) \in \Theta \times \Gamma$ we have

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

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

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

(1.2)
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 \( \gamma \). Here we use \( \text{vec}(A) \) to denote vectorization, which maps the \( k \times p \) matrix \( A \) to a \( kp \times 1 \) vector. We further assume that \( \Sigma_{\theta \gamma} \) and \( \Sigma_{\theta \theta} \) are continuous functions of \( \gamma \) and are thus consistently estimable. We will generally suppress the dependence of the terms in the limit problem on the parameters \((\theta, \gamma)\) when there is no loss of clarity from doing so, writing simply \( m, \mu, \) 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 \( H_1 : m \in \mathcal{M}(\mu)\setminus\{0\}, \mu \in \mathbb{M} \).

Limit problems of the form (1.2) 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 the supplementary appendix. In the supplement we also show that general weakly identified GMM models give rise to limiting problems of the form (1.2).

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

\[
Y = Z\pi\beta + V_1 \\
X = Z\pi + V_2
\]

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. We are interested in testing a hypothesis \( H_0 : \beta = \beta_0 \) about the scalar coefficient \( \beta \). As elsewhere in the literature (see e.g. AMS) we can accommodate additional exogenous regressors, but omit such variables here to simplify the exposition.

The 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 view linear IV as a special case of GMM with moment condition

\[
f_t(\beta) = (Y_t - X_t\beta)Z_t
\]
and identifying assumption $E_\theta [f_i(\beta)] = 0$ (where $E_\theta [X]$ denotes the expectation of $X$ under true parameter value $\theta$). For fixed $\pi \neq 0$ it is straightforward to construct consistent, asymptotically normal GMM estimates based on (1.4) and to use these estimates to test hypotheses about $\beta$. As is now well understood, however, the standard asymptotic approximations to the distribution of estimators and test statistics may be quite poor if $\pi$ 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 $\pi$ as changing with the sample size, taking $\pi_T = \frac{c}{\sqrt{T}}$ for a fixed vector $c \in \mathbb{R}^k$. Staiger and Stock show that the (1949) Anderson-Rubin test for $H_0 : \beta = \beta_0$ controls size under these asymptotics when the data $(Y_t, X_t, Z'_t)$ are independent across $t$ and the errors $(V_1, V_2, \ldots)$ are homoskedastic, and a large subsequent literature including Stock and Wright (2000), Kleibergen (2002), Moreira (2003), Kleibergen (2005), Oele (2012), Andrews and Cheng (2012), and MM has extended these results to more general models and alternative identification-robust tests.

To derive the limit problem (1.2) for this model, define $f_T(\beta) = \frac{1}{\sqrt{T}} \sum f_i(\beta)$ and let $\Omega$ be the asymptotic variance matrix of $\sqrt{T} \left( f_T(\beta_0)' - \frac{\partial}{\partial \beta} f_T(\beta_0)' \right)'$, 

$$\Omega = \begin{pmatrix} \Omega_{ff} & \Omega_{f\beta} \\ \Omega_{\beta f} & \Omega_{\beta\beta} \end{pmatrix} = \lim_{T \to \infty} Var \left( \sqrt{T} \left( \begin{pmatrix} f_T(\beta_0) \\ -\frac{\partial}{\partial \beta} f_T(\beta_0) \end{pmatrix} \right) \right). \quad (1.5)$$

We assume that $\Omega_{ff}$ is full-rank. For $\hat{\Omega}$ a consistent estimator of $\Omega$, 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 $\hat{\gamma} = vec(\hat{\Omega})$. For $\theta = \beta$, $\Theta = \mathbb{R}$, $\gamma = vec(\Omega)$, and $\Gamma$ the set of of values $\gamma$ such that $\Omega(\gamma)$ is symmetric and positive definite, for all $(\theta, \gamma) \in \Theta \times \Gamma$, under mild conditions

$$\begin{pmatrix} g_T(\beta_0) \\ \Delta g_T(\beta_0) \end{pmatrix} \to_d \begin{pmatrix} g \\ \Delta g \end{pmatrix} \sim N \left( \begin{pmatrix} m \\ \mu \end{pmatrix}, \begin{pmatrix} I & \Sigma_{\theta\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix} \right) \quad (1.6)$$

so (1.1) and (1.2) hold here with $m = \Omega_{ff}^{-\frac{1}{2}} Q_{zc} (\beta - \beta_0)$, $\mu = \Omega_{ff}^{-\frac{1}{2}} Q_{zc} c \in \mathcal{M} = \mathbb{R}^k$, $\Sigma_{\theta\theta} = \Omega_{ff}^{-\frac{1}{2}} \Omega_{\beta\beta} \Omega_{ff}^{-\frac{1}{2}}$, and $\Sigma_{\theta g} = \Omega_{ff}^{-\frac{1}{2}} \Omega_{\beta\beta} \Omega_{ff}^{-\frac{1}{2}}$. Note that for any $\mu$, $m \in \mathcal{M}(\mu) = \{b \cdot \mu : b \in \mathbb{R} \}$ and $m = 0$ when $\beta = \beta_0$. To derive this limit problem we have imposed

18
very little structure on the data generating process, and so can easily accommodate heteroskedastic, clustered, or serially correlated data 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 (2009), 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) 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 et al. (1993) and simulated method of moments as in McFadden (1989) and much of the subsequent literature.

In minimum distance or moment-matching models, for $\theta$ a $p \times 1$ vector of structural parameters and $\eta$ 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 $\eta$ that, together with an estimator $\hat{\Omega}_\eta$ for the variance of $\hat{\eta}$, satisfies $\hat{\Omega}_\eta^{-\frac{1}{2}}(\hat{\eta} - \eta) \overset{d}{\to} N(0, I)$. Under strong identification asymptotics, $\hat{\eta} - \eta = O_p \left( \frac{1}{\sqrt{T}} \right)$ and we have the usual asymptotic distribution for the structural parameter estimates $\hat{\theta} = \arg \min_{\theta} (\hat{\eta} - f(\theta))^T \hat{\Omega}_\eta^{-1} (\hat{\eta} - f(\theta))$ and the standard test statistics. As Canova and Sala (2009) highlight, however, if there is limited information about the structural parameters $\theta$ 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 $\Omega_\eta$ non-degenerate, which implies that $\hat{\eta}$ is not consistent for $\eta$. 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 $\hat{\gamma} = vec(\hat{\Omega}_\eta)$. For $\gamma = vec(\Omega_\eta)$ and $\Gamma$ again the set of $\gamma$ values corresponding to
symmetric positive definite matrices, we have that under \((\theta, \gamma) \in \Theta \times \Gamma, \hat{\gamma} \to_p \gamma\) and 
\[
\begin{pmatrix}
g_T(\theta_0) \\
\Delta g_T(\theta_0)
\end{pmatrix} \to_d 
\begin{pmatrix}
g \\
vec(\Delta g)
\end{pmatrix} \sim N\left(\begin{pmatrix}
m \\
vec(\mu)
\end{pmatrix}, \begin{pmatrix}
I & 0 \\
0 & 0
\end{pmatrix}\right)
\]
\(1.7\)
where \(m \in \mathcal{M} = \left\{\Omega^{-\frac{1}{2}} \left(f(\theta) - f(\theta_0)\right) : \theta \in \Theta\right\}\) and \(\mu = \Omega^{-\frac{1}{2}} \frac{\partial}{\partial \theta} f(\theta_0)\) (see the supplementary appendix for details).\(\Box\)

As these examples highlight, limit problems of the form (1.2) arise in a wide variety of econometric models with weak identification. In the supplementary appendix we show that general GMM models that are weakly identified in the sense of Stock and Wright (2000) generate limit problems of the form (1.2), and Example I could be viewed as a special case of this result. As Example II illustrates, however, the limit problem (1.2) is more general. The supplementary appendix provides another non-GMM example, considering a weakly identified generalized minimum distance model studied by Magnusson and Mavroeidis (2010).\(^2\)

Since the limit problem (1.2) appears in a wide range of weakly identified contexts, for the next several sections we focus on tests in this limit problem. Similar to Mueller (2011) we consider the problem of testing \(H_0 : m = 0, \mu \in \mathcal{M}\) against \(H_1 : m \in \mathcal{M}(\mu) \setminus \{0\}, \mu \in \mathcal{M}\) with the limiting random variables \((g, \Delta g, \gamma)\) observed and seek to derive tests with good properties. In Section 1.6 we return to the original problem, and argue that under mild assumptions results for the limit problem (1.2) can be viewed as asymptotic results along sequences of models satisfying (1.1).

\(^{2}\)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 (1.2) 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 for theorems 4 and 6) show that such convergence also holds in weakly identified Generalized Empirical Likelihood (GEL) models with independent data, both with and without strongly identified nuisance parameters. (Guggenberger et al., 2012, proofs for theorems 3.2 and 4.2) extend these results to time series GEL applications, further highlighting the relevance of the limit problem (1.2).
1.2 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 even when identification is weak. 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 Kleiber-gen (2005), or their 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.

When testing $H_0 : m = 0, \mu \in \mathbb{M}$ in (1.2), a natural statistic is

$$ S = g'g \sim \chi_k^2 (m'm). \tag{1.8} $$

Under the null $S$ is $\chi^2$ distributed with $k$ degrees of freedom, while under the alternative it is non-central $\chi^2$ distributed with non-centrality parameter $m'm = \|m\|^2$. Statistics asymptotically equivalent to (1.8) for appropriately defined $g_T$ have been suggested in a number of contexts by a wide range of papers, including Anderson and Rubin (1949) for linear IV, Stock and Wright (2000) for GMM, Magnusson and Mavroeidis (2010) for minimum distance models, and Ramalho and Smith (2004), Guggenberger and Smith (2005), Otsu (2006), Guggenberger and Smith (2008), and Guggenberger et al. (2012) for generalized empirical likelihood (GEL) models.

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 (1.2) generally arises when we consider local alternatives to the null hypothesis, in which case $\Delta g$ is typically non-random (so $\Sigma_{\theta \theta} = \Sigma_{a \theta} = 0$) and $m = \Delta g(\theta - \theta_0)$.\footnote{For discussion of this point in a GMM context see the supplementary appendix. For a more extensive analysis see Newey and McFadden (1994), Section 9.} Hence, under strong identification tests based on $S$ test the parametric restriction $\theta = \theta_0$ together with the over-identifying restriction $m' (I - P_{\Delta g}) m = 0$, and so are inefficient if we
only want to test $\theta = \theta_0$. To avoid this problem Moreira (2001) and Kleibergen (2002) propose a weak identification-robust score statistic for linear IV models which is efficient under strong identification, and Kleibergen (2005) generalizes this statistic to GMM. Following Kleibergen (2005) define $D$ as the $k \times p$ matrix such that

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

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 $vec(\mu_D) = vec(\mu) - \Sigma_{\delta g} m$, $\mu_D \in M_D$, $\Sigma_D = \Sigma_{\theta \theta} - \Sigma_{\delta g} \Sigma_{g \theta}$, and $m \in M_D(\mu_D)$,

$$M_D(\mu_D) = \{ m : m \in M(\mu) \text{ for } vec(\mu) = vec(\mu_D) + \Sigma_{\delta g} m \}.$$  

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

Kleibergen defines the $K$ statistic as

$$K = g' D (D'D)^{-1} D' g = g' P_{D} g. \quad (1.9)$$

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 $\chi^2_p$ as well. In the strongly identified limit problem $\Delta g = D$ is non-random and one can show that tests based on the $K$ statistic are efficient for this
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 developed related statistics in a number of other settings, all of which yield the K statistic (1.9) in the appropriately defined limit problem. In particular, Magnusson and Mavroeidis (2010) propose such a statistic for weakly identified generalized minimum distance models, while Ramalho and Smith (2004), Guggenberger and Smith (2005), Guggenberger and Smith (2008), and Guggenberger et al. (2012) discuss analogs of K for GEL models.

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. Likewise, except in special cases weighted average power optimal tests based on (g, Δg), and in particular the tests proposed by MM for linear IV, will depend on g through more than just S and K and so fall outside this class.

1.2.1 Other Asymptotically Pivotal Statistics

A number of other identification-robust test statistics have been created using S, K, and D. Since some of these will play an important role later in our analysis we briefly introduce them here.

Kleibergen (2005) defines J as the difference between the S and K statistics

\[ J = S - K = g'(I - D(D'D)^{-1}D)g = g'(I - PD)g \]

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

Moreira (2003) considers the problem of testing hypotheses on the parameter $\beta$ in weak IV (Example I) when the instruments $Z$ are fixed and the errors $V$ are normal and homoskedastic with known variance. Moreira derives a conditional likelihood ratio statistic which, for $p = 1$ and $r(D) = D' \Sigma_D^{-1} D$, is

$$
\frac{1}{2} \left( K + J - r(D) + \sqrt{(K + J + r(D))^2 - 4J \cdot r(D)} \right). \tag{1.10}
$$

Under the null the CLR statistic has distribution

$$
\frac{1}{2} \left( \chi_p^2 + \chi_{k-p}^2 - r(d) + \sqrt{\left( \chi_p^2 + \chi_{k-p}^2 + r(d) \right)^2 - 4\chi_{k-p}^2 \cdot r(d)} \right) \tag{1.11}
$$

conditional on $D = d$, where $\chi_p^2$ and $\chi_{k-p}^2$ are independent $\chi^2$ random variables with $p$ and $k-p$ degrees of freedom, respectively. The size $\alpha$ CLR test then rejects when the CLR statistic (1.10) exceeds $q_\alpha(r(D))$, the $1 - \alpha$ quantile of (1.11) for $d = D$.

Given this definition, it is natural to consider the class of quasi-CLR (QCLR) tests obtained by using other functions $r : D \to \infty$, where for $r(D) = \infty$ we define the QCLR statistic (1.10), denoted by $QCLR_r$, to equal $K$. This class nests the quasi-CLR tests of Kleibergen (2005), Smith (2007), and Guggenberger et al. (2012).

1.2.2 Distribution of $J$ and $K$ 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 $K|D = d \sim \chi_p^2(\tau_K(d))$ and $J|D = d \sim \chi_{k-p}^2(\tau_J(d))$ where

$$
\tau_K(D) = m' P_D m, \quad \tau_J(D) = m' (I - P_D) m. \tag{1.12}
$$

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,
which is why the K test

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

is optimal in this case.\footnote{For a (non-randomized) test, we take \( \phi = 1 \) to denote rejection and \( \phi = 0 \) failure to reject.} 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.

\textbf{Example II: Minimum Distance (Continued)} We established in (1.7) that \( \Delta g \) is non-random, so \( D = \Delta g = \mu = \mu_D \). To simplify the exposition, assume for this section that \( \Omega_\eta = I \). Since \( \mathcal{M} = \{(f(\theta) - f(\theta_0)) : \theta \in \Theta \} \) we have that under alternative \( \theta \) the non-centrality parameters in the \( J \) and \( K \) statistics are

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

Since \( \mu = \frac{\partial}{\partial \theta} f(\theta_0) \), this means that under alternative \( \theta \) the non-centrality parameter \( \tau_K \) is the squared length of \( f(\theta) - f(\theta_0) \) projected onto the model’s tangent space at the null parameter value, while \( \tau_J \) 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 \( \mathcal{M} = \left\{ \frac{\partial}{\partial \theta} f(\theta_0) : b \in \mathbb{R}^p \right\} \). \( \tau_J \equiv 0 \) and the \( K \) test \( \phi_K \) will be uniformly most powerful in the class of tests based on \( (S, K, D) \). As argued in Andrews and Mikusheva (2012), under strong identification minimum distance models are approximately linear, confirming the desirable properties of the \( K \) statistic in this case. Under weak identification, however, non-linearity of \( f(\theta) \) may remain important even asymptotically. To take an extreme case, if there is some \( \theta \in \Theta \) 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 \( \theta \) depends on \( J \) alone.\footnote{For \( \phi \) a (non-randomized) test, we take \( \phi = 1 \) to denote rejection and \( \phi = 0 \) failure to reject.}
Example I: Weak IV (Continued) In the limit problem (1.6) $\Delta g$ is random and may be correlated with $g$, so $D / \Delta g$ and $\mu_D = \mu - \Sigma g m$. Since $m = \mu(\beta - \beta_0)$,

$$\mu_D = \mu - \Sigma g \mu(\beta - \beta_0) = (I - \Sigma g (\beta - \beta_0)) \mu.$$ 

Note that if $\mu$ is proportional to an eigenvector of $\Sigma g$ corresponding to a non-zero eigenvalue $\lambda$, then for $(\beta - \beta_0) = \lambda^{-1}$ we have that $\mu_D = 0$. Hence for some $(\Sigma g, \mu)$ combinations, while $\mu$ may be quite large relative to both $\Sigma g$ and $\Sigma g \theta$, there will be some alternatives $\beta$ under which $\mu_D = 0$. When this occurs the direction of the vector $D$ bears no relation to the direction of $m$ or $\mu$ 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, $\Omega$ as defined in (1.5) has Kronecker product structure, so $\Omega = A \otimes B$ for a $2 \times 2$ 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 $\mu$. This is precisely what occurs at the point $\beta_{AR}$ discussed by AMS, where they show that the test

$$\phi_S = 1 \left\{ S > \chi_{k,1-a}^2 \right\}$$  

(1.14)

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 $\Omega$ has Kronecker product structure is extreme in that $\mu_D = 0$ at alternative $\beta_{AR}$ regardless of the true value $\mu$. 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 $\mu_D \neq \mu$ in the general (non-Kronecker) case. An extreme version of this issue arises if there is some value $\beta^*$ such that $(I - \Sigma g (\beta^* - \beta_0)) \mu \neq 0$ but $\mu' (I - \Sigma g (\beta^* - \beta_0)) \mu = 0$. For this value of $\beta^*$ we have that $\mu_D \neq 0$ but $\mu_D m = 0$, and hence the $K$ statistic tends to focus on directions that yield low power against alternative $\beta^*$. □
To summarize, while tests rejecting for large values of the \( K \) statistic are efficient under strong identification, they can have low power when identification is weak. In contrast, the \( S \) test (1.8) is inefficient under strong identification but has power that depends only on \( ||m|| \) 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 to use the information contained in \( D \) to combine the \( S \) and \( K \) statistics to retain the advantages of each while ameliorating their deficiencies.

### 1.3 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 \( a : \mathcal{D} \to [0,1] \) 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:

\[
\phi_{a(D)} = 1 \{(1 - a(D)) \cdot K + a(D) \cdot S > c_a(a(D))\} = 1 \{K + a(D) \cdot J > c_a(a(D))\}.
\]

We take the conditional critical value \( c_a(a) \) 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 \( a : \mathcal{D} \to [0,1] \) the test \( \phi_{a(D)} \) defined in (1.15) is conditionally similar with \( E_{m=\hat{m},\mu_D} \left[ \phi_{a(D)} \right] = \alpha \) almost surely for all \( \mu_D \in \mathcal{M}_D \). 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 \( c_a(a) \) 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 \( r : \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\} \) define the quasi-CLR statistic
QCLR, as in (1.10) and let \( q_{\alpha}(r(d)) \) be the \( 1 - \alpha \) quantile of (1.11). Then:

**Theorem 2** For any function \( r : \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\} \) 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 \( \phi_{QCLR_r} \equiv \phi_{\tilde{a}(D)} \). Conversely, for any \( a : \mathcal{D} \to [0,1] \) there exists an \( \tilde{r} : \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\} \) such that \( \phi_{a(D)} \equiv \phi_{QCLR_r} \). Hence, the class of CLC tests for \( a : \mathcal{D} \to [0,1] \) is precisely the same as the class of QCLR tests for \( r : \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\} \).

Theorem 2 shows that the QCLR test \( \phi_{QCLR_r} \) 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.

### 1.3.1 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.

Conditional on \( D = d \) (for \( d \) full rank), \( J \) and \( K \) are independent and distributed \( \chi_{k-p}^2(\tau_J(d,m)) \) and \( \chi_p^2(\tau_K(d,m)) \), respectively, for \( \tau_J \) and \( \tau_K \) as defined in (1.12). Conditional on \( D = d \) the non-centrality parameters \( \tau_J \) and \( \tau_K \) are fixed, though unknown, values and our null hypothesis \( H_0 : m = 0, \mu \in \mathcal{M} \) 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 \( (\tau_J, \tau_K) \) under the alternative \( H_1 \). Let \( \mathcal{M}_D(d) \) denote
the set of values \( \mu_D \in M_D \) such that \( d \) is in the support of \( D \). Letting \( \hat{M}(d) = \bigcup_{\mu_D \in M_D(d)} M(\mu_D) \), \( m \) may take any value in \( \hat{M}(d) \) and still be consistent with both \( m \in M(\mu_D) \) and \( d \) lying in the support of \( D \). Hence, the non-centrality parameters \((\tau_J, \tau_K)\) may take any value in the set

\[
\mathcal{T}(d) = \bigcup_{m \in \hat{M}(d)} (\tau_J(d, m), \tau_K(d, m)).
\]

Conditional on \( D = d \) our problem becomes one of testing \( H_0 : \tau_J = \tau_K = 0 \) against the alternative \( H_1 : (\tau_J, \tau_K) \in \mathcal{T}(d) \setminus \{0\} \) based on observing \((J, K) \sim \left(\chi^2_{k-p}(\tau_J), \chi^2_{p}(\tau_K)\right)\).

**CLC Tests are Admissible in the Conditional Problem**

We say that a test \( \phi \) is admissible if there is no other test \( \tilde{\phi} \) with size less than or equal to \( \phi \) and power greater than or equal to \( \phi \) at all points and strictly higher power (or smaller size) at some point (that is, if there is no test \( \tilde{\phi} \) that dominates \( \phi \)). 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 \( \phi \) be a test of \( H_0 : \tau_J = \tau_K = 0 \) against \( H_1 : (\tau_J, \tau_K) \in \mathbb{R}^2_+ \setminus \{0\} \). \( \phi \) is admissible in the conditional problem if and only if it is almost surely equal to 1 \{\( (\sqrt{J}, \sqrt{K}) \notin C_d \)\} for some set \( C_d \) which is

1. closed and convex
2. monotone decreasing: i.e. \( x \in C_d \) and \( y, \leq x \) \( \forall i \) implies \( y \in C_d \)

Thus, a test \( \phi \) is admissible in the conditional problem if and only if its acceptance region in \( (\sqrt{J}, \sqrt{K}) \) 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 \( a : \mathcal{D} \to [0, 1] \) and all \( d \).

---

5If \( \Sigma_D \) is full rank then \( M_D(d) = M_D \), since the support of \( D \) is the same for all \( \mu_D \in M_D \), but if \( \Sigma_D \) is reduced rank (e.g. \( \Sigma_D = 0 \)) then we may have \( M_D(d) \subset M_D \).
**Corollary 1** For all weight functions $\alpha : \mathcal{D} \to [0, 1]$ the CLC test $\phi_{\alpha(D)}$ is admissible in the problem conditional on $D = d$ for all $d$.

**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 class of tests satisfying the conditions of Theorem 3 for all realizations of $D$ is quite large. 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.

**Theorem 4** *(Monti and Sen (1976), Koziol and Perlman (1978))* Fix a conditional linear combination test $\phi_{\alpha(D)}$ and a value $d$. Let $\Phi_\alpha(d)$ denote the class of tests which have size $\alpha$ conditional on $D = d$, $E_{(\tau_J, \tau_K)}[\phi|D = d] = \alpha \forall \phi \in \Phi_\alpha(d)$.

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 $0 < \lambda < \bar{\lambda}$,

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

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

$$\phi_{\alpha(D)} \in \arg \max_{\phi \in \Phi_\alpha(d)} \int E_{(\tau_J, \tau_K)}[\phi_{\alpha(D)}|D = d] \cdot dF_{t_J, t_K}(\tau_J, \tau_K).$$

Theorem 4 follows immediately from results in Monti and Sen (1976) and Koziol and Perlman (1978) on the optimal combination of independent non-central $\chi^2$ statistics. Theorem 4(1) shows that conditional on $D = d$ the CLC test $\phi_{\alpha(D)}$ 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 $\phi$ 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 $\chi^2$ 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 $D = d$, but that the direction in which these tests have power depends critically on the weight function $a(D)$. In the next section we discuss how to select this function.

1.4 Optimal CLC Tests

For any weight function $a : \mathcal{D} \to [0, 1]$ we can define a CLC test $\phi_a(D)$ for $H_0$ against $H_1$ using (1.15). 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 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 $a : \mathcal{D} \to [0, 1]$. $\beta^*_{m,\mu_D}$ gives the highest attainable power against alternative $(m, \mu_D)$ in the class of CLC tests and, as we vary $(m, \mu_D)$, 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 $(m, \mu_D)$ 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

\[
a_{MMR} \in \arg \min_{a \in \mathcal{A}} \sup_{(m, \mu_D) \in \mathcal{H}_1} \left( \beta^*_m \mu_D - E_{m, \mu_D} [\phi_a(D)] \right).
\]

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 CLC 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.

Given the way that MMR tests are defined, it is not obvious that an MMR test exists, i.e. that \( \inf_{a \in \mathcal{A}} \sup_{(m, \mu_D) \in \mathcal{H}_1} \left( \beta^*_m \mu_D - E_{m, \mu_D} [\phi_a(D)] \right) \) is achieved by any \( a \in \mathcal{A} \). Theorem 5 establishes that an MMR test \( \phi_{MMR} \) always exists.

**Theorem 5** For any non-empty set of alternatives \( \mathcal{H}_1 : m = \mathcal{M}_D(\mu_D) \setminus \{0\}, \mu_D \in \mathcal{M}_D \) in the limit problem (1.2) there exists a weight function \( a^* \in \mathcal{A} \) such that

\[
\sup_{(m, \mu_D) \in \mathcal{H}_1} \left( \beta^*_m \mu_D - E_{m, \mu_D} [\phi_{a^*}(D)] \right) = \inf_{a \in \mathcal{A}} \sup_{(m, \mu_D) \in \mathcal{H}_1} \left( \beta^*_m \mu_D - E_{m, \mu_D} [\phi_a(D)] \right).
\]

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

### 1.4.1 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, so solving for \( \phi_{MMR} \) 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 $\phi_{MMR}$ 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 which make calculating $\phi_{MMR}$ straightforward. First, rather than optimizing over the space of 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, imagine first that $\mu_D$ is known. Let us restrict attention to unconditional linear combination tests with $a(D) = a(\mu_D) \in [0,1]$, where we write $a$ as a function of $\mu_D$ to emphasize its dependence on this parameter. The power envelope for this class of unconditional linear combination tests is $\beta_{a,\mu_D}^u = \sup_{\alpha \in [0,1]} E_{m,\mu_D}[\phi_a]$. A minimax regret unconditional (MMRU) test $\phi_{MMRU} = \phi_{a_{MMRU}(\mu_D)}$ then uses

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

Just as when we derived $\phi_{MMR}$ 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 $\phi_{MMRU}$ we assumed that $\mu_D$ 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

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

is a CLC test and so controls size by Theorem 1. Moreover, to calculate this test we need only 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 $\phi_{PI}$ is not in general a true MMR test. First, $\phi_{PI}$ treats the estimated value $\hat{\mu}_D$ as the true value, and hence does not account for any uncertainly in the estimation of $\mu_D$. Second, even taking the value $\mu_D$ as given $\phi_{PI}$
restricts attention to unconditional linear combination tests, which represent a strict subset of the possible functions \( a \in A \). Despite these potential shortcomings, we find that PI tests perform quite well in simulation, and show in Section 1.6 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 \( \mu_D \) based on \( D \), \( \hat{\mu}_D = D \), is a natural choice we may be able to do better in many cases. 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 \( (J, K, D'S_D^{-1}D) \) is a maximal invariant under rotations of the instruments, where \( D'S_D^{-1}D \sim \chi^2_k \left( \mu'_D \Sigma_D^{-1} \mu_D \right) \).\(^6\) AMS show that the distribution of \( (J, K, D'S_D^{-1}D) \) depends on \( c = \sqrt{T} \pi_T \) only through the non-centrality parameter \( r = \mu'_D \Sigma_D^{-1} \mu_D \).

Note that the MLE \( \hat{\mu}_D = D \) for \( \mu_D \) based on \( D \) implies a severely biased estimator for \( r \), \( \hat{r} = D'S_D^{-1}D \), with \( E [\hat{r}] = E \left[ D'S_D^{-1}D \right] = r + k \). The problem of estimating \( r \) relates to the well-studied problem of estimating the non-centrality parameter of a non-central \( \chi^2 \) distribution, and a number of different estimators have been proposed 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 \( \hat{r} - k \),\(^7\) 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 et al. (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 1.5 below that estimators \( \hat{\mu}_D \) corresponding to all three non-centrality estimators \( \hat{r}_{MLe} \),

\(^6\)It suffices to note that \( (J, K, D'S_D^{-1}D) \) is a one-to-one transformation of \( Q \) as defined in AMS.

\(^7\)\( \hat{r}_{PP} \) has been shown to dominate \( \hat{r}_{MLe} \) in terms of mean squared error but is itself inadmissible (Saxena and and Alam, 1982).
\( \tilde{r}_{PP} \) and \( \tilde{r}_{KRS} \) yield PI tests \( \phi_{PI} \) with good power properties. □

### 1.4.2 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 \( \mu_D \), 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 the supplementary appendix.

Calculating a good approximation to the weight function \( a_{MMRU} \) is straightforward when \( \mathcal{M}_D(\mu_D) \) is compact. In particular, since \( \mu_D \) and \( \Sigma_D \) are both known in the MMRU problem, the only unknown parameter that affects the power of linear combination tests \( \phi_a \) 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 \( \mathcal{M}_D(\mu_D) \), for a sufficiently fine grids of values \( M \subset \mathcal{M}_D(\mu_D) \) and \( A \subset [0,1] \) the approximate value for \( a_{MMRU}(\mu_D) \)

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

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

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

To calculate \( a^*_ {MMRU}(\mu_D) \), however, it suffices to evaluate \( E_{m,\mu_D}[\phi_a] \) for all \((a,m) \in A \times M\), which can be done by simulation.\(^8\)

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 \( \phi_S \) attains power less than some pre-specified level \( \beta \). This amounts to considering \( \widetilde{\mathcal{M}}_D(\mu_D) = \mathcal{M}_D(\mu_D) \cap B_\beta \) where

\(^8\)For this step it is helpful to note that \( E_{m,\mu_D}[\phi_a] = \int E_{r_{r(D),\tau_K(D)}}[\phi_a] dF_D \), so we can tabulate \( E_{r_{r},\tau_K}[\phi_a] \) in advance and calculate the integral by simulation.
$B_C = \{m \in \mathbb{R}^k : \|m\|^2 < C\}$ and $\bar{C}$ solves $\Pr \{\chi^2_k(\bar{C}) > \chi^2_{k,1-a}\} = \beta$. Once we restrict attention to $\tilde{M}_D(\mu_D)$ we can approximate $a_{M_{M,RU}}(\mu_D)$ as discussed above.

1.5 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 $\Omega$. 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.

1.5.1 Homoskedastic Linear IV

AMS consider the linear IV model Example I under the additional restriction that the errors $V$ are independent of $Z$, normal, and iid across $t$ with $\text{corr}(V_{1,t}, V_{2,t}) = \rho$. 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 (1.6) with the additional restriction that $\Omega = A \otimes B$ for $A$ and $B$ symmetric positive-definite matrices of dimension $2 \times 2$ and $k \times k$, respectively. As Mueller notes, matrices $\Omega$ 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 of $\phi_{PI}$. In particular, by Theorem 2 the CLR test is a CLC test. Hence, if our plug-in minimax regret approach 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 estimators for $r = \mu_D' \Sigma_D^{-1} \mu_D$ described in Section 1.4.1. 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.

**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** For $A$ and $B$ symmetric positive-definite matrices of dimension $2 \times 2$ and $k \times k$, respectively, the function $a_{MMR}(\mu_D)$ in the limit problem (1.6) with $\Omega = A \otimes B$ can be taken to depend on $\mu_D$ only through $r = \mu_D' \Sigma_D^{-1} \mu_D$.

Since the weights of both the CLR test and the plug-in approaches discussed in Section 1.4.1 depend on $\hat{r}$ alone, in Figure 1-1 we plot the values of $a_{CLR}(\hat{r})$, $a_{MMR}(\hat{r})$, $a_{MMR}(\hat{r}_{MLE})$, $a_{MMR}(\hat{r}_{PP})$, and $a_{MMR}(\hat{r}_{KHS})$ as functions of $\hat{r}$ for $k = 5$. All the weight functions exhibit similar qualitative behavior, placing large weight on $S$ for small values of $\hat{r}$ and increasing the weight on $K$ as $\hat{r}$ grows, but there are some notable differences. Perhaps most pronounced, $a_{MMR}(\hat{r})$ is lower than any of the other functions, as is intuitively reasonable given that $\hat{r}$ tends to overestimate $r$. As previously noted both $\hat{r}_{MLE}$ and $\hat{r}_{PP}$ are zero for a range of strictly positive values $\hat{r}$. Also notable, we have that $a_{MMR}(0) < 1$: this reflects the fact that the MMRU test down-weights $S$ even when $r = 0$, due to its higher degrees of freedom.

**Power Simulation Results**

To study the power of the PI tests, we follow the simulation design of AMS and consider a homoskedastic normal model with a known reduced-form covariance matrix. Like AMS we consider models with five instruments, reduced-form error correlation
Figure 1-1: Weight functions $a_{CLR}(\hat{r})$ for CLR and $a_{MMRU}(\hat{r}(\hat{r}))$ for PI tests with different estimators $\hat{r}$ of $r$ discussed in Section 1.4.1 for linear IV with five instruments and homoskedastic errors.

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

Table 1.1: 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. 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 $a_{MMRU}(\hat{r})$, $a_{MMRU}(\hat{r}_{MLE})$, $a_{MMRU}(\hat{r}_{PP})$, and $a_{MMRU}(\hat{r}_{KRS})$, respectively. AR is the Anderson Rubin test (equivalent to the S test) and K is Kleibergen’s (2002) K test.
\( p \) 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, as
in AMS we also consider models with two and ten instruments, in each case fixing
\( p \) equal to 0.5 and letting \( \lambda \) equal 5 and 20. The resulting power plots (based on
10,000 simulations) are reported in the supplementary appendix. For brevity, here
we report only the maximal power shortfall of each test, measured as 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. These values can be viewed as a
measure of maximum regret relative to this restricted set of tests. As Table 1.1 makes
clear, the PI tests considered largely match the near-optimal performance of the CLR
test. The one exception is the PI test using the badly biased estimator \( \hat{r} \) for \( r \), which
systematically overweights 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.

1.5.2 Linear IV with Unrestricted Covariance Matrix

The near-optimal performance of PI tests in linear IV models where \( \Omega \) has Kronecker
product structure is promising, but is of limited relevance for empirical work. Eco-
nomic data frequently exhibit of heteroskedasticity, serial dependence, clustering, and
other features that render a Kronecker structure assumption for \( \Omega \) implausible. It is
natural to ask whether PI tests continue to have good power properties in this more
general case. As an alternative CLC test, 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
(\( S \)) tests.

In addition to these CLC tests, we also consider the MM1-SU and MM2-SU tests
of MM. These tests maximize weighted average power, for weights which depend on
the covariance matrix \( \Sigma \), over a class of similar tests satisfying a sufficient condition
for local unbiasedness (their SU tests). We show in the supplementary appendix that
all conditional linear combination tests are SU tests, and thus that the MM1-SU and MM2-SU tests have, by construction, weighted average power at least as great as the other tests considered under their respective weight functions. MM present extensive simulation results which show that these tests perform very well in models where $\Omega$ has Kronecker structure, as well as in some examples with non-Kronecker structure.

There are a multitude of ways in which $\Omega$ may depart from Kronecker structure, however, and it is far from clear ex-ante how the power of the tests we consider may be expected to compare under different departures. To assess the relative performance of all the tests we consider at parameter values relevant for empirical practice, we calibrate our simulations 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. Yogo notes that both the strength of identification and the degree of heteroskedasticity appear 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 $a_{\Omega}(D)$ may depend on all of $D$, rather than only on the scalar $\hat{r}$. Consequently visual comparison of $a_{\Omega}(D)$ to the weight function $a_{QCLR}(D)$ implied by the QCLR test is impractical and we move directly to our simulation results.

Power Simulation Results

We simulate the behavior of tests in the weak IV limit problem (1.6), and so require estimates for $\mu$ and $\Omega$. 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

---

9 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.
using three lags. A detailed description of this estimation procedure, together with the implementation of all tests considered, is given in the supplementary appendix. 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 for the non-MM tests, and 5,000 simulations for the MM tests) are plotted in Figures 1-2-1-4. Since for many countries the power curves are difficult to distinguish visually, in Table 1.2 we list the maximum regret for each test relative to the other tests studied, 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, MM, 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, the QCLR test does

<table>
<thead>
<tr>
<th></th>
<th>PI</th>
<th>QCLR</th>
<th>AR</th>
<th>K</th>
<th>MM1-SU</th>
<th>MM2-SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>0.53%</td>
<td>0.28%</td>
<td>17.85%</td>
<td>0.95%</td>
<td>9.32%</td>
<td>2.46%</td>
</tr>
<tr>
<td>Canada</td>
<td>1.73%</td>
<td>1.44%</td>
<td>19.07%</td>
<td>2.22%</td>
<td>4.37%</td>
<td>2.27%</td>
</tr>
<tr>
<td>France</td>
<td>1.70%</td>
<td>1.84%</td>
<td>21.33%</td>
<td>1.60%</td>
<td>4.03%</td>
<td>3.55%</td>
</tr>
<tr>
<td>Germany</td>
<td>3.64%</td>
<td>3.03%</td>
<td>17.76%</td>
<td>11.09%</td>
<td>2.58%</td>
<td>2.62%</td>
</tr>
<tr>
<td>Italy</td>
<td>1.69%</td>
<td>3.81%</td>
<td>14.59%</td>
<td>6.44%</td>
<td>9.26%</td>
<td>3.20%</td>
</tr>
<tr>
<td>Japan</td>
<td>8.45%</td>
<td>42.02%</td>
<td>11.23%</td>
<td>85.22%</td>
<td>7.17%</td>
<td>14.84%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1.76%</td>
<td>1.45%</td>
<td>19.07%</td>
<td>7.43%</td>
<td>2.66%</td>
<td>2.20%</td>
</tr>
<tr>
<td>Sweden</td>
<td>1.02%</td>
<td>1.13%</td>
<td>18.34%</td>
<td>1.73%</td>
<td>6.36%</td>
<td>0.81%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1.33%</td>
<td>0.44%</td>
<td>17.53%</td>
<td>1.53%</td>
<td>5.16%</td>
<td>4.90%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>7.38%</td>
<td>28.33%</td>
<td>14.09%</td>
<td>38.73%</td>
<td>10.85%</td>
<td>4.82%</td>
</tr>
<tr>
<td>United States</td>
<td>2.74%</td>
<td>8.53%</td>
<td>13.84%</td>
<td>11.09%</td>
<td>24.10%</td>
<td>6.54%</td>
</tr>
</tbody>
</table>

Table 1.2: Maximal point-wise power shortfall relative to other tests considered, for simulations calibrated to match data in Yogo (2004). QCLR denotes the quasi-CLR test of Kleibergen (2005) while PI is the plug-in test discussed in Section 1.5.2. AR is the Anderson Rubin (or S) test, K is Kleibergen (2005)'s K test, and MM1-SU and MM2-SU are the weighted average power optimal SU tests of Moreira and Moreira (2013).

\(^{10}\)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.

\(^{11}\)In implementing the MM tests, we encountered some minor numerical issues which may have a small effect on the results—see supplementary appendix section D.4 for details.
Figure 1.2: Power functions for QCLR, K, AR (or S), PI, and MM tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 1.5.2.
Figure 1-3: Power functions for QCLR, K, AR (or S), PI, and MM tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 1.5.2.
Figure 1-4: Power functions for QCLR, K, AR (or S), PI, and MM tests in simulation calibrated to Yogo (2004) data with four instruments, discussed in Section 1.2.
not fully resolve these issues. Instead, in cases 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 substantial 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 MM1-SU and MM2-SU tests both have high power in most calibrations, but in some cases exhibit larger maximal power shortfalls than does the PI test, with maximal shortfalls of 24.1% and 14.84% respectively, while the maximal power shortfall of the PI test is 8.45%.

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 worse performance when \( \Sigma \) lacks Kronecker product structure. In contrast, the PI test takes both the structure of \( \Sigma \) 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.

The relative performance of PI and MM-SU tests is as we might hope given their derivation. The PI test, while not a true minimax-regret test, is motivated from a minimax-regret perspective, so it is encouraging that it has a smaller overall maximal power shortfall than do the MM-SU tests. The power shortfalls of the PI test in particular calibrations also tend to be smaller, with the PI shortfall smaller than the MM1-SU shortfall in nine of the eleven countries considered, and smaller than the MM2-SU shortfall in eight of the eleven.\(^\text{12}\) By contrast the MM-SU tests are designed to maximize weighted average power, and simulations in the supplementary appendix confirm that the MM tests have higher weighted average power than the other tests considered under their respective weight functions.

\(^\text{12}\)Note that even were the PI test a true MMR test this would not be guaranteed, since the class of SU tests larger than the class of CLC tests.
Unlike in the homoskedastic case we can see that none of the tests considered is even approximately uniformly most powerful, but the PI test nonetheless delivers powerful, stable performance over the range of parameter values considered.

1.6 Asymptotic Properties of CLC Tests

The results of Sections 1.2-1.5 treat the limiting random variables \((g, \Delta g, \gamma)\) as observed and consider the problem of testing \(H_0 : m = 0, \mu \in \mathcal{M}\) against \(H_1 : m \in \mathcal{M}(\mu) \setminus \{0\}, \mu \in \mathcal{M}\). In this section, we show that under mild assumptions our results for the limit problem (1.2) imply asymptotic results along sequences of models satisfying (1.1). 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.

1.6.1 Postmultiplication Invariant Weight Functions

Let us introduce finite-sample analogs \(S_T, D_T, K_T, J_T\) to the limiting random variables \(S, D, K, J\), where \(S_T = g_T g_T\) and so on. We previously wrote the weight functions \(a\) of CLC tests as functions of \(D\) alone, since in the limit problem the parameter \(\gamma\) is fixed and known. In this section, however, it is helpful to instead write \(a(D, \gamma)\). Likewise, since the estimator \(\hat{\mu}_D\) used in plug-in tests may depend on \(\gamma\), in this section we will write it as \(\hat{\mu}_D(D, \gamma)\).

Our weak convergence assumption (1.1), 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 \(\Delta g_T\), which may seem like an obstacle to applying our approach. In the linear IV model for instance, the appropriate definition for \(\Delta g_T\) will depend on the strength of identification.

Example I: Weak IV (Continued) In Section 1.1 we assumed that the instruments were weak, with \(\pi_T = \frac{e}{\sqrt{T}}\), and showed that \(\Delta g_T = \sqrt{T} \hat{\Omega}^{-\frac{1}{2}} \partial_\beta f_T(\beta_0)\) con-
verged in distribution. If on the other hand the instruments are strong, \( \pi_T = \pi_1 \) and \( \|\pi_1\| > 0 \), then \( \frac{\partial}{\partial \beta} f_T(\beta) \to_p E [X_t Z_t] \neq 0 \) so \( \sqrt{T} \Omega^{-\frac{1}{2}} f_T(\beta_0) \) diverges and we should instead take \( \Delta g_T = \hat{\Omega}^{-\frac{1}{2}} f_T(\beta_0) \).

This apparent dependence on normalization is not typically a problem, however, since many CLC tests are 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

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

\[
h_{\Sigma} (\Sigma; A) = \left( \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} \otimes I_k \right)^t \Sigma \left( \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} \otimes I_k \right)
\]

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

Let

\[
h (g_T, \Delta g_T, \hat{\gamma}; A) = (g_T, h_{\Delta g} (\Delta g_T; A), h, (\hat{\gamma}; A))
\]

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 \( a(D, \gamma) \) which is invariant, the CLC test \( \phi_{a(D, \gamma)} \) will be invariant as well. 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

\[
a(D, \gamma) = a (h_{\Delta g} (D; A), h, (\gamma; A)),
\]

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 useful since to obtain results for invariant tests based on \((\tilde{g}_T, \Delta \tilde{g}_T, \hat{\gamma})\) it suffices that there exist some sequence \( A_T \) such that

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

satisfies the weak convergence assumption (1.1), without any need to know the cor-
rect sequence $A_T$ for a given application. Thus, in the linear IV example discussed above we can take $\Delta g_T$ as originally defined and make use of results derived under the convergence assumption (1.6) without knowing 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. In particular we can choose the minimax regret weight function $a_{MIR}$ 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 $h_{\Delta g}(\hat{\mu}_D(D, \gamma); A) = \mu_D(h_{\Delta g}(D; A), h_{\gamma}(\gamma; A))$, the plug-in weight function $a_{PI}$ will be invariant as well.

1.6.2 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 the marginal distribution for $\gamma$ in the limit problem is a point mass. Since we have assumed that $D$ is full rank almost surely, $J$ and $K$ are $F$-almost-everywhere continuous functions of $(g, \Delta g, \gamma)$ and the continuous mapping theorem implies

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

To obtain asymptotic size control for the CLC test

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

all we require is that $a$ be almost-everywhere continuous. 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 (1.1) and let $a(D, \gamma)$ be $F(g, \Delta g, \gamma)$-almost-everywhere continuous for $(\theta_0, \gamma) \in \{\theta_0\} \times \Gamma$. Then under $(\theta_0, \gamma)$ we have that

$$\lim_{T \rightarrow \infty} E_{T,(\theta_0, \gamma)} \left[ \phi_{a(D_T, \hat{\gamma})} \right] = \alpha. \quad (1.17)$$

48
Moreover, for $\mathcal{F}$ the set of bounded functions $f(D)$ which are $F(g, \Delta g, \gamma)$-almost-everywhere continuous under $(\theta_0, \gamma),$

$$\lim_{T \to \infty} E_T(\theta_0, \gamma) \left[ (\phi_{a(D_T, \gamma)} - \alpha) f(D_T) \right] = 0 \forall f \in \mathcal{F}. \quad (1.18)$$

It is important to note that Proposition 1 only establishes sequential size control, and depending on the underlying model establishing uniform size control over some base parameter space may require substantial further restrictions. In Example I, however, we can use results from (Andrews et al., 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 with a single endogenous regressor.

Example I: Weak IV (Continued) Define $\hat{\Omega}$ and $\hat{\Sigma}$ in the usual way (detailed in the proof of Proposition 2 in the supplementary appendix). Define a parameter space $\Lambda$ of null distributions as in ACG Section 3, noting that $\gamma$ 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 $\phi_{a(D_T, \gamma)}$ based on a postmultiplication-invariant weight function $a(D, \gamma)$ which is continuous in $D$ and $\gamma$ at all points with $||D|| > 0$ and satisfies

$$\lim_{\delta \to 0} \sup_{(D, \gamma): ||D|| > \epsilon, \text{max eig}(\gamma) \leq \delta} a(D, \gamma) = \lim_{\delta \to 0} \inf_{(D, \gamma): ||D|| > \epsilon, \text{max eig}(\gamma) \leq \delta} a(D, \gamma) = a_0 \quad (1.19)$$

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

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

The assumption (1.19), together with the assumed postmultiplication invariance of $a(D, \gamma)$ and the restrictions on the parameter space $\Lambda$, ensures that under sequences with $\sqrt{T}||\pi_T|| \to \infty$ we have that $a(D_T, \gamma) \to_p a_0$ asymptotically, and hence that under all strongly identified sequences the test converges to the linear combina-
tion test $\phi_{a_0}$. 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 this condition for PI tests $\phi_{PI}$ 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 $(D, \gamma)$ and taking $\tilde{a}$ to continuously interpolate between these values.\footnote{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 (1.16).}

Power results in the limit problem (1.2) also imply asymptotic power results under (1.1). In particular, for $a(D, \gamma)$ almost everywhere continuous with respect to $F(g, \Delta g, \gamma)$, the asymptotic power of $\phi_{a(D)}$ is simply the power of $\phi_{a(D, \gamma)}$ in the limit problem.

**Proposition 3** Assume $(g_T, \Delta g_T, \hat{\gamma})$ satisfies the weak convergence assumption (1.1) and let $a(D, \gamma)$ be $F(g, \Delta g, \gamma)$-almost-everywhere continuous for some $(\theta, \gamma) \in \Theta \times \Gamma$. Then under $(\theta, \gamma)$

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

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

Thus, under mild continuity conditions on $a(D, \gamma)$, the asymptotic size and power of tests under (1.1) 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.

### 1.6.3 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 1.2, under strong identification we typically have that \( \Sigma_{\theta \theta} = \Sigma_{\theta \gamma} = 0 \), that \( \mu \) 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 \((\theta, \gamma)\) if \((g_T, \Delta g_T, \hat{\gamma}) \to_d (g, \Delta g, \gamma)\) for

\[
\begin{pmatrix}
  g \\
  \text{vec}(\Delta g)
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
  \mu \cdot b \\
  \text{vec}(\mu)
\end{pmatrix},
\begin{pmatrix}
  I & 0 \\
  0 & 0
\end{pmatrix}
\right),
\]

(1.20)

where \( \mu \) is full rank and \( b \in \mathbb{R}^p \). We show in the supplementary appendix 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 (1.20), 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 (2012): the first stage converges to zero, but at a sufficiently slow rate that many standard asymptotic results are preserved. Let \( \hat{\Omega} \) be a consistent estimator for \( \lim_{T \to \infty} \text{Var} \left( \frac{1}{\sqrt{T}} f_T(\beta_0)'r_T^{-1}f_T(\beta_0)' \right) \) and define \( \tilde{g}_T(\beta) = \sqrt{T} \hat{\Omega}_{ff}^{-\frac{1}{2}} f_T(\beta) \) and \( \tilde{\gamma} = \text{vec}(\tilde{\Omega}) \) as before. Consider sequences of local alternatives with \( \beta_T = \beta_0 + \frac{b^*}{r_T \sqrt{T}} \) and let \( \Delta \tilde{g}_T = r_T^{-1} \hat{\Omega}_{ff}^{-\frac{1}{2}} \frac{\partial}{\partial \beta} f_T(\beta) \). As \( T \to \infty \), \((\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\gamma})\) converges to the Gaussian shift limit problem (1.20) with \( \mu = E[Z_tZ_t'] c \) and \( b = b^*, \square \)

In the Gaussian shift limit problem (1.20), the Neyman Pearson Lemma implies that the uniformly most powerful level \( \alpha \) test based on \((J, K, D)\) is \( \phi_K \) as defined in (1.13). Further, under the weak convergence assumption (1.1) for \((g, \Delta g)\) as in (1.20) the test \( \phi_{K_T} = 1 \{ K_T > \chi_{p,1-\alpha}^2 \} \) is asymptotically efficient in the sense of Mueller.
Under strong identification \( \phi_{K_T} = 1 \left\{ K_T > \chi^2_{p,1-\alpha} \right\} \) is also generally equivalent to the usual Wald tests, though we will need conditions beyond (1.1) to establish this. It is straightforward to show that a CLC test based on the weight function \( a(D, \gamma) \) will share these properties, and so be asymptotically efficient under sequences converging to (1.20), if and only if \( a(D_T, \gamma) \to p 0 \) under such sequences.

**Proposition 4** Denote by \( \mathcal{A} \), the class of weight functions functions \( a(D, \gamma) \) that are continuous in both \( D \) and \( \gamma \) for all full-rank \( D \). Fix \((\theta, \gamma) \in \Theta \times \Gamma \) with \( \theta \neq \theta_0 \) and suppose that \((g_T, \Delta g_T, \gamma)\) converges weakly to the Gaussian shift limit problem (1.20) with \( b \neq 0 \). For \( a(D, \gamma) \) almost-everywhere continuous with respect to the limiting measure \( F(g, \Delta g, \gamma) \) under \((\theta, \gamma)\),

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

if and only if \( a(D, \gamma) = 0 \) almost surely with respect to \( F(g, \Delta g, \gamma) \). Thus

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

Using this proposition, it is easy to see that the condition (1.19) 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 \((\mu_D, \gamma) \in M_D \times \Gamma \) with \( \mu_D \) full-rank and \( \Sigma_{\theta g} (\gamma) = \Sigma_{g\theta} (\gamma) = 0 \), for all \( C > 0 \) and all sequences \((\mu_{D_n}, \gamma_n) \in M_D \times \Gamma \) such that

\[\text{Formally, in the limit problem } \mu = \Delta g \text{ is known so to derive weighted average power optimal tests we need only consider weights on } b. \text{ For any weights } G(b) \text{ with density } g(b) \text{ that depends on } b \text{ only through } \|mb\|, \text{ so that } g(b) \propto g(\|mb\|), \phi_K \text{ is weighted average power maximizing in the limit problem and by Mueller (2011) } \phi_{K_T} \text{ is asymptotically optimal over the class of tests with correct size under (1.1) and (1.20).} \]
\((\mu_{D,n}, \gamma_n) \to (\mu_D, \gamma)\) we have

\[ d_H (M_D (\mu_{D,n}, \gamma_n) \cap B_C, \{\mu_D \cdot b : b \in \mathbb{R}^p\} \cap B_C) \to 0 \]

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 \(\beta_{m_{D,n}, \gamma_n}^u = \sup_{\alpha \in [0,1]} E_{m_{D,n}, \gamma_n} [\phi_{a}]\) and all \((\mu_{D,n}, \gamma_n) \to (\mu_D, \gamma)\) the MMR weight

\[ a_{MMRU} (\mu_{D,n}, \gamma_n) = \arg \min_{\alpha \in [0,1]} \sup_{\mu \in M_D (\mu_{D,n}, \gamma_n)} \left( \beta_{m_{D,n}, \gamma_n}^u - E_{m_{D,n}, \gamma_n} [\phi_{a}] \right) \]

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

Using Theorem 6 we can show 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 \(\pi_T = r_TC\) 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.

**Corollary 2** 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. \(\square\)

**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 \(\hat{\gamma} = \text{vec} (\hat{\Omega}_\eta)\) and note that \(\mathcal{M} (\hat{\gamma}) = \left\{ \hat{\Omega}_\eta^{-\frac{1}{2}} (f (\theta) - f (\theta_0)) \right\}\). Defining \(g_T (\theta) = \hat{\Omega}_\eta^{-\frac{1}{2}} (\hat{\gamma} - f (\theta))\) and
\[ \Delta g_T(\theta) = \frac{\partial}{\partial \theta} g_T(\theta) = \hat{\Omega}_T^{-\frac{1}{2}} \frac{\partial}{\partial \theta} f(\theta) \] as before, a global identification assumption yields that PI tests are asymptotically efficient.

**Corollary 3** Assume that \( \theta \) is in the interior of \( \Theta \) and that for all \( \delta > 0 \) there exists \( \varepsilon(\delta) > 0 \) such that \( \| f(\hat{\theta}) - f(\theta) \| < \varepsilon(\delta) \) implies \( \| \hat{\theta} - \theta \| < \delta \). Provided \( r_T \to 0 \), 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 \( \phi_{PI} \) is asymptotically efficient under strong and semi-strong identification.

### 1.7 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, compare favorably to alternative approaches in simulations calibrated to match an IV model with heteroskedastic time-series data, and are efficient in linear IV and minimum distance models with strong instruments.

### 1.8 Appendix

This appendix proves Theorems 1 and 2. Proofs for all other results, further details on our examples and simulations, and simulation results in a new Keynesian Phillips curve model, can be found in the supplementary appendix.
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

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

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

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

for $F_D$ the distribution of $D$, proving the theorem.

Proof of Theorem 2

We first argue that conditional on $D = d$ the test $\phi_{QCLR_r}$ is exactly equivalent to the level $\alpha$ 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 $r(d) = \infty$. For $r(d) < \infty$ and $K > 0$ or $J - r(D) > 0$, 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 $D = d$,

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

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

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

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

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

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

Since we've taken \( K = L - \frac{L}{L + r(D)} \cdot J \) and we know \( K \geq 0 \), we have that \( J \leq L + r(d) \).

Thus \( L + r(d) - \frac{r(d)}{L + r(d)} \cdot J \geq 0 \) and we can open the square root and collect terms to obtain \( QCLR_r = L \) on the set \( \{(J, K) \in \mathbb{R}_+^2 : L = K + \frac{L}{L + r(d)} \cdot J\} \), as we claimed.

Conditional on \( D = d \) the rejection region of \( \phi_{QCLR} \) is

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

Since \( J \) and \( K \) are pivotal under the null,

\[
P_{F_m = 0, \mu_D} \left\{ q_a(r(d)) < K + \frac{q_a(r(d))}{q_a(r(d)) + r(d)} \cdot J \middle| D = d \right\} = \alpha,
\]

so since \( K + \frac{q_a(r(d))}{q_a(r(d)) + r(d)} \cdot J \) is continuously distributed with support equal \( \mathbb{R}_+ \), \( q_a(r(d)) \) must be the \( 1 - \alpha \) quantile of this random variable. Hence, if we define the test \( \phi_{\tilde{a}(D)} \) as in (1.15) with \( \tilde{a}(D) = \frac{q_a(r(D))}{q_a(r(D)) + r(D)} \), we can see that \( c_a(\tilde{a}(d)) = q_a(d) \) and thus that \( \phi_{QCLR} = \phi_{\tilde{a}(d)} \) conditional on \( D = d \). Since this hold for all \( d \), \( \phi_{QCLR} \equiv \phi_{\tilde{a}(D)} \).

Thus, for any function \( r : D \rightarrow \mathbb{R}_+ \cup \{\infty\} \) there is a function \( \tilde{a} : D \rightarrow [0, 1] \) such that \( \phi_{QCLR} \equiv \phi_{\tilde{a}(D)} \).

To prove the converse, that for any CLC test \( \phi_{a(D)} \) for \( a : D \rightarrow [0, 1] \) we can find a function \( r : D \rightarrow \mathbb{R}_+ \cup \{\infty\} \) yielding the same test, fix the function \( a(D) \) and note that \( q_a(r(D)) \) is a continuous function of \( r(D) \) which is deceasing in \( r(D) \) and is bounded below by \( \chi_{p,1-\alpha}^2 \) and above by \( \chi_{k,1-\alpha}^2 \) (see Moreira (2003)). Hence for any value \( d \), as \( r(d) \) goes from zero to infinity \( \frac{q_a(r(d))}{q_a(r(d)) + r(d)} \) varies continuously between zero
and one, with \( \lim_{r(d) \to 0} \frac{q_0(r(d))}{q_0(r(d)) + r(d)} = 1 \) and \( \lim_{r(d) \to \infty} \frac{q_0(r(d))}{q_0(r(d)) + r(d)} = \frac{q_0(\infty)}{q_0(\infty) + \infty} = 0 \). If \( a(d) = 0 \) define \( \tilde{r}(d) = \infty \). If \( a(d) > 0 \), note that there exists a value \( r^* < \infty \) such that \( a(d) > \frac{q_0(r^*)}{q_0(r^*) + r^*} \), so by the intermediate value theorem we can pick \( \tilde{r}(d) \in [0, r^*] \) such that \( a(d) = \frac{q_0(r(d))}{q_0(r(d)) + \tilde{r}(d)} \). Repeating this exercise for all values \( d \) we can construct a function \( \tilde{r} : \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\} \) such that \( \varphi_{a(D)} = \varphi_{QLR_r} \), completing the proof.
Appendix A provides further details on the derivation of the limit problems discussed in Section 1.1 of the paper. Appendix B shows that general nonlinear GMM models which are weakly identified in the sense of Stock and Wright (2000) give rise to limiting problems of the form (1.2). Appendix C collects the proofs for all results save Theorems 1 and 2, for which the proofs are included with the paper. Appendix D concerns our linear IV simulations and gives power plots for PI tests in linear IV with homoskedastic errors, provides further information on our simulation design, and discusses our implementation of the MM1-SU, MM2-SU, and PI tests. Finally, Appendix E discusses simulation results in a nonlinear new Keynesian Phillips curve model.

Appendix A: Derivation of Limit Problems for Examples

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

Example I: Weak IV

Re-writing our moment condition we have that

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

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

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

we can see that provided that $\frac{1}{T} \sum Z_t Z' \rightarrow_p Q_Z$ for $Q_Z$ positive definite and $\frac{1}{\sqrt{T}} \sum Z_t V_{1,t}$ and $\frac{1}{\sqrt{T}} \sum Z_t V_{2,t}$ converge in distribution to jointly normal random vectors, the weak-
instruments sequence $\pi_T = \frac{1}{\sqrt{T}}$ implies that under true parameter value $\beta$,

$$
\sqrt{T} \left( \begin{array}{c} f_T(\beta_0) \\ - \frac{\partial}{\partial \beta} f_T(\beta_0) \end{array} \right) \rightarrow_d N \left( \begin{array}{c} \Omega_{ZC}(\beta - \beta_0) \\ Q_{ZC} \end{array} \right), \Omega(\beta_0) \right).$

Combined with the consistency of $\hat{\Omega}_{ff}$, this immediately yields (1.6).

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

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

where by assumption the first term converges to a $N(0, I_k)$ distribution and the second term converges to $\Omega^{-\frac{1}{2}} (f(\theta) - f(\theta_0))$ by the Continuous Mapping Theorem and the assumed consistency of $\hat{\Omega}_\eta$. The consistency of $\Delta g_T(\theta)$ for $\Omega^{-\frac{1}{2}} \frac{\partial}{\partial \theta} f(\theta_0)$ follows similarly, immediately implying (1.7).

**Appendix B: Limit Problem for Weak GMM Models**

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

$$E_\psi [f_t(\psi_0)] = \hat{h}_1(\psi_1) + \frac{1}{\sqrt{T}} \hat{h}_2(\psi_1, \psi_2)$$
for \( \tilde{h}_1, \tilde{h}_2 \) continuously differentiable. Letting \( \psi \) denote the true parameter value, for sample size \( T \) let us reparametrize in terms of \( \theta = \theta_T = (\sqrt{T}(\psi - \psi_{1.0}) , \psi_2) \) 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 \( \psi_1 \), 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 \( \theta \) is

\[
E_{\theta} [g_t(\theta_0)] = \tilde{h}_1 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}} \right) + \frac{1}{\sqrt{T}} \tilde{h}_2 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}}, \theta_2 \right)
\]

\[
= \tilde{h}_1 \left( \psi_{1.0} - \frac{\theta_1}{\sqrt{T}} \right) + \frac{1}{\sqrt{T}} \tilde{h}_2 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}}, \theta_2 \right)
\]

\[
= \tilde{h}_1 (\psi_{1.0}) + \frac{1}{\sqrt{T}} \frac{\partial}{\partial \psi_1} \tilde{h}_1 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}} \right) \theta_1 + \frac{1}{\sqrt{T}} \tilde{h}_2 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}}, \theta_2 \right)
\]

where in the last step we have taken a mean value expansion in \( \theta_1 \) with intermediate value \( \tilde{\theta}_1 \). Note, however, that the identifying assumption for GMM implies that \( \tilde{h}_1 (\psi_{1.0}) = 0 \) while under our continuity assumptions

\[
\frac{\partial}{\partial \psi_1} \tilde{h}_1 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}} \right) \rightarrow \frac{\partial}{\partial \psi_1} \tilde{h}_1 (\psi_{1.0})
\]

and

\[
\tilde{h}_2 \left( \psi_{1.0} + \frac{\theta_1}{\sqrt{T}}, \theta_2 \right) \rightarrow \tilde{h}_2 (\psi_{1.0}, \theta_2).
\]

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

\[
h(\theta) = \frac{1}{\sqrt{T}} \frac{\partial}{\partial \psi_1} \tilde{h}_1 (\psi_{1.0}) \theta_1 + \frac{1}{\sqrt{T}} \tilde{h}_2 (\psi_{1.0}, \theta_2) .
\]

Note that the strongly identified parameters \( \theta_1 \) enter \( h(\theta) \) linearly while the weakly identified parameters \( \theta_2 \) may enter non-linearly. Suppose that for our original moment
functions \( f_t(\theta) \), we have that under the sequence of alternatives \( \psi_T = (\psi_{1,0} - \frac{1}{\sqrt{T}} \theta_1, \psi_2) \)

\[
\frac{1}{\sqrt{T}} \left( \begin{array}{c}
\sum f_t(\psi_0) - E_{\psi_T} [f_t(\psi_0)] \\
vec \left( \sum \frac{\partial}{\partial \psi_0} f_t(\psi_0) - E_{\psi_T} \left[ \frac{\partial}{\partial \psi_0} f_t(\psi_0) \right] \right)
\end{array} \right) \rightarrow_d N (0, \Omega_f)
\]

where \( \Omega_f \) is consistently estimable and \( \Omega_{ff} \), the upper-left block of \( \Omega_f \), is full rank. Since alternative \( \theta \) in the new parametrization corresponds to this sequence of alternatives in the original parametrization, this implies that under \( \theta \) we have

\[
\frac{1}{\sqrt{T}} \left( \begin{array}{c}
\sum g_t(\theta_0) - E_\theta [g_t(\theta_0)] \\
vec \left( \sum \frac{\partial}{\partial \theta_0} g_t(\theta_0) - E_\theta \left[ \frac{\partial}{\partial \theta_0} g_t(\theta_0) \right] \right)
\end{array} \right) \rightarrow_d N (0, \Omega)
\]

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

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

and

\[
\Delta g_T(\theta_0) = \frac{1}{\sqrt{T}} \hat{\Omega}_{gg}^{-\frac{1}{2}} \sum \frac{\partial}{\partial \theta_0} g_t(\theta_0)
\]

note that

\[
\left( \begin{array}{c}
g_T(\theta_0) \\
\Delta g_T(\theta_0)
\end{array} \right) \rightarrow_d N \left( \begin{pmatrix} m \\ \mu \end{pmatrix}, \begin{pmatrix} I & \Sigma_{g\theta} \\ \Sigma_{\theta g} & \Sigma_{\theta\theta} \end{pmatrix} \right)
\]

where \( \mu = \lim_{T \to \infty} E_\theta \left[ \frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \theta_0} g_t(\theta_0) \right] \) 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 \( \Delta g_T \) using the Jacobian of the original moment function, \( \frac{\partial}{\partial \psi_0} f_t(\psi_0) \), post-multiplied by an appropriate sequence of normalizing matrices \( A_T \), as in Section 1.6.1.
We can say a bit more regarding the strongly identified parameters $\theta_1$. Note that by the definition of $\theta$, $\frac{\partial}{\partial \theta} g_t(\theta_0) = \frac{1}{\sqrt{T}} \frac{\partial}{\partial \psi_1} f_1(\psi_0)$. Hence, $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \psi_1} f_1(\psi_0) = \frac{1}{T} \sum \frac{\partial}{\partial \psi_1} f_1(\psi_0)$ and we can re-write $\mu$ as $\lim_{T \to \infty} E_{\theta} \left[ \frac{1}{T} \sum \frac{\partial}{\partial \psi_1} f_1(\psi_0) \right]$.

Further, the central limit theorem we have assumed for $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \psi_1} f_1(\psi_0)$ implies that $\frac{1}{\sqrt{T}} \sum \frac{\partial}{\partial \psi_1} g_t(\theta_0) \to_p \mu_1 = \lim_{T \to \infty} E_{\theta} \left[ \frac{1}{T} \sum \frac{\partial}{\partial \psi_1} f_1(\psi_0) \right]$. Together with (1.21) this implies that under standard regularity conditions (see e.g. Newey and McFadden (1994)) $h(\theta_1, \theta_2, 0) = \mu_1 \cdot \theta_1$ and hence that in the special case where all parameters are strongly identified we obtain the Gaussian shift limiting problem

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

Appendix C: Proofs

Proof of Corollary 1

Conditional on $D = d$ the CLC test $\phi_{d(D)}$ 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

$$C_d = \left\{ (\sqrt{J_1}, \sqrt{K_1}) : K + a(d) \cdot J \leq c_\alpha(a(d)) \right\}.$$

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 $(\sqrt{J_1}, \sqrt{K_1})$, $(\sqrt{J_2}, \sqrt{K_2}) \in C_d$ 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 Theorem 4

(1) follows 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 \(\phi\) any size \(\alpha\) test for \(H_0 : \tau_A = \tau_B = 0\) based on \((A, B)\) there exists some \(\tilde{\lambda} > 0\) such that for \(0 < \lambda < \tilde{\lambda}\),

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

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 by the fact that \((J, K) \mid D = d \sim \left(\chi^2_{k-p}(\tau_I), \chi^2_p(\tau_K)\right)\).

Establishing statement (2) is similarly straightforward. In particular for \(F_{t_{i,j}}\) 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 \(\alpha\) test based on \((A, B) \sim \left(\chi^2_{k-p}(\tau_A), \chi^2_p(\tau_B)\right)\) is \(\phi^*_p = 1 \left\{\frac{t_k}{t_{k+1}} A + \frac{t_l}{l_{l+1}} B > c\right\}\), where \(c\) is the \(1 - \alpha\) quantile of a \(\frac{t_k}{t_{k+1}} \chi^2_{k-p} + \frac{t_l}{l_{l+1}} \chi^2_p\) distribution. In particular, for \(\Phi_\alpha\) the class of level \(\alpha\) tests based on \((A, B)\),

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

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

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

\[
\lim_{n \to \infty} \sup_{(m, \mu_D) \in \mathcal{H}_1} \left(\beta^*_n - E_{m, \mu_D}[\phi_{a_n}(D)]\right) = \inf_{a \in \mathcal{A}} \sup_{(m, \mu_D) \in \mathcal{H}_1} \left(\beta^*_a - E_{m, \mu_D}[\phi_a(D)]\right).
\]

(1.22)

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 $\nu$ Lebesgue measure on $\mathbb{R}^{kp+2}$

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

(1.23)

for all $\nu$-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 $f_{JKD}(m, \mu)$ with respect to Lebesgue measure $\nu$. Note that for any bounded, continuous function $n$ of $(J, K, D)$, $n(J, K, D)f_{JKD}(m, \mu)$ is an integrable function with respect to Lebesgue measure $\nu$. Hence, (1.23) 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

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

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

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

64
for some $\varepsilon > 0$, which since $E_{m^*\mu^*_D} [\phi_{a_1(D)}] = \lim_{n \to \infty} E_{m^*\mu^*_D} [\phi_{a_1(D)}]$ implies that

$$\lim_{n \to \infty} \sup_{(m,\mu_D) \in H_1} \left( \beta_{m,\mu_D}^* - E_{m,\mu_D} [\phi_{a_1(D)}] \right) \geq \inf_{a \in A} \sup_{(m,\mu_D) \in H_1} \left( \beta_{m,\mu_D}^* - E_{m,\mu_D} [\phi_{a(D)}] \right) + \varepsilon$$

which contradicts (1.22).

**Proof of Lemma 1**

Note that $\Omega = A \otimes B$ implies that $\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix} \otimes I$. To prove the result, it is easier to work with the formulation of the problem discussed in AMS. In particular, consider $k \times 1$ random vectors $\tilde{S}$ and $\tilde{T}$ (denoted by $S$ and $T'$ in AMS) with

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

were $c_\beta$ ranges over $\mathbb{R}$ for different true values of $\beta$. AMS (Theorem 1) show that the maximal invariant to rotations of the instruments is $(\tilde{S}'\tilde{S}, \tilde{S}'\tilde{T}, \tilde{T}'\tilde{T})$, and note that the $S$ statistic can be written $S = \tilde{S}'\tilde{S}$, while the $K$ statistic is $K = \frac{\tilde{S}'\tilde{T}}{\tilde{T}'\tilde{T}}$. 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) $\tilde{T}'\tilde{T} = D'\Sigma_D^{-1}D'$, where $\Sigma_D = I \left( \begin{bmatrix} A_{12}/A_{11} \\ A_{22}/A_{11} \end{bmatrix} \right)^2$. Hence, in the limit problem (1.6) with $\Sigma = \begin{bmatrix} 1 & A_{12}/A_{11} \\ A_{12}/A_{11} & A_{22}/A_{11} \end{bmatrix} \otimes I$, the maximal invariant under rotations of the instruments $(\tilde{S}'\tilde{S}, \tilde{S}'\tilde{T}, \tilde{T}'\tilde{T})$ is a one-to-one transformation of $(J, K, D'\Sigma_D D)$.

By the imposed invariance to rotations of the instruments, it is without loss of generality to assume that $d_\beta \mu_\pi = e_1 \cdot \sqrt{r}$, where $e_1 \in \mathbb{R}^k$ has a one in its first entry and zeros everywhere else. Hence, $\tilde{T}'\tilde{T} = D'\Sigma_D D \sim \chi^2_r (r)$. 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 $\phi_a$ can be written as a function of $||m||$ and $r$, the power envelope for unconditional linear combination
tests is defined by $\beta_{m\mid r} = \sup_{\alpha \in [0,1]} E_{m\mid r} [\phi_\alpha]$, and the maximum regret for any unconditional linear combination test (taking $\mu_D$ and hence $r$ to be known) is

$$\sup_{m\in R} \left( \beta_{m\mid r} - E_{m\mid r} [\phi_\alpha] \right)$$

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

Proof of Proposition 1

The discussion preceding Proposition 1 establishes that under $(\theta, \gamma), (J_T, K_T, D_T) \to_d (J, K, D)$ 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 implies that

$$Pr_{T,\theta,\gamma} \{ K_T + a(D_T, \hat{\gamma}) J_T > c_\alpha(a(D_T, \hat{\gamma})) \} \to \alpha$$

$$Pr_{m=0,\mu_D} \{ K + a(D, \gamma) J > c_\alpha(a(D, \gamma)) \} = \alpha$$

which proves (1.17). To prove (1.18) note that the results above establish that $\phi_{a(D, \gamma)}$ is almost-everywhere continuous with respect to $F$, and hence for $f \in F$

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

Since the left hand side is bounded, convergence in distribution implies convergence in expectation, proving (1.18).
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

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

These choices imply that our \( S_T \) and \( K_T \) coincide exactly with \( AR \) and \( LM \) in ACG, and that our \( D_T \) is \( \sqrt{T} \hat{\Omega}_{ff}^{-\frac{1}{2}} \hat{\Delta} \) for \( \hat{\Delta} \) as in ACG. To prove the proposition we will rely heavily on their results. ACG consider two cases: sequences \( \lambda_T \) 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 \( (LM, AR, \hat{\Delta}) \) converges in distribution to \( \left( \chi^2_1, \chi^2_{k-1}, \hat{\Delta} \right) \) where all three random variables are independent and \( \hat{\Delta} \) has a non-degenerate Gaussian distribution. Since \( \hat{\Omega}_{ff} \rightarrow_p \Omega_{ff} \) which is full-rank by assumption, this proves that \( (K_T, S_T, D_T) \rightarrow_d (\chi^2_1, \chi^2_{k-1}, \hat{\Delta}) \) where again all the variables on the RHS are mutually independent and \( \hat{\Delta} \) has a non-degenerate Gaussian distribution. Thus, by the Continuous Mapping Theorem and consistency of \( \hat{\Sigma}_{\theta^0} \) and \( \hat{\Sigma}_{\theta^0} \), which under the null follows from (6.7) and (6.9) in ACG, we have that

\[
(1 - a(D_T, \hat{\gamma})) K_T + a(D_T, \hat{\gamma}) S_T - c_\alpha (a(D_T, \hat{\gamma})) \rightarrow_d
\]

\[
(1 - a(D, \gamma)) K + a(D, \gamma) S - c_\alpha (a(D, \gamma))
\]

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

Next, consider the case where \( \sqrt{T} ||\pi_T|| \) diverges. Let

\[
(\tilde{g}_T, \Delta \tilde{g}_T, \hat{\gamma}) = h \left( g_T, \Delta g_T, \hat{\gamma}; ||\pi_T||^{-1} \right),
\]

67
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 $a(\tilde{D}_T, \gamma)$ thus implies that $a(\tilde{D}_T, \gamma) \to_p a_0$. Since ACG establish the convergence in distribution of $(LM, AR)$ under sequences of this type, we have that

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

and thus that the CLC test $\phi_{a(\tilde{D}_T, \gamma)}$ has asymptotic rejection probability equal to $\alpha$ under these sequences. By the assumed invariance the postmultiplication, however, this implies that $\phi_{a(D_T, \gamma)}$ has asymptotic rejection probability $\alpha$ 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, $\phi_K$ is efficient in the limit problem (1.20) by the Neyman-Pearson Lemma, and $\phi_{K_T} = \phi_{\bar{a}(D_T, \gamma)}$ for $\bar{a}(D, \gamma) \equiv 0$, $\bar{a} \in A_c$, so

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

follows from Proposition 3.

If $a(D, \gamma) = 0$ almost surely, then we have that $\lim_{T \to \infty} E_{T,(\theta, \gamma)} \left[ \phi_{a(D_T, \gamma)} \right] = \lim_{T \to \infty} E_{T,(\theta, \gamma)} [\phi_{K_T}]$ 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 $\phi_{a^*}$ 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 \( \phi_K \) in the limit problem, which together with Proposition 3 implies that \( \lim_{T \to \infty} E_{T, (\theta, \gamma)} [\phi_a(D_{T, \gamma})] < \lim_{T \to \infty} E_{T, (\theta, \gamma)} [\phi_K] \).

**Proof of Theorem 6**

Define \( \mathcal{M}_L = \{ \mu_D \cdot b : b \in \mathbb{R}^n \} \). 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 \( \phi_K \) is UMP over the class of tests depending on \( (J, K, D) \) against \( m \in \mathcal{M}_L \) for \( \Sigma_D = 0 \), we can see that for \( \beta_{m, \mu_D, \gamma}^u = \sup_{a \in [0, 1]} E_{m, \mu_D, \gamma} [\phi_a] \) we have \( \beta_{m, \mu_D, \gamma}^u = E_{m, \mu_D, \gamma} [\phi_K] \) \( \forall m \in \mathcal{M}_L \). Thus,

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

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 \( \zeta = \frac{\varepsilon(a)}{2}, B_C = \{ m : |m| \leq C \}, \) and \( A = \mathcal{M}_L \cap B_{C_\zeta}^C \) (for \( B_{C_\zeta}^C \) the complement of \( B_{C_\zeta} \))

\[
\sup_{m \in A} \left( E_{m, \mu_D, \gamma} [\phi_K] - E_{m, \mu_D, \gamma} [\phi_a] \right) \leq 1 - \frac{\varepsilon(a)}{2}.
\]
by the definition of $C_\zeta$. Thus, for $\tilde{A} = \mathcal{M}_L \cap B_{C_\zeta}$,

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

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 at some $m^* \in \tilde{A}$.

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

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

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 $\varepsilon(\delta) > 0$ such that $\varepsilon(a) < \varepsilon(\delta)$ only if $a < \delta$. Further, by the intermediate value theorem there exists $a(\delta) > 0$ such that $\varepsilon(a(\delta)) = \varepsilon(\delta) / 2$.

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

$$
\arg \min_{a \in [0, 1]} \sup_{m \in \mathcal{M}_L (\mu_D, \gamma_n, \gamma)} \left( \beta_{m,\mu_D,\gamma_n} - E_{m,\mu_D,\gamma_n} [\phi_a]\right) < \nu.
$$

Fixing $\nu$, let $\varepsilon^* = \varepsilon(\nu)$, $a^* = a(\nu)$, for $\varepsilon(\cdot)$ and $a(\cdot)$ as defined above. Let $\zeta^* = \varepsilon^* / 4$, and take $C^*$ to be such that

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

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

$$
\inf_{a \in [0, 1]} \sup_{m \in \mathcal{M}_L (\mu_D, \gamma_n) \cap B_{C^*}} \left( \beta_{m,\mu_D,\gamma_n} - E_{m,\mu_D,\gamma_n} [\phi_a]\right) > 3\zeta^*$$

70
while
\[
\sup_{m \in \mathcal{M}_D(\mu_{D,n}, \gamma_n) \cap B_{C^*}} \left( \beta_{m,\mu_{D,n},\gamma_n}^u - E_{m,\mu_{D,n},\gamma_n}[\phi_{u^*}] \right) < 3\zeta^* \]

and
\[
\sup_{m \in \mathcal{M}_D(\mu_{D,n}, \gamma_n) \cap B_{C^*}} \left( \beta_{m,\mu_{D,n},\gamma_n}^u - E_{m,\mu_{D,n},\gamma_n}[\phi_{u^*}] \right) < 2\zeta^*.
\]

Thus, for \( n > N \) we have
\[
\sup_{m \in \mathcal{M}_D(\mu_{D,n}, \gamma_n)} \left( \beta_{m,\mu_{D,n},\gamma_n}^u - E_{m,\mu_{D,n},\gamma_n}[\phi_{u^*}] \right) < \inf_{\nu \geq 0} \sup_{m \in \mathcal{M}_D(\mu_{D,n}, \gamma_n) \cap B_{C^*}} \left( \beta_{m,\mu_{D,n},\gamma_n}^u - E_{m,\mu_{D,n},\gamma_n}[\phi_{u^*}] \right)
\]
and thus that \( a(\mu_{D,n}, \gamma_n) < \nu \) since \( a^* < \nu \). Since we can repeat this argument for all \( \nu > 0 \) we obtain that \( a(\mu_{D,n}, \gamma_n) \to 0 \) as desired.

**Proof of Corollary 2**

Let \((\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\gamma}) = h (g_T, \Delta g_T, \tilde{\gamma}, r_T^{-1}/\sqrt{T})\) for \( h \) as defined in (1.16), and note that this is the same definition of \((\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\gamma})\) given near the beginning of Section 1.6.3. 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 \( \hat{\mu}_D = \tilde{D}_T \) will be the same as those based on \((g_T, \Delta g_T, \tilde{\gamma})\) with estimate \( \mu_D = D_T \). 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 \( \mu = E [Z_i Z_i'] c \) and \( b = b^* \). Note that in linear IV we have
\[
\mathcal{M}_D(\mu_D, \gamma) = \{(I - \Sigma_{\beta g} \cdot b)^{-1} \mu_D \cdot b : b \in \mathbb{R}\}.
\]
Hence, for any sequence \((\mu_{D,n}, \gamma_n)\) with \( \mu_{D,n} \to \mu, ||\mu|| > 0, \) and \( \Sigma_{\beta g}(\gamma_n) \to 0 \) we can see that for any \( C > 0 \)
\[
d_H(\mathcal{M}_D(\mu_{D,n}, \gamma_n) \cap B_C, \{\mu_D \cdot b : b \in \mathbb{R}^p\} \cap B_C) \to 0,
\]
so by Theorem 6 we have that \( a_{Pl}(\mu_{D,n}, \gamma_n) \to 0 \). Note, however, that under our
assumptions $(\mu_D, \gamma) \rightarrow_p (\mu, \gamma)$ with $\|\mu\| > 0$ and $\Sigma_{\beta\theta} (\gamma) = \Sigma_D (\gamma) = 0$. Thus, the Continuous Mapping Theorem yields that $a_{PL} (\mu_D, \gamma) \rightarrow_p 0$.

**Proof of Corollary 3**

Note that

$$M (\gamma) = \left\{ \Omega_{\eta, T}^{-\frac{1}{2}} (f (\theta) - f (\theta_0)) : \theta \in \Theta \right\} = r_T^{-\frac{1}{2}} \left\{ \left( r_T^{-1} \Omega_{\eta, T} \right)^{-\frac{1}{2}} (f (\theta) - f (\theta_0)) : \theta \in \Theta \right\}.$$

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

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

From the definition of differentiability, we know that

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

Thus, for any sequence $\delta_T \to 0$,

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

Moreover, by our identifiability assumption on $\theta_0$ we know that for any constant $K > 0$,

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

Combined with the previous equation, this implies that

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

which in turn shows that for any $C > 0$, provided $\theta$ belongs to the interior of $\Theta$

$$d_H \left( \left\{ r_T^{-\frac{1}{2}} \Omega_{\eta, 0}^{-\frac{1}{2}} (f (\theta) - f (\theta_0)) : \theta \in \Theta \right\} \cap B_C, r_T^{-\frac{1}{2}} \left\{ \Omega_{\eta, 0}^{-\frac{1}{2}} \frac{\partial}{\partial \theta} f (\theta_0) : b \in \mathbb{R}^p \right\} \cap B_C \right) \to 0.$$
Thus, we see that for any $r_T^{-1} \Omega_{nT} \to \Omega_{0,0}$ 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 $a_{MMR}(\gamma) \to_{p} 0$.

One can show that sequences of local alternatives of the form $\theta_T = \theta_0 + \frac{1}{T} b^*$ yield Gaussian Shift limit problems in this model. The fact that $a_{MMR}(\gamma) \to_{p} 0$ 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: Additional Details of Weak IV Simulation**

In Section 1.5.2 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: Power Plots for Homoskedastic Model**

Figures 1-5 and 1-6 plot power curves for the CLR, K, AR, and PI tests in the linear IV calibrations discussed in Section 1.5.1 of the paper.

**Appendix D.2: Estimation of Parameters for the Limit Problem**

The behavior of $(g, \Delta g)$ in the weak IV limit problem (1.6) is determined entirely by $(m, \mu, \Omega)$. The set $\mathcal{M}(\mu)$ of possible values $m$ given $\mu$ 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 $\mu$ and $\Omega$. 

73
Figure 1.5: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with five instruments, discussed in Section 1.5.1.
Figure 1-6: Power functions of CLR, AR (or S), K, and PI tests in homoskedastic linear IV with two instruments and ten instruments, discussed in Section 1.5.1.
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 \(\Delta c\)), \(X_t\) is the real interest rate, and \(Z_t\) is a \(4 \times 1\) vector of instruments which following Yogo we take to be lagged values of the nominal interest rate, inflation, consumption growth, and the log dividend-price ratio. We focus on the case with \(X_t\) the risk-free rate since this is the case for which Yogo finds the strongest relationship between the instruments and the endogenous regressor (see Table 1 of Yogo (2004)). All data is de-meaned prior to beginning the analysis.

For country \(i\) we estimate \(\mu\) 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 \(\beta\), and let \(\hat{\Omega}_i\) be the Newey-West covariance estimator for \(Var \left( (f_i(\hat{\beta}_i)', Z_t X_t') \right)\) based on 3 lags of all variables. These estimates will not in general be consistent for the parameters of the limit problem under weak-instrument asymptotics, but give us empirically reasonable values for our simulations.

**Appendix D.3: Simulation Design**

For each country \(i\) we consider the problem of testing \(H_0: \beta = \beta_0\) in the limit problem.

For true parameter value \(\beta\), in simulation runs \(b = 1, ..., B\) we draw

\[
\left( \begin{array}{c} g_b \\ \Delta g_b \end{array} \right) \sim N \left( \begin{array}{c} \hat{\mu}_i (\beta - \beta_0) \\ \hat{\mu}_i \end{array} \right), \hat{\Sigma}_i
\]

where

\[
\hat{\Sigma}_i = \begin{bmatrix} I & \hat{\Sigma}_{g0,i} \\ \hat{\Sigma}_{g0,i} & \hat{\Sigma}_{00,i} \end{bmatrix} = \begin{bmatrix} I & \hat{\Omega}_{f0,i}^{\frac{1}{2}} \hat{\Omega}_{f0,i} \hat{\Omega}_{f0,i}^{\frac{1}{2}} \\ \hat{\Omega}_{f0,i}^{\frac{1}{2}} \hat{\Omega}_{f0,i} \hat{\Omega}_{f0,i}^{\frac{1}{2}} & \hat{\Omega}_{f0,i}^{\frac{1}{2}} \hat{\Omega}_{f0,i} \hat{\Omega}_{f0,i}^{\frac{1}{2}} \end{bmatrix}
\]

76
Note that this is the limiting distribution (1.6) of the normalized moment condition and Jacobian \((g_T, \Delta g_T)\) in a weak IV problem with true parameters \(\beta, \Omega = \hat{\Omega}_t,\) and \(\mu = \hat{\mu}_t.\) We then calculate the S and K tests \(\phi_{S,b}, \phi_{K,b}\) as in (1.14) and (1.13). We define the QCLR test as in Theorem 2 and, following Kleibergen (2005), take \(r = \frac{D^T_b \hat{\Sigma}^{-1}_{D,i} D_b}{\sqrt{D^T_b \hat{\Sigma}^{-1}_{D,i} D_b}}\) for \(\hat{\Sigma}_{D,i} = \hat{\Sigma}_{g_{0,i}} - \hat{\Sigma}_{g_{0,i}} \hat{\Sigma}_{g_{0,i}}.\) Details on the implementation of the MM-SU tests are discussed in the next section. Finally, to calculate the PI test we take

\[
\hat{\mu}_{D,b} = D_b \cdot \frac{\max \{ D^T_b \hat{\Sigma}^{-1}_{D,i} D_b - k, 0\} }{\sqrt{D^T_b \hat{\Sigma}^{-1}_{D,i} D_b}}
\]

which is a generalization of the positive-part estimator \(\hat{r}_{PP}\) to the non-Kronecker case, and consider \(\phi_{PL,b} = \phi_{MMRI} \hat{\mu}_{D,b} \). Details on calculation of the PI test are given in the section D.6.

For each value \(\beta\) in a grid we estimate the power of each test against this alternative by averaging over \(B,\) e.g. estimating the power of \(\phi_K\) by \(\frac{1}{\hat{\theta}} \sum \phi_{K,b}\) (where we take \(B = 5,000\) for the MM tests and \(B = 10,000\) for the others), and repeat this exercise for each of the eleven countries considered.

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

The MM1-SU and MM2-SU procedures of MM maximize weighted average power against particular weights on \((\beta, \mu)\) over a class of tests satisfying a sufficient condition for local unbiasedness. To relate the results of MM in our context, we can take \((Z'Z)^{-1} Z'Y\) (in the notation of MM) to equal \(\begin{pmatrix} g & \Delta g \end{pmatrix}\) and then derive their statistics \(S\) and \(T\) as they describe, noting that \(S\) as defined in MM is, up to rotation, equal to \(g\) as defined here. MM calculate their weights using the \(2 \times 2\) and \(k \times k\) symmetric positive definite matrices \(\Omega^*\) and \(\Phi^*\) solving \(\min \| \Sigma - \Omega \otimes \Phi \| \). (see MM for the weights). Thus, since we estimate different covariance matrices for each of the 11 countries in the Yogo data, the MM tests use different weight functions for each country. For each pair \((\Omega^*, \Phi^*)\) MM consider two different weight functions, which they label MM1 and MM2 respectively. Each of these weight functions features a tuning parameter (which MM call \(\sigma\) and \(\zeta\) for the MM1 and MM2 weights, respec-
tively). Following the choice made by MM in their simulations we set $\sigma = \zeta = 1$ for the simulations reported in the paper.

MM consider several different tests based on their weights. They find in simulation that weighted average power optimal conditionally similar tests can have some undesirable power properties in non-homoskedastic linear IV models, in particular exhibiting substantial bias. To remedy this they impose further restrictions on the class of tests, considering first locally unbiased (LU) tests, which satisfy $\frac{\partial}{\partial \beta} E_{\beta_0, \mu}[\phi] = 0$ for all $\mu \in \mathcal{M}$. They then consider the class of strongly unbiased (SU) tests which satisfy the condition $E_{\beta_0, \mu}[\phi g] = 0$ for all $\mu \in \mathcal{M}$, and which they show are a subset of the LU tests. They find that weighted average power optimal tests within this class (based on the MM1 and MM2 weights) have good power in their simulations, and it is these tests which we consider here.

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

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

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

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

To complement the power simulation results in the main text (which consider the power of the MM-SU and CLC tests at parameter values estimated from the data used by Yogo (2004)), Tables 1.3 and 1.4 report the weighted average power of the CLC and MM-SU tests under the MM1 and MM2 weights, respectively (the CLC tests considered here are the same as in Section 1.5.2), based on the covariance matrices $\Sigma$ estimated from the data. Table 1.3 shows that the MM1-SU test has higher power than the PI, QCLR, S and K tests under the MM1 weights, as expected. Likewise, Table 1.4 shows that the MM2-SU test has power greater than the PI, QCLR, S and K tests under the MM2 weights, up to simulation error (the power of the MM2-SU test appears lower than that of the PI and QCLR tests in the calibration to German data, but the difference is statistically insignificant at the 10% level).
### Table 1.4: Weighted average power of tests under MM2 weights of Moreira and Moreira (2013), with covariance $\Sigma$ calibrated to data from Yogo (2004). Based on 5,000 simulations.

<table>
<thead>
<tr>
<th>Country</th>
<th>MM2-SU</th>
<th>PI</th>
<th>QCLR</th>
<th>AR</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>37.93%</td>
<td>35.70%</td>
<td>35.33%</td>
<td>31.33%</td>
<td>32.73%</td>
</tr>
<tr>
<td>Canada</td>
<td>17.87%</td>
<td>17.22%</td>
<td>16.20%</td>
<td>15.56%</td>
<td>15.04%</td>
</tr>
<tr>
<td>France</td>
<td>15.93%</td>
<td>14.63%</td>
<td>14.39%</td>
<td>12.98%</td>
<td>13.35%</td>
</tr>
<tr>
<td>Germany</td>
<td>8.59%</td>
<td>8.77%</td>
<td>8.83%</td>
<td>8.01%</td>
<td>8.24%</td>
</tr>
<tr>
<td>Italy</td>
<td>36.35%</td>
<td>34.43%</td>
<td>33.89%</td>
<td>30.24%</td>
<td>30.75%</td>
</tr>
<tr>
<td>Japan</td>
<td>20.94%</td>
<td>17.97%</td>
<td>18.25%</td>
<td>15.83%</td>
<td>16.12%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>12.57%</td>
<td>12.19%</td>
<td>12.25%</td>
<td>11.52%</td>
<td>10.95%</td>
</tr>
<tr>
<td>Sweden</td>
<td>34.23%</td>
<td>32.91%</td>
<td>33.36%</td>
<td>29.87%</td>
<td>28.89%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>18.99%</td>
<td>16.78%</td>
<td>16.45%</td>
<td>14.94%</td>
<td>15.21%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>35.59%</td>
<td>31.20%</td>
<td>30.49%</td>
<td>27.38%</td>
<td>28.35%</td>
</tr>
<tr>
<td>United States</td>
<td>38.21%</td>
<td>36.82%</td>
<td>36.51%</td>
<td>32.36%</td>
<td>32.84%</td>
</tr>
</tbody>
</table>

### Appendix D.5: Implementation of PI Test

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

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

so the critical task is evaluating $a_{MMRU}(\tilde{\mu}_{D,b})$. As discussed in Section 1.4 above, $a_{MMRU}(\tilde{\mu}_{D,b})$ solves a minimization problem which depends on $\tilde{\mu}_{D,b}$ and on $\hat{\Sigma}_i$.

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

$$Pr \left\{ \chi_p^2(\tau_K) + a \cdot \chi_p^2(\tau_J) > c_a(a) \right\}$$

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

We next consider a grid $B$ of 41 values for the alternative $\beta_h$. For each value $\beta_h$, we solve for

$$\tilde{\mu}_{h,h} = \left( I - \hat{\Sigma}_{0,i} (\beta_h - \beta_0) \right)^{-1} \hat{\mu}_{D,b}$$
which gives us the value \( \mu \) for which \( D \) would have mean \( \hat{\mu}_{D,b} \) under alternative \( \beta_h \). Note that the mean \( m \) of \( y \) under \( \beta_h \) is then \( m_{b,h} = \hat{\mu}_{b,h} (\beta_h - \beta_0) \). We take draws \( l = 1, \ldots, L = 10,000 \) from \( D_{bj} \sim N (\hat{\mu}_{D,b}, \hat{\Sigma}_{D,b}) \) and for each \((h, l)\) pair we calculate \( \tau_{K,b,h,l} = m_{b,h}' P_{D_{bj}, m_{b,h}} \) and \( \tau_{J,b,h,l} = m_{b,h}' M_{D_{bj}, m_{b,h}} \).

We could estimate the power of the linear combination test with weight \( a \) against alternative \( \beta_h \) by

\[
\hat{E} [\phi_a | \beta = \beta_h] = \frac{1}{L} \sum Pr \{ \chi_p^2 (\tau_{K,b,h,l}) + a \cdot \chi_p^2 (\tau_{J,b,h,l}) > c_a(a) \}.
\]

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 \((\tau_K, \tau_J)\) values given the by average of its endpoints. This gives us pairs \((\bar{\tau}_{K,q}, \bar{\tau}_{J,q})\) for \( q \in \{1, \ldots, 1000\} \), and we estimate

\[
\hat{E} [\phi_a | \beta = \beta_h] = \frac{1}{1000} \sum Pr \{ \chi_p^2 (\bar{\tau}_{K,q}) + a \cdot \chi_p^2 (\bar{\tau}_{J,q}) > c_a(a) \}
\]

where by using \((\bar{\tau}_{K,q}, \bar{\tau}_{J,q})\) we need only calculate power 1000 times rather than 10000.

To further speed computation, we approximate \( Pr \{ \chi_p^2 (\bar{\tau}_{K,q}) + a \cdot \chi_p^2 (\bar{\tau}_{J,q}) > c_a(a) \} \) by interpolating using our stored values for \( Pr \{ \chi_p^2 (\tau_K) + a \cdot \chi_p^2 (\tau_J) > c_a(a) \} \).

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

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

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

E.1 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 et al. (2014) review the empirical literature on the role of expectations in the NKPC and find that parameter estimates are extremely sensitive to model specification and, conditional on correct specification, suffer from weak identification. To address these weak identification issues Magnusson and Mavroeidis (2010) (henceforth MM) propose identification-robust $S$ and $K$ statistics for testing hypotheses on NKPC parameters using a minimum distance approach. These statistics will form the basis for our analysis.

MM study a simple new Keynesian Phillips curve model

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

(1.24)

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

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

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, $\varphi$ is the vector of $4n$ unknown VAR parameters, and $\varepsilon_t$ are VAR innovations with
$E[\epsilon_{t+1}|I_t] = 0$. For $\epsilon_{\pi}$ and $\epsilon_x$ unit vectors such that $\epsilon_{\pi}'z_t = \pi_t$, $\epsilon_x'z_t = x_t$ and $\theta = (\nu, \rho)$, define the 2n-dimensional distance function $f(\varphi, \theta)$ as

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

MM show that the NKPC model (3.14) implies that the true parameter values $\varphi$ and $\theta$ satisfy $f(\varphi, \theta) = 0$, and propose testing $H_0 : \theta = \theta_0$ using an identification-robust minimum distance approach.

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

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

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

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

To model weak identification in this context MM assume that $\frac{\partial}{\partial \varphi'} f(\varphi_T, \theta) = \frac{1}{\sqrt{T}} C$ for a fixed matrix $C$, with the result that $\sqrt{T} \frac{\partial}{\partial \varphi'} f(\varphi_T, \theta)$ 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 $\frac{\partial}{\partial \varphi'} f(\varphi_T, \theta)$ 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, $\Delta$-method-based estimator for $\Omega_{ff} = \lim_{T \to \infty} T \cdot Var \left( f(\hat{\varphi}, \theta_0)' \right)$.  

83
We can then define
\[ g_T(\theta) = \sqrt{T} \hat{\Delta}_{gT}^{1/2} \hat{f}(\hat{\varphi}, \theta) \]
\[ \Delta g_T(\theta) = \hat{\Delta}_{gT}^{1/2} \frac{\partial}{\partial \theta} \hat{f}(\hat{\varphi}, \theta) A_T. \]

for \( A_T \) a sequence of full-rank normalizing matrices which may depend on the sequence of true VAR parameters \( \varphi_T \). Under sequences of true parameter values \( \theta_T \) such that \( g_T(\theta_0) \) converges in distribution, corresponding to local alternatives for strongly identified parameters and fixed alternatives for weakly identified ones, arguments discussed in section E.3 below yield the weak convergence
\[
\left( \begin{array}{c} g_T(\theta_0) \\ \Delta g_T(\theta_0) \end{array} \right) \Rightarrow_d \left( \begin{array}{c} g \\ \Delta g \end{array} \right) \sim N \left( \left( \begin{array}{c} m \\ \mu \end{array} \right), \left( \begin{array}{cc} I & \Sigma_{g\theta} \\ \Sigma_{g\theta} & \Sigma_{\theta \theta} \end{array} \right) \right). \tag{1.25}
\]

where \( \Delta g \) is full rank almost surely, \( m \in \mathcal{M}(\mu, \gamma) \) for \( \mathcal{M}(\mu, \gamma) \) appropriately defined, \( \Sigma \) is consistently estimable, and details on all terms may be found below. Hence, this model falls into the class considered in the paper. While \( \Delta g_T \) 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 \( (\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\varphi}) \) \( h \left( g_T, \Delta g_T, \tilde{\varphi}; A_T^{-1} \right) \). For \( \tilde{\varphi} \) as defined below the statistics \( S_T \) and \( K_T \) based on \( (\tilde{g}_T, \Delta \tilde{g}_T, \tilde{\varphi}) \) are equivalent to the \( MD - AR \) and \( MD - K \) statistics discussed in MM.

### E.2 Coverage Simulations

After assuming that \((\pi_t, x_t)\) follows a VAR(3), MM apply their approach to create confidence sets for the parameter \( \theta \) 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
\[ \phi_{S_T} = 1 \left\{ S_T > \chi_6^2 \right\} \]
and
\[ \phi_{JK_T} = \max \left\{ 1 \left\{ K_T > \chi_{2,1-\alpha_K}^2 \right\}, 1 \left\{ J_T > \chi_{3,1-\alpha_J}^2 \right\} \right\}, \]

84
Table 1.5: Size of nominal 5% tests in NKPC simulation example based on 10,000 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

<table>
<thead>
<tr>
<th></th>
<th>PI</th>
<th>JK</th>
<th>S</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>9.34%</td>
<td>10.52%</td>
<td>12.28%</td>
<td>8.74%</td>
</tr>
</tbody>
</table>

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

\[
\phi_{PI_t} = 1 \left\{ (1 - a_{PI} (D_T, \hat{\gamma})) \cdot K_T + a_{PI} (D_T, \hat{\gamma}) \cdot S_T > c_\alpha(a_{PI}(D_T, \hat{\gamma})) \right\}
\]

where as before

\[
a_{PI}(D, \gamma) = \arg \min_{\alpha \in [0,1]} \sup_{m \in M_D(\hat{\mu}_D, \gamma)} (\beta_m^\alpha - E_m(\hat{\mu}_D, \gamma) | \phi_0))
\]

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 \( \epsilon_t \) (see below for details). We calculate the true size of nominal 5% tests, based on 10,000 simulations, and report the results in Table 1.5. We find that all the tests over-reject, which is unsurprising given the non-linearity of the model together with the small sample size, but that only the JK and S tests has true size exceeding 10%.

Next, we simulate false coverage probabilities for confidence sets formed by inverting these tests. In particular we calculate the rejection rates for PI, JK, S, and K tests of hypotheses \( H_0 : \theta = \theta_0 \) for \( \theta_0 \) not equal to the true parameter value. Table 1.6

---

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

16 We focus on calculating false coverage probabilities rather than power because there are many
Table 1.6: Maximal point-wise differences in false coverage probability of nominal 5% tests in NKPC example. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test 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.

<table>
<thead>
<tr>
<th></th>
<th>PI</th>
<th>JK</th>
<th>S</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>*</td>
<td>3.0%</td>
<td>4.2%</td>
<td>1.0%</td>
</tr>
<tr>
<td>JK</td>
<td>6.4%</td>
<td>*</td>
<td>2.0%</td>
<td>6.0%</td>
</tr>
<tr>
<td>S</td>
<td>31.2%</td>
<td>26.6%</td>
<td>*</td>
<td>30.0%</td>
</tr>
<tr>
<td>K</td>
<td>17.6%</td>
<td>17.6%</td>
<td>17.4%</td>
<td>*</td>
</tr>
</tbody>
</table>

For example, the second entry of the first row of Table 1.6 reports

$$\sup_{\theta_0 \in (0,1)^2} E_{\tilde{\theta}} [\phi_{JK,\theta_0} - \phi_{PI,\theta_0}]$$

where \(\phi_{PI,\theta_0}\) and \(\phi_{JK,\theta_0}\) 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 1.7 reports results analogous

---

reduced-form parameter values \(\phi\) compatible with a given structural parameter value \(\theta^*\), and the power of tests of \(H_0 : \theta = \theta_0\) against \(\theta^*\) will generally depend on \(\phi\). Hence, to simulate the power function we must either adopt some rule to pick \(\phi\) based on \(\theta^*\) or calculate power on a 12-dimensional space, whereas to calculate false coverage probabilities it suffices to consider a 2-dimensional space of values \(\theta\).

17For computational reasons, our simulations use a discretized version of this parameter space—see below.
Table 1.7: Maximal point-wise differences in false coverage probability of size corrected 5% tests in NKPC example. The entry in row i, column j lists the maximum extent to which the rejection probability of test i falls short of the rejection probability of test j. For example, the largest margin by which the simulated rejection probability of the PI test 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.

<table>
<thead>
<tr>
<th></th>
<th>PI</th>
<th>JK</th>
<th>S</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>*</td>
<td>1.6%</td>
<td>2.0%</td>
<td>8.2%</td>
</tr>
<tr>
<td>JK</td>
<td>11.4%</td>
<td>*</td>
<td>1.4%</td>
<td>16.4%</td>
</tr>
<tr>
<td>S</td>
<td>42.0%</td>
<td>33.2%</td>
<td>*</td>
<td>46.0%</td>
</tr>
<tr>
<td>K</td>
<td>20.4%</td>
<td>21.6%</td>
<td>21.8%</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 1.8: Expected area of 95% confidence sets formed by inverting tests in NKPC example, based on 500 simulations. PI is plug-in test, while JK, S, and K are MD-KJ, MD-AR, and MD-K tests of Magnusson and Mavroeidis (2010), respectively.

As to those of Table 1.6 based on (infeasible) size-corrected versions of all four tests. As in Table 1.6, we can see that the PI test offers the best performance, followed by the JK test.\(^{18}\)

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 1.8. As we can see, using size-corrected tests increases the area of all confidence sets. In each case the PI test produces confidence sets with the smallest expected area, while the S test yields confidence sets with the largest expected area.

\(^{18}\)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 \(H_0 : \theta = \bar{\theta}\). To size-correct the PI test we consider

\[
\phi_{PI_t} = 1 \{ (1 - a_{PI}(DT, \bar{\gamma})) \cdot KT + a_{PI}(DT, \bar{\gamma}) \cdot ST \cdot c_0(a_{PI}(DT, \bar{\gamma})) > c^* \}
\]

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

\[
\phi_{JK_t} = 1 \{ \max \{ KT - \chi^2_{2,1, \alpha}, J_T - \chi^2_{4,1, \alpha} \} > c^* \}
\]

for \(c^*\) chosen to ensure correct size for testing \(H_0 : \theta = \bar{\theta}\). Note that if we instead take \(c^* = 0\), these coincide with the non-size-corrected PI and JK tests.
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. These results confirm that the PI test outperforms the other tests considered.

E.3 Details of NKPC Example

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

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{ff} & \hat{\Omega}_{f\theta} \\ \hat{\Omega}_{f\theta} & \hat{\Omega}_{\theta\theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) \hat{\Sigma}_{\varphi \varphi} \frac{\partial}{\partial \varphi} f(\hat{\varphi}, \theta_0)' \\ \frac{\partial}{\partial \varphi'} \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) A_T / \sqrt{T} \right) \hat{\Sigma}_{\varphi \varphi} \frac{\partial}{\partial \varphi} f(\hat{\varphi}, \theta_0)' \\ \frac{\partial}{\partial \theta^2} f(\hat{\varphi}, \theta_0) \hat{\Sigma}_{\varphi \varphi} \frac{\partial}{\partial \varphi'} \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) A_T / \sqrt{T} \right)' \\ \frac{\partial}{\partial \varphi'} \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) A_T / \sqrt{T} \right) \hat{\Sigma}_{\varphi \varphi} \frac{\partial}{\partial \varphi} f(\hat{\varphi}, \theta_0)' \end{pmatrix}$$

and note that given our assumptions this will be consistent for

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

To derive the weak convergence (1.25), as well as the form of the matrices $A_T$, note that since we have assumed $\varphi_T \to \varphi$ and $\sqrt{T} (\hat{\varphi} - \varphi_T) \to_d N(0, \Sigma_{\varphi \varphi})$, the $\Delta$-method (Theorem 3.1 in Van der Vaart (2000)) yields that

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

\footnote{Note that there is no guarantee that confidence sets formed by inverting the PI test will be non-empty.}
\[
\sqrt{T} \cdot \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) - \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right) \rightarrow_d \\
\frac{\partial}{\partial \varphi} \left( \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) \right) \right) \Sigma_{\varphi} \frac{\partial}{\partial \varphi} \left( \text{vec} \left( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0) \right) \right)'.
\]

We can see that the assumed convergence of \( g_T(\theta_0) = \sqrt{T} \hat{\Omega}_{\theta \theta}^{-\frac{1}{2}} f(\hat{\varphi}, \theta_0) \) 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 \( A_T(\theta_0) \), which may in turn depend on the sequence of true VAR parameters \( \varphi_T \). 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 \( \nu \) and the other in weak identification for \( \rho \).

**E.3.1 Possible Sources of Weak Identification**

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

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

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

Fix a true parameter value \( \theta_0 \). Identification of \( \nu \) fails if \( \varphi_{2i} = 0 \) for all \( i \in \{1, \ldots, 6\} \). In this case we have that \( A(\varphi)'e_x = 0 \), with the consequence that \( \nu \) does not enter the distance function \( f(\varphi, \theta) \) and \( \frac{\partial}{\partial \nu} f(\varphi, \theta) = 0 \). To model \( \nu \) as weakly identified, fix \( \varphi_{1i,T} = \varphi_{1i} \) for \( i \in \{1, \ldots, 6\} \) at values such that \( f(\varphi_T, \theta) = 0 \) when \( \varphi_{2i} = 0 \) for \( i \in \{1, \ldots, 6\} \). We can take sequences of true VAR parameter values \( \varphi_T \) such that \( \varphi_{1i,T} = \frac{1}{\sqrt{T}} c_{1i} + o \left( \frac{1}{\sqrt{T}} \right) \), \( \varphi_{2i,T} = \frac{1}{\sqrt{T}} c_{2i} + o \left( \frac{1}{\sqrt{T}} \right) \) and \( f(\varphi_T, \theta) = 0 \) \( \forall T \), which will imply that \( \sqrt{T} \frac{\partial}{\partial \nu} f(\varphi_T, \theta) \rightarrow C_\nu \) for a 6 \( \times \) 1 vector \( C_\nu \). Thus, if we take \( A_T = \begin{bmatrix} \sqrt{T} & 0 \\ 0 & 1 \end{bmatrix} \) we will have that the first column of \( \frac{\partial}{\partial \theta} f(\hat{\varphi}, \theta_0)A_T \) converges in distribution to a non-
degenerate random vector. Provided the values $\varphi_{1i,T}$ are such that $\frac{\partial}{\partial \rho} f(\varphi_T, \theta) \to C_\rho$
for a non-zero vector $C_\rho$, then $\Omega^{-1/2}_{gg} \frac{\partial}{\partial \rho} f(\hat{\varphi}, \theta) A_T \to_d \Delta g$ for a matrix $\Delta g$ which is full rank almost surely, as we assumed.

The parameter $\rho$ may also be weakly identified. In particular, note that

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

so if

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

then $\frac{\partial}{\partial \rho} f(\varphi, \theta) = 0$ for all values of $\rho$, so $\rho$ is unidentified. In the same manner as above, for any pair $(\varphi, \nu)$ satisfying this restriction we can take $\nu$ fixed and construct a sequence $\varphi_T$ converging to $\varphi$ at a $\sqrt{T}$ rate such that $\Omega^{-1/2}_{gg} \frac{\partial}{\partial \rho} f(\hat{\varphi}, \theta) A_T \to_d \Delta g$ for $\Delta g$ full rank almost-surely with $A_T = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{T} \end{bmatrix}$.

E.3.2 Derivation of the Limit Problem

To derive the form of the limit problem (1.25) we need to understand the behavior of $g_T$ and $\Delta g_T$ under alternatives. Note that for alternative $\theta_T$ and true reduced-form parameter value $\varphi_T$, 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

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

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

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 $M(\mu)$, which characterizes the behavior of $m$ under various alternatives, note that
\[
\frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) = \left[ A(\varphi_T)' \left( \frac{1 - \nu_0^2}{\nu_0^2 (1 + \rho_0)} \right) e_x \right] A(\varphi_T)' \left\{ \left[ \frac{1 - \nu_0^2}{\nu_0^2 (1 + \rho_0)} A(\varphi_T)' \right] e_x + \frac{(1 - \nu_0)^2}{\nu_0 (1 + \rho_0)} e_x \right\} - \frac{1}{(1 + \rho_0)^2} e_x
\]

and hence
\[
A(\varphi_T) e_x = \left( \frac{1 - \nu_0^2}{\nu_0^2 (1 + \rho_0)} \right)^{-1} \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_1 = h_1(\theta_0) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_1
\]
for \( h_1(\theta_0) = \left( \frac{1 - \nu_0^2}{\nu_0^2 (1 + \rho_0)} \right)^{-1} \), and
\[
A(\varphi_T)' A(\varphi_T)' e_x = (1 + \rho_0)^3 \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_2 - \frac{(1 - \nu_0)^2}{\nu_0} A(\varphi_T)' e_x + e_x = h_2(\theta_0) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_2 - h_3(\theta_0) h_1(\theta_0) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_1 + e_x,
\]
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
\[
m(\varphi_T, \theta_T) = \left( \frac{1}{1 + \rho_T} - \frac{1}{1 + \rho_0} \right) \left( h_2(\theta_0) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_2 - h_3(\theta_0) h_1(\theta_0) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_1 + e_x \right) + \left( \frac{(1 - \nu_T)^2}{\nu_T (1 + \rho_T)} - \frac{(1 - \nu_0)^2}{\nu_0 (1 + \rho_0)} \right) e_x = h_4(\theta_0, \theta_T) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_1 + h_5(\theta_0, \theta_T) \left[ \frac{\partial}{\partial \theta} f(\varphi_T, \theta_0) \right]_2
\]
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 \( \theta \) 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 \( \mathcal{M}_T(\mu_T) = \{ \sqrt{T} m(\varphi_T, \theta) : \theta \in \Theta \} \).

For a given convergent sequence \( \varphi_T \), 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 \( \frac{\partial}{\partial \theta} f(\varphi, \theta_0) \) 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.3.3 NKPC Simulation Details

The assumption that \((\pi_t, x_t)\) follows a 3rd order VAR means that, once we make a distributional assumption on the driving shocks \( \varepsilon_t \), we can simulate data from the NKPC model discussed above for any combination of parameters \((\varphi, \theta)\) 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 \((\hat{\varphi}, \hat{\theta})\) solving

\[
(\hat{\varphi}, \hat{\theta}) = \arg \min_{(\varphi, \theta) : f(\varphi, \theta) = 0} (\hat{\varphi} - \varphi)' \Sigma_{\varphi \varphi}^{-1} (\hat{\varphi} - \varphi).
\]

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

\[
\hat{\Sigma}_\varepsilon = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{\varepsilon}_t - \frac{1}{T} \sum_{s=1}^{T} \hat{\varepsilon}_s \right) \left( \hat{\varepsilon}_t - \frac{1}{T} \sum_{s=1}^{T} \hat{\varepsilon}_s \right)'.
\]

Taking \( \varepsilon_t \) to be normally distributed, to conduct our simulations we then generate samples of 100 observations from the the model with true parameter values \((\hat{\varphi}, \hat{\theta})\)
and true covariance matrix $\hat{\Gamma}_t$ for $\epsilon_t$.

For computational purposes, when calculating PI tests and simulating coverage probabilities we discretize the parameter space, considering grids of values in both $\nu$ and $\rho$. For both parameters we consider grids ranging from 0.005 to 0.995, with grid points spaced 0.03 apart.
Chapter 2

Two Step Confidence Sets, and the Trouble with the First Stage F-Statistic

In contexts where weak identification is a concern, empirical researchers in economics frequently begin by calculating some statistic intended to measure identification strength. If this preliminary step indicates that identification is not “too” weak, researchers then proceed as usual and calculate non-robust confidence sets, while if weak identification seems to be an issue researchers may calculate identification-robust confidence sets, look for a different specification, or simply decline to report results. Pretests based on the first stage F-statistic in linear instrumental variables (IV) models are the most common example of this practice. Specifically, researchers often calculate the F-statistic for the hypothesis that the model is unidentified and report two-stage least squares confidence sets provided this statistic exceeds the Staiger and Stock (1997) “rule of thumb” cutoff of 10.

We can think of these procedures as constructing a confidence set in two steps, where the first step assesses identification strength and the second step chooses what confidence set to report based on this initial assessment. Unfortunately, unless carefully constructed two-step confidence sets of this sort can exhibit large coverage distortions, in the sense that they may cover the true parameter value with probability
substantially smaller than we intend. Early work by Hall et al. (1996) highlights this issue in the weak IV context. In this paper we present two main results. First, we show numerically that pretests based on the first stage F-statistic fail to control coverage distortions in linear IV with heteroskedastic errors, even when we use heteroskedasticity-robust forms of the F-statistic. Second, we propose a novel approach to detecting weak identification which yields two-step confidence sets with bounded coverage distortions under mild conditions. The idea behind our approach is intuitive: if identification is strong, non-robust confidence sets should closely resemble appropriately-chosen robust confidence sets. Thus, to assess whether the usual asymptotic approximations are reliable, we can simply compare robust and non-robust confidence sets. This approach is no harder computationally than calculating a robust confidence set and indicates strong identification with probability tending to one under the usual strong-identification asymptotics. Our results apply to general non-linear GMM models, and so allow us to reliably assess identification strength in a wide range of settings. Our approach performs well in linear IV simulations, and produces novel results when applied to two empirical examples.

To formally describe two-step confidence sets, let us denote the first-step diagnostic for identification strength, which following D. Andrews and Cheng (2012) we will call an identification category selection (ICS) statistic, by $\phi_{ICS}$. We assume that $\phi_{ICS} \in \{0, 1\}$ where $\phi_{ICS} = 0$ is interpreted as evidence for strong identification while $\phi_{ICS} = 1$ is interpreted as evidence for weak identification. The rule of thumb for the first stage F-statistic, for example, takes $\phi_{ICS}$ to be an indicator function for $F < 10$. In the second step we use a non-robust confidence set if $\phi_{ICS} = 0$ and a robust confidence set if $\phi_{ICS} = 1$. Denoting the robust and non-robust confidence sets by $CS_R$ and $CS_{NR}$ respectively, this results in the two-step confidence set $CS_{2S}$

$$CS_{2S} = \begin{cases} CS_{NR} & \text{if } \phi_{ICS} = 0 \\ CS_R & \text{if } \phi_{ICS} = 1 \end{cases} \quad (2.1)$$

As noted above, rather than calculating a robust confidence set $CS_R$ when $\phi_{ICS} = 1$,
in practice many researchers instead search for a new specification or simply do not report their results. These practices can lead to extremely poor properties for reported confidence sets, so here we focus on “complete” procedures which report a robust confidence set if $\phi_{ICS} = 1$.

If we have an ICS statistic which can perfectly detect weak identification, setting $\phi_{ICS} = 1$ if and only if the coverage of $CS_{NR}$ is below $1 - \alpha$, then if $CS_R$ has coverage at least $1 - \alpha$ we can see that $CS_{2S}$ will have coverage at least $1 - \alpha$ as well. In practice, however, we must estimate the strength of identification from the data, so our ICS procedures must be chosen carefully to ensure reasonable performance for two-step procedures. Absent a perfect ICS procedure we will typically pay a price for pretesting, often in the form of lower coverage for $CS_{2S}$ than for $CS_R$ at some parameter values. Thus our focus will not be on whether a nominal level $1 - \alpha$ confidence set has exactly correct coverage but rather on the extent to which it controls coverage distortions in the sense of having coverage at least $1 - \alpha - \gamma$ for some known $\gamma$.

Pretests based on the first-stage F-statistic in linear IV are the most common ICS procedure. The well-known Staiger and Stock (1997) rule of thumb for linear IV models with a single endogenous regressor declares instruments weak if the first stage F-statistic falls below 10. As discussed by Stock et al. (2002) this cutoff controls the bias of two stage least squares in linear IV with homoskedastic errors, but unfortunately gives no assurance of controlling coverage distortions for two-step confidence sets, as we illustrate numerically in the next section. Stock and Yogo (2005) (henceforth SY) show that it is possible to construct cutoffs for the first stage F-statistic which control coverage distortions for two-step confidence sets, but these cutoffs will depend on both the estimator used and the number of instruments and, moreover, are derived under the assumption that the data are independent and homoskedastic.

Our first main result, demonstrated numerically in the next section, is that if we allow heteroskedasticity the cutoffs developed by SY no longer control coverage distortions for two-step confidence sets, even when we use heteroskedasticity-robust forms of the F-statistic. Hence, even in the well-studied case of linear IV commonly-
used pretesting procedures fail to control coverage distortions for two-step confidence sets. This is to our knowledge the first demonstration that pretests based on the heteroskedasticity-robust first-stage F-statistic fail to control coverage distortions in heteroskedastic linear IV: while recent work by Bun and de Haan (2010) and Olea and Pfeuger (2013) shows that the first stage F-statistic does not give a reliable guide to the bias of two stage least squares in linear IV with non-homoskedastic errors, and Olea and Pfeuger (2013) provide an alternative approach to gauge bias in IV models with general covariance structures, neither of these papers addresses questions of testing or confidence set construction.

Our second main result is a procedure for calculating two-step confidence sets in general GMM models which controls coverage distortions under mild assumptions. Our approach is based on the observation that for the commonly-used level $1 - \alpha$ non-robust confidence sets, for any $\gamma > 0$ we can construct a level $1 - \alpha - \gamma$ robust confidence set which will be contained in the non-robust confidence set with probability tending to one under conventional strong-identification asymptotics. Thus, taking $\phi_{ICS}$ to be an indicator for the event that the robust confidence set is not contained in the non-robust one, $\phi_{ICS} \rightarrow_p 0$ under conventional asymptotics. Provided we use an appropriate robust confidence set when $\phi_{ICS} = 1$, however, our two-step confidence set will always contain our level $1 - \alpha - \gamma$ robust confidence set, and thus has coverage at least $1 - \alpha - \gamma$ regardless of identification strength. Thus under strong identification two-step confidence sets based on our approach coincide with conventional non-robust confidence sets, while under weak identification our two-step confidence sets will control coverage up to an arbitrarily small user-specified maximal distortion $\gamma$.

To illustrate the utility of our results, we consider two empirical examples. First, we revisit the well-studied Angrist and Krueger (1991) analysis of the relationship between earnings and years of schooling. We find that, like pretests based on the first-stage F-statistic, the approach developed in this paper suggests that the usual confidence sets based on two stage least squares and continuous updating GMM are reliable in this application when the number of instruments is small, but become
much less so when we use large instrument sets. Unlike available results based on the
first-stage F-statistic, however, the approach of this paper also finds that the many-
weak-instrument-robust Wald confidence sets of Newey and Windmeijer (2009) and
Hausman et al. (2012) appear fairly reliable in this application, even when the number
of instruments is large. Our results also confirm the previous finding by Cruz and
Moreira (2005) and Hansen et al. (2008) that while two-stage least squares confidence
sets may be unreliable in the Angrist and Krueger (1991) context, some identification-
robust confidence sets deliver substantively similar conclusions.

Second, we examine recent panel IV results from Acemoglu et al. (2008) and
Cervellati et al. (2014) on the relationship between income and democracy. Ace-
emoglu et al. (2008) argue that once one controls for time-invariant country effects,
panel regressions based on international data find no statistically significant relation-
ship between per-capita income and various measures of democracy. Cervellati et al.
(2014) argue that this null effect masks considerable heterogeneity across different
groups of countries, and find a negative relationship between income and demo-
cracy among former European colonies. Using an instrumental variables specification
studied in both papers we find that despite large first-stage F-statistics the usual
strong-identification approximations are not reliable in this context. Indeed, for all
specifications considered robust confidence sets contain all economically plausible pa-
rameter values. 

Why Study Two-Step Confidence Sets? Two-step confidence sets can be unap-
ppealing from a theoretical perspective. As already noted, unless carefully constructed
tests and confidence sets that rely on pretests may suffer from substantial distortions.
Even if we choose our pretests to control coverage or size, in contexts where efficient
tests are available pretesting can only reduce power relative to an efficient procedure.
In linear IV with a single endogenous regressor and homoskedastic normal errors, for
example, Andrews et al. (2006) show that the conditional likelihood ratio test of Mor-
eira (2003) is efficient under strong identification and nearly uniformly most powerful

\footnote{The IV approach we study here is only one of many specifications studied in these papers, so this finding does not invalidate their broader conclusions.}
under weak identification, so from the perspective of power there is little to be gained from pretesting in that context. Provided we have robust tests or confidence sets that perform well under both weak and strong identification there is a compelling theoretical case for using these procedures.

Despite this theoretical point, the operational properties of two-step confidence sets are nonetheless of considerable practical interest. Empirical researchers rarely use identification-robust procedures absent prior evidence of weak identification in a given context. Of the 38 empirical papers calculating robust tests or confidence sets based on Moreira (2003) which themselves have over 20 citations, for example, 35 report the first stage F-statistic, 33 have a first stage F smaller than 15, and 29 have a first stage F smaller than 10.² There are a number of potential reasons for the prevalence of pretesting, ranging from the ubiquity of the “point estimate, standard error” format for reporting empirical results to the sometimes counter-intuitive forms taken by robust confidence sets.³

Whatever the reason, this reality of current empirical practice highlights the importance of understanding the properties of two-step confidence sets. Moreover the ubiquity of two-step confidence sets in makes it essential to have pretesting procedures that control coverage distortions. One encouraging implication of our results is that it is straightforward to construct two-step confidence sets which have arbitrarily small coverage distortions under weak identification and thus that, as far as coverage is concerned, the cost of pretesting for identification strength can be made quite small.

**Alternative Pretests for Weak Identification** In addition to those already mentioned, a number of additional papers suggest evaluating instrument strength based on the first-stage F-statistic or, relatedly, the scaled first-stage $R^2$. Examples include

²Based on a Google Scholar search on 3/4/2014 and the author’s tabulation. For papers which report results based on multiple specifications, here we consider the smallest first stage F across specifications.
³Mikusheva (2010) notes, for example, that confidence sets formed by inverting the conditional likelihood ratio test may take the three forms: an interval $(a, b)$, the whole line $(-\infty, \infty)$, or the whole line excluding an interval $(-\infty, a) \cup (b, \infty)$. 

Another group of procedures, including those of Wright (2003) and Arellano et al. (2012), resembles the first stage F-statistic in testing the null hypothesis of identification failure, but consider other test statistics. Since these procedures test the null of under-identification rather than weak identification, if we use them to construct two-step confidence sets their power may be too high. Specifically, they may frequently reject under-identification, and so set $\phi_{ICS} = 0$, even in contexts where non-robust confidence sets are highly unreliable. SY show that one can overcome this problem in linear IV with homoskedastic errors by increasing the critical values for the first stage F-statistic but whether a similar adjustment is possible in general models, let alone what appropriate critical values might be, remains an open question. Recent work by Antoine and Renault (2012) takes a related approach, testing the null hypothesis of "nearly weak" identification (in a sense that they define) against the alternative that the model is well-identified, but also gives no results bounding coverage distortions for two-step confidence sets when identification may be weak.

Other procedures, including those of Hahn and Hausman (2002), Inoue and Rossi (2011), and Wright (2010), takes the opposite approach and tests the null hypothesis of strong identification. For these pretests the concern is that their power may be too low and thus that they may fail to reject strong identification even when non-robust tests are unreliable. In particular, we are unaware of any results guaranteeing that the power of these tests suffices to control coverage distortions for two-step confidence sets under weak identification. To the contrary, Hausman et al. (2005) document that conventional non-robust confidence sets may exhibit large distortions even when the Hahn-Hausman test has only a low probability of detecting weak instruments, implying that if we base $\phi_{ICS}$ on the Hahn-Hausman test for weak instruments the resulting two-step confidence set will in some cases exhibit large coverage distortions.

\footnote{Like the present paper, Wright (2010) compares robust and non-robust confidence sets. Wright considers different robust confidence sets than we do, however, and his approach does not guarantee bounded coverage distortions for two-step confidence sets.}
Finally, there are some diagnostics for identification strength proposed in the literature that, together with appropriately chosen robust and non-robust confidence sets, yield two-step confidence sets that control coverage in particular contexts. Specifically, one can view the critical value selection procedures of D. Andrews and Cheng (2012), the pretest-based procedures of Andrews and Mikusheva (2012), and the "switching" procedures of Elliott et al. (2012) as particular two-step procedures where the ICS statistics, together with the robust and non-robust tests and confidence sets, are chosen in such a way that the procedure as a whole controls coverage. Unfortunately, however, these procedures are limited in their range of application and, in the case of Andrews and Cheng (2012) and Elliott et al. (2012), may be prohibitively computationally demanding for many cases of economic interest.

In the next section we show numerically that pretests based on the first stage F-statistic and known critical values fail to control coverage distortions for two-step confidence sets. Section 2.2 introduces a robust two-step confidence set. This section focuses on the definition and implementation of our robust procedures, as well as some intuition for their properties, while formal derivation of their asymptotic properties under weak and strong identification asymptotics is deferred to Sections 2.5 and 2.6. Section 2.3 demonstrates via simulation that our approach controls coverage distortions in linear IV with heteroskedasticity and is competitive with F-statistic-based approaches under homoskedasticity. Section 2.4 applies our results to the Angrist and Krueger (1991) and Acemoglu et al. (2008) and Cervellati et al. (2014) datasets. Section 2.5 shows that our approach can be used to construct robust two-step confidence sets, including confidence sets with uniformly bounded coverage, whenever an appropriate robust confidence set is available, and Section 2.6 shows that our approach is valid in general nonlinear GMM models. All proofs may be found in the Supplementary Appendix.
2.1 The Trouble with the First Stage F-Statistic

In this section we demonstrate numerically that two-step confidence sets based on the first stage F-statistic with SY’s cutoffs fail to control coverage distortions in linear IV with heteroskedastic errors. To frame our discussion, we focus on the linear IV model with a single endogenous regressor, where we assume that either there are no additional exogenous regressors or that any such regressors have already been partialled out.\(^5\) The model, written in reduced form, is

\[
Y = Z_0 \beta_0 + V_1 \\
X = Z_0 + V_2
\]

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, where we assume that \(E[V_{1,t}Z_t] = E[V_{2,t}Z_t] = 0\) for \(Z_t\) the transpose of row \(t\) of \(Z\).

We are interested in constructing confidence sets for the scalar coefficient \(\beta\), treating the \(k \times 1\) vector of first-stage parameters \(\pi\) as nuisance parameters. A common nominal level \(1 - \alpha\) confidence set in empirical practice is the two stage least squares (2SLS) Wald confidence set

\[
CS_{2SLS} = \left[ \hat{\beta}_{2SLS} - c_{\alpha/2} \frac{\hat{\sigma}_{2SLS}}{\sqrt{T}}, \hat{\beta}_{2SLS} + c_{\alpha/2} \frac{\hat{\sigma}_{2SLS}}{\sqrt{T}} \right]
\]

(2.2)

where \(c_{\alpha/2}\) is the \(1 - \alpha/2\) quantile of a standard normal distribution, \(\hat{\beta}_{2SLS}\) is the 2SLS estimator for \(\beta_0\), and \(\hat{\sigma}_{2SLS}\) is an estimator of the standard deviation of \(\sqrt{T}\hat{\beta}_{2SLS}\). We can likewise construct Wald confidence sets based on other estimators \(\hat{\beta}\), for example limited information maximum likelihood (LIML) or, in the heteroskedastic case, efficient two-step GMM (2SGMM) or continuous updating GMM (CUGMM). Confidence sets based on all these estimators may exhibit large coverage distortions when \(\pi\) is small.

The common approach to assessing whether \(\pi\) is “too” small is based on the first-

\(^5\)That is, for exogenous controls \(W\) and initial data \((\tilde{Y}, \tilde{X}, \tilde{Z}, W)\) \(Y = M_W \tilde{Y}, X = M_W \tilde{X}, Z = M_W \tilde{Z}\), where \(M_W = I - W (W'W)^{-1} W'\).
stage F-statistic for testing the hypothesis that $\pi = 0$. If we let $\hat{\pi}$ be the OLS estimate of $\pi_0$, $\hat{\pi} = (Z'Z)^{-1} Z'X$, the first stage F-statistic is

$$F = \frac{T}{k} \frac{\hat{\pi}^T \hat{\Sigma}_\pi^{-1} \hat{\pi}}$$

for $\hat{\Sigma}_\pi$ an estimator for the variance of $\sqrt{T} (\hat{\pi} - \pi_0)$. The conventional first-stage F-statistic, which assumes the errors are homoskedastic, uses

$$\hat{\Sigma}_\pi = \left( \frac{1}{T-k} \hat{\Sigma}_2 \right)^2 \frac{1}{T} Z'Z$$

where $\hat{\Sigma}_2 = Y - Z\hat{\pi}$. Unfortunately, this statistic relies critically on the assumption of homoskedasticity, and may behave erratically if $\text{Var} (V_{2t}|Z_t)$ depends on $Z_t$. By contrast the heteroskedasticity-robust F-statistic, based on the White (1980) covariance matrix estimator, uses

$$\hat{\Sigma}_\pi = \frac{1}{T} \sum_i \hat{\Sigma}_2^2 Z_i Z_t - \left( \frac{1}{T} \sum_i \hat{\Sigma}_2 Z_i \right) \left( \frac{1}{T} \sum_i \hat{\Sigma}_2 Z_t \right)$$

The robust and non-robust F-statistics are asymptotically equivalent and distributed $\frac{1}{k} \chi^2_k$ asymptotically when the data are homoskedastic, but under heteroskedasticity the robust F-statistic continues to have a scaled $\chi^2$ distribution asymptotically while the non-robust F-statistic does not. One can likewise define serial correlation and clustering-robust F-statistics by using appropriate robust estimators $\hat{\Sigma}_\pi$. Unsurprisingly, two-step procedures based on the non-robust first stage F-statistic are unreliable when used with heteroskedastic data, so going forward we focus on the heteroskedasticity-robust F-statistic.

To construct a two-step confidence set as in (2.1) we need to define an appropriate robust confidence set $CS_R$. Here we consider confidence sets based on the $S$ statistic of Stock and Wright (2000) (a generalization of Anderson Rubin statistic),

$$S(\beta) = T g_T (\beta)' \hat{\Sigma}_\pi (\beta)^{-1} g_T (\beta)$$  \hspace{1cm} (2.3)
where $g_T(\beta) = \frac{1}{T} \sum Z_t (Y_t - \beta X_t)$ and $\hat{\Sigma}_g(\beta)$ is the usual heteroskedasticity-robust variance estimator for $\sqrt{T} g_T(\beta)$,

$$\hat{\Sigma}_g(\beta) = \frac{1}{T} \sum_t (Y_t - X_t \beta)^2 Z_t Z_t' - g_T(\beta) g_T(\beta)' .$$

The $S$ statistic evaluated at the true parameter value will be approximately $\chi^2_k$ distributed in large samples regardless of identification strength, so the level $1 - \alpha$ confidence set for $\beta$ based on this statistic is

$$CS_S = \{ \beta : S(\beta) < \chi^2_{k,1-\alpha} \}$$

for $\chi^2_{k,1-\alpha}$ the $1 - \alpha$ quantile of a $\chi^2_k$ distribution.

In the rest of this section, we focus on two-step confidence sets $CS_{2S}$ defined as in (2.1), where we take $\phi_{IC} = 1 \{ F < c \}$ for some cutoff $c$, let $CS_{NR}$ be some Wald confidence set, and take $CS_R = CS_S$. We study the coverage of these confidence sets, $Pr \{ \beta_0 \in CS_{2S} \}$, for different choices of Wald confidence set and cutoff. We first highlight that the conventional rule of thumb cutoff $c = 10$ does not control coverage for two-step confidence sets even in homoskedastic models, while the SY cutoffs do. We then provide what is, to our knowledge, the first demonstration in the literature that the SY cutoffs fail to control coverage distortions under heteroskedasticity, even when using the heteroskedasticity-robust F-statistic.

For our simulations we set $\beta_0 = 0$ and assume that $Z_t$ is a collection of dummy variables for different values of a categorical instrument $\tilde{Z}_t \in \{ C, ..., k \}$. We take $k \in \{ 5, 10, 20 \}$ and for each $k$ consider two calibrations: one with a moderate degree of endogeneity (denoted by M) and the other with a very high degree of endogeneity (denoted by H). In each calibration we consider a wide range of values for identification strength (as measured by $\| \pi \|$), where we hold $\pi / \| \pi \|$ fixed) ranging from non-identification to very strong identification, and report the smallest coverage probability for each confidence set over these different values, $\min_{\| \pi \|} Pr_{\pi, \beta_0} \{ \beta_0 \in CS \}$. All simulations are based on samples of 10,000 observations. For further details on our
simulation design see the Supplementary Appendix.

2.1.1 The First Stage F-Statistic Under Homoskedasticity

Before considering the heteroskedastic case, we briefly examine the performance of nominal 95\% two-step confidence sets based on the F-statistic together with different cutoffs $c$ and Wald confidence sets $CS_{NR}$ under homoskedasticity. We begin with the usual LIML and 2SLS confidence sets, which we can view as two-step confidence sets with $c = 0$. Next, we consider rule of thumb confidence sets which take $c = 10$. Finally we use cutoffs based on critical values from SY, specifically $c = 26.87$, $c = 38.54$, and $c = 62.30$ for $k = 5$, $k = 10$, and $k = 20$ respectively when we take $CS_w$ to be the 2SLS confidence set, and $c = 5.44$, $c = 3.68$, and $c = 3.21$ when we take $CS_w$ to be the LIML confidence set. The results of SY imply that in models with homoskedastic errors this choice of cutoffs ensures coverage distortions no larger than 10\%, and so coverage no less than 85\%, for two-step confidence sets with nominal coverage 95\%.

The results of this exercise are reported in Table 2.1. As these make clear, the rule of thumb cutoff of 10 does not ensure any fixed level of coverage for two-step confidence sets: while 2SLS confidence sets based on the rule of thumb have coverage distortions less than 10\% in the M calibrations, they exhibit more substantial distortions in the H calibrations, and the degree of distortion is increasing in the number of instruments $k$. In contrast, two-step confidence sets based on the cutoffs of SY have distortions not exceeding 10\% (and thus coverage not less than 85\%) in all cases, as expected.

2.1.2 The First Stage F-Statistic Under Heteroskedasticity

As already noted, coverage control for two-step procedures based on the first stage F-statistic with cutoffs from SY relies critically on the assumption of homoskedastic-

\footnote{We obtain these cutoffs by taking $\phi_{ICS} = 1$ when the 5\% F-test of SY cannot reject the hypothesis that the nominal 5\% Wald test of interest has true size exceeding 10\%.}

\footnote{Our focus on maximal 10\% distortions is not an endorsement of this choice of $\gamma$, but rather due to the fact that this is the smallest $\gamma$ achievable using the critical values published in SY. One could, of course, use the approach of SY to derive pretests implying smaller values of $\gamma$ in the homoskedastic case.}
Table 2.1: Minimal coverage for nominal level 95% confidence sets in homoskedastic IV simulations with 10,000 observations, based on 10,000 simulations. LIML CS and 2SLS CS are the usual Wald confidence sets based on LIML and 2SLS, while the rule of thumb confidence sets are two-step confidence sets using the rule-of-thumb cutoff $c = 10$ and the robust S (or Anderson-Rubin) confidence set $CS_S$. Finally, the SY confidence sets use the SY cutoffs discussed in the text along with the robust S (or Anderson-Rubin) confidence set $CS_S$, and have asymptotic coverage at least 85% in models with homoskedastic errors.

As in the homoskedastic case neither Wald confidence sets nor two-step confidence sets based on the rule of thumb cutoffs control coverage distortions. Unlike in the homoskedastic case, two-step confidence sets using the SY cutoffs also fail to control coverage distortions, regardless of whether we use efficient or inefficient estimators. More generally, we see that in many cases heteroskedasticity gives rise to far more pronounced coverage shortfalls than we observed under homoskedasticity.

The central problem with two-step confidence sets based on the SY cutoffs under heteroskedasticity is that the first stage F-statistic is no longer a reliable indicator of identification strength, at least when used with conventional cutoffs. While this point

---

8We also considered CUGMM and 2SGMM in the homoskedastic case but, unsurprisingly given the large sample size, their behavior is indistinguishable from that of LIML and 2SLS respectively.
Table 2.2: Minimal coverage for nominal level 95% confidence sets in heteroskedastic IV simulations with 10,000 observations, based on 10,000 simulations. LIML CS, 2SLS CS, CUGMM CS, and 2SGMM CS are the usual Wald confidence sets based on LIML, 2SLS, CUGMM, and 2SGMM, while the rule of thumb confidence sets are two-step confidence sets using the rule-of-thumb cutoff $c = 10$ and the robust S (or Anderson-Rubin) confidence set $CS_S$. Finally, the SY confidence sets use the SY cutoffs discussed in the text together with the robust S confidence set $CS_S$, and have asymptotic coverage at least 85% in models with homoskedastic errors.

has previously been highlighted with regard to 2SLS bias by Bun and de Haan (2010) and Olea and Pfeueger (2013), the issue appears especially stark when considering coverage. In Figure 2-1 we plot the coverage of Wald confidence sets against the mean of the first-stage F-statistic for the model with ten instruments in the medium endogeneity (M) calibration as we vary $\pi$, noting that $E[F]$ is a strictly increasing function of $\pi$. As this figure makes clear, even when the mean of the first stage F-statistic is 500, many nominal 95% Wald confidence sets exhibit coverage distortions exceeding 15%. A still more extreme version of this issue arises in the H calibration, where the 2SLS confidence set has a 15% coverage distortion even when the mean of the first stage F-statistic is 100,000. Given these large distortions, it is unsurprising that two-step confidence sets based on the first-stage F-statistic and known cutoffs

---

9 An important point to note here is that while we see similarly poor behavior for non-robust confidence sets under our other calibrations, the particular ranking across confidence sets (i.e. which Wald confidence set exhibits the largest distortion) varies across calibrations.
fail to generate reliable two-step confidence sets in models with heteroskedastic data.

![Graph](image)

Figure 2-1: Coverage of Wald confidence sets plotted against the mean of the first stage F-statistic as we vary $\|\pi\|$ in heteroskedastic linear IV M calibration with $k = 10$.

2.2 A Simple Two-Step Approach: A User’s Guide

In this section we develop a simple approach to constructing two-step confidence sets which control coverage in general models. Our technique is based on the observation that one can construct a preliminary weak identification-robust confidence set $CS_{R,P}$ which is a strict subset of the conventional non-robust confidence sets $CS_{NR}$ with probability tending to one under the usual strong identification asymptotics. Consequently, if we take $\phi_{ICS}$ to be an indicator for the event that $CS_{R,P} \not\subseteq CS_{NR}$ we have that $\phi_{ICS} \to_p 0$ under strong identification. On the other hand provided our robust confidence set $CS_R$ contains $CS_{R,P}$, $CS_{R,P} \subseteq CS_R$, we can see that $CS_{2S}$ will always contain the preliminary confidence set and thus will have coverage at least equal to that of $CS_{R,P}$. Hence, this approach yields two-step confidence sets which are equivalent to the usual, non-robust confidence set under strong identification while also controlling the maximal coverage distortion under weak identification.
This section focuses on the definition and implementation of our robust procedures, as well as some intuition for their properties. Formal discussion of the asymptotic properties of these procedures under weak- and strong-identification asymptotics is deferred to Sections 2.5 and 2.6 below.

### 2.2.1 GMM Model and Test Statistics

We consider a general GMM setting with \( k \times 1 \)-dimensional moment condition \( g_t (\theta) \) which we assume is mean zero when the \( m \)-dimensional parameter \( \theta \) is equal to its true value \( \theta_0 \). In the linear IV model, for example, \( \theta = \beta \) and \( g_t (\theta) = Z_t (Y_t - X_t \theta) \).

Suppose we are interested in inference on a \( p \)-dimensional parameter \( (p < m) \beta = f (\theta) \) for \( f \) a continuously differentiable function such that \( \frac{\partial}{\partial \theta} f (\theta_0) \) is full-rank. For example, we may be interested in constructing a confidence set for the \( i \)-th element of the structural parameter vector and so take \( f (\theta) = \theta_i \).

Let \( g_T (\theta) = \frac{1}{T} \sum_t g_t (\theta) \) be the sample average of \( g_t (\theta) \), and let \( \hat{\Sigma}_g, \hat{\Sigma}_{gg}, \) and \( \hat{\Sigma}_g \) be consistent estimators for \( \text{Var} \left( \sqrt{T} g_T (\theta) \right), \text{Cov} \left( \sqrt{T} \text{vec} \left( \frac{\partial}{\partial \theta} g_T (\theta) \right), \sqrt{T} g_T (\theta) \right), \) and \( \text{Var} \left( \sqrt{T} \text{vec} \left( \frac{\partial^2}{\partial \theta^2} g_T (\theta) \right) \right) \) respectively, where \( \text{vec} (A) \) denotes the vectorization of the matrix \( A \), achieved by stacking its columns. In models where the data is independent across observations \( t \) we can use the usual heteroskedasticity-robust covariance estimators

\[
\begin{align*}
\hat{\Sigma}_g (\theta) &= \frac{1}{T} \sum_t g_t (\theta) g_t (\theta)' - g_T (\theta) g_T (\theta)' \\
\hat{\Sigma}_{gg} (\theta) &= \frac{1}{T} \sum_t \text{vec} \left( \frac{\partial}{\partial \theta} g_t (\theta) \right) g_t (\theta)' - \text{vec} \left( \frac{\partial}{\partial \theta} g_T (\theta) \right) g_T (\beta)'
\end{align*}
\]

\[
\hat{\Sigma}_g (\theta) = \frac{1}{T} \sum_t \text{vec} \left( \frac{\partial}{\partial \theta} g_t (\theta) \right) \text{vec} \left( \frac{\partial^2}{\partial \theta^2} g_t (\theta) \right)' - \text{vec} \left( \frac{\partial}{\partial \theta} g_T (\theta) \right) \text{vec} \left( \frac{\partial^2}{\partial \theta^2} g_T (\theta) \right)',
\]

(2.5)

but all of our results apply more broadly, allowing for clustering- or serial correlation-robust estimators. Define the \( S \) statistic as in (2.3), noting that (2.4) is simply \( \hat{\Sigma}_g \) as defined in (2.5) specialized to the linear IV model.

Assume we have some estimator \( \hat{\theta} \) for \( \theta \) which under strong identification is first-order equivalent to

\[
\hat{\theta} = \arg \min_{\theta} g_T (\theta)' \hat{\Omega} (\theta) g_T (\theta).
\]

(2.6)
Here $\hat{\Omega}(\theta)$ is a symmetric positive-definite weighting matrix which we will assume converges uniformly in probability to a full-rank matrix-valued function $\Omega(\theta)$ under strong identification. Note that by taking $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$ this definition nests CUGMM as a special case. Since 2SGMM and CUGMM are first-order equivalent under strong identification this means we can also take $\hat{\theta}$ to be the efficient 2SGMM estimator. Inefficient estimators are also allowed by (2.6), for instance the 2SLS estimator (which takes $\hat{\Omega}(\theta) = \left(\frac{1}{T}Z'Z\right)^{-1}$) applied to linear IV models with heteroskedastic errors.

The most common non-robust confidence set for $\theta$ is based on the Wald statistic

$$W(\beta) = T(\hat{\beta} - \beta)'\hat{\Sigma}_{\beta}^{-1}(\hat{\beta} - \beta)$$

(2.7)

where $\hat{\Sigma}_{\beta}$ is an estimator for the variance of $\sqrt{T}\hat{\beta}$, e.g.

$$\hat{\Sigma}_{\beta} = \frac{\partial}{\partial \theta} f(\theta)'\left(G_T(\theta)'\hat{\Omega}(\theta)G_T(\theta)\right)^{-1}G_T(\theta)'\hat{\Omega}(\theta)\hat{\Sigma}_g(\theta)\hat{\Omega}(\theta)G_T(\theta) \times$$

$$\left(G_T(\theta)'\hat{\Omega}(\theta)G_T(\theta)\right)^{-1}\frac{\partial}{\partial \theta} f(\theta)$$

for $G_T(\theta) = \frac{\partial}{\partial \theta} g_T(\theta)$ the Jacobian of $g_T(\theta)$ with respect to $\theta$. Under strong identification standard assumptions yield that $W(\beta_0) \rightarrow_d \chi^2_p$.

### 2.2.2 A Generalized Weak Identification-Robust Test Statistic

To construct two-step confidence sets it will be useful to have a weak identification-robust test statistic which is locally asymptotically equivalent to the Wald statistic (2.7) under strong identification. In this section we give a heuristic derivation such a statistic, whose properties are formally developed in Section 2.6.

A major difficulty with the Wald statistic (2.7) under weak identification is that the estimator $\hat{\theta}$ will typically be inconsistent with an unknown, non-standard distribution. Thus, rather than basing inference on $\hat{\theta}$, robust statistics instead rely on the properties of the moment condition and its derivatives evaluated at a given parameter
value $\theta$, which can easily be studied even when identification is weak. We would like to follow this route and construct a statistic that depends only on the behavior of $g_T(\theta)$ and $G_T(\theta)$, but that is also equivalent to $W(\beta)$ when identification is strong.

Well known results (see e.g. Newey and McFadden (1994), Section 3.4) establish that the estimator $\hat{\theta}$ is asymptotically equivalent a one-step estimator starting at any value in a $\sqrt{T}$ neighborhood of $\theta_0$. Specifically, note that the first order condition for (2.6) implies that

$$
\frac{\partial}{\partial \theta} \left( g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) \right) \bigg|_{\theta=\hat{\theta}} = 0
$$

which, taking a mean-value expansion and collecting lower-order terms, yields that

$$
G_T(\hat{\theta})' \hat{\Omega}(\hat{\theta}) \left( g_T(\theta) + G_T(\theta^*) \left( \hat{\theta} - \theta \right) \right) + o_p \left( \frac{1}{\sqrt{T}} \right) = 0
$$

where $\theta^*$ is a value between $\hat{\theta}$ and $\theta$ that may vary across rows of $G_T(\theta^*)$ and $o_p \left( \frac{1}{\sqrt{T}} \right)$ denotes terms that converge in probability to zero at a rate faster than $\sqrt{T}$ under strong identification. Rearranging and imposing standard strong-identification assumptions (see Assumptions 5 and 6 below) this gives us that for $\theta$ in a $\sqrt{T}$-neighborhood of $\theta_0$,

$$
\hat{\theta} - \theta = - \left( G_T(\theta)' \hat{\Omega}(\theta) G_T(\theta) \right)^{-1} G_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) + o_p \left( \frac{1}{\sqrt{T}} \right)
$$

where the leading term on the right hand side is referred to as a one-step estimator. Note that this term depends only on terms evaluated at $\theta$, and so avoids issues arising from non-standard behavior for the estimator. As noted in Kleibergen (2005), however, under weak identification $G_T(\theta)$ will in general be noisy and correlated with $g_T(\theta)$, which results in intractable behavior for the one-step estimator.

To solve this issue, following Kleibergen (2005) define a modified estimator for the mean of $\frac{\partial}{\partial \theta} g_T(\theta)$,

$$
D_T(\theta) = \left[ \frac{\partial}{\partial \theta_1} g_T(\theta) - \hat{\Sigma}_{\theta g}(\theta) \hat{\Sigma}_g(\theta)^{-1} g_T(\theta), ..., \frac{\partial}{\partial \theta_p} g_T(\theta) - \hat{\Sigma}_{\theta g}(\theta) \hat{\Sigma}_g(\theta)^{-1} g_T(\theta) \right]
$$
where \( \hat{\Sigma}_{\theta}^{(0)}(\theta) \) is the \( k \times k \) block of \( \Sigma_{\theta}^{(0)}(\theta) \) corresponding to \( \theta \). Note that \( \text{vec} \left( D_T^{(0)}(\theta) \right) = \text{vec} \left( \frac{\partial}{\partial \theta} g_T(\theta) \right) - \hat{\Sigma}_{\theta}^{(0)}(\theta) \hat{\Sigma}_{g}^{(0)}(\theta)^{-1} g_T(\theta) \). \( D_T^{(0)}(\theta) \) is asymptotically equivalent to \( G_T(\theta) \) when identification is strong, but is asymptotically independent of \( g_T(\theta) \) when identification is weak. Thus, if we substitute \( D_T^{(0)}(\theta) \) for \( G_T(\theta) \) in the expression for the one-step estimator, the resulting term remains well-behaved even under weak identification. Using the \( \Delta \)-method approximation \( f(\hat{\theta}) - f(\theta) = \frac{\partial}{\partial \theta} f(\theta) (\hat{\theta} - \theta) + o_p \left( \frac{1}{\sqrt{T}} \right) \) we then obtain a modified one-step estimator for \( f(\hat{\theta}) - f(\theta) \):

\[
 f(\hat{\theta}) - f(\theta) = - \frac{\partial}{\partial \theta} f(\theta) \left( D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \right)^{-1} D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) g_T(\theta) + o_p \left( \frac{1}{\sqrt{T}} \right). \tag{2.8}
\]

To construct our robust test statistic, we simply substitute the one step estimator in (2.8) into the expression for the Wald statistic (2.7), evaluate the covariance estimator \( \hat{\Sigma}_{\theta}^{(0)}(\theta) \) at \( \theta \) rather than \( \hat{\theta} \), and replace \( G_T(\theta) \) with \( D_T^{(0)}(\theta) \) in all expressions. This yields the generalized robust statistic

\[
 K_{\Omega f}^{(0)}(\theta) = T g_T(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \left( D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \right)^{-1} \frac{\partial}{\partial \theta} f(\theta)^{\prime} \times \left( \frac{\partial}{\partial \theta} f(\theta) \left( D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \right)^{-1} D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) \times \hat{\Sigma}_{g}^{(0)}(\theta) \right)^{-1} \times \hat{\Omega}(\theta) D_T^{(0)}(\theta) \left( D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \right)^{-1} \frac{\partial}{\partial \theta} f(\theta)^{\prime} \times \left( D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) D_T^{(0)}(\theta) \right)^{-1} D_T^{(0)}(\theta)^{\prime} \hat{\Omega}(\theta) g_T(\theta) \right). \tag{2.9}
\]

As our heuristic derivation highlights, we can think of \( K_{\Omega f}^{(0)}(\theta) \) as a Wald statistic based on a one-step estimator. Since this is one interpretation for \( C(\alpha) \) statistics, \( K_{\Omega f}^{(0)}(\theta) \) is closely related to the GMM \( C(\alpha) \) statistics discussed by Lee (2005). If we take \( \hat{\Omega}(\theta) \) to be the efficient GMM weighting matrix, \( K_{\Omega f}^{(0)}(\theta) \) simplifies to the K statistic of Kleibergen (2005) when we test the full parameter vector, while for \( f(\theta) \) which selects a subvector of \( \theta \) (e.g. the first parameter alone) \( K_{\Omega f}^{(0)} \) is equivalent to the \( LM_{e,f} \) statistic proposed by Chaudhuri and Zivot (2011).

We show in Section 2.6 that when evaluated at the true parameter value \( \theta_0 \), \( K_{\Omega f}^{(0)}(\theta_0) \) converges to a \( \chi^2_\nu \) distribution under both weak and strong identification, while \( S(\theta_0) - K_{\Omega f}^{(0)}(\theta_0) \) converges to a \( \chi^2_{k-p} \) distribution and is asymptotically in-
dependent of $K_{\Omega,f}(\theta_0)$. Critically, under strong identification this statistic will be asymptotically equivalent to $W(f(\theta))$ on $\sqrt{T}$ neighborhoods of the true parameter value $\theta_0$.

### 2.2.3 A Robust Two-Step Confidence Set

Given that $K_{\Omega,f}$ is asymptotically equivalent to a Wald statistic when identification is strong, it is natural to gauge the reliability of Wald confidence sets by comparing them to $K_{\Omega,f}$ confidence sets. Since $K_{\Omega,f}$ and $W(f(\theta))$ are equivalent only local to $\theta_0$, to justify this approach we need to restrict attention to a neighborhood of the true parameter value under strong identification. In well-identified models the $S$ statistic diverges to infinity outside of shrinking neighborhoods of $\theta_0$, so one way to localize the problem is by taking a linear combination of the $S$ and $K_{\Omega,f}$ statistics, similar to Andrews (2013).

In particular, for $a > 0$ define the preliminary robust confidence set as the set of values $\beta$ such that there exists a value $\theta$ with $f(\theta) = \beta$ and $K_{\Omega,f}(\theta) + a \cdot S(\theta) \leq \chi^2_{p,1-\alpha}$,

$$
CS_{R,P} = \left\{ \beta : \min_{\theta : f(\theta) = \beta} (K_{\Omega,f}(\theta) + a \cdot S(\theta)) \leq \chi^2_{p,1-\alpha} \right\}.
$$

For an estimator $\hat{\beta}$ which is first-order equivalent to (2.6) under strong identification, let $W(\beta)$ be the Wald statistic (2.7). Let $CS_{NR}$ be the Wald confidence set

$$
CS_{NR} = \left\{ \beta : W(\beta) \leq \chi^2_{p,1-\alpha} \right\},
$$

noting that for $\hat{\beta}$ 1-dimensional this gives us $t$-statistic confidence sets of the form (2.2). Let $\phi_{ICS}$ be an indicator function for $CS_{R,P} \nsubseteq CS_{NR}$

$$
\phi_{ICS} = 1 \{CS_{R,P} \nsubseteq CS_{NR}\}.
$$

Denote by $H(x; a, k, p)$ the cumulative distribution function for a $(1+a)\chi^2_p + a \cdot \chi^2_{k-p}$ distribution and by $H^{-1}(1-\alpha; a, k, p)$ the $1-\alpha$ quantile of this distribution. Our
level $1 - \alpha$ robust confidence set is

$$CS_R = \left\{ \beta : \min_{\theta; \beta = f(\theta)} (K_{\Omega, f}(\theta) + a \cdot S(\theta)) \leq H^{-1}(1 - \alpha; a, k, p) \right\}. \quad (2.13)$$

We obtain the following as an immediate corollary of Theorems 7-9 below:

**Corollary 4** For $CS_{R,P}$, $CS_{NR}$, $\phi_{ICS}$, and $CS_R$ as defined in (2.10)-(2.13), the two step confidence set (2.1) has the following properties:

1. Under Assumptions 2-4 below, for $\gamma = 1 - \alpha - H \left( \chi^2_{p,1-\alpha}; a, k, p \right)$,

$$\liminf_{T \to \infty} Pr \{ \beta_0 \in CS_{2S} \} \geq H \left( \chi^2_{p,1-\alpha}; a, k, p \right) = 1 - \alpha - \gamma$$

under weak identification.

2. Under strong identification (and Assumptions 2, 5, and 6 below), $\phi_{ICS} \to_p 0$,

$$Pr \{ CS_{2S} = CS_{NR} \} \to 1, \text{ and } Pr \{ \beta_0 \in CS_{2S} \} \to_p 1 - \alpha.$$

Corollary 4 establishes that the recommended two-step confidence set has asymptotic coverage at least $1 - \alpha - \gamma$ regardless of identification strength. Moreover, $CS_{2S}$ coincides with the non-robust confidence set $CS_{NR}$ with probability tending to one under strong identification, and thus has asymptotic coverage $1 - \alpha$ in this case. Thus, unlike two-step procedures based on the first-stage F-statistic, our approach controls coverage distortions in general models, including linear IV with heteroskedastic, serially correlated, or clustered errors. The linear-combination procedures considered here are only one of many ways to generate robust confidence sets satisfying our requirements. Supplementary Appendix C details one alternative approach, which has the advantage of requiring only $\chi^2$ critical values.

### 2.2.4 Choosing the Value $\alpha$ and Computing Critical Values

The parameter $\gamma = 1 - \alpha - H \left( \chi^2_{p,1-\alpha}; a, k, p \right)$ measures the maximal coverage distortion for $CS_{2S}$. This is the price paid (in terms of coverage) for using this two-step
procedure, rather than using a fully robust confidence set like $CS_R$. Choosing a smaller value of $\gamma$ reduces the maximal possible coverage distortion but also makes our preliminary confidence set $CS_{R,P}$ larger, which in turn increases $Pr\{\phi_{ICS} = 1\}$ for any finite sample size and thus makes our ICS procedure more stringent, in the sense that it requires stronger evidence before concluding that $CS_{NR}$ is reliable.

Since $H\left(\chi^2_{p,1-\alpha}; a, k, p\right)$ is decreasing and continuous in $a$, converges to $1 - \alpha$ as $a \to 0$, and converges to 0 as $a \to \infty$, we can set $a$ to achieve any value $\gamma$ between 0 and $1 - \alpha$. In practice, we suggest choosing $\gamma$ and then selecting $a$ accordingly. In particular, given $k = \dim (g_t(\theta))$ and $p = \dim (\beta)$ we recommend the following:

1. Select the coverage $\alpha$ for the non-robust confidence set and the maximal permitted coverage shortfall $\gamma$ for the two-step procedure.

2. For $M$ a large number, for each $m \in \{1, \ldots, M\}$ draw $(A_m, B_m) \sim (\chi^2_{p,1-\alpha}, \chi^2_{k-p})$ independently. Let $\hat{a}(\gamma)$ solve

$$\frac{1}{M} \sum_{m=1}^{M} 1\left\{ (1 + \hat{a}(\gamma)) A_m + \hat{a}(\gamma) B_m > \chi^2_{p,1-\alpha} \right\} = 1 - \alpha - \gamma.$$ 

3. Let $\hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ be the $1 - \alpha$ quantile of $(1 - \hat{a}(\gamma)) A_m + \hat{a}(\gamma) B_m$.

4. Construct the two-step confidence set $CS_{2S}$ as above, using $a = \hat{a}(\gamma)$ and approximating $H^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ by $\hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$.

For $M$ large this procedure will deliver two-step confidence sets with asymptotic coverage at least $1 - \alpha - \gamma$ regardless of identification strength, and asymptotic coverage exactly $1 - \alpha$ under strong identification. This procedure is easy to implement and the only parameter which must be selected by the user is the permitted level of coverage distortion $\gamma$. For ease of use, values for $\left(\hat{a}(\gamma), \hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)\right)$ for $p \in \{1, 2, 3\}$ and $\alpha \in \{1\%, 5\%, 10\%\}$ are reported in Supplementary Appendix D.
2.2.5 Reporting Results

Our discussion above assumes that a researcher will report either $CS_{NR}$ or $CS_R$, but not both. In empirical papers, however, robust tests and confidence sets have typically been used to supplement, rather than replace, non-robust ones. For instance, of the 38 empirical papers calculating robust tests or confidence sets based on Moreira (2003) which themselves have over 20 citations, 37 also report non-robust Wald tests or confidence sets. Different readers may also be comfortable with different levels of coverage distortion $\gamma$, making the choice of this parameter potentially difficult. A simple modification of the procedure described above addresses both of these concerns.

To implement this modified procedure, the researcher specifies some value $\gamma_{\min} \geq 0$. For $\gamma \geq \gamma_{\min}$, define a family of preliminary robust confidence sets

$$
CS_{R,P}(\gamma) = \left\{ \beta : \min_{\theta : \beta = f(\theta)} (K_{\Omega,f}(\theta) + a(\gamma) \cdot S(\theta)) \leq \chi^2_{p,1-\alpha} \right\}
$$

for $a(\gamma)$ as defined above, which may again be approximated by $\tilde{a}(\gamma)$. Note that these confidence sets are decreasing in $\gamma$, in the sense that $\gamma \geq \gamma$ implies $CS_{R,P}(\gamma) \subseteq CS_{R,P}(\gamma')$. Define $\gamma'$ is the smallest value such that $CS_{R,P}(\gamma') \subseteq CS_{NR}$.

$$
\gamma' = \min \{ \gamma \geq \gamma_{\min} : CS_{R,P}(\gamma) \subseteq CS_{NR} \}.
$$

$\gamma'$ is the smallest maximal distortion such that our ICS statistic $\phi_{ICS}$ will indicate strong identification, $\phi_{ICS} = 0$, so we will refer to $\gamma'$ as the maximal distortion cutoff.

Define our robust confidence set $CS_R$ such that it contains $CS_{R,P}(\gamma_{\min})$, for example taking

$$
CS_R = \left\{ \beta : \min_{\theta : \beta = f(\theta)} (K_{\Omega,f}(\theta) + a(\gamma_{\min}) \cdot S(\theta)) \leq H^{-1}(1 - \alpha ; a(\gamma_{\min}), k, p) \right\}.
$$

A researcher can then report $(CS_{NR}, CS_R, \gamma')$. If a reader adopts the rule that they will focus on the non-robust confidence set when $\gamma' \leq \gamma$ and the robust confidence set when $\gamma' > \gamma$, it follows immediately from Corollary 4 that the two-step confidence set used by this reader will have coverage at least $1 - \alpha - \gamma$. Thus, reporting $\gamma'$ allows
the construction of a variety of two-step confidence sets, and allows readers decide for themselves whether they are comfortable relying on the non-robust confidence sets.

2.3 Two-Step Confidence Sets in Linear IV

In this section, we apply the two-step confidence sets developed in Section 2.2 to the linear IV model. We show in simulation that our approach has performance competitive with pretests based on the first stage F-statistic in models with homoskedastic data but also controls coverage distortions for $CS_{2S}$ even under heteroskedasticity.

2.3.1 Two Step Confidence Sets Under Homoskedasticity

We return to the homoskedastic IV model considered in Section 2.1.1 and simulate the coverage of the robust confidence sets $CS_R$ and $CS_{R,P}$, as well as the two-step confidence sets constructed in Section 2.2. For comparability with the simulations in Section 2.1.1, in all cases we set $\alpha = 5\%$ and $\gamma = 10\%$. By construction the robust confidence set $CS_R$ has asymptotic coverage 95\% under both weak and strong identification, while $CS_{R,P}$ has asymptotic coverage 85\% and the two step confidence sets have minimal asymptotic coverage at least 85\%. The simulation performance of these confidence sets, reported in Table 2.3, supports our theoretical results, showing that the simulated coverage of $CS_R$ and $CS_{R,P}$ is quite close to their theoretical coverage while the minimal coverage of the two-step confidence sets is in all cases at least 85\%.

Since pretests based on the critical values of SY (discussed in Section 2.3) also guarantee coverage at least 85\% under homoskedasticity, it is interesting to compare their behavior to that of our ICS statistic (2.12). In Figure 2-2 we plot the mean of our ICS statistics $E[\phi_{ICS}]$ together with $E[\phi_{ICS,SY}]$, the mean of the ICS statistics based on the first stage F-statistic with SY's critical values, against the mean of the first-stage F-statistic as we vary $\|\pi\|$ in the moderate endogeneity (M) calibration with $k = 10$. As we can see, our ICS procedure for LIML behaves quite similarly to that of SY, while our ICS procedure for 2SLS indicates strong identification with
Table 2.3: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations, based on 2,500 simulations. $CS_R$ and $CS_{R,P}$ are robust 95% and 85% confidence sets, respectively, calculated as suggested in Section 2.2 for $\alpha = 5\%$ and $\gamma = 10\%$ based on the two-stage least squares weight $\hat{\Omega} (\theta) = \left( \frac{1}{\tau} Z'Z \right)^{-1}$. $CS_{2S}$ LIML and $CS_{2S}$ 2SLS are two-step confidence sets (2.1) based on LIML and 2SLS, calculated as described in Section 2.2 for $\alpha = 5\%$ and $\gamma = 10\%$.

substantially higher probability than that of SY. Repeating this exercise for the other M calibrations we find similar results (not shown), while when we consider the high endogeneity H calibrations we find no general ordering between our ICS procedures and those of SY.

### 2.3.2 Two Step Confidence Sets Heteroskedasticity

We now turn to the performance of our robust and two-step confidence sets in the heteroskedastic linear IV calibrations introduced in Section 2.1.2. In particular, we consider the robust confidence sets $CS_R$ and $CS_{R,P}$ based on both the inefficient 2SLS weight matrix $\hat{\Omega} (\theta) = \left( \frac{1}{\tau} Z'Z \right)^{-1}$ and the efficient weight matrix $\hat{\Omega} (\theta) = \hat{\Sigma}_g (\theta)^{-1}$, as well as two-step confidence sets based on LIML, 2SLS, CUGMM, and 2SGMM constructed as discussed in Section 2.2. For consistency with Section 2.1.2 we take $\alpha = 5\%$ and $\gamma = 10\%$ in all cases. We can see that as in the homoskedastic case discussed above our robust confidence sets $CS_R$ and $CS_{R,P}$ have minimal coverage quite close to their theoretical coverage of 95% and 85%, respectively. Unlike the procedures based on the first stage F-statistic our two-step confidence sets have minimal coverage at least 85%, consistent with our theoretical results. Thus, our simulation results confirm that our approach delivers two-step confidence sets $CS_{2S}$ which control coverage distortions.
2.4 Empirical Illustration

To illustrate the utility of our approach, we apply the methods developed in Section 2.2 to two empirical examples. We first consider the well-studied Angrist and Krueger (1991) IV regression of wages on years of schooling, and then consider recent analyses by Acemoglu et al. (2008) and Cervellati et al. (2014) of the relationship between income and democracy.

2.4.1 Returns to Schooling

Angrist and Krueger (1991) study the relationship between workers’ year of schooling and their labor market earnings. Using the structure of US compulsory schooling laws, they argue that the quarter in which a worker is born provides a valid instrument for years of education. They then interact quarter of birth with a worker’s year and
Table 2.4: Minimal coverage for confidence sets in homoskedastic IV simulations with 10,000 observations, based on 2,500 simulations. \( CS_R \) and \( CS_{R,P} \) Inefficient are robust 95% and 85% confidence sets (2.13) based on the two-stage least squares weight matrix \( \hat{\Omega}(\theta) = \left( \frac{1}{T} \tilde{Z}'\tilde{Z} \right)^{-1} \), calculated as suggested in Section 2.2 for \( \alpha = 5\% \) and \( \gamma = 10\% \). \( CS_R \) and \( CS_{R,P} \) Efficient are robust confidence sets with \( \hat{\Omega}(\theta) = \tilde{\Sigma}_q(\theta)^{-1} \) calculated as suggested in Section 2.2 \( \alpha = 5\% \) and \( \gamma = 10\% \). \( CS_{2S} \) LIML, \( CS_{2S} \) 2SLS, \( CS_{2S} \) CUGMM, and \( CS_{2S} \) 2SGMM are two-step confidence sets (2.1) based on LIML, 2SLS, CUGMM, and 2SGMM, calculated as described in Section 2.2 for \( \alpha = 5\% \) and \( \gamma = 10\% \).

State of birth to generate a large number of instruments. As observed in Bound et al. (1995) and Staiger and Stock (1997), the first-stage F-Statistic is quite small for some of the specifications considered in Angrist and Krueger (1991), and Monte-Carlo evidence suggests that the usual 2SLS confidence sets may be highly unreliable.

We consider four specifications studied by Staiger and Stock (1997). The total number of instruments ranges from 3 to 178, and we focus on the 1930-1939 cohort- see Angrist and Krueger (1991) and Staiger and Stock (1997) for further information on the data and specification. We find strong evidence of heteroskedasticity in the first-stage relationship between quarter of birth and years of schooling (the p-value for a White (1980) test for heteroskedasticity is zero to over ten decimal places) and so focus on heteroskedasticity-robust procedures. This high level of significance seems more the result of high power due to the large sample size than of substantively important heteroskedasticity, however, and the results reported here are largely similar to those obtained from the corresponding non-heteroskedasticity-robust procedures. Much as in Staiger and Stock (1997) we find that the first-stage F-statistic varies considerably
across specifications, ranging from a high of 30.6 to a low of 1.58.

We consider Wald statistics based on 2SLS and CUGMM with the usual standard errors. Hansen et al. (2008) argue that the main difficulty in this application is the large number of instruments rather than the strength of identification as such, so we also consider many-weak-instrument robust Wald tests. These tests require that the first stage parameter $\pi$ not be too small, but have correct size even when the number of instruments is large relative to the sample size. Specifically, we consider CUGMM with the many-weak-moment-robust standard errors of Newey and Windmeijer (2009), and the many-weak-instrument-robust HLIM and HFUL estimators and standard errors of Hausman et al. (2012). Confidence sets based on these procedures will not ensure correct coverage when the data is highly uninformative, but can be expected to perform substantially better than 2SLS in cases where (i) the data contains a limited but non-trivial amount of information about model parameters and (ii) the number of instruments is large.

To construct two-step confidence sets, we compare these Wald statistics to the appropriate robust confidence sets. In particular, the usual CUGMM standard errors and the Newey and Windmeijer (2009) standard errors are asymptotically equivalent under strong identification, so we compare these confidence sets to the robust confidence set $CS_{R,F}$ as in (2.10) based on the efficient weighting $\hat{\Omega} = \hat{\Sigma}_y^{-1}$.10 HLIM and HFUL, on the other hand, are asymptotically equivalent to 2SLS under the usual asymptotics, so we compare the confidence sets based on these estimators and 2SLS to the robust confidence set $CS_{R,F}$ obtained using the 2SLS weighting $\hat{\Omega} = \left(\frac{1}{n}Z'Z\right)^{-1}$. Rather than specifying a maximal coverage distortion, we report the maximal distortion cutoff $\hat{\gamma}$ for 95% confidence sets as defined in Section 2.2.5 and set $\gamma_{\text{min}} = 0$, so in the event that $\phi_{ICS} = 1$ we will consider the 95% confidence set $CS_R = \left\{\beta : K_{\Omega,F}(\beta) \leq \chi^2_{1,0.95}\right\}$ with the appropriate weighting for $K_{\Omega,F}$ (efficiently weighted for CUGMM, 2SLS weighted for 2SLS, HLIM, and HFUL). For purposes of comparison, we also report confidence sets based on the $S$ statistic.

10Newey and Windmeijer (2009) show that their Wald statistic is asymptotically equivalent to Kleibergen’s efficiently-weighted $K_{\Omega,F}$ statistic at the true parameter value under their many-weak-moment condition asymptotics, further justifying this choice.
To calculate robust confidence sets we restrict attention to $\beta \in [-0.5, 0.5]$, which includes (and substantially exceeds) the range of economically plausible values. In particular, $\beta = -0.5$ means that an additional year of schooling decreases expected weekly earnings by nearly 40%, while $\beta = 0.5$ means it increases them by over 60%. This restriction is important for the 180 instrument specification, since without this we find that many robust confidence sets are disjoint, consisting of a small interval of economically plausible values, together with a wide range of economically implausible ones.

Our results are reported in Table 2.5. While the maximal distortion cutoffs $\hat{\gamma}$ are small for all Wald confidence sets in specification I where $k = 3$, as we increase the number of instruments the robust confidence sets differ substantially from the usual 2SLS and CUGMM confidence sets, which is reflected by large values of $\hat{\gamma}$. Strikingly, 95% Wald confidence sets based on CUGMM with Newey and Windermierer (2009) standard errors and the HLIM and HFUL estimators are close to the appropriate robust confidence sets in all specifications save III, with $\hat{\gamma}$ below 4%. Hence, if we are willing to tolerate 5% coverage distortions, the two-step approach developed in this paper would select the Wald confidence set in all specifications but III. Even in specification III, $\hat{\gamma}$ for these confidence sets is still under 10%. Thus, applying our two-step approach to this example we find that 95% Wald confidence sets with many-weak-instrument-robust standard errors are similar to robust confidence sets, even though 2SLS and CUGMM with the usual standard errors are not.

If we used the SY critical values together with the first-stage F-statistic, setting $\gamma = 10\%$, we would use the 2SLS confidence set in specification I and a robust confidence set in all other cases. If we used the SY cutoffs for LIML to assess the performance of CUGMM, we would likewise choose the usual CUGMM confidence set for specifications I and II and a robust confidence set in specifications III and IV. There are no available analogs of the SY cutoffs for Wald tests based on the Newey and Windermierer (2009) standard errors or for HLIM and HFUL, however, so unlike our approach existing techniques based on the F-statistic give us no way to gauge the
reliability of these procedures.\textsuperscript{11}

For all specifications other than III, our robust optimally-weighted $K$ statistic gives confidence sets which, while often wider than 2SLS confidence sets, are still reasonably informative. This is consistent with previous findings from Cruz and Moreira (2005) and Hansen et al. (2008) that while the confidence sets reported in Angrist and Krueger (1991) based on a large number of instruments are unreliable, some robust methods deliver substantively similar conclusions. We find that the 2SLS-weighted $K$ statistic gives informative confidence sets for specifications I and II, and an informative upper bound in specification IV, while the $S$ confidence set is for the most part less informative that the efficiently-weighted confidence set.

\subsection*{2.4.2 Income and Democracy}

Acemoglu et al. (2008) are interested in the causal relationship between per-capita income and democracy. They examine a wide range of specifications and argue that once one controls for time-invariant country effects, panel regressions using international data find no statistically significant relationship between per-capita income and various measures of democracy at the country level. Since per-capita GDP may be endogenous even after controlling for country fixed-effects, one of their specifications instruments this variable using the trade-weighted income of a country's trading partners. Cervellati et al. (2014) argue that the zero-effect finding of Acemoglu et al. (2008) masks considerable heterogeneity across subgroups of countries, and in particular find that for former European colonies the relationship between income and democracy is negative and significant. Here, we focus on a five-year panel specification based on the trade-weighted income instrument, corresponding to Table 6 Column 5 in Acemoglu et al. (2008) and Table 3 Column 3 in Cervellati et al. (2014). See Acemoglu et al. (2008) and Cervellati et al. (2014) for further information on the data and specification.

\textsuperscript{11}One could in principal use the same approach as in SY to derive $F$-statistic cutoffs appropriate for these estimators in the homoskedastic case, but whether these results can be extended to the non-homoskedastic case is far from clear.
<table>
<thead>
<tr>
<th>Specification</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>s.e.</td>
<td>$\hat{\gamma}$</td>
<td>$\hat{\beta}$</td>
</tr>
<tr>
<td>2SLS</td>
<td>0.0990</td>
<td>0.0207</td>
<td>2.5%</td>
<td>0.0806</td>
</tr>
<tr>
<td>CUGMM</td>
<td>0.1000</td>
<td>0.0208</td>
<td>1.9%</td>
<td>0.0857</td>
</tr>
<tr>
<td>CUGMM, NW</td>
<td>0.1000</td>
<td>0.0213</td>
<td>1.3%</td>
<td>0.0857</td>
</tr>
<tr>
<td>HLIM</td>
<td>0.0999</td>
<td>0.0213</td>
<td>1.3%</td>
<td>0.0838</td>
</tr>
<tr>
<td>HFUL</td>
<td>0.0995</td>
<td>0.0210</td>
<td>1.8%</td>
<td>0.0836</td>
</tr>
</tbody>
</table>

First Stage F | 30.5822 | 4.6245 | 1.5788 | 1.8232 |

<table>
<thead>
<tr>
<th>Controls</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Age, Age²</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>SOB</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instruments</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>QOB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>QOB*YOB</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>QOB*SOB</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td># instruments</td>
<td>3</td>
<td>30</td>
<td>28</td>
</tr>
<tr>
<td>Observations</td>
<td>329,509</td>
<td>329,509</td>
<td>329,509</td>
</tr>
</tbody>
</table>

Table 2.5: Results for Angrist and Krueger (1991) data. Specifications as in Staiger and Stock (1997): $Y = \text{log weekly wages}$, $X = \text{years of schooling}$, instruments $Z$ and exogenous controls as indicated. Base controls are Race, MSA, Married, Region, and Year of Birth. QOB, YOB, and SOB are quarter, year, and state of birth dummies, respectively. CUGMM, NW is CUGMM together with Newey and Windermier (2009) standard errors, while HLIM and HFUL are estimators and standard errors proposed in Hausman et al. (2012). The maximal distortion cutoffs $\hat{\gamma}$ are based on comparing nominal 95% Wald confidence sets to the appropriate robust confidence sets $CS_{R,P}$ as described in Section 2.2.5. Efficiently weighted K statistic confidence sets with $\hat{\gamma} = \hat{\Sigma}^{-1} Z'$ are used to calculate $\hat{\gamma}$ for CUGMM with usual and NW standard errors, while 2SLS-weighted K statistic confidence sets with $\hat{\gamma} = \left( \frac{1}{\hat{\gamma}^T Z' Z} \right)^{-1}$ is used for other estimators. For non-convex confidence sets we report the convex hull.
As in Acemoglu et al. (2008) and Cervellati et al. (2014), all results are clustered at the country level. Since this clustering allows dependence across observations, the Newey and Windmeijer (2009) and Hausman et al. (2012) Wald confidence sets discussed above are not applicable to this context. Moreover, the model is exactly identified so 2SLS and the continuous updating estimator are the same. Thus we consider only the 95% Wald confidence set based on 2SLS with the usual standard errors.

To construct two-step confidence sets, we compare the Wald confidence set to the robust confidence sets $CS_{R,P}$ (2.10). We again report the maximal distortion cutoff $\gamma$ as defined in Section 2.2.5 and set $\gamma_{\text{min}} = 0$, so that if $\phi_{ICS} = 0$ we will use the 95% confidence set $CS_R = \{\beta : S(\beta) \leq \chi^2_{1,0.95}\}$. Here the confidence sets $CS_{R,P}$ and $CS_R$ depend only on $S$ since $S \equiv K_{\theta,f}$ in just-identified models with $f(\theta) = \theta$, like the present case.

As with the Angrist Krueger model, it is helpful here to restrict the range of parameter values considered when computing robust confidence sets. Here we focus on $\beta \in [-1, 1]$, motivated by the fact that the outcome variable $Y$ in this example (a measure of democracy from Freedom House) ranges from zero to one, and that the standard deviation of log income per capita (the endogenous regressor $X$) is about one. Thus, $\beta = 1$ means that a one standard-deviation increase in log per-capita income suffices to move a country’s expected level of democracy from the lower bound to the upper bound, while $\beta = -1$ implies the same for a one standard-deviation decrease in log income per capita.

The results of our analysis are reported in Table 2.6. Most strikingly, for all specifications considered the 95% $S$ confidence set covers the whole parameter space $[-1, 1]$, highlighting that specifications based on the weighted world income instrument are not very informative. In spite of this, in specifications I and III the first-stage F-statistic substantially exceeds the SY cutoff of 16.38 (for $k = 1$, $\gamma = 10\%$), so that two-step confidence sets based on the first-stage F-statistic would use Wald confidence sets. Indeed, Cervellati et al. (2014) state that for this specification “the instrument works well, with an F-statistic substantially exceeding the conventional
critical value.” By contrast, we see that in specifications I and III \( \hat{\gamma} \) exceeds 20%, showing that strong-identification approximations appear unreliable in this context. Thus, applied to this example, our two-step approach selects the robust confidence set \( CS_R \) in all three specifications. If we were to rely on the first stage F-statistic to judge the strength of identification in this application, we would use the 2SLS confidence set in specifications I and III, and so be at risk of drawing highly misleading conclusions.

Substantively, the \([-1, 1]\) robust confidence set for specification II is similar in interpretation to the Wald confidence set in this context, since the latter also covers all economically plausible parameter values. In specification I, the \([-1, 1]\) robust confidence set is consistent with the Acemoglu et al. (2008) finding of no relationship between income and democracy, but is so imprecise that it adds little support one way or the other. The \([-1, 1]\) confidence set is specification III, by contrast, shows that the IV specification is much too noisy to support conclusion of Cervellati et al. (2014) that the relationship between income and democracy in former European colonies is negative.\(^{12}\)

<table>
<thead>
<tr>
<th>Specification</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>(-0.12)</td>
<td>(-1.37)</td>
<td>(-0.16)</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.11</td>
<td>23.03</td>
<td>0.07</td>
</tr>
<tr>
<td>( \hat{\gamma} )</td>
<td>20.2%</td>
<td>0%</td>
<td>57.9%</td>
</tr>
<tr>
<td>First Stage F</td>
<td>26.5</td>
<td>0.0</td>
<td>57.6</td>
</tr>
<tr>
<td>S CS</td>
<td>([-1,1])</td>
<td>([-1,1])</td>
<td>([-1,1])</td>
</tr>
<tr>
<td>Sample Restriction</td>
<td>None</td>
<td>Non-colonies</td>
<td>Former Colonies</td>
</tr>
<tr>
<td>Observations</td>
<td>895</td>
<td>218</td>
<td>718</td>
</tr>
</tbody>
</table>

Table 2.6: Results for Acemoglu et al. (2008) and Cervellati et al. (2014) specifications and data: \( Y = \) Freedom House democracy measure, \( X = \) lagged log of per-capita GDP, \( Z = \) trade-weighted income of trading partners. Exogenous controls for all specifications are country and time fixed effects and lagged democracy. Column I matches Table 6 Column 5 of Acemoglu et al. (2008), Columns II and III match Table 3 Column 3 of Cervellati et al. (2014) , panels 1 and 2 respectively. The maximal distortion cutoffs \( \hat{\gamma} \) are based on two-step procedures comparing the nominal 95% the Wald confidence set to the robust confidence sets \( CS_{R,P} \) as described in Section 2.2.5. For non-convex confidence sets we report the convex hull.

\(^{12}\) This is only one of many specifications considered by Cervellati et al. (2014) and so does not undermine their broader point.
2.5 A Simple Two-Step Confidence Set: Theory

Our results apply to contexts where we have a robust confidence set which is contained within the non-robust confidence set of interest with probability tending to one under strong identification. We show that whenever this condition holds we can construct a two-step confidence set which controls coverage distortions under weak identification and is asymptotically equivalent to the non-robust confidence set when identification is strong. These results can be used to create two-step confidence sets with uniformly bounded asymptotic coverage distortions when an appropriate uniformly correct robust confidence set is available. As a preliminary step we begin by defining several different notions of limiting coverage probability, including asymptotic coverage and coverage under particular sequences of parameter values.

We assume that in sample $T$ we observe data with distribution $F_T (\beta_{0,T}, \psi_{0,T})$, where $\beta \in B \subseteq \mathbb{R}^p$ is a parameter of interest and $\psi \in \Psi$ is some, potentially infinite-dimensional, nuisance parameter. We allow the true parameter values $\beta_{0,T}$ and $\psi_{0,T}$ to change with the sample size, and denote the sequence of true parameter values by

$$\xi_0 = \{(\beta_{0,T}, \psi_{0,T})\}_{T=1}^\infty \in \Xi = \prod_{T=1}^\infty (B \times \Psi).$$

We typically want confidence sets to control asymptotic coverage probability, defined as the limit of the minimal probability that $CS$ contains the true value

$$ACP (CS) = \lim_{T \to \infty} \inf_{(\beta_{0,T}, \psi_{0,T}) \in B \times \Psi} \inf \Pr_{T, (\beta_{0,T}, \psi_{0})} \{\beta_0 \in CS\}.$$  

In this paper, we will also be interested in less stringent notions of coverage. In particular, define the (asymptotic) sequential coverage probability of confidence set $CS$ under the sequence of true parameter values $\xi_0$ as

$$SCP (CS, \xi_0) = \lim_{T \to \infty} \inf P_{T, \xi_0} \{\beta_{0,T} \in CS\}.$$  

Likewise, define the sequential coverage probability of confidence set $CS$ under the set
of sequences $\Xi^* \subseteq \Xi$ as the minimal sequential coverage probability under sequences in this set,

$$SCP(CS, \Xi^*) = \inf_{\xi \in \Xi} SCP(CS, \xi).$$

Note that sequential coverage probability under $\Xi$ is simply the asymptotic coverage probability

$$SCP(CS, \Xi) = ACP(CS).$$

We will be particularly concerned with sequential coverage probability under “strongly identified sequences” which we will denote by $\Xi_S \subseteq \Xi$, and under “weakly identified sequences” which we will denote by $\Xi_W \subseteq \Xi$.

As above, suppose we have non-robust and robust confidence sets $CS_{NR}$ and $CS_R$. We assume that our non-robust confidence set has sequential coverage at least $1 - \alpha$ under strong identification

$$SCP(CS_{NR}, \Xi_S) \geq 1 - \alpha. \quad (2.14)$$

but impose no restriction on the performance of this confidence set under weak identification. By contrast, we will assume that the robust confidence set $CS_R$ has coverage at least $1 - \alpha$ under both weak and strong identification

$$\min \{SCP(CS_R, \Xi_S), SCP(CS_R, \Xi_W)\} \geq 1 - \alpha. \quad (2.15)$$

For two-step confidence sets $CS_2S$ as in (2.1), we can see that the sequential coverage of $CS_{2S}$ will depend on the limiting behavior of $\phi_{ICS}$. Some bounds hold for all choices of $\phi_{ICS}$, specifically

**Lemma 2** Under (2.14) and (2.15),

1. $SCP(CS_{2S}, \Xi_W) \geq 1 - \alpha - \sup_{\xi \in \Xi_W} \limsup_{T \to \infty} Pr_{T,\xi} \{\phi_{ICS} = 0\}$

2. $SCP(CS_{2S}, \Xi_S) \geq 1 - \alpha - \min \{\alpha, \sup_{\xi \in \Xi_S} \limsup_{T \to \infty} Pr_{T,\xi} \{\phi_{ICS} = 1\}\}$.

These inequalities are tight, in the sense that one cannot obtain a sharper bound without additional restrictions on the joint behavior of $CS_{NR}$, $CS_R$, and $\phi_{ICS}$. 129
Our approach to constructing two-step confidence sets gives us such additional structure. In particular, we make the following assumption:

**Assumption 1** We have a preliminary confidence set \( CS_{R,P} \) such that:

1. \( SCP(CS_{R,P}, \Xi_w) \geq 1 - \alpha - \gamma \)

2. \( Pr_{T,\xi} \{ CS_{R,P} \subseteq CS_R \} = 1 \) for all \( T \) and \( \xi \in \Xi \)

3. \( \inf_{\xi_0 \in \Xi} \lim \inf_{T \to \infty} Pr_{T,\xi_0} \{ CS_{R,P} \subseteq CS_{N,R} \} = 1 \).

This assumption requires the existence of a preliminary confidence set that (1) has sequential coverage at least \( 1 - \alpha - \gamma \) when identification is weak, (2) is contained in \( CS_R \) with probability one, and (3) is contained in \( CS_{N,R} \) with probability tending to one under all strongly identified sequences. While this might seem quite demanding, we show in the next section that for \( CS_{N,R} \) a Wald confidence set and \( CS_R, CS_{R,P} \) as defined in Section 2.2.3, this condition holds in general GMM models under appropriate sequences. Under Assumption 1 we can easily establish that the two-step confidence sets \( CS_{2s} \) defined using \( \phi_{ICS} \) as in (2.12) have bounded coverage distortions under weak identification and correct coverage under strong identification.

**Theorem 7** Under Assumption 1 together with (2.14), for \( \phi_{ICS} \) as defined in (2.12) the two step confidence set \( CS_{2s} \) has the following properties:

1. \( SCP(CS_{2s}, \Xi_w) \geq 1 - \alpha - \gamma \)

2. \( SCP(CS_{2s}, \Xi_s) \geq 1 - \alpha \)

3. \( \inf_{\xi_0 \in \Xi_s} \lim \inf_{T \to \infty} Pr_{T,\xi_0} \{ CS_{2s} = CS_{N,R} \} = 1 \).

Further, \( \sup_{\xi_0 \in \Xi_s} \lim \sup_{T \to \infty} Pr_{T,\xi_0} \{ \phi_{ICS} = 1 \} = 0 \).

Theorem 7 shows that whenever we have a preliminary robust confidence set satisfying Assumption 1 we can create a two-step confidence set which controls coverage distortions and is equal to our non-robust confidence set with probability tending...
to one under strong identification. This result highlights the value of identification-
robust test statistics which are locally asymptotically equivalent to the conventional
non-robust (e.g. Wald) statistics when identification is strong, since given such statis-
tics it is typically straightforward to construct a preliminary confidence set satisfying
Assumption 1 and thus to construct robust two-step confidence sets.

Uniform Confidence Sets for Linear IV with One Endogenous Regressor
If there exists a preliminary confidence set $CS_{R,P}$ satisfying Assumption 1 which has
uniformly correct coverage (one with $ACP(CS_{R,P}) \geq 1 - \alpha - \gamma$), by Theorem 7
(taking $\Xi_W = \Xi$) we can use it to construct two-step confidence sets with uniformly
correct coverage. Thus, since Proposition 2 of Andrews (2013) implies that $CS_{R,P}$
as defined in (2.10) based on the efficient weighting matrix $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$ has
asymptotic coverage probability $1 - \alpha - \gamma$ in linear IV with a single endogenous
regressor and heteroskedastic errors, Theorem 7 shows that the two-step confidence
set developed in Section 2.2.3 has asymptotic coverage at least $1 - \alpha - \gamma$ in this case,
$ACP(CS_{2S}) \geq 1 - \alpha - \gamma.$

2.6 Two-Step Confidence Sets for GMM

As in Section 2.2, suppose we have a GMM model with moment condition $g_i(\theta)$
which is mean zero when evaluated at the true parameter value. To accommodate the
drifting sequences of parameter values introduced in Section 2.5, let $E_{T,\xi} \{ \cdot \}$ denote an
expectation in sample size $T$ under sequence of true parameter values $\xi$. We begin by
giving conditions under which the robust confidence sets $CS_R$ and $CS_{R,P}$ have correct
sequential coverage even in weakly identified models, and then turn to establishing
their properties in strongly identified models.

---

13To obtain this result, we impose the same restrictions on the nuisance parameter space $\Psi$ as in
2.6.1 Behavior Under Weak Identification

As noted above, existing results establish the uniform validity of robust confidence sets in heteroskedastic linear IV with a single endogenous regressor. Unfortunately, similar uniformity results are not yet available for the general GMM case, and proving such results is beyond the scope of this paper. Thus our focus is on sequential coverage under particular sequences of models. The generality of Theorem 7, however, means that once uniformity results are available for appropriate GMM confidence sets we can directly use them to construct robust two step confidence sets. 14

As in Kleibergen (2005) we assume that a central limit theorem holds for the moment condition and its Jacobian.

**Assumption 2** For all \( \xi_0 \in \Xi_W \cup \Xi_S \), under \( \xi_0 \) we have that for

\[
J_{T,\xi}(\theta) = E_{T,\xi} \left[ \frac{\partial}{\partial \theta} g_T(\theta) \right],
\]

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( g_t(\theta_{0,T}) \right) \rightarrow_d \left( \psi_\theta \right) \sim N \left( 0, \left( \begin{array}{cc} \Sigma_g & \Sigma_g \theta \\ \Sigma_g \theta & \Sigma_\theta \end{array} \right) \right)
\]

where \( \Sigma_g \) is positive definite and

\[
\left( \begin{array}{cc} \Sigma_g & \Sigma_g \theta \\ \Sigma_g \theta & \Sigma_\theta \end{array} \right) = \lim_{T \to \infty} Var_{T,\xi_0} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \begin{array}{c} g_t(\theta_{0,T}) \\ \text{vec} \left( \frac{\partial}{\partial \theta} g_t(\theta_{0,T}) \right) \end{array} \right) \right).
\]

Under this assumption we have that \( \sqrt{T} g_T(\theta_{0,T}) = \frac{1}{\sqrt{T}} \sum_t g_t(\theta_{0,T}) \) and its recentered Jacobian converge jointly to a normal distribution. To construct test statistics with a known limiting distribution, we further need to assume that we can estimate the asymptotic variance matrix of \( \left( \sqrt{T} g_T(\theta_{0,T}), \sqrt{T} \text{vec} \left( \frac{\partial}{\partial \theta} g_T(\theta_{0,T}) \right) \right) \) and that the weight matrix \( \hat{\Omega}(\theta) \) converges to some non-random positive-definite limit.

**Assumption 3** For all \( \xi_0 \in \Xi_W \) we have estimators \( \hat{\Sigma}_g(\theta_{0,T}), \hat{\Sigma}_g(\theta_{0,T}) \) and \( \hat{\Sigma}_\theta(\theta_{0,T}) \) which converge in probability to fixed \( \Sigma_g, \Sigma_g \theta, \) and \( \Sigma_\theta \). Further, evaluated at the true

---

14In this direction, note that by subsequence arguments as in D. Andrews et al. (2011), for \( \Xi_W \) satisfying the assumptions below we can easily extend our results to prove correct coverage \( \Xi_W \), the set of all sequences which have subsequences in \( \Xi_W \). The key to proving uniformity is then proving that \( \Xi_W = \Xi \) for a reasonable base parameter space \( \Theta \times \Psi \).
parameter value the weight matrix $\hat{\Omega}(\theta)$ satisfies

$$\hat{\Omega}(\theta_{0,T}) \rightarrow_p \Omega$$

for a non-stochastic symmetric positive-definite matrix $\Omega$.

Provided Assumptions 2 and 3 hold, Lemma 1 of Kleibergen (2005) gives us a joint limiting distribution for the appropriately scaled and re-centered $g_T(\theta_{0,T})$ and $D_T(\theta_{0,T})$.

**Lemma 3** Under Assumptions 2 and 3 we have that for all $\xi_0 \in \Xi_W$,

$$\left( \begin{array}{c} \sqrt{T} g_T(\theta_{0,T}) \\ \sqrt{T} \text{vec}(D_T(\theta_{0,T}) - J_{T,X}(\theta_{0,T})) \end{array} \right) \rightarrow_d \begin{pmatrix} \psi_g \\ \psi_D \end{pmatrix} \sim N \left( \begin{array}{c} 0 \\ \begin{bmatrix} \Sigma_g & 0 \\ 0 & \Sigma_D \end{bmatrix} \end{array} \right)$$

where $\Sigma_D = \Sigma_\theta - \Sigma_{\theta\theta} \Sigma_g^{-1} \Sigma_{g\theta}$.

To derive the limiting distribution for $K_{0,f}$ as defined in (2.9), we need a final assumption to ensure that (i) we can normalize $D_T(\theta_{0,T})$ such that it converges to a limiting normal random variable which is non-singular almost surely and (ii) there exists a corresponding normalization for $\frac{\partial}{\partial \theta} f(\theta_{0,T})$ which ensures that this term converges to a non-singular limit.

**Assumption 4** There exist sequences of full-rank normalizing matrices $\Lambda_{1,T}$ and $\Lambda_{2,T}$ of dimension $m$ and $p$, respectively, such that

1. $D_T(\theta) \Lambda_{1,T} \rightarrow_d D$ for a Gaussian random matrix $D$ which is full rank almost surely, but whose distribution may be degenerate

2. $\Lambda_{2,T} \frac{\partial}{\partial \theta} f(\theta_{0,T}) \Lambda_{1,T} \rightarrow F$ for a full-rank matrix $F$

Further, the elements of $\Lambda_{1,T}$ are of order $O(\sqrt{T})$.\(^{15}\)

\(^{15}\)That is, they are bounded above in absolute value by $C\sqrt{T}$ for some constant $C$. 

133
This assumption is rather high-level, but can easily be verified in many leading cases. For example, Kleibergen (2005) considers the case where $\sqrt{T}J_{T,\xi}$ converges to a finite matrix $J$, in which case we can take $\Lambda_{1,T} = \sqrt{T}I_m$ and $\Lambda_{2,T} = \frac{1}{\sqrt{T}}I_p$. More broadly, this assumption holds under the commonly-used weakly-identified GMM embedding of Stock and Wright (2000). Given these assumptions, it is easy to establish that $K_{\Omega,f}(\theta_0)$ and $S(\theta_0) - K_{\Omega,f}(\theta_0)$ have a well-behaved limiting distribution even under weak identification.

**Theorem 8** Under Assumptions 2, 3, and 4, under all $\xi_0 \in \Xi_W$, 

$$(K_{\Omega,f}(\theta_0,T), S(\theta_0,T) - K_{\Omega,f}(\theta_0,T)) \rightarrow_d (\chi_p^2, \chi_{k-p}^2)$$

and $K_{\Omega,f}(\theta_0,T)$ and $S(\theta_0,T) - K_{\Omega,f}(\theta_0,T)$ are asymptotically independent.

For tests of the full parameter vector $(f(\theta) = \theta)$ based on the efficient weighting matrix $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$, Theorem 8 follows from results in Kleibergen (2005). Likewise, in the case where we are interested in testing a subset of the parameter vector (e.g. $f(\theta) = \theta_1$) and use the efficient weighting matrix $\hat{\Omega}(\theta) = \hat{\Sigma}_g(\theta)^{-1}$, Theorem 8 follows from results in Chaudhuri and Zivot (2011). The primary innovation of Theorem 8 is thus the fact that it allows for alternative weighting matrices $\hat{\Omega}(\theta) \neq \hat{\Sigma}_g(\theta)^{-1}$. Empirical researchers frequently use estimators based on inefficient weights (e.g. two-stage least squares applied to heteroskedastic data), so to construct two-step confidence sets based on such estimators it is important that we allow for general $\hat{\Omega}(\theta)$.

Note that Theorem 8 implies that

$$SCP(CS_{R,P}, \Xi_W) = 1 - \alpha - \gamma$$

and

$$SCP(CS_{R}, \Xi_W) = 1 - \alpha$$

for $CS_{R,P}$ and $CS_{R}$ as defined in (2.10) and (2.13). Thus, since $CS_{R,P} \subseteq CS_{R}$ by definition, Assumption 1(1) and (2) hold. Hence, to verify Assumption 1 all
that remains is to show that $CS_{R,P} \subseteq CS_{NR}$ with high probability under strong identification.

### 2.6.2 Behavior Under Strong Identification

We now examine the properties of the preliminary confidence set $CS_{R,P}$ under strong identification. Specifically, we will define the set of strongly identified sequences $\Xi_S$ as the set of sequences such that conventional GMM assumptions (closely related to those of e.g. Newey and McFadden (1994)) are satisfied. To simplify the exposition, we assume that for all $\xi \in \Xi_S$, $\theta_{0,T} = \theta_0$ for all $T$ so that the true value $\theta_0$ is not changing with the sample size. To obtain the usual (point-wise) strong identification asymptotics, we could further fix $\psi_{0,T} = \psi_0$, and so hold the data generating process fixed while taking the sample size to infinity. This additional restriction does not substantially simplify the results, however, so we do not impose it.

Following Newey and McFadden we consider two main groups of assumptions. The first group implies the consistency of the estimator

$$\hat{\theta} = \arg \min_\theta g_T (\theta)' \hat{\Omega} (\theta) g_T (\theta),$$

while the second group implies its asymptotic normality.\(^{16}\)

**Assumption 5** For all $\xi_0 \in \Xi_S$ the following conditions hold:

1. $g_T(\theta) \rightarrow_p \lim_{T \to \infty} E_{T,\xi_0} [g_T(\theta)]$ uniformly over the compact parameter space $\Theta$ for $\theta$

2. $E_{T,\xi_0} [g_T(\theta_0)] = 0 \forall T$

3. $\hat{\Omega} (\theta) \rightarrow_p \Omega (\theta)$ uniformly over $\Theta$ for $\Omega(\theta)$ continuous and everywhere positive definite with a uniformly bounded maximal eigenvalue and minimal eigenvalue bounded away from zero

---

\(^ {16}\)For simplicity we assume that we are working with $CS_{NR}$ based directly on $\hat{\theta}$, but all results also apply to any $\hat{\theta}$ which is first-order equivalent to $\hat{\theta}$, that is $\hat{\theta}$ such that $\| \hat{\theta} - \theta \| = o_p \left( \frac{1}{\sqrt{T}} \right)$ under strong identification.
4. For all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left( \lim_{T \to \infty} E_T, \xi_0 \left[ g_T(\theta) \right] \right)^T \Omega(\theta) \left( \lim_{T \to \infty} E_T, \xi_0 \left[ g_T(\theta) \right] \right) < \delta$$

only if $\|\theta - \theta_0\| < \varepsilon$.

Assumption 5(1) requires that the sample average of the moment condition $g_T(\theta_0)$ be uniformly close to its mean in large samples, while Assumption 5(3) requires that the weighting matrix be well-behaved. Assumptions 5(2) and (4) are identification conditions, which ensure that the population objective function is small if and only if evaluated in a neighborhood of the true parameter value. Assumption 5(4) will generally fail in contexts where weak or partial-identification issues arise. Provided these conditions hold, standard arguments yield the consistency of $\hat{\theta}$.

**Lemma 4** Under Assumption 5, we have that $\hat{\theta} \to_p \theta_0$ under all $\xi_0 \in \Xi_S$.

Next, we consider an assumption yielding asymptotic normality of $\hat{\theta}$ with the usual limiting distribution.

**Assumption 6** The following conditions hold for all $\xi_0 \in \Xi_S$

1. $\theta_0$ belongs to the interior of $\Theta$

2. $g_T(\theta)$ and $\hat{\Omega}(\theta)$ are almost surely continuously differentiable on some open ball $B(\theta_0)$ around $\theta_0$

3. For

$$J(\theta) = \lim_{T \to \infty} J_{T, \xi_0}(\theta) = \lim_{T \to \infty} E_T, \xi_0 \left[ \frac{\partial}{\partial \theta} g_T(\theta) \right],$$

$J(\theta)$ is continuous at $\theta_0$, $G_T(\theta) = \frac{\partial}{\partial \theta} g_T(\theta) \to_p J(\theta)$ uniformly on $B(\theta_0)$, and $J(\theta_0)$ is full-rank

4. $\sup_{\theta \in B(\theta_0)} \left\| \text{vec}\left( \frac{\partial^2 \hat{\Omega}(\theta)}{\partial \theta^2} \right) \right\| = O_p(1)$

5. $\hat{\Sigma}_g(\theta) \to_p \Sigma_g(\theta)$ uniformly on $B(\theta_0)$, and $\Sigma_g(\theta) = \lim_{T \to \infty} \text{Var}_{T, \xi_0}(\sqrt{T} g_T(\theta))$

is continuous in $\theta$ and everywhere positive-definite on $B(\theta_0)$
Assumption 6(1) rules out cases where the true parameter value lies near the boundary of the parameter space. Assumption 6(2) requires that the moment condition and weight function both be smooth, while (3) and (4) require that their derivatives be well-behaved. Finally, Assumption 6(5) requires that we have a uniformly consistent estimator for $\Sigma_g(\theta)$ on a neighborhood of $\theta_0$. Assumptions 2, 5, and 6 together yield the usual limiting distribution for the estimator $\hat{\theta}$ and the Wald statistic $W(\hat{\beta}_0) = W(f(\theta_0))$. In particular, for $W(\beta)$ as in (2.7), standard arguments yield $W(f(\theta_0)) \to d \chi^2_p$. Critically, these assumptions also imply the local asymptotic equivalence of the statistics $K_{\Omega,f}(\theta)$ and $W(f(\theta))$. Formally,

**Lemma 5** Let $\{A_{\theta,T}\}$ be a sequence of random sets such that

$$\limsup_{T \to \infty} \Pr\{A_{\theta,T} = \emptyset\} < 1$$

and

$$\sup_{\theta \in A_{\theta,T}} \|\theta - \theta_0\| = O_p\left(\frac{1}{\sqrt{T}}\right)$$

(where we define the sup to be zero if $A_{\theta,T}$ is empty). Under Assumptions 2, 5, and 6, under all $\xi \in \Xi_S$ we have

$$\sup_{\theta \in A_{\theta,T}} \|W(f(\theta)) - K_{\Omega,f}(\theta)\| = o_p(1).$$

Lemma 5 establishes that $K_{\Omega,f}(\theta)$ and $W(f(\theta))$ are locally asymptotically equivalent under strong identification, as suggested by our heuristic derivation in Section 2.2.2. Using this lemma, it is easy to verify Assumption 1(3).

**Theorem 9** Under Assumptions 2, 5, and 6, for $CSR_P$ as defined in (2.10) and $CS_{NR}$ as in (2.11) we have that

$$\inf_{\xi \in \Xi_S} \liminf_{T \to \infty} \Pr_{T,\xi}\{CSR_P \subseteq CS_{NR}\} = 1.$$ 

Thus, under strong identification Assumption 1(3) holds. Since one can easily verify (2.14) under Assumptions 2-6, the conditions of Theorem 7 hold for the approach
suggested in Section 2.2 in general GMM models.

2.7 Conclusion

In this paper we highlight that commonly-used approaches to detecting weak identification and constructing confidence sets can lead to highly misleading inferences. In particular we demonstrate that in linear IV two-step confidence sets based on the first-stage F-statistic can exhibit severe coverage distortions under heteroskedasticity. To remedy this issue we suggest an approach to detecting weak identification and constructing two-step confidence sets which we show controls coverage distortions in general GMM models. Applied to linear IV our approach yields two-step confidence sets that are competitive with those based on the first stage F-statistic under homoskedasticity, but which unlike conventional approaches also control coverage distortions under heteroskedasticity. While our results focus on the case where the non-robust component of our two-step confidence set is based on the Wald statistic, by the equivalence of the trinity of classical tests one can extend these results to Quasi-Likelihood ratio or distance metric confidence sets.

While our simulations focus on linear IV, which is the most common context in which two-step procedures are currently applied, our theoretical results apply far more broadly and offer a novel approach to assessing identification strength in a wide range of nonlinear and dynamic contexts where no alternative which controls coverage distortions for two-step confidence sets is currently available.
2.8 Appendix

Supplementary Appendix A provides proofs for the results stated in the paper. Supplementary Appendix B details the simulation design for the results reported in Sections 2.1 and 2.3. Supplementary Appendix C discusses an alternative approach to creating robust confidence sets which requires only $\chi^2$ critical values. Finally, Supplementary Appendix D tabulates weights $\tilde{a}(\gamma)$ and critical values $\tilde{H}^{-1}(1 - \alpha; \tilde{a}(\gamma), k, p)$ for calculating the robust confidence sets $CS_{R,p}$ and $CS_R$ defined in (2.10) and (2.13) in the main text.

Supplementary Appendix A: Proofs

Proof of Lemma 2

To prove (1), note that for any $\xi_0 \in \Xi$ and any $T$, we have

$$Pr_{T,\xi_0}\{\beta_{0,T} \in CS_{2S}\} \geq Pr_{T,\xi_0}\{\beta_{0,T} \in CS_R\} - Pr_{T,\xi_0}\{\phi_{ICS} = 0\}.$$ 

By (2.15) $SCP(CS_R, \Xi_W) \geq 1 - \alpha$, so Lemma 2(1) follows immediately from the definition of sequential coverage probability.

To prove (2), note that

$$Pr_{T,\xi_0}\{\beta_{0,T} \in CS_{2S}\} \geq Pr_{T,\xi_0}\{\beta_{0,T} \in CS_{NR}\} - Pr_{T,\xi_0}\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\},$$

and

$$Pr_{T,\xi_0}\{\beta_{0,T} \notin CS_R \text{ and } \phi_{ICS} = 1\} \leq \min\{Pr_{T,\xi_0}\{\beta_{0,T} \notin CS_R\}, Pr_{T,\xi_0}\{\phi_{ICS} = 1\}\}.$$ 

By (2.14) $SCP(CS_{NR}, \Xi_S) \geq 1 - \alpha$ so

$$SCP(CS_{2S}, \Xi_S) \geq 1 - \alpha - \sup_{\xi_0 \in \Xi_S} \lim_{T \to \infty} \sup_{T} \min\{Pr_{T,\xi_0}\{\beta_{0,T} \notin CS_R\}, Pr_{T,\xi_0}\{\phi_{ICS} = 1\}\}$$

but $\sup_{\xi_0 \in \Xi_S} \lim_{T \to \infty} Pr_{T,\xi_0}\{\beta_{0,T} \notin CS_R\} \leq \alpha$ by assumption, implying the result.
Proof of Theorem 7

To establish (1), note that by Assumption 1(2), \( Pr_{T,\xi_0} \{ CS_{R,P} \subseteq CS_R \} = 1 \) for all \( T \) and \( \xi_0 \in \Xi \). Thus by the definition of \( CS_{2S} \), \( Pr_{T,\xi_0} \{ CS_{R,P} \subseteq CS_{2S} \} = 1 \) for all \( T \) and \( \xi_0 \in \Xi \). Consequently, \( Pr_{T,\xi_0} \{ \beta_{0,T} \in CS_{R,P} \} \leq Pr_{T,\xi_0} \{ \beta_{0,T} \in CS_{2S} \} \), so (1) follows immediately from Assumption 1(1). (2) follows immediately from Lemma 2(2) and Assumption 1(3). (3) is implied by

\[
\sup_{\xi_0 \in \Xi_S} \lim_{T \to \infty} \sup_{\xi_0} Pr_{T,\xi_0} \{ \phi_{ICS} = 1 \} = 0,
\]

which is an immediate consequence of Assumption 1(3).

Proof of Lemma 3  Follows immediately from Lemma 1 of Kleibergen (2005).

Proof of Theorem 8  Note that we can re-write \( K_{\omega,f} \) as

\[
K_{\omega,f} (\theta) =
T g_T (\theta)' \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \left( \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \right)^{-1} \Lambda_{1,T} \frac{\partial}{\partial \theta} f (\theta) \Lambda_{2,T} \times
\left( \Lambda_{2,T} \frac{\partial}{\partial \theta} f (\theta) \Lambda_{1,T} \left( \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \right)^{-1} \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) \Sigma (\theta) \times
\hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \left( \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \right)^{-1} \Lambda_{1,T} \frac{\partial}{\partial \theta} f (\theta)' \Lambda_{2,T} \right)^{-1}
\Lambda_{2,T} \frac{\partial}{\partial \theta} f (\theta) \Lambda_{1,T} \left( \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \right)^{-1} \Lambda_{1,T} D_T (\theta)' \hat{\Omega} (\theta) g_T (\theta).
\]

By Lemma 3, \( \left( \sqrt{T} g_T (\theta_{0,T}) , \sqrt{T} \text{vec} (D_T (\theta_{0,T}) - J_{T,\xi}) \right) \) converges to \( (\psi_g , \psi_D) \) which are mutually independent. Since we have assumed that the elements of \( \Lambda_{1,T} \) are of order \( O(1) \), note that \( \frac{1}{\sqrt{T}} \Lambda_{1,T} = O(1) \), so \( \left( \sqrt{T} g_T (\theta_{0,T}) , D_T (\theta_{0,T}) \Lambda_{1,T} \right) \) will be asymptotically independent as well. In particular, \( \left( \sqrt{T} g_T (\theta_{0,T}) , D_T (\theta_{0,T}) \Lambda_{1,T} \right) \to_d (\psi_g , D) \) where \( \psi_g | D \sim N(0, \Sigma_g) \).

We can further re-write \( K_{\omega,f} (\theta) \) as

\[
T g_T (\theta)' \hat{\Sigma}_g (\theta) \left( \frac{1}{2} P \left( \hat{\Sigma}_g (\theta) \frac{1}{2} \hat{\Omega} (\theta) D_T (\theta) \Lambda_{1,T} \left( \Lambda_{1,T} D_T (\theta)' \right) \right) \Lambda_{1,T} \frac{\partial}{\partial \theta} f (\theta)' \Lambda_{2,T} \right)^{-\frac{1}{2}} g_T (\theta).
\]

140
where \( P(X) = X(X'X)^{-1}X' \) denotes the projection matrix onto \( X \). By Assumptions 3 and 4 and the Continuous Mapping Theorem,

\[
\hat{\Sigma}_g(\theta_0,T)^{1/2} \hat{\Omega}(\theta_0,T)|D_T(\theta_0,T)\Lambda_1,T \left( \Lambda'_1,T \right)^{-1} \hat{\Sigma}_g(\theta_0,T)^{1/2} \hat{\Omega}(\theta_0,T)D_T(\theta_0,T)\Lambda_2,T \twoheadrightarrow \sigma_g^2 \Omega D \left( D'\Omega \sigma_g \Omega D \right)^{-1} F
\]

where the sole random component on the right hand side is \( D \), and the right hand side has rank \( p \) almost surely. Together with the fact that \( \sigma_g^{-1/2} \psi_g |D \sim N(0, I_k) \), this implies by the Continuous Mapping Theorem that \( (K_{\Omega,f}(\theta_0,T), D_T(\theta_0,T)\Lambda_1,T) \twoheadrightarrow_d \left( \tilde{K}_{\Omega,f}, D \right) \) where \( \tilde{K}_{\Omega,f} |D \sim \chi^2_p \), since conditional on \( D \) \( \tilde{K}_{\Omega,f} \) is a quadratic form in a standard-normal random vector and a rank-\( p \) projection matrix.

We can handle \( S(\theta_0,T) - K_{\Omega,f}(\theta_0,T) \) in a similar manner. In particular, note that

\[
S(\theta) - K_{\Omega,f}(\theta) = T g_T(\theta)' \hat{\Sigma}_g(\theta)^{-1/2} \left( I - P \left( \hat{\Sigma}_g(\theta)^{1/2} \hat{\Omega}(\theta) D_T(\theta)\Lambda_1,T \left( \Lambda'_1,T \right)^{-1} \hat{\Sigma}_g(\theta)^{1/2} \hat{\Omega}(\theta) D_T(\theta)\Lambda_2,T \right) \right) \hat{\Sigma}_g(\theta)^{-1/2} g_T(\theta).
\]

so

\[
(K_{\Omega,f}(\theta_0,T), S(\theta_0,T) - K_{\Omega,f}(\theta_0,T), D_T(\theta_0,T)\Lambda_1,T) \twoheadrightarrow_d \left( \tilde{K}_{\Omega,f}, \tilde{S} - \tilde{K}_{\Omega,f}, D \right)
\]

where \( \left( \hat{K}_{\Omega,f}, \hat{S} - \hat{K}_{\Omega,f} \right) |D \sim \left( \chi^2_p, \chi^2_{k-p} \right) \) and \( \left( \hat{K}_{\Omega,f}, \hat{S} - \hat{K}_{\Omega,f} \right) \) are independent conditional on \( D \). Thus we have that \( \left( \hat{K}_{\Omega,f}, \hat{S} - \hat{K}_{\Omega,f} \right) \) and independent and distributed \( \left( \chi^2_p, \chi^2_{k-p} \right) \) unconditionally as well, which establishes the result.

**Proof of Lemma 4** The proof is standard but is included for completeness. By Assumption 5(1), (2), and (3), \( g_T(\theta_0)' \hat{\Omega} g_T(\theta_0) \rightarrow_p 0 \). Consider any \( \varepsilon > 0 \), and note that by Assumption 5(4),

\[
\inf_{||\theta' - \theta_0|| \geq \varepsilon} \left( \lim_{T \rightarrow \infty} E_{T,\xi_0} \left[ g_T(\theta) \right] \right)' \Omega(\theta) \left( \lim_{T \rightarrow \infty} E_{T,\xi_0} \left[ g_T(\theta) \right] \right)^{-1} > \delta
\]

for some \( \delta > 0 \). By Assumption 5(3), the minimal eigenvalue of \( \Omega(\theta) \) is uniformly bounded below by some \( \underline{\Lambda} > 0 \), so the minimal eigenvalue of \( \hat{\Omega}(\theta) \) is uniformly bounded below by some \( \hat{\underline{\Lambda}} \rightarrow_p \underline{\Lambda} \). Since we can make an analogous bounding argument
with respect to the maximal eigenvalue, by the continuous mapping theorem we have that

\[ Pr_{T,\xi} \left\{ \sup_{\theta \in \Theta} \left( \lim_{T \to \infty} E_{T,\xi} [g_t(\theta)] \right)' \Omega(\theta) \left( \lim_{T \to \infty} E_{T,\xi} [g_t(\theta)] \right) - g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) > \frac{\Delta}{3} \right\} \]

converges to zero. This in turn implies that

\[ Pr_{T,\xi} \left\{ \inf_{\|\theta_0\| \geq \varepsilon} g_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) < \frac{2\delta}{3} \right\} \to 0 \]

while by point-wise convergence

\[ Pr_{T,\xi} \left\{ g_T(\theta_0)' \hat{\Omega}(\theta_0) g_T(\theta_0) > \frac{\delta}{3} \right\} \to 0. \]

Hence, we have that \( Pr_{T,\xi} \{ \| \hat{\theta} - \theta_0 \| \geq \varepsilon \} \to 0. \) Since we can repeat this argument for any \( \varepsilon > 0, \) consistency follows immediately.

**Proof of Lemma 5** The proof is standard, but is included for completeness. If \( A_{\theta,T} \) is empty we have defined \( \sup_{\theta \in A_{\theta,T}} \| W(f(\theta)) - K_{\Omega,f}(\theta) \| = 0. \) Hence, we restrict attention to non-empty realizations of \( A_{\theta,T} \) and condition on \( A_{\theta,T} \neq \emptyset \) for the remainder of the analysis. We know that for \( B(\theta_0) \) as in Assumption 6, \( Pr \{ A_{\theta,T} \subset B(\theta_0) \} \to 1. \) Hence, by our assumptions, \( \sup_{\theta \in A_{\theta,T}} \| G_T(\theta) - J(\theta_0) \| = o_p(1), \)

\[ \sup_{\theta \in A_{\theta,T}} \| D_T(\theta) - J(\theta_0) \| = o_p(1) \]

and \( \sup_{\theta \in A_{\theta,T}} \| \hat{\Sigma}_g(\theta) - \Sigma_g(\theta_0) \| = o_p(1). \) By a mean value expansion

\[ \left( D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) = \left( D_T(\theta)' \hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \hat{\Omega}(\theta) (g_T(\theta_0) + G_T(\theta^*)(\theta - \theta_0)) \]

for \( \theta^* \) a value between \( \theta \) and \( \theta_0 \) which can vary across rows. By the Continuous Mapping Theorem and the fact that \( J(\theta) \) and \( \Omega(\theta) \) are continuous and full rank at
\[ \theta_0: \]
\[
\sup_{\theta \in A_{\theta,T}} \left\| \begin{pmatrix} D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \end{pmatrix}^{-1} D_T(\theta)'\hat{\Omega}(\theta) - \\
(J(\theta_0)'\Omega(\theta_0) J(\theta_0))^{-1} J(\theta_0)'\Omega(\theta_0) \right\| = o_p(1)
\]

while
\[
\sup_{\theta \in A_{\theta,T}} \left\| \begin{pmatrix} D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \end{pmatrix}^{-1} D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta^*) - I \right\| = o_p(1).
\]

Hence,
\[
\sup_{\theta \in A_{\theta,T}} \left\| \begin{pmatrix} D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \end{pmatrix}^{-1} D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta) \right. \\
- (J(\theta_0)'\Omega(\theta_0) J(\theta_0))^{-1} J(\theta_0)'\Omega(\theta_0) g_T(\theta_0) - (\theta - \theta_0) \left\| = o_p \left( \frac{1}{\sqrt{T}} \right) .
\]

Similar arguments establish that
\[
\sup_{\theta \in A_{\theta,T}} \left\| (J(\theta_0)'\Omega(\theta_0) J(\theta_0))^{-1} J(\theta_0)'\Omega(\theta_0) g_T(\theta_0) + (\hat{\theta} - \theta_0) \right\| = o_p \left( \frac{1}{\sqrt{T}} \right)
\]

and hence, by the triangle inequality, that
\[
\sup_{\theta \in A_{\theta,T}} \left\| \begin{pmatrix} D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \end{pmatrix}^{-1} D_T(\theta)'\hat{\Omega}(\theta) g_T(\theta) - (\hat{\theta} - \theta) \right\| = o_p \left( \frac{1}{\sqrt{T}} \right) .
\]

Since
\[
\sup_{\theta \in A_{\theta,T}} \left\| \begin{pmatrix} D_T(\hat{\theta})'\hat{\Omega}(\hat{\theta}) \hat{\Sigma}_g(\hat{\theta}) \hat{\Omega}(\hat{\theta}) D_T(\hat{\theta}) \end{pmatrix}^{-1} - \\
\left( G_T(\hat{\theta})'\hat{\Omega}(\hat{\theta}) \hat{\Sigma}_g(\hat{\theta}) \hat{\Omega}(\hat{\theta}) G_T(\hat{\theta}) \right)^{-1} \right\| = o_p(1)
\]

under our assumptions, this suffices to establish the desired equivalence for tests of the full parameter vector, which take \( f(\hat{\theta}) = \theta \). To complete the proof, we need only show that the same result holds for general \( f(\cdot) \), which follows from \( \Delta \)-method arguments. In particular, as noted in Van der Vaart (2000) Theorem 3.1, under our assumptions
\[
\left\| \sqrt{T} \left( f(\hat{\theta}) - f(\theta_0) \right) - \sqrt{T} \frac{\partial}{\partial \theta} f(\theta_0) (\hat{\theta} - \theta_0) \right\| = o_p(1).
\]

143
Since \( \sup_{\theta \in A_{\theta}, T} \| \theta - \theta_0 \| = o_p(1) \), by the definition of differentiability

\[
\sup_{\theta \in A_{\theta}, T} \left\| \frac{f(\theta) - f(\theta_0) - \frac{\partial}{\partial \theta_0} f(\theta_0) (\theta - \theta_0)}{\| \theta - \theta_0 \|} \right\| = o_p(1)
\]

which implies that

\[
\sup_{\theta \in A_{\theta}, T} \left\| \sqrt{T} \left( f(\theta) - f(\theta_0) \right) - \sqrt{T} \frac{\partial}{\partial \theta_0} f(\theta_0) (\theta - \theta_0) \right\| = o_p(1)
\]

and hence by the triangle inequality

\[
\left\| \sqrt{T} \left( f(\theta) - f(\theta_0) \right) - \sqrt{T} \frac{\partial}{\partial \theta_0} f(\theta_0) (\theta - \theta_0) \right\| = o_p(1),
\]

yielding the statement:

\[
\left\| \sqrt{T} \left( f(\theta) - f(\theta_0) \right) + \sqrt{T} \frac{\partial}{\partial \theta_0} f(\theta_0) \left( D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \hat{\Omega}(\theta) g_T(\theta) \right\| = o_p(1).
\]

Since \( \sup_{\theta \in A_{\theta}, T} \left\| \frac{\partial}{\partial \theta_0} f(\theta_0) - \frac{\partial}{\partial \theta_0} f(\theta_0) \right\| = o_p(1) \) and \( \frac{\partial}{\partial \theta_0} f(\theta_0) \) is full rank,

\[
\sup_{\theta \in A_{\theta}, T} \left\| \left( \frac{\partial}{\partial \theta_0} f(\theta) \left( D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \right)^{-1} D_T(\theta)' \right)^{-1} - \right.
\]

\[
\left. \hat{\Sigma}_g(\theta)' \hat{\Omega}(\theta) \hat{D}_T(\theta) \left( D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \right)^{-1} \frac{\partial}{\partial \theta_0} f(\theta)' \right\| = o_p(1)
\]

by the Continuous Mapping Theorem, and by the triangle inequality

\[
\sup_{\theta \in A_{\theta}, T} \left\| \left( \frac{\partial}{\partial \theta_0} f(\theta) \left( G_T(\theta)'\hat{\Omega}(\theta) G_T(\theta) \right)^{-1} G_T(\theta)' \hat{\Omega}(\theta) \right)' - \right.
\]

\[
\left. \hat{\Sigma}_g(\theta)' \hat{\Omega}(\theta) \hat{G}_T(\theta) \left( G_T(\theta)'\hat{\Omega}(\theta) G_T(\theta) \right)^{-1} \frac{\partial}{\partial \theta_0} f(\theta)' \right\| = o_p(1)
\]

\[
\left. \hat{\Omega}(\theta) D_T(\theta) \left( D_T(\theta)'\hat{\Omega}(\theta) D_T(\theta) \right)^{-1} d_T(\theta)' \hat{\Omega}(\theta) \right\| = o_p(1).
\]

Hence \( \sup_{\theta \in A_{\theta}, T} \| W(\theta) - K_{\Omega, f}(\theta) \| = o_p(1) \), so the \( K_{\Omega, f} \) and Wald statistics are first-order equivalent on \( A_{\theta, T} \) as we wanted to show.
Proof of Theorem 9  For \( S(\theta) \) as in (2.3), note that Assumption 5 implies that for any \( \varepsilon > 0, \)

\[
\inf_{\|\theta - \theta_0\| \geq \varepsilon} S(\theta) \to_p \infty.
\]

Thus, if we define \( A_{\theta,T} = \left\{ \theta : a \cdot S(\theta) \leq \chi^2_{p,1-\alpha} \right\}, \) we can see that \( \sup_{\theta \in A_{\theta,T}} \|\theta - \theta_0\| = o_p(1). \) A mean-value expansion yields that \( g_T(\theta) = g_T(\theta_0) + G_T(\theta^*)(\theta - \theta_0). \) Since \( \sup_{\theta \in B(\theta_0)} \|G_T(\theta) - J(\theta)\| = o_p(1), \) and \( \sup_{\theta \in B(\theta_0)} \|\hat{\Sigma}_g(\theta) - \Sigma_g(\theta)\| = o_p(1) \) for an open ball \( B(\theta_0) \) around \( \theta_0 \) as in Assumption 6 and \( J(\theta) \) and \( \Sigma_g(\theta) \) are continuous in \( \theta, \) we can see that

\[
\sup_{\theta \in A_{\theta,T}} \left| S(\theta) - T(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0))' \Sigma_g(\theta_0)^{-1}(g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)) \right| = o_p(1).
\]

Thus, for any \( \varepsilon > 0 \) we can see that for \( A \) the minimal eigenvalue of \( \Sigma_g(\theta_0)^{-1}, \)

\[
Pr_{T,\xi} \left\{ \inf_{\theta \in A_{\theta,T}} \left( S(\theta) - \Delta T \|g_T(\theta_0) + J(\theta_0)(\theta - \theta_0)\|^2 \right) > -\varepsilon \right\} \to 1.
\]

Since \( \sqrt{T} g_T(\theta_0) = O_p(1) \) by Assumption 2, this implies that \( \sup_{\theta \in A_{\theta,T}} \|\theta - \theta_0\| = o_p\left(\frac{1}{\sqrt{T}}\right). \) Thus, we have established that \( A_{\theta,T} = \left\{ \theta : a \cdot S(\theta) \leq \chi^2_{p,1-\alpha} \right\} \) shrinks towards \( \theta_0 \) at rate \( \sqrt{T}. \)

Next, note that \( K_{\Omega,f}(\theta) \geq 0 \) by construction, so \( K_{\Omega,f}(\theta) + a \cdot S(\theta) \geq a \cdot S(\theta) \) and \( CS_{R,P} \subseteq A_{\theta,T}. \) By standard results on the distribution of tests for over-identifying restrictions \( \inf_{\theta} S(\theta) \to_d \chi^2_{k-p_w}, \) so since

\[
K_{\Omega,f}(\theta) + a \cdot S(\theta) \geq K_{\Omega,f}(\theta) + a \cdot \inf_{\theta} S(\theta)
\]

and by Lemma 5 we know that \( \sup_{\theta \in A_{\theta,T}} |K_{\Omega,f}(\theta) - W(f(\theta))| = o_p(1), \) we obtain

\[
Pr_{T,\xi} \left\{ \inf_{\theta \in A_{\theta,T}} (K_{\Omega,f}(\theta) + a \cdot S(\theta) - W(f(\theta))) > 0 \right\} \to 1
\]

with the consequence that \( Pr_{T,\xi} \{CS_{R,P} \subseteq CS_{NR}\} \to 1, \) as we wanted to show.
Supplementary Appendix B: IV Simulation Design

Heteroskedastic Case To examine the behavior of two-step confidence sets in simulation we need to specify the process generating \((Z, V_1, V_2)\). Our focus is on heteroskedasticity, so we consider models where \((Z_t, V_{1,t}, V_{2,t})\) are independent across \(t\) but where \(\text{Var}(V_{1,t}, V_{2,t}|Z_t)\) may depend on \(Z_t\). We consider a categorical instrument and let \(Z_t\) be a collection of dummy variables for different values of \(Z_t \in \{1, \ldots, k\}\), so \(Z_t \in \{e_1, \ldots, e_k\}\) where \(e_i\) is the \(k \times 1\) vector with 1 in the \(i\)th entry and zeros everywhere else. We take \(Z_t\) to be uniformly distributed so that \(\Pr \{Z_t = e_i\} = \frac{1}{k}\) for all \(i \in \{1, \ldots, k\}\) and set the true parameter value \(\beta_0 = 0\).

Since the support of \(Z_t\) is finite we can model \(\text{Var}(V_{1,t}, V_{2,t}|Z_t)\) fully flexibly and take

\[
\begin{pmatrix}
V_{1,t} \\
V_{2,t}
\end{pmatrix}
|Z_t \sim N(0, \Sigma_V(Z_t)).
\]

To explore the behavior of the model for different parameter values we drew at random many values of \(\Sigma_V(Z_t)\) and the direction of the first stage \(\pi/\|\pi\|\). For each draw we considered a large range of values for \(\|\pi\|\), ranging from non-identification to very strong identification, and for our simulations we focus on particular draws of \(\Sigma_V(Z_t)\) and \(\pi/\|\pi\|\) that generate large coverage distortions for some values of \(\|\pi\|\).

We study models with five, ten, and twenty instruments \((k \in \{5, 10, 20\})\) and in each case consider two calibrations, one with a very high degree of endogeneity as measured by the correlation between the errors \(V_{1,t}\) and \(V_{2,t}\) (H), and the other with more moderate endogeneity (M). The space of possible covariance structures is extremely large, however, so there likely exist parameter values generating much more pathological behavior for non-robust procedures than we report here. Consequently, our results give only lower bounds for possible coverage distortions. In all cases we simulate samples of 10,000 observations.

To give a sense of the parameter values used in our simulations, in Table 2.7 we report the (unconditional) correlation between \(V_{1,t}\) and \(V_{2,t}\) as well as

\[\text{Stdev} \left(\text{Stdev}(V_{1,t}|Z_t)\right) / \text{Stdev}(V_{1,t})\],

which is a natural measure for the degree of heteroskedasticity.
Table 2.7: Summary of linear IV calibration values. Note that $\text{Corr} (V_{1,t}, V_{2,t})$ is in all cases strictly less than one in absolute value, but the reported value is rounded to nearest 0.01.

<table>
<thead>
<tr>
<th></th>
<th>Medium Endogeneity (M)</th>
<th>High Endogeneity (H)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 5$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td>$\text{Corr} (V_{1,t}, V_{2,t})$</td>
<td>-0.66</td>
<td>-0.59</td>
</tr>
<tr>
<td>$\frac{\text{Stdev} (\text{Stdev}(V_{1,t}, Z_t))}{\text{Stdev}(V_{1,t})}$</td>
<td>0.40</td>
<td>0.53</td>
</tr>
<tr>
<td>$\frac{\text{Stdev} (\text{Stdev}(V_{2,t}, Z_t))}{\text{Stdev}(V_{2,t})}$</td>
<td>0.63</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Homoskedastic Case

For the homoskedastic case, we consider the same simulation calibrations described above, except that in each case we eliminate heteroskedasticity by taking $V_{1,t}$, $V_{2,t}$ to be independent of $Z_t$ with

$$
\begin{pmatrix}
V_{1,t} \\
V_{2,t}
\end{pmatrix} \sim N(0, \Sigma_V(Z_t)).
$$

Supplementary Appendix C: Alternative Robust Confidence Sets

The robust confidence sets $CS_{R,p}$ and $CS_R$ defined in (2.10) and (2.13) represent only one of many ways to construct robust confidence sets which satisfy Assumption 1 and that can thus be used to construct two-step confidence sets. While we find that this approach performs quite well in simulation, it has the disadvantage of requiring that we calculate appropriate weights $u(\gamma)$ and critical values $H^{-1}(1 - \alpha; \hat{u}(\gamma), k, p)$ (though we tabulate these for many values of $(\alpha, \gamma, k, p)$ in Supplementary Appendix D). In this section, we discuss an alternative approach to constructing robust confidence sets, based on a generalization of the JK tests suggested by Kleibergen (2005), which requires only $\chi^2$ critical values.

Like the confidence sets $CS_{R,p}$ and $CS_R$ discussed in the main text, JK confidence sets are based on the fact that, evaluated at the true parameter value, $K_{\Omega,f}(\theta_0)$ and $S(\theta_0) - K_{\Omega,f}(\theta_0)$ are asymptotically independent and distributed $\chi^2_k$ and $\chi^2_{k-p}$. To construct the JK confidence set for $\beta$, define $J_{\Omega,f}(\theta) = S(\theta) - K_{\Omega,f}(\theta)$ and for
constants $\alpha_K, \alpha_J \in (0, 1)$ let

$$CS_{JK} = \{\beta : \exists \theta \text{ s.t. } \beta = f(\theta), K_{\alpha_J}(\theta) \leq \chi^2_{1-\alpha_J}, \text{ and } J_{\alpha_J}(\theta) \leq \chi^2_{1-\alpha_J} \}$$

or, equivalently

$$CS_{JK} = \{\beta : \min_{\theta: \beta=f(\theta)} \max\{K_{\alpha_J}(\theta) - \chi^2_{1-\alpha_J}, J_{\alpha_J}(\theta) - \chi^2_{1-\alpha_J} \} \leq 0 \}.$$

Under Assumptions 2-6 one can show using the proofs of Theorems 8 and 9 that the asymptotic coverage of this robust confidence set is at least $(1 - \alpha_J)(1 - \alpha_K) = 1 - \alpha_J - \alpha_K + \alpha_J \alpha_K$ under both weak and strong identification.

To construct a preliminary confidence set $CS_{JK,P}$ such that $CS_{R,P} = CS_{JK,P}$ will satisfy Assumption 1, let $CS_{JK,P}$ be the JK confidence set based on $(\alpha_{J,P}, \alpha_{K,P})$ satisfying $(1 - \alpha_{J,P})(1 - \alpha_{K,P}) = 1 - \alpha - \gamma$ and $\alpha_{K,P} > \alpha$. $CS_{JK,P}$ will have sequential coverage at least $1 - \alpha - \gamma$ under weak identification,

$$SCP(CS_{JK,P}, \Xi_W) \geq 1 - \alpha - \gamma$$

and will be contained in the non-robust confidence set with probability tending to one under strong identification

$$\inf_{\zeta \in \Xi_s} \Pr_{\tau, \zeta} \{CS_{JK,P} \subseteq CS_{NR} \} \to 1$$

for $CS_{NR}$ as in (2.11). Thus, this choice satisfies Assumption 1(1) and (3). If we define $CS_R = CS_{JK}$ for $\alpha_K \leq \alpha_{K,P}$, $\alpha_J \leq \alpha_{J,P}$, and $(1 - \alpha_J)(1 - \alpha_K) = 1 - \alpha$, we can see that $CS_{JK,P} \subseteq CS_{JK}$ for all realizations of the data, so Assumption 1(2) is satisfied as well. Thus, under the assumptions of Theorem 9 the choice $CS_{R,P} = CS_{JK,P}$ and $CS_R = CS_{JK}$ satisfies Assumption 1 and thus yields two-step confidence sets which control coverage distortions. Simulation results based on the same calibrations studied in Section 2.3 (available from the author upon request) show that two-step procedures based on JK confidence sets with $\alpha_K = 0.8 \cdot \alpha$ and $\alpha_{K,P} = 0.8 (\alpha + \gamma)$.
perform comparably to the approach developed in the main text for $\alpha = 5\%$ and $\gamma = 10\%$.

**Supplementary Appendix D: Critical Values**

Tables 2.8-2.16 report weights $\hat{a}(\gamma)$ for constructing robust confidence sets $CS_{R,P}$ and $CS_R$ as in (2.10) and (2.13), together with critical values $\hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ for calculating $CS_R$. These values are calculated following the approach discussed in Section 2.2.4 based on ten million simulation draws ($M = 10^7$). Each table corresponds to a given pair $(\alpha, p)$ for $\alpha \in \{0.01, 0.05, 0.1\}$ and $p \in \{1, 2, 3\}$ and reports values for number of moment conditions $k \in \{1, 2, \ldots, 30\}$ and maximal distortion $\gamma \in \{0.01, 0.05, 0.1, 0.15\}$. 
\[ Y = 0.01 \]
\[ -y = 0.05 \]
\[ Y = 0.1 \]
\[ -y = 0.15 \]

\[ k \quad \hat{a}(\gamma) \quad \text{Crit Val} \quad \hat{a}(\gamma) \quad \text{Crit Val} \quad \hat{a}(\gamma) \quad \text{Crit Val} \quad \hat{a}(\gamma) \quad \text{Crit Val} \]

<table>
<thead>
<tr>
<th></th>
<th>[ \gamma = 0.01 ]</th>
<th>[ \gamma = 0.05 ]</th>
<th>[ \gamma = 0.1 ]</th>
<th>[ \gamma = 0.15 ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.221</td>
<td>8.15</td>
<td>0.871</td>
<td>12.48</td>
</tr>
<tr>
<td>2</td>
<td>0.185</td>
<td>8.10</td>
<td>0.623</td>
<td>11.63</td>
</tr>
<tr>
<td>3</td>
<td>0.159</td>
<td>8.08</td>
<td>0.498</td>
<td>11.24</td>
</tr>
<tr>
<td>4</td>
<td>0.140</td>
<td>8.06</td>
<td>0.419</td>
<td>10.99</td>
</tr>
<tr>
<td>5</td>
<td>0.124</td>
<td>8.03</td>
<td>0.363</td>
<td>10.82</td>
</tr>
<tr>
<td>6</td>
<td>0.113</td>
<td>8.01</td>
<td>0.321</td>
<td>10.67</td>
</tr>
<tr>
<td>7</td>
<td>0.102</td>
<td>7.99</td>
<td>0.289</td>
<td>10.58</td>
</tr>
<tr>
<td>8</td>
<td>0.094</td>
<td>7.99</td>
<td>0.263</td>
<td>10.52</td>
</tr>
<tr>
<td>9</td>
<td>0.087</td>
<td>7.98</td>
<td>0.241</td>
<td>10.45</td>
</tr>
<tr>
<td>10</td>
<td>0.081</td>
<td>7.97</td>
<td>0.223</td>
<td>10.40</td>
</tr>
<tr>
<td>11</td>
<td>0.076</td>
<td>7.96</td>
<td>0.207</td>
<td>10.34</td>
</tr>
<tr>
<td>12</td>
<td>0.071</td>
<td>7.95</td>
<td>0.193</td>
<td>10.30</td>
</tr>
<tr>
<td>13</td>
<td>0.067</td>
<td>7.95</td>
<td>0.182</td>
<td>10.26</td>
</tr>
<tr>
<td>14</td>
<td>0.064</td>
<td>7.95</td>
<td>0.171</td>
<td>10.23</td>
</tr>
<tr>
<td>15</td>
<td>0.060</td>
<td>7.94</td>
<td>0.162</td>
<td>10.20</td>
</tr>
<tr>
<td>16</td>
<td>0.057</td>
<td>7.94</td>
<td>0.154</td>
<td>10.18</td>
</tr>
<tr>
<td>17</td>
<td>0.055</td>
<td>7.94</td>
<td>0.146</td>
<td>10.15</td>
</tr>
<tr>
<td>18</td>
<td>0.053</td>
<td>7.94</td>
<td>0.140</td>
<td>10.13</td>
</tr>
<tr>
<td>19</td>
<td>0.050</td>
<td>7.93</td>
<td>0.134</td>
<td>10.12</td>
</tr>
<tr>
<td>20</td>
<td>0.048</td>
<td>7.93</td>
<td>0.128</td>
<td>10.10</td>
</tr>
<tr>
<td>21</td>
<td>0.046</td>
<td>7.93</td>
<td>0.123</td>
<td>10.10</td>
</tr>
<tr>
<td>22</td>
<td>0.045</td>
<td>7.92</td>
<td>0.118</td>
<td>10.08</td>
</tr>
<tr>
<td>23</td>
<td>0.043</td>
<td>7.93</td>
<td>0.114</td>
<td>10.07</td>
</tr>
<tr>
<td>24</td>
<td>0.041</td>
<td>7.93</td>
<td>0.110</td>
<td>10.06</td>
</tr>
<tr>
<td>25</td>
<td>0.040</td>
<td>7.93</td>
<td>0.106</td>
<td>10.05</td>
</tr>
<tr>
<td>26</td>
<td>0.039</td>
<td>7.93</td>
<td>0.102</td>
<td>10.04</td>
</tr>
<tr>
<td>27</td>
<td>0.038</td>
<td>7.93</td>
<td>0.099</td>
<td>10.04</td>
</tr>
<tr>
<td>28</td>
<td>0.036</td>
<td>7.92</td>
<td>0.096</td>
<td>10.03</td>
</tr>
<tr>
<td>29</td>
<td>0.035</td>
<td>7.92</td>
<td>0.093</td>
<td>10.02</td>
</tr>
<tr>
<td>30</td>
<td>0.034</td>
<td>7.92</td>
<td>0.090</td>
<td>10.01</td>
</tr>
</tbody>
</table>

Table 2.8: Weights \( \hat{a}(\gamma) \) and critical values \( \hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p) \) for \( p = 1 \) and \( \alpha = 0.01 \).
<table>
<thead>
<tr>
<th>$k$</th>
<th>$\hat{a}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{a}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{a}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{a}(\gamma)$</th>
<th>Crit Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.085</td>
<td>4.17</td>
<td>0.419</td>
<td>5.45</td>
<td>0.852</td>
<td>7.12</td>
<td>1.338</td>
<td>8.99</td>
</tr>
<tr>
<td>2</td>
<td>0.066</td>
<td>4.17</td>
<td>0.289</td>
<td>5.30</td>
<td>0.510</td>
<td>6.48</td>
<td>0.710</td>
<td>7.60</td>
</tr>
<tr>
<td>3</td>
<td>0.054</td>
<td>4.16</td>
<td>0.226</td>
<td>5.23</td>
<td>0.378</td>
<td>6.24</td>
<td>0.504</td>
<td>7.12</td>
</tr>
<tr>
<td>4</td>
<td>0.046</td>
<td>4.16</td>
<td>0.186</td>
<td>5.18</td>
<td>0.304</td>
<td>6.11</td>
<td>0.396</td>
<td>6.86</td>
</tr>
<tr>
<td>5</td>
<td>0.040</td>
<td>4.16</td>
<td>0.159</td>
<td>5.15</td>
<td>0.256</td>
<td>6.02</td>
<td>0.329</td>
<td>6.70</td>
</tr>
<tr>
<td>6</td>
<td>0.035</td>
<td>4.16</td>
<td>0.138</td>
<td>5.13</td>
<td>0.221</td>
<td>5.96</td>
<td>0.282</td>
<td>6.59</td>
</tr>
<tr>
<td>7</td>
<td>0.031</td>
<td>4.16</td>
<td>0.123</td>
<td>5.11</td>
<td>0.196</td>
<td>5.92</td>
<td>0.247</td>
<td>6.52</td>
</tr>
<tr>
<td>8</td>
<td>0.028</td>
<td>4.15</td>
<td>0.111</td>
<td>5.10</td>
<td>0.175</td>
<td>5.88</td>
<td>0.221</td>
<td>6.46</td>
</tr>
<tr>
<td>9</td>
<td>0.026</td>
<td>4.15</td>
<td>0.101</td>
<td>5.09</td>
<td>0.159</td>
<td>5.86</td>
<td>0.199</td>
<td>6.41</td>
</tr>
<tr>
<td>10</td>
<td>0.024</td>
<td>4.15</td>
<td>0.092</td>
<td>5.08</td>
<td>0.146</td>
<td>5.84</td>
<td>0.182</td>
<td>6.37</td>
</tr>
<tr>
<td>11</td>
<td>0.022</td>
<td>4.15</td>
<td>0.085</td>
<td>5.07</td>
<td>0.134</td>
<td>5.82</td>
<td>0.167</td>
<td>6.34</td>
</tr>
<tr>
<td>12</td>
<td>0.021</td>
<td>4.15</td>
<td>0.079</td>
<td>5.06</td>
<td>0.125</td>
<td>5.80</td>
<td>0.155</td>
<td>6.32</td>
</tr>
<tr>
<td>13</td>
<td>0.019</td>
<td>4.15</td>
<td>0.074</td>
<td>5.06</td>
<td>0.116</td>
<td>5.79</td>
<td>0.145</td>
<td>6.30</td>
</tr>
<tr>
<td>14</td>
<td>0.018</td>
<td>4.15</td>
<td>0.069</td>
<td>5.05</td>
<td>0.109</td>
<td>5.78</td>
<td>0.136</td>
<td>6.28</td>
</tr>
<tr>
<td>15</td>
<td>0.017</td>
<td>4.15</td>
<td>0.065</td>
<td>5.05</td>
<td>0.103</td>
<td>5.77</td>
<td>0.127</td>
<td>6.26</td>
</tr>
<tr>
<td>16</td>
<td>0.016</td>
<td>4.15</td>
<td>0.062</td>
<td>5.04</td>
<td>0.097</td>
<td>5.76</td>
<td>0.120</td>
<td>6.25</td>
</tr>
<tr>
<td>17</td>
<td>0.015</td>
<td>4.15</td>
<td>0.059</td>
<td>5.04</td>
<td>0.092</td>
<td>5.75</td>
<td>0.114</td>
<td>6.24</td>
</tr>
<tr>
<td>18</td>
<td>0.015</td>
<td>4.15</td>
<td>0.056</td>
<td>5.04</td>
<td>0.087</td>
<td>5.75</td>
<td>0.108</td>
<td>6.23</td>
</tr>
<tr>
<td>19</td>
<td>0.014</td>
<td>4.15</td>
<td>0.053</td>
<td>5.04</td>
<td>0.083</td>
<td>5.74</td>
<td>0.103</td>
<td>6.22</td>
</tr>
<tr>
<td>20</td>
<td>0.013</td>
<td>4.15</td>
<td>0.051</td>
<td>5.03</td>
<td>0.079</td>
<td>5.74</td>
<td>0.098</td>
<td>6.21</td>
</tr>
<tr>
<td>21</td>
<td>0.013</td>
<td>4.15</td>
<td>0.048</td>
<td>5.03</td>
<td>0.076</td>
<td>5.73</td>
<td>0.094</td>
<td>6.20</td>
</tr>
<tr>
<td>22</td>
<td>0.012</td>
<td>4.15</td>
<td>0.046</td>
<td>5.03</td>
<td>0.073</td>
<td>5.73</td>
<td>0.090</td>
<td>6.20</td>
</tr>
<tr>
<td>23</td>
<td>0.012</td>
<td>4.15</td>
<td>0.045</td>
<td>5.03</td>
<td>0.070</td>
<td>5.72</td>
<td>0.087</td>
<td>6.19</td>
</tr>
<tr>
<td>24</td>
<td>0.011</td>
<td>4.15</td>
<td>0.043</td>
<td>5.02</td>
<td>0.067</td>
<td>5.72</td>
<td>0.083</td>
<td>6.18</td>
</tr>
<tr>
<td>25</td>
<td>0.011</td>
<td>4.15</td>
<td>0.041</td>
<td>5.02</td>
<td>0.065</td>
<td>5.71</td>
<td>0.080</td>
<td>6.18</td>
</tr>
<tr>
<td>26</td>
<td>0.011</td>
<td>4.15</td>
<td>0.040</td>
<td>5.02</td>
<td>0.063</td>
<td>5.71</td>
<td>0.078</td>
<td>6.18</td>
</tr>
<tr>
<td>27</td>
<td>0.010</td>
<td>4.15</td>
<td>0.039</td>
<td>5.02</td>
<td>0.060</td>
<td>5.71</td>
<td>0.075</td>
<td>6.17</td>
</tr>
<tr>
<td>28</td>
<td>0.010</td>
<td>4.15</td>
<td>0.037</td>
<td>5.02</td>
<td>0.058</td>
<td>5.70</td>
<td>0.072</td>
<td>6.17</td>
</tr>
<tr>
<td>29</td>
<td>0.010</td>
<td>4.15</td>
<td>0.036</td>
<td>5.02</td>
<td>0.057</td>
<td>5.70</td>
<td>0.070</td>
<td>6.16</td>
</tr>
<tr>
<td>30</td>
<td>0.009</td>
<td>4.15</td>
<td>0.035</td>
<td>5.02</td>
<td>0.055</td>
<td>5.70</td>
<td>0.068</td>
<td>6.16</td>
</tr>
</tbody>
</table>

Table 2.9: Weights $\hat{a}(\gamma)$ and critical values $\tilde{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ for $p = 1$ and $\alpha = 0.05$. 

151
<table>
<thead>
<tr>
<th></th>
<th>$\gamma = 0.01$</th>
<th></th>
<th>$\gamma = 0.05$</th>
<th></th>
<th>$\gamma = 0.1$</th>
<th></th>
<th>$\gamma = 0.15$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$\hat{a}(\gamma)$</td>
<td>Crit Val</td>
<td>$\hat{a}(\gamma)$</td>
<td>Crit Val</td>
<td>$\hat{a}(\gamma)$</td>
<td>Crit Val</td>
<td>$\hat{a}(\gamma)$</td>
<td>Crit Val</td>
</tr>
<tr>
<td>1</td>
<td>0.058</td>
<td>2.87</td>
<td>0.305</td>
<td>3.53</td>
<td>0.647</td>
<td>4.46</td>
<td>1.043</td>
<td>5.53</td>
</tr>
<tr>
<td>2</td>
<td>0.042</td>
<td>2.86</td>
<td>0.196</td>
<td>3.46</td>
<td>0.359</td>
<td>4.13</td>
<td>0.510</td>
<td>4.79</td>
</tr>
<tr>
<td>3</td>
<td>0.033</td>
<td>2.86</td>
<td>0.147</td>
<td>3.43</td>
<td>0.258</td>
<td>4.02</td>
<td>0.352</td>
<td>4.55</td>
</tr>
<tr>
<td>4</td>
<td>0.027</td>
<td>2.86</td>
<td>0.118</td>
<td>3.41</td>
<td>0.204</td>
<td>3.96</td>
<td>0.273</td>
<td>4.43</td>
</tr>
<tr>
<td>5</td>
<td>0.023</td>
<td>2.86</td>
<td>0.099</td>
<td>3.40</td>
<td>0.170</td>
<td>3.93</td>
<td>0.224</td>
<td>4.36</td>
</tr>
<tr>
<td>6</td>
<td>0.020</td>
<td>2.86</td>
<td>0.085</td>
<td>3.39</td>
<td>0.145</td>
<td>3.90</td>
<td>0.191</td>
<td>4.31</td>
</tr>
<tr>
<td>7</td>
<td>0.017</td>
<td>2.86</td>
<td>0.075</td>
<td>3.38</td>
<td>0.127</td>
<td>3.89</td>
<td>0.166</td>
<td>4.27</td>
</tr>
<tr>
<td>8</td>
<td>0.016</td>
<td>2.86</td>
<td>0.067</td>
<td>3.38</td>
<td>0.114</td>
<td>3.87</td>
<td>0.148</td>
<td>4.25</td>
</tr>
<tr>
<td>9</td>
<td>0.014</td>
<td>2.86</td>
<td>0.061</td>
<td>3.38</td>
<td>0.102</td>
<td>3.86</td>
<td>0.133</td>
<td>4.23</td>
</tr>
<tr>
<td>10</td>
<td>0.013</td>
<td>2.86</td>
<td>0.055</td>
<td>3.37</td>
<td>0.093</td>
<td>3.86</td>
<td>0.121</td>
<td>4.22</td>
</tr>
<tr>
<td>11</td>
<td>0.012</td>
<td>2.86</td>
<td>0.051</td>
<td>3.37</td>
<td>0.086</td>
<td>3.85</td>
<td>0.111</td>
<td>4.21</td>
</tr>
<tr>
<td>12</td>
<td>0.011</td>
<td>2.86</td>
<td>0.047</td>
<td>3.37</td>
<td>0.079</td>
<td>3.84</td>
<td>0.103</td>
<td>4.20</td>
</tr>
<tr>
<td>13</td>
<td>0.010</td>
<td>2.86</td>
<td>0.044</td>
<td>3.37</td>
<td>0.074</td>
<td>3.84</td>
<td>0.095</td>
<td>4.19</td>
</tr>
<tr>
<td>14</td>
<td>0.010</td>
<td>2.86</td>
<td>0.041</td>
<td>3.36</td>
<td>0.069</td>
<td>3.83</td>
<td>0.089</td>
<td>4.18</td>
</tr>
<tr>
<td>15</td>
<td>0.009</td>
<td>2.86</td>
<td>0.038</td>
<td>3.36</td>
<td>0.065</td>
<td>3.83</td>
<td>0.084</td>
<td>4.18</td>
</tr>
<tr>
<td>16</td>
<td>0.008</td>
<td>2.86</td>
<td>0.036</td>
<td>3.36</td>
<td>0.061</td>
<td>3.83</td>
<td>0.079</td>
<td>4.17</td>
</tr>
<tr>
<td>17</td>
<td>0.008</td>
<td>2.86</td>
<td>0.034</td>
<td>3.36</td>
<td>0.058</td>
<td>3.82</td>
<td>0.075</td>
<td>4.17</td>
</tr>
<tr>
<td>18</td>
<td>0.008</td>
<td>2.86</td>
<td>0.032</td>
<td>3.36</td>
<td>0.055</td>
<td>3.82</td>
<td>0.071</td>
<td>4.16</td>
</tr>
<tr>
<td>19</td>
<td>0.007</td>
<td>2.86</td>
<td>0.031</td>
<td>3.36</td>
<td>0.052</td>
<td>3.82</td>
<td>0.067</td>
<td>4.16</td>
</tr>
<tr>
<td>20</td>
<td>0.007</td>
<td>2.86</td>
<td>0.029</td>
<td>3.36</td>
<td>0.050</td>
<td>3.82</td>
<td>0.064</td>
<td>4.16</td>
</tr>
<tr>
<td>21</td>
<td>0.007</td>
<td>2.86</td>
<td>0.028</td>
<td>3.36</td>
<td>0.047</td>
<td>3.82</td>
<td>0.061</td>
<td>4.15</td>
</tr>
<tr>
<td>22</td>
<td>0.006</td>
<td>2.86</td>
<td>0.027</td>
<td>3.36</td>
<td>0.045</td>
<td>3.81</td>
<td>0.059</td>
<td>4.15</td>
</tr>
<tr>
<td>23</td>
<td>0.006</td>
<td>2.86</td>
<td>0.026</td>
<td>3.36</td>
<td>0.043</td>
<td>3.81</td>
<td>0.056</td>
<td>4.15</td>
</tr>
<tr>
<td>24</td>
<td>0.006</td>
<td>2.86</td>
<td>0.025</td>
<td>3.35</td>
<td>0.042</td>
<td>3.81</td>
<td>0.054</td>
<td>4.15</td>
</tr>
<tr>
<td>25</td>
<td>0.006</td>
<td>2.86</td>
<td>0.024</td>
<td>3.35</td>
<td>0.040</td>
<td>3.81</td>
<td>0.052</td>
<td>4.14</td>
</tr>
<tr>
<td>26</td>
<td>0.005</td>
<td>2.86</td>
<td>0.023</td>
<td>3.35</td>
<td>0.039</td>
<td>3.81</td>
<td>0.050</td>
<td>4.14</td>
</tr>
<tr>
<td>27</td>
<td>0.005</td>
<td>2.86</td>
<td>0.022</td>
<td>3.35</td>
<td>0.037</td>
<td>3.81</td>
<td>0.048</td>
<td>4.14</td>
</tr>
<tr>
<td>28</td>
<td>0.005</td>
<td>2.86</td>
<td>0.021</td>
<td>3.35</td>
<td>0.036</td>
<td>3.80</td>
<td>0.047</td>
<td>4.14</td>
</tr>
<tr>
<td>29</td>
<td>0.005</td>
<td>2.86</td>
<td>0.021</td>
<td>3.35</td>
<td>0.035</td>
<td>3.80</td>
<td>0.045</td>
<td>4.14</td>
</tr>
<tr>
<td>30</td>
<td>0.005</td>
<td>2.86</td>
<td>0.020</td>
<td>3.35</td>
<td>0.034</td>
<td>3.80</td>
<td>0.044</td>
<td>4.14</td>
</tr>
</tbody>
</table>

Table 2.10: Weights $\hat{a}(\gamma)$ and critical values $\hat{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ for $p = 1$ and $\alpha = 0.1$. 

152
\[
Y = 0.01 \\
\gamma = 0.05 \\
\gamma = 0.1 \\
\gamma = 0.15
\]

\[
k & \hat{d}(\gamma) & \text{Crit Val} & \hat{d}(\gamma) & \text{Crit Val} & \hat{d}(\gamma) & \text{Crit Val} & \hat{d}(\gamma) & \text{Crit Val} \\
2 & 0.176 & 10.85 & 0.637 & 15.10 & 1.086 & 19.25 & 1.513 & 23.18 \\
3 & 0.155 & 10.82 & 0.523 & 14.68 & 0.838 & 18.06 & 1.110 & 21.01 \\
4 & 0.138 & 10.79 & 0.447 & 14.41 & 0.690 & 17.36 & 0.889 & 19.81 \\
5 & 0.125 & 10.78 & 0.392 & 14.21 & 0.590 & 16.87 & 0.746 & 19.01 \\
6 & 0.114 & 10.76 & 0.349 & 14.06 & 0.518 & 16.53 & 0.646 & 18.45 \\
7 & 0.105 & 10.75 & 0.316 & 13.94 & 0.462 & 16.26 & 0.570 & 18.03 \\
8 & 0.097 & 10.74 & 0.288 & 13.85 & 0.418 & 16.05 & 0.512 & 17.69 \\
9 & 0.091 & 10.73 & 0.266 & 13.76 & 0.381 & 15.87 & 0.464 & 17.42 \\
10 & 0.085 & 10.72 & 0.246 & 13.70 & 0.351 & 15.72 & 0.426 & 17.20 \\
11 & 0.080 & 10.71 & 0.230 & 13.64 & 0.326 & 15.60 & 0.393 & 17.01 \\
12 & 0.076 & 10.71 & 0.215 & 13.58 & 0.304 & 15.49 & 0.365 & 16.84 \\
13 & 0.072 & 10.70 & 0.203 & 13.53 & 0.285 & 15.39 & 0.341 & 16.70 \\
14 & 0.068 & 10.70 & 0.191 & 13.49 & 0.268 & 15.31 & 0.321 & 16.58 \\
15 & 0.065 & 10.69 & 0.181 & 13.45 & 0.253 & 15.24 & 0.302 & 16.47 \\
16 & 0.062 & 10.69 & 0.172 & 13.42 & 0.240 & 15.17 & 0.286 & 16.37 \\
17 & 0.059 & 10.69 & 0.164 & 13.39 & 0.228 & 15.11 & 0.271 & 16.29 \\
18 & 0.057 & 10.68 & 0.157 & 13.37 & 0.218 & 15.06 & 0.258 & 16.22 \\
19 & 0.054 & 10.68 & 0.150 & 13.34 & 0.208 & 15.01 & 0.246 & 16.15 \\
20 & 0.052 & 10.67 & 0.144 & 13.32 & 0.199 & 14.97 & 0.235 & 16.09 \\
21 & 0.050 & 10.67 & 0.138 & 13.30 & 0.191 & 14.93 & 0.225 & 16.03 \\
22 & 0.049 & 10.67 & 0.133 & 13.28 & 0.183 & 14.90 & 0.216 & 15.98 \\
23 & 0.047 & 10.67 & 0.128 & 13.27 & 0.176 & 14.87 & 0.208 & 15.93 \\
24 & 0.045 & 10.67 & 0.124 & 13.25 & 0.170 & 14.84 & 0.200 & 15.88 \\
25 & 0.044 & 10.66 & 0.120 & 13.23 & 0.164 & 14.80 & 0.193 & 15.84 \\
26 & 0.043 & 10.66 & 0.116 & 13.22 & 0.159 & 14.78 & 0.186 & 15.81 \\
27 & 0.041 & 10.66 & 0.112 & 13.20 & 0.153 & 14.75 & 0.180 & 15.77 \\
28 & 0.040 & 10.66 & 0.109 & 13.19 & 0.149 & 14.73 & 0.174 & 15.74 \\
29 & 0.039 & 10.65 & 0.105 & 13.18 & 0.144 & 14.71 & 0.169 & 15.71 \\
30 & 0.038 & 10.65 & 0.102 & 13.17 & 0.140 & 14.69 & 0.164 & 15.68 \\

Table 2.11: Weights $\hat{d}(\gamma)$ and critical values $H^{-1}(1 - \alpha; \hat{d}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.01.$
<table>
<thead>
<tr>
<th>$k$</th>
<th>$\hat{\alpha}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{\alpha}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{\alpha}(\gamma)$</th>
<th>Crit Val</th>
<th>$\hat{\alpha}(\gamma)$</th>
<th>Crit Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.065</td>
<td>6.38</td>
<td>0.301</td>
<td>7.80</td>
<td>0.579</td>
<td>9.46</td>
<td>0.861</td>
<td>11.16</td>
</tr>
<tr>
<td>3</td>
<td>0.055</td>
<td>6.38</td>
<td>0.243</td>
<td>7.72</td>
<td>0.441</td>
<td>9.16</td>
<td>0.623</td>
<td>10.51</td>
</tr>
<tr>
<td>4</td>
<td>0.047</td>
<td>6.38</td>
<td>0.204</td>
<td>7.66</td>
<td>0.359</td>
<td>8.98</td>
<td>0.494</td>
<td>10.15</td>
</tr>
<tr>
<td>5</td>
<td>0.042</td>
<td>6.37</td>
<td>0.177</td>
<td>7.63</td>
<td>0.305</td>
<td>8.86</td>
<td>0.412</td>
<td>9.92</td>
</tr>
<tr>
<td>6</td>
<td>0.038</td>
<td>6.37</td>
<td>0.156</td>
<td>7.60</td>
<td>0.265</td>
<td>8.77</td>
<td>0.354</td>
<td>9.76</td>
</tr>
<tr>
<td>7</td>
<td>0.034</td>
<td>6.37</td>
<td>0.140</td>
<td>7.58</td>
<td>0.235</td>
<td>8.71</td>
<td>0.311</td>
<td>9.64</td>
</tr>
<tr>
<td>8</td>
<td>0.031</td>
<td>6.37</td>
<td>0.126</td>
<td>7.56</td>
<td>0.212</td>
<td>8.66</td>
<td>0.278</td>
<td>9.54</td>
</tr>
<tr>
<td>9</td>
<td>0.029</td>
<td>6.37</td>
<td>0.116</td>
<td>7.54</td>
<td>0.192</td>
<td>8.61</td>
<td>0.252</td>
<td>9.47</td>
</tr>
<tr>
<td>10</td>
<td>0.027</td>
<td>6.37</td>
<td>0.106</td>
<td>7.53</td>
<td>0.176</td>
<td>8.58</td>
<td>0.230</td>
<td>9.41</td>
</tr>
<tr>
<td>11</td>
<td>0.025</td>
<td>6.37</td>
<td>0.099</td>
<td>7.52</td>
<td>0.163</td>
<td>8.55</td>
<td>0.211</td>
<td>9.35</td>
</tr>
<tr>
<td>12</td>
<td>0.023</td>
<td>6.37</td>
<td>0.092</td>
<td>7.51</td>
<td>0.151</td>
<td>8.52</td>
<td>0.196</td>
<td>9.31</td>
</tr>
<tr>
<td>13</td>
<td>0.022</td>
<td>6.37</td>
<td>0.086</td>
<td>7.50</td>
<td>0.141</td>
<td>8.50</td>
<td>0.183</td>
<td>9.27</td>
</tr>
<tr>
<td>14</td>
<td>0.021</td>
<td>6.37</td>
<td>0.081</td>
<td>7.49</td>
<td>0.133</td>
<td>8.48</td>
<td>0.171</td>
<td>9.24</td>
</tr>
<tr>
<td>15</td>
<td>0.019</td>
<td>6.37</td>
<td>0.077</td>
<td>7.48</td>
<td>0.125</td>
<td>8.47</td>
<td>0.161</td>
<td>9.21</td>
</tr>
<tr>
<td>16</td>
<td>0.018</td>
<td>6.37</td>
<td>0.073</td>
<td>7.48</td>
<td>0.118</td>
<td>8.45</td>
<td>0.152</td>
<td>9.19</td>
</tr>
<tr>
<td>17</td>
<td>0.018</td>
<td>6.37</td>
<td>0.069</td>
<td>7.47</td>
<td>0.112</td>
<td>8.44</td>
<td>0.144</td>
<td>9.16</td>
</tr>
<tr>
<td>18</td>
<td>0.017</td>
<td>6.37</td>
<td>0.066</td>
<td>7.47</td>
<td>0.107</td>
<td>8.42</td>
<td>0.137</td>
<td>9.14</td>
</tr>
<tr>
<td>19</td>
<td>0.016</td>
<td>6.36</td>
<td>0.063</td>
<td>7.46</td>
<td>0.102</td>
<td>8.41</td>
<td>0.130</td>
<td>9.13</td>
</tr>
<tr>
<td>20</td>
<td>0.015</td>
<td>6.36</td>
<td>0.060</td>
<td>7.46</td>
<td>0.097</td>
<td>8.40</td>
<td>0.124</td>
<td>9.11</td>
</tr>
<tr>
<td>21</td>
<td>0.015</td>
<td>6.36</td>
<td>0.057</td>
<td>7.46</td>
<td>0.093</td>
<td>8.39</td>
<td>0.119</td>
<td>9.09</td>
</tr>
<tr>
<td>22</td>
<td>0.014</td>
<td>6.36</td>
<td>0.055</td>
<td>7.45</td>
<td>0.089</td>
<td>8.38</td>
<td>0.114</td>
<td>9.08</td>
</tr>
<tr>
<td>23</td>
<td>0.014</td>
<td>6.36</td>
<td>0.053</td>
<td>7.45</td>
<td>0.086</td>
<td>8.38</td>
<td>0.109</td>
<td>9.07</td>
</tr>
<tr>
<td>24</td>
<td>0.013</td>
<td>6.36</td>
<td>0.051</td>
<td>7.45</td>
<td>0.082</td>
<td>8.37</td>
<td>0.105</td>
<td>9.06</td>
</tr>
<tr>
<td>25</td>
<td>0.013</td>
<td>6.36</td>
<td>0.049</td>
<td>7.45</td>
<td>0.079</td>
<td>8.36</td>
<td>0.101</td>
<td>9.05</td>
</tr>
<tr>
<td>26</td>
<td>0.012</td>
<td>6.36</td>
<td>0.048</td>
<td>7.44</td>
<td>0.077</td>
<td>8.36</td>
<td>0.098</td>
<td>9.03</td>
</tr>
<tr>
<td>27</td>
<td>0.012</td>
<td>6.36</td>
<td>0.046</td>
<td>7.44</td>
<td>0.074</td>
<td>8.35</td>
<td>0.094</td>
<td>9.02</td>
</tr>
<tr>
<td>28</td>
<td>0.011</td>
<td>6.36</td>
<td>0.044</td>
<td>7.44</td>
<td>0.072</td>
<td>8.35</td>
<td>0.091</td>
<td>9.02</td>
</tr>
<tr>
<td>29</td>
<td>0.011</td>
<td>6.36</td>
<td>0.043</td>
<td>7.44</td>
<td>0.069</td>
<td>8.34</td>
<td>0.088</td>
<td>9.01</td>
</tr>
<tr>
<td>30</td>
<td>0.011</td>
<td>6.36</td>
<td>0.042</td>
<td>7.44</td>
<td>0.067</td>
<td>8.34</td>
<td>0.086</td>
<td>9.00</td>
</tr>
</tbody>
</table>

Table 2.12: Weights $\hat{\alpha}(\gamma)$ and critical values $\hat{H}^{-1}(1 - \alpha; \hat{\alpha}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.05$. 

154
Table 2.13: Weights $\hat{a}(\gamma)$ and critical values $H^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ for $p = 2$ and $\alpha = 0.1$. 
\[
\begin{array}{c|c|c|c|c|c|c|c|c|c}
\gamma = 0.01 & \gamma = 0.05 & \gamma = 0.1 & \gamma = 0.15 \\
\hline
k & \ddot{a}(\gamma) & \text{Crit Val} & \ddot{a}(\gamma) & \text{Crit Val} & \ddot{a}(\gamma) & \text{Crit Val} & \ddot{a}(\gamma) & \text{Crit Val} \\
3 & 0.153 & 13.08 & 0.531 & 17.38 & 0.880 & 21.34 & 1.196 & 24.92 \\
4 & 0.138 & 13.06 & 0.459 & 17.10 & 0.732 & 20.58 & 0.965 & 23.58 \\
5 & 0.126 & 13.04 & 0.406 & 16.89 & 0.630 & 20.05 & 0.814 & 22.68 \\
6 & 0.116 & 13.03 & 0.364 & 16.73 & 0.555 & 19.66 & 0.707 & 22.03 \\
7 & 0.107 & 13.01 & 0.331 & 16.60 & 0.497 & 19.35 & 0.626 & 21.53 \\
8 & 0.100 & 13.00 & 0.303 & 16.49 & 0.451 & 19.11 & 0.562 & 21.13 \\
9 & 0.093 & 12.99 & 0.280 & 16.41 & 0.412 & 18.91 & 0.512 & 20.83 \\
10 & 0.088 & 12.98 & 0.260 & 16.33 & 0.381 & 18.73 & 0.469 & 20.55 \\
11 & 0.083 & 12.98 & 0.243 & 16.26 & 0.354 & 18.59 & 0.434 & 20.33 \\
12 & 0.079 & 12.97 & 0.229 & 16.20 & 0.330 & 18.47 & 0.404 & 20.14 \\
13 & 0.075 & 12.97 & 0.216 & 16.15 & 0.310 & 18.36 & 0.377 & 19.97 \\
14 & 0.071 & 12.96 & 0.204 & 16.11 & 0.292 & 18.26 & 0.354 & 19.83 \\
15 & 0.068 & 12.96 & 0.194 & 16.06 & 0.276 & 18.18 & 0.334 & 19.70 \\
16 & 0.065 & 12.95 & 0.184 & 16.03 & 0.262 & 18.10 & 0.316 & 19.58 \\
17 & 0.062 & 12.95 & 0.176 & 15.99 & 0.249 & 18.03 & 0.300 & 19.48 \\
18 & 0.059 & 12.94 & 0.168 & 15.96 & 0.238 & 17.96 & 0.286 & 19.38 \\
19 & 0.057 & 12.94 & 0.161 & 15.93 & 0.227 & 17.91 & 0.273 & 19.29 \\
20 & 0.055 & 12.94 & 0.155 & 15.91 & 0.217 & 17.86 & 0.261 & 19.22 \\
21 & 0.053 & 12.93 & 0.149 & 15.89 & 0.209 & 17.81 & 0.250 & 19.15 \\
22 & 0.051 & 12.93 & 0.143 & 15.86 & 0.201 & 17.77 & 0.240 & 19.09 \\
23 & 0.050 & 12.93 & 0.138 & 15.84 & 0.193 & 17.72 & 0.231 & 19.03 \\
24 & 0.048 & 12.92 & 0.133 & 15.82 & 0.186 & 17.68 & 0.222 & 18.97 \\
25 & 0.047 & 12.92 & 0.129 & 15.80 & 0.180 & 17.65 & 0.214 & 18.92 \\
26 & 0.045 & 12.92 & 0.125 & 15.79 & 0.174 & 17.62 & 0.207 & 18.87 \\
27 & 0.044 & 12.92 & 0.121 & 15.77 & 0.168 & 17.58 & 0.200 & 18.83 \\
28 & 0.043 & 12.92 & 0.117 & 15.76 & 0.163 & 17.55 & 0.194 & 18.79 \\
29 & 0.041 & 12.92 & 0.114 & 15.74 & 0.158 & 17.53 & 0.188 & 18.75 \\
30 & 0.040 & 12.91 & 0.111 & 15.73 & 0.153 & 17.50 & 0.182 & 18.71 \\
\end{array}
\]

Table 2.14: Weights \( \ddot{a}(\gamma) \) and critical values \( H^{-1}(1 - \alpha; \ddot{a}(\gamma), k, p) \) for \( p = 3 \) and \( \alpha = 0.01 \).
| $k$ | $\gamma = 0.01$ | | $\gamma = 0.05$ | | $\gamma = 0.1$ | | $\gamma = 0.15$ |
|---|---|---|---|---|---|---|
| | $\hat{a}(\gamma)$ | Crit Val | $\hat{a}(\gamma)$ | Crit Val | $\hat{a}(\gamma)$ | Crit Val | $\hat{a}(\gamma)$ | Crit Val |
| 3  | 0.055 | 8.24 | 0.250 | 9.77 | 0.470 | 11.49 | 0.683 | 13.16 |
| 4  | 0.048 | 8.24 | 0.213 | 9.71 | 0.387 | 11.29 | 0.547 | 12.74 |
| 5  | 0.043 | 8.24 | 0.186 | 9.67 | 0.331 | 11.15 | 0.459 | 12.47 |
| 6  | 0.039 | 8.24 | 0.165 | 9.64 | 0.289 | 11.04 | 0.396 | 12.27 |
| 7  | 0.035 | 8.24 | 0.149 | 9.61 | 0.258 | 10.96 | 0.349 | 12.13 |
| 8  | 0.033 | 8.23 | 0.135 | 9.59 | 0.232 | 10.90 | 0.312 | 12.01 |
| 9  | 0.030 | 8.23 | 0.124 | 9.57 | 0.212 | 10.85 | 0.283 | 11.91 |
| 10 | 0.028 | 8.23 | 0.115 | 9.56 | 0.195 | 10.81 | 0.259 | 11.84 |
| 11 | 0.026 | 8.23 | 0.107 | 9.55 | 0.180 | 10.77 | 0.239 | 11.77 |
| 12 | 0.025 | 8.23 | 0.100 | 9.54 | 0.168 | 10.74 | 0.221 | 11.71 |
| 13 | 0.023 | 8.23 | 0.094 | 9.52 | 0.157 | 10.71 | 0.207 | 11.67 |
| 14 | 0.022 | 8.23 | 0.088 | 9.51 | 0.147 | 10.69 | 0.194 | 11.62 |
| 15 | 0.021 | 8.23 | 0.084 | 9.51 | 0.139 | 10.67 | 0.182 | 11.58 |
| 16 | 0.020 | 8.23 | 0.079 | 9.50 | 0.132 | 10.65 | 0.172 | 11.55 |
| 17 | 0.019 | 8.23 | 0.076 | 9.50 | 0.125 | 10.63 | 0.163 | 11.52 |
| 18 | 0.018 | 8.23 | 0.072 | 9.49 | 0.119 | 10.62 | 0.155 | 11.50 |
| 19 | 0.017 | 8.23 | 0.069 | 9.49 | 0.113 | 10.60 | 0.148 | 11.47 |
| 20 | 0.017 | 8.23 | 0.066 | 9.48 | 0.108 | 10.59 | 0.141 | 11.45 |
| 21 | 0.016 | 8.23 | 0.063 | 9.48 | 0.104 | 10.58 | 0.135 | 11.43 |
| 22 | 0.015 | 8.23 | 0.061 | 9.47 | 0.100 | 10.56 | 0.129 | 11.41 |
| 23 | 0.015 | 8.23 | 0.058 | 9.47 | 0.096 | 10.56 | 0.124 | 11.40 |
| 24 | 0.014 | 8.23 | 0.056 | 9.47 | 0.092 | 10.55 | 0.119 | 11.38 |
| 25 | 0.014 | 8.23 | 0.054 | 9.46 | 0.089 | 10.54 | 0.115 | 11.37 |
| 26 | 0.013 | 8.23 | 0.053 | 9.46 | 0.086 | 10.53 | 0.111 | 11.35 |
| 27 | 0.013 | 8.23 | 0.051 | 9.46 | 0.083 | 10.53 | 0.107 | 11.34 |
| 28 | 0.013 | 8.23 | 0.049 | 9.46 | 0.080 | 10.52 | 0.104 | 11.33 |
| 29 | 0.012 | 8.23 | 0.048 | 9.45 | 0.078 | 10.51 | 0.100 | 11.32 |
| 30 | 0.012 | 8.23 | 0.046 | 9.45 | 0.075 | 10.51 | 0.097 | 11.31 |

Table 2.15: Weights $\hat{a}(\gamma)$ and critical values $\tilde{H}^{-1}(1 - \alpha; \hat{a}(\gamma), k, p)$ for $p = 3$ and $\alpha = 0.05$. 
<table>
<thead>
<tr>
<th>$k$</th>
<th>$\gamma = 0.01$</th>
<th>$\gamma = 0.05$</th>
<th>$\gamma = 0.1$</th>
<th>$\gamma = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{a}(\gamma)$</td>
<td>Crit Val</td>
<td>$\bar{a}(\gamma)$</td>
<td>Crit Val</td>
</tr>
<tr>
<td>3</td>
<td>0.036</td>
<td>6.48</td>
<td>0.176</td>
<td>7.35</td>
</tr>
<tr>
<td>4</td>
<td>0.031</td>
<td>6.48</td>
<td>0.147</td>
<td>7.32</td>
</tr>
<tr>
<td>5</td>
<td>0.027</td>
<td>6.48</td>
<td>0.126</td>
<td>7.31</td>
</tr>
<tr>
<td>6</td>
<td>0.024</td>
<td>6.48</td>
<td>0.111</td>
<td>7.29</td>
</tr>
<tr>
<td>7</td>
<td>0.022</td>
<td>6.47</td>
<td>0.099</td>
<td>7.28</td>
</tr>
<tr>
<td>8</td>
<td>0.020</td>
<td>6.47</td>
<td>0.089</td>
<td>7.27</td>
</tr>
<tr>
<td>9</td>
<td>0.018</td>
<td>6.47</td>
<td>0.081</td>
<td>7.26</td>
</tr>
<tr>
<td>10</td>
<td>0.017</td>
<td>6.47</td>
<td>0.074</td>
<td>7.26</td>
</tr>
<tr>
<td>11</td>
<td>0.015</td>
<td>6.47</td>
<td>0.069</td>
<td>7.25</td>
</tr>
<tr>
<td>12</td>
<td>0.014</td>
<td>6.47</td>
<td>0.064</td>
<td>7.25</td>
</tr>
<tr>
<td>13</td>
<td>0.013</td>
<td>6.47</td>
<td>0.060</td>
<td>7.24</td>
</tr>
<tr>
<td>14</td>
<td>0.013</td>
<td>6.47</td>
<td>0.056</td>
<td>7.24</td>
</tr>
<tr>
<td>15</td>
<td>0.012</td>
<td>6.47</td>
<td>0.053</td>
<td>7.24</td>
</tr>
<tr>
<td>16</td>
<td>0.011</td>
<td>6.47</td>
<td>0.050</td>
<td>7.23</td>
</tr>
<tr>
<td>17</td>
<td>0.011</td>
<td>6.47</td>
<td>0.048</td>
<td>7.23</td>
</tr>
<tr>
<td>18</td>
<td>0.010</td>
<td>6.47</td>
<td>0.045</td>
<td>7.23</td>
</tr>
<tr>
<td>19</td>
<td>0.010</td>
<td>6.47</td>
<td>0.043</td>
<td>7.23</td>
</tr>
<tr>
<td>20</td>
<td>0.009</td>
<td>6.47</td>
<td>0.041</td>
<td>7.23</td>
</tr>
<tr>
<td>21</td>
<td>0.009</td>
<td>6.47</td>
<td>0.040</td>
<td>7.22</td>
</tr>
<tr>
<td>22</td>
<td>0.009</td>
<td>6.47</td>
<td>0.038</td>
<td>7.22</td>
</tr>
<tr>
<td>23</td>
<td>0.008</td>
<td>6.47</td>
<td>0.036</td>
<td>7.22</td>
</tr>
<tr>
<td>24</td>
<td>0.008</td>
<td>6.47</td>
<td>0.035</td>
<td>7.22</td>
</tr>
<tr>
<td>25</td>
<td>0.008</td>
<td>6.47</td>
<td>0.034</td>
<td>7.22</td>
</tr>
<tr>
<td>26</td>
<td>0.007</td>
<td>6.47</td>
<td>0.033</td>
<td>7.22</td>
</tr>
<tr>
<td>27</td>
<td>0.007</td>
<td>6.47</td>
<td>0.032</td>
<td>7.22</td>
</tr>
<tr>
<td>28</td>
<td>0.007</td>
<td>6.47</td>
<td>0.030</td>
<td>7.22</td>
</tr>
<tr>
<td>29</td>
<td>0.007</td>
<td>6.47</td>
<td>0.029</td>
<td>7.21</td>
</tr>
<tr>
<td>30</td>
<td>0.007</td>
<td>6.47</td>
<td>0.029</td>
<td>7.21</td>
</tr>
</tbody>
</table>

Table 2.16: Weights $\bar{a}(\gamma)$ and critical values $H_{1}(1-\alpha; \bar{a}(\gamma), k, p)$ for $p = 3$ and $\alpha = 0.1$. 

158
Chapter 3

A Geometric Approach to Weakly Identified Econometric Models

with Anna Mikusheva

Empirical researchers in Economics frequently find that even in large samples the data provides little information about some model parameters. In such cases, known as weakly identified, the usual asymptotic approximations to the behavior of estimators and test statistics may be quite poor, making standard approaches to inference unreliable. Weak identification has been detected in a wide range of non-linear estimation contexts, including estimation of the new Keynesian Phillips curve (Dufour, Khalaf, and Kichian (2006), Kleibergen and Mavroeidis (2009), Mavroeidis (2005), Nason and Smith (2008)), monetary policy rules (Mavroeidis (2010)), Dynamic Stochastic General Equilibrium (DSGE) Models (Ruge-Murcia (2007), Canova and Sala (2009), Iskrev (2010), I. Andrews and Mikusheva (2013), Guerron-Quintana, Inoue and Kilian (2013)), and Euler equations (Yogo (2004)). The need for more reliable procedures robust to weak identification in non-linear contexts has inspired a large literature in econometrics - for a survey, see Dufour (2003) and Stock, Wright, and Yogo (2002).

A number of different testing procedures have been proposed in this literature, most of which address two situations: the case in which one is interested in testing the full parameter vector, and the case in which one is interested in testing only a
subset of parameters but the parameters not under test (the nuisance parameters) are strongly identified. Examples of such tests include those of Stock and Wright (2000), Guggenberger and Smith (2005, 2008), Kleibergen (2005, 2007), I. Andrews and Mikusheva (2013), and Qu (2013). The literature to date has, however, been largely silent about the case in which part of the nuisance parameter vector may be weakly identified. A notable exception is the recent paper by D. Andrews and Cheng (2012).

Our paper directly addresses the question of inference with weakly identified nuisance parameters in the context of minimum distance estimation. We suggest a fully robust testing procedure which controls size without any assumption on the strength of identification of the parameters. Further, if the nuisance parameters are strongly identified, our procedure is asymptotically equivalent to the “concentrated out” S-test suggested by Stock and Wright (2000) for hypotheses with strongly identified nuisance parameters.

Our procedure is based on novel finite-sample bounds on the distribution of the test statistic under the null, derived under the assumption that the reduced-form parameter estimate is exactly normally distributed with a known covariance matrix. We show that tests based on these bounds control asymptotic size uniformly over a broad class of models. For our asymptotic results, we assume that the reduced-form parameter estimates are uniformly asymptotically normal with consistently estimable variance. Our results hold uniformly over a large class of link functions relating the structural and reduced-form parameters and do not rely on any particular asymptotic embedding, such as those used by Stock and Wright (2000) or D. Andrews and Cheng (2012), to model weak identification.

The bounds we derive rely on techniques from differential geometry which are new in the econometrics literature. Our starting point is the observation that hypotheses in non-linear models with strongly identified nuisance parameters are asymptotically linear in a geometrical sense. In contrast, hypotheses with weakly identified nuisance parameters need not be asymptotically linear and can exhibit substantial curvature even in large samples, leading to the breakdown of the usual asymptotic approxima-
tions. Our bounds can be viewed as a strengthening of the usual approximations, where rather than appealing to asymptotic linearity of the null hypothesis we quantify the maximal deviation of the null from linearity and use it to construct stochastic bounds. As a result, under strong-identification asymptotics our bounds recover the usual approximations. The test we suggest uses a standard minimum-distance statistic paired with easy-to-simulate robust critical values derived using these geometric bounds. The bounds we derive are also of potential interest for a range of other applications, including testing nonlinear hypotheses and inference in highly non-linear models. Our approach differs from the statistical geometry literature initiated by Efron (1975) in that we produce finite sample bounds on the distribution of the test statistic, whereas the statistical geometry literature is primarily concerned with higher-order asymptotic approximations.

To date the dominant recommendation for testing hypotheses with weakly identified nuisance parameters has been the projection method (see Dufour and Jasiak (2001), Dufour and Taamouti (2005), Dufour, Khalaf, and Kichian (2006)). The strength of the projection method is that it requires no assumptions beyond the validity of the test for the full parameter vector. It is in general conservative, however, and may be extremely so in cases where the nuisance parameter is high-dimensional and/or strongly identified. Our approach is an improvement over the projection method, in that it pairs the minimum distance statistic with smaller critical values while still maintaining size.

If one knows that part of the nuisance parameter vector is strongly identified, it has been proved that in many cases one can obtain a more powerful test by concentrating out the nuisance parameter as in e.g. Stock and Wright (2000). Maintaining correct size in such cases, however, relies critically on the strong identification assumption on the nuisance parameter. In contrast, our approach requires no assumption of strong identification but, in the event that the nuisance parameters are strongly identified, is asymptotically equivalent to concentrating them out. In this sense, our robust critical values can be viewed as providing a continuous transition between projecting over and concentrating out the nuisance parameters, depending on the strength of
identification.

To illustrate our approach, we revisit the question of weak identification-robust inference on new Keyensian Phillips curve parameters previously studied by Magnusson and Mavroeidis (2010). Applying our method to their empirical example we find that there is a substantial amount of curvature. Using our robust approach we construct confidence sets which both have better coverage than those developed by Magnusson and Mavroeidis and are smaller in the empirical application. We also apply our approach to a small-scale DSGE model and find evidence of substantial curvature. We consider the problem of testing composite hypotheses about model parameters, and show that our robust critical values are substantially smaller than those used by the projection method while still controlling size.

The paper is structured as follows. In Section 3.1 we show that hypotheses with strongly identified nuisance parameters are asymptotically linear, while weakly identified nuisance parameters may cause non-trivial curvature of the null hypothesis. In Section 3.2 we derive our geometric and stochastic bounds. In Section 3.3 we introduce our finite-sample test as well as several modifications, under the assumption that the reduced-form parameters are normally distributed with known variance. In Section 3.4 we relax the finite-sample normality assumption and discuss the uniform asymptotic validity of our procedure. We also compare our testing procedures with existing methods. Section 3.5 presents empirical results from applying our approach to a new Keyensian Phillips curve example, and Section 3.6 presents simulation results from a small-scale DSGE model. Most may be found in the Appendix, while proofs of secondary importance and the details of our empirical and simulation examples are given in the Supplementary Appendix, which can be found on Anna Mikusheva’s website.

Throughout the paper we use the following notation: $\hat{\alpha}$ is the derivative of the function $\alpha$, $\hat{\alpha}$ is the second derivative, $B_R(x_0) = \{x \in \mathbb{R}^k : \|x - x_0\| \leq (1 + \sqrt{2})R\}$ is a $k$-dimensional ball of radius $(1 + \sqrt{2})R$ with center $x_0$. Let $D_C = \{x = (x^{(1)}, x^{(2)}) : \|x^{(1)}\| \leq C, \|x^{(2)}\| \leq C, x^{(1)} \in \mathbb{R}^p, x^{(2)} \in \mathbb{R}^{k-p}\} \subset \mathbb{R}^k$, which is a natural generalization.

\[\text{economics.mit.edu/files/9271}\]
of a cylinder. We take $N_Z$ to equal $I - Z(Z'Z)^{-1}Z'$.

### 3.1 Model setting

In many empirical settings researchers are interested in statistical inference on a subset of structural parameters, for example constructing confidence sets for individual parameters in multi-parameter models. Suppose we are interested in testing a hypothesis $H_0: \alpha = \alpha_0$ about a structural parameter $\alpha$ but that the model also has a $p$-dimensional nuisance parameter (i.e. parameter not under test) $\beta$ which can vary freely under the null. This paper is concerned with inference in cases where we are not willing assume that the nuisance parameter $\beta$ is strongly identified.

We explore this issue in a minimum-distance context, and in particular assume that we have some $k$-dimensional ($k > p$) reduced-form parameter $\theta$ which is known to be strongly identified. Here by strong identification we mean that there is an estimator $\hat{\theta}$ of the unknown true value $\theta_0$ such that, for some estimator $\hat{\Sigma}$ for the variance of $\hat{\theta}$, $\hat{\Sigma}^{-\frac{1}{2}}(\hat{\theta} - \theta_0)$ is uniformly asymptotically normally distributed. In time-series models $\theta$ could consist of auto-covariances or reduced-form VAR parameters, while in linear IV it could contain reduced-form regression coefficients. For the main part of our analysis we make the stronger assumption that our estimator $\hat{\theta}$ is exactly normally distributed, $\hat{\theta} - \theta_0 \sim N(0, \Sigma)$ for a known covariance matrix $\Sigma$. This allows us to develop exact finite-sample results. We then relax our assumptions to require that $\hat{\theta}$ be uniformly asymptotically normal and its asymptotic variance be uniformly consistently estimable, and show that our finite-sample results imply uniform asymptotic results.

Let $\theta(\alpha, \beta)$ be a link function connecting the structural and reduced-form parameters. For many models the link function and its derivatives can easily be calculated for any values $\alpha$ and $\beta$. In such cases, we consider a test based on the minimum-distance statistic

$$MD = \min_{\beta}(\hat{\theta} - \theta(\alpha_0, \beta))')\Sigma^{-1}(\hat{\theta} - \theta(\alpha_0, \beta)).$$

(3.1)
In some contexts the link function may only be defined implicitly, for example by the restriction that the true parameter values $\theta_0$ and $\alpha_0$ must satisfy $g(\theta_0, \alpha_0) = 0$ for a $(k - p)$-dimensional function $g$. This corresponds to having an implicit $p$-dimensional nuisance parameter, and in such models we define the minimum-distance statistic as

$$MD = \min_{\theta: g(\theta, \alpha_0) = 0} (\hat{\theta} - \theta)'\Sigma^{-1}(\hat{\theta} - \theta).$$

Our empirical example, inference on New Keynesian Phillips curve parameters, falls into this class.

To proceed, it is useful for us to introduce the random vector $\xi = \Sigma^{-1/2}(\hat{\theta} - \theta_0) \sim N(0, I_k)$ and the $p$-dimensional manifold $S = \{x : x = \Sigma^{-1/2}(\theta(\alpha_0, \beta) - \theta_0), \beta \in \mathbb{R}^p\}$. Note that under the null the manifold $S$ is known up to a location shift determined by the true value $\beta_0$, since $\theta_0 = \theta(\alpha_0, \beta_0)$. Thus, we know the shape of $S$ and, moreover, know that it passes through the origin. In the case of an implicitly defined link function we have $S = \{x : x = \Sigma^{-1/2}(\theta - \theta_0), g(\theta, \alpha_0) = 0\}$. The minimum distance statistics defined above are simply the squared distance between $\xi$ and $S$:

$$MD = \min_{x \in S} (\xi - x)'(\xi - x) = \rho^2(\xi, S).$$

The distribution of $\rho^2(\xi, S)$ is in general non-standard and depends on the unknown $\beta_0$. The central statistical issue of this paper is how to define critical values such that tests based on $\rho^2(\xi, S)$ control size.

A well-known property of the normal distribution is that if $S$ is a $p$-dimensional linear sub-space in $\mathbb{R}^k$ then the squared distance $\rho^2(\xi, S)$ has a $\chi^2_{k-p}$ distribution. Note that in this special case the distribution does not depend on $\beta_0$. Numerous classical results on testing in the presence of nuisance parameters are based on this fact. Indeed, most of the classical statistics literature deals with testing hypotheses that are either linear or asymptotically linear, in the sense that $S$ is either a linear subspace or is arbitrarily well-approximated by one in large samples. Below, we argue that testing in the presence of strongly identified nuisance parameters is asymptotically equivalent to testing a linear hypothesis, while testing in the presence
of *weakly* identified nuisance parameters tends to result in asymptotically *non-linear* null hypotheses. Our main results construct bounds on the distribution of $\rho^2(\xi, S)$ and define critical values based on a measure of the non-linearity of $S$, specifically its maximal curvature.

There is a less informative bound that can be placed on $\rho^2(\xi, S)$ without any assumptions, namely that $\rho^2(\xi, S)$ is dominated by a $\chi_k^2$ distribution. Indeed, since $0 \in S$,

$$
\rho(\xi, S)^2 = \min_{x \in S} (\xi - x)'(\xi - x) \leq (\xi - 0)'(\xi - 0) \sim \chi_k^2.
$$

(3.3)

This bound is precisely the one used by the “projection method”, which is currently the main approach available for testing with weakly identified nuisance parameters, see Dufour and Jasiak (2001), Dufour and Taamouti (2005), and Dufour, Khalaf, and Kichian (2006).

### 3.1.1 Strength of identification and linearity.

In this section we argue that testing with strongly identified nuisance parameters is asymptotically equivalent to the case when the manifold $S$ is linear. In contrast, if there are weakly identified nuisance parameters the manifold corresponding to the null hypothesis need not converge to a linear subspace, so $S$ may remain non-linear even in large samples. To fix ideas, in this section we use the weak identification asymptotic framework introduced by Stock and Wright (2000). It is important to emphasize that the discussion in this section is solely for motivation and that the validity of our method does not rely on this or any other device for modeling weak identification.

Consider a GMM model in which the moment function is additively separable in the data. In particular, assume that we observe a sample $\{x_i\}$ of size $n$ consisting of identically and independently distributed observations such that

$$
E(h(x_i) - \theta(\alpha, \beta)) = 0 \quad \text{for} \quad \alpha = \alpha_0, \ \beta = \beta_0.
$$

(3.4)
Here \( \theta_0 = Eb(x_i) \) is a \( k \)-dimensional reduced-form parameter, while \( \alpha \) and \( \beta \) are \( p_\alpha \times 1 \) and \( p_\beta \times 1 \) vectors respectively, with \( p_\alpha + p_\beta \leq k \). Assume that \((\alpha_0, \beta_0)\) is the unique point at which the moment condition (3.4) is satisfied, so the model is point identified. As in Stock and Wright (2000), we allow the function \( \theta(\alpha, \beta) \) to change as the sample size grows. In particular, let

\[
\theta(\alpha, \beta) = \theta_n(\alpha, \beta) = \tilde{M}(\alpha) + \frac{1}{\sqrt{n}} M^*(\alpha, \beta),
\]

where \( \tilde{M}(\alpha) \) and \( M^*(\alpha, \beta) \) are fixed twice-continuously-differentiable functions with full-rank Jacobians. In this setting \( \alpha \) is strongly identified while \( \beta \) is weakly identified, because information about \( \beta \) does not accumulate as the sample size grows.

Suppose we are interested in testing hypotheses about the structural parameters \( \alpha \) and \( \beta \). Consider first the problem of testing the hypothesis \( H_0 : \beta = \beta_0 \) with strongly identified nuisance parameter \( \alpha \). The appropriate minimum distance test statistic is

\[
MD(\beta_0) = \min_{\alpha} \left( \frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha, \beta_0) \right)' \Sigma^{-1} \left( \frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha, \beta_0) \right),
\]

where \( \Sigma \) is the covariance matrix of random vector \( h(x_i) \) (which we take to be nonsingular) or a consistent estimate thereof. Stock and Wright (2000) prove that under the null \( MD(\beta_0) \Rightarrow \chi^2_{k-p_\beta} \). Interested readers may find a full proof of this result in Stock and Wright (2000): here, we instead show that this testing problem is asymptotically equivalent to a testing problem with linear \( S \).

Define \( \xi_n = \sqrt{n} \Sigma^{-1/2} (\frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha_0, \beta_0)) \). By the central limit theorem, \( \xi_n \Rightarrow \xi \sim N(0, I_k) \). Let the manifold \( S_n \) be the image of the function

\[
m_n(\alpha) = n \Sigma^{-1/2} (\theta_n(\alpha, \beta_0) - \theta_n(\alpha_0, \beta_0)) =
\]

\[
= n \Sigma^{-1/2} (\tilde{M}(\alpha) - \tilde{M}(\alpha_0)) + \Sigma^{-1/2} (M^*(\alpha, \beta_0) - M^*(\alpha_0, \beta_0)) =
\]

\[
= n \Sigma^{-1/2} (\tilde{M}(\alpha) - \tilde{M}(\alpha_0)) + O(||\alpha - \alpha_0||).
\]

Then \( MD(\beta_0) = \rho^2(\xi_n, S_n) \). Under standard conditions for global identification, the
value of $\hat{M}(\alpha)$ is in a small neighborhood of $\hat{M}(\alpha_0)$ only if $\alpha$ is close to $\alpha_0$. Under such conditions one can easily show that the range of values of $\alpha$ such that $m_n(\alpha) \in S_n \cap B$ is of order $1/\sqrt{n}$ for any bounded set $B$ containing zero. Consequently, Taylor approximation shows that the intersection $S_n \cap B$ converges to the intersection of $B$ with the $p_\alpha$-dimensional linear sub-space $S$ spanned by the columns of the Jacobian of $\hat{M}(\alpha)$ at point $\alpha_0$. Informally, we may say that due to the factor $\sqrt{n}$ in equation (3.6), as the sample size increases we zoom in on an infinitesimal neighborhood of the true value $\alpha_0$ of the strongly identified nuisance parameter. Any regular manifold, however, is arbitrarily well approximated by its tangent space on an infinitesimal neighborhood of a regular point. As a result, it is easy to show that $\rho^2(\xi_n, S_n) \Rightarrow \rho^2(\xi, S) \sim \chi^2_{p_\alpha}$, where the last step uses the fact discussed at the beginning of this section that the squared distance from a standard normal vector to a linear subspace is $\chi^2$-distributed.

Tests for hypotheses with weakly identified nuisance parameters behave quite differently. In particular, the curvature of a null hypothesis with a weakly identified nuisance parameter does not in general vanish asymptotically. To illustrate this point, assume that the hypothesis of interest is $H_0 : \alpha = \alpha_0$, so that $\beta$ is a weakly identified nuisance parameter. Again, we consider the appropriate minimum distance statistic:

$$MD(\alpha_0) = \min_{\beta} n \left( \frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha_0, \beta) \right) \Sigma^{-1} \left( \frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha_0, \beta) \right).$$

Define $\xi_n = \sqrt{n} \Sigma^{-1/2}(\frac{1}{n} \sum_i h(x_i) - \theta_n(\alpha_0, \beta_0))$ as before, and let $S_n$ be the image of

$$m_n(\beta) = \sqrt{n} \Sigma^{-1/2}(\theta_n(\alpha_0, \beta) - \theta_n(\alpha_0, \beta_0)) = \Sigma^{-1/2}(M^*(\alpha_0, \beta) - M^*(\alpha_0, \beta_0)).$$

By construction, $S_n$ is a $p_\beta$-dimensional manifold in $k$-dimensional Euclidean space. In contrast to the strongly identified case, however, here $S_n$ does not change with the sample size so we may denote it by $S$. Hence, if $S_n$ is nonlinear for a given sample size, it remains nonlinear in the limit. As a result, we have that

$$MD(\alpha_0) = \rho^2(\xi_n, S) \Rightarrow \rho^2(\xi, S).$$
where $\xi \sim N(0, I_k)$ and $S$ is a $p_\beta$-dimensional manifold, which is not in general a linear sub-space.

**Linearity vs strength of identification.** We showed above that the problem of testing a hypothesis with strongly identified nuisance parameters is asymptotically equivalent to testing a linear hypothesis in a Gaussian model. In contrast, if there are weakly identified nuisance parameters the manifold $S$ corresponding to the null hypothesis need not converge to a linear subspace and may remain non-linear even in large samples. Rather than focusing on strength of identification, we may view the key distinction here as between linearity and non-linearity. In particular, while strong identification guarantees that the null hypothesis will correspond to a linear subspace in the limit, even with weakly identified nuisance parameters if $S$ (the set of parameter values satisfying the null) happens to be a linear subspace, the usual limiting distribution will be correct asymptotically. Hence, in models where the nuisance parameters enter the link function $\theta$ linearly, the usual (strong-identification) critical values for the minimum distance statistic will yield asymptotically valid tests regardless of the strength of identification. Asymptotic linearity, not strong identification as such, is the essential condition. On a related note, Andrews and Mikusheva (2013) show that in a parametric model, a score test which concentrates out nuisance parameters provided they enter the log-likelihood function linearly.

**Weak Instrumental Variable regression.** A natural example of a model with nuisance parameters that enter linearly is linear IV. In particular, consider the linear IV model written in reduced-form:

$$
\begin{cases}
Y = \beta \pi Z + v \\
X = \pi Z + u
\end{cases}
$$

where $Y$ is a $n \times 1$ vector of realizations of the dependent variable, $X$ is an endogenous regressor of interest, and $Z$ is a set of $k$ instrumental variables. The hypothesis of
interest is $H_0 : \beta = \beta_0$, where the nuisance parameter $\pi$ takes values in $\mathbb{R}^k$.

To construct a minimum-distance statistic we match reduced-form coefficients. Let $\theta(\beta, \pi) = (\beta \pi', \pi')'$ be the link function between the $2k \times 1$-dimensional reduced form parameter $\theta$ and structural parameters $\beta$ and $\pi$. We take

$$\hat{\theta} = vcc((Z'Z)^{-1}Z'(Y, X))$$

to be the OLS estimator of the reduced-form coefficients. Under mild assumptions $\hat{\theta}$ is consistent and $\sqrt{n}$-asymptotically normal with consistently estimable variance matrix $\Sigma$. Algebraic manipulation shows that the minimum distance statistic in this case is a version of the well-known Anderson-Rubin (1949) statistic. The null manifold

$S = \{ \sqrt{n}\Sigma^{-1/2}(\theta(\beta_0, \pi) - \theta_0), \pi \in \mathbb{R}^k \}$

is a $k$-dimensional linear subspace in $\mathbb{R}^{2k}$. Thus, by the argument above we can use $\chi^2_k$ critical values. Hence, in this case our approach gives us the Anderson-Rubin test, which is known to be robust to weak instruments. This example happens to be trivial from a theoretical perspective, as the nuisance parameter $\pi$ enters linearly.

### 3.2 Geometry

In the previous section we argued that the distribution of the statistic $MD$ depends on the shape of the null hypothesis manifold $S$. In this section, we begin by introducing a number of geometrical concepts. Using these tools, and in particular the curvature of $S$, we bound $MD$ from above for each realization of the data. From this realization-by-realization bound we then derive an upper bound on the distribution of $MD$ that depends only on the dimension and curvature of $S$.

#### 3.2.1 Manifolds, tangent spaces, curvature

In this paper we focus on regular manifolds embedded in $k$-dimensional Euclidean space with the usual Euclidean norm $\| \cdot \|$. A subset $S \subset \mathbb{R}^k$ is called a $p$-dimensional regular manifold if for each point $q \in S$ there exists a neighborhood $V$ in $\mathbb{R}^k$ and a
twice-continuously-differentiable map \( x : U \to V \cap S \) from an open set \( U \subseteq \mathbb{R}^p \) onto \( V \cap S \subseteq \mathbb{R}^k \) such that (i) \( x \) is a homeomorphism, which is to say it has a continuous inverse and (ii) the Jacobian \( dx_q \) has full rank. A mapping \( x \) which satisfies these conditions is called a parametrization or a system of local coordinates, while the set \( V \cap S \) is called a coordinate neighborhood.

Note that the manifold \( S \) is defined as a set, rather than as a map. In keeping with this spirit, many of the statements below will be invariant to parametrization. Hence, if we have different parameterizations for the same manifold, which of them we use is entirely a matter of convenience. In some problems it may be the case that there does not exist a global parametrization, that is a fixed mapping \( x \) satisfying the conditions above such that \( S \) is the image of \( x \). For instance, suppose that as discussed in Section 3.1, \( S \) is defined as the set of points \( q \) satisfying the \( k - p \)-dimensional restriction \( g(q) = 0 \). If we assume that the twice continuously differentiable function \( g \) has a Jacobian of rank \( k - p \) at all points, the Implicit Function Theorem guarantees the existence of a local parametrization in a neighborhood of each point \( q \in S \) but a global parametrization need not exist.

We begin by developing some geometrical concepts for the special case of a regular 1-dimensional manifold, also known as a curve. In particular, let \( S \) be a curve given by \( \alpha : (t_0, t_1) \to \mathbb{R}^k \) where \( \alpha \) is twice continuously differentiable and \( (t_0, t_1) \) is an interval in \( \mathbb{R} \). The arc length is defined as \( s(t) = \int_{t_0}^{t} \| \dot{\alpha}(\tau) \| \, d\tau \). Without loss of generality, we can take \( \alpha \) to be parameterized by arc length \( s \), in which case at all points \( \| \dot{\alpha}(s) \| = 1 \) and the vector \( \ddot{\alpha}(s) \) is perpendicular to \( \dot{\alpha}(s) \). The vector \( \dot{\alpha}(s) \) is called the tangent vector to \( S \) at \( q = \alpha(s) \), while \( \kappa_q(S) = \| \ddot{\alpha}(s) \| \) is called the curvature at \( q \). The curvature measures how quickly the curve \( S \) deviates from its tangent line local to \( q \), and the scaling is such that a circle of radius \( C \) has curvature \( 1/C \) at all points.

The change of variables from the arbitrary parametrization \( t \) to arc length \( s \) is not necessary for the calculation of curvature. In particular, as before let \( \dot{\alpha}(t) \) and \( \ddot{\alpha}(t) \) denote the first and second derivatives of \( \alpha \), now with respect to \( t \). If we let \( (\ddot{\alpha}(t))_{\perp} \) be the part of \( \ddot{\alpha}(t) \) orthogonal to \( \dot{\alpha}(t) \), then the curvature at \( q = \alpha(t) \) is \( \kappa_q(S) = \frac{\| (\ddot{\alpha}(t))_{\perp} \|}{\| \dot{\alpha}(t) \|^2} \). One can show that this definition of curvature is invariant to parametrization, and
hence that in the special case of a curve parameterized by arc length it reduces to the definition given above.

These concepts can all be extended to general regular manifolds. Fixing a $p$-dimensional manifold $S$, for any curve $\alpha : (-\varepsilon, \varepsilon) \to S$ on $S$ which passes through the point $q = \alpha(0) \in S$ the tangent vector $\dot{\alpha}(0)$ is called a tangent vector to $S$ at $q$. For $x$ a system of local coordinates at $q$, the set of all tangent vectors to $S$ at $q$ coincides with the linear space spanned by the Jacobian $dx_q$ and is called the tangent space to $S$ at $q$ (denoted $T_q(S)$). While we have defined the tangent space using the local coordinates $x$, as one would expect from its geometrical interpretation $T_q(S)$ is independent of the parametrization.

To calculate the curvature at $q$, consider a curve $\alpha : (t_0, t_1) \to S$ which lies in $S$ and passes through $q = \alpha(0)$. Taking $T_q^\perp$ to be the $k - p$-dimensional linear space orthogonal to $T_q(S)$, define

$$\kappa_q(\alpha, S) = \frac{\|\langle \dot{\alpha}(0) \rangle^\perp \|}{\|\dot{\alpha}(0)\|^2},$$

where $(W)^\perp$ stands for the projection of $W$ onto the space $T_q^\perp$. One can show that $\kappa_q(\alpha, S)$ depends on the curve $\alpha$ only through $\dot{\alpha}(0)$, so for two curves $\alpha$ and $\alpha^*$ in $S$ with $\alpha(0) = \alpha^*(0) = q$ and $\dot{\alpha}(0) = \dot{\alpha}^*(0)$ we have $\kappa_q(\alpha, S) = \kappa_q(\alpha^*, S)$. We can also show that for any $X \in T_q(S)$ one can find a curve $\alpha$ in $S$ with property that $\alpha(0) = q$ and $\dot{\alpha}(0) = X$. The measure of curvature we consider is

$$\kappa_q(S) = \sup_{X \in T_q(S), \dot{\alpha}(0) = X} \kappa_q(\alpha, S) = \sup_{X \in T_q(S), \dot{\alpha}(0) = X} \frac{\|\langle \dot{\alpha}(0) \rangle^\perp \|}{\|\dot{\alpha}(0)\|^2}. \quad (3.7)$$

This measure of curvature is closely related to the Second Fundamental Tensor (we refer the interested reader to Kobayashi and Nomizu (1969, v.2, ch. 7)), and is equal to the maximal curvature over all curves passing through the point $q$. As with the curvature measure discussed for curves, (3.7) is invariant to the parametrization. Also analogous to the 1-dimensional case, if $S$ is a $p$-dimensional sphere of radius $C$ then for each $q \in S$ we have $\kappa_q(S) = 1/C$. Finally, if $S$ is a linear subspace its curvature
is zero at all points.

**How to calculate curvature in practice.** Let $S$ be a $p$-dimensional manifold in $\mathbb{R}^k$, and let $x$ be a local parametrization at a point $q$, $q = x(y^*)$. Denote the derivatives of $x$ at $q$ by $v_i = \frac{\partial x}{\partial y_i}(y^*)$. By the definition of a local parametrization, we know that the Jacobian $Z = (v_1, ..., v_p)$ is full rank, so the tangent space $T_q(S) = \text{span}\{v_1, ..., v_p\}$ is $p$-dimensional. As before, for any vector $W \in \mathbb{R}^k$ let $W^\perp$ denote the part of $W$ orthogonal to $T_q(S)$, that is, $W^\perp = N_Z W = (I - Z(Z'Z)^{-1}Z')W$. Finally, denote the $p^2$ vectors of second derivatives $V_{ij} = \frac{\partial^2}{\partial y_i \partial y_j} x(y^*)$. The curvature can then be written as

$$K_q(S) = \sup_{u = (u_1, ..., u_p) \in \mathbb{R}^p} \left\| \sum_{i,j=1}^{p} u_i u_j V_{ij} \right\| = \sup_{(w_1, ..., w_p) \in \mathbb{R}^p} \left( \frac{\sum_{i,j=1}^{p} w_i w_j V_{ij}}{\left\| \sum_{i=1}^{p} w_i v_i \right\|^2} \right). \quad (3.8)$$

Notice that here we calculate maximal curvature over all directions in the tangent space; the possibility of calculating curvature over only a subset of directions is discussed in Section 3.3.2.

**3.2.2 Geometric bounds**

In this section we establish a bound on the distance in $\mathbb{R}^k$ from a random vector $\xi \sim N(0, I_k)$ to a $p$-dimensional non-random manifold $S$ that contains zero. Our bound depends on the maximal curvature $\kappa_q(S)$ over all relevant points in the manifold $S$. Our bound will be based on global properties of the manifold, in the sense of properties that hold on a fixed bounded set, but we abstract from the behavior of the manifold at infinity as irrelevant. In what follows, we restrict attention to a connected part of the manifold that lies inside of a finite cylinder centered at zero.

We derive our bound in two steps: first, we construct an envelope for the manifold $S$ using a collection of $p$-dimensional spheres. We show that the distance from any point $\xi$ to $S$ is bounded above by the distance from $\xi$ to the most distant sphere in
the collection we consider. Second, we show that our geometric construction implies a bound on the distribution of $\rho^2(\xi, S)$ and hence on the distribution of the minimum distance statistic. To provide intuition for our main statement we walk the reader through two simple cases in which the construction of the envelope can be easily visualized.

**Case 1 (k=2, p=1): A curve in $\mathbb{R}^2$.** Consider a curve $S$ passing through zero (i.e. $(0,0) \in S$). Suppose that the curvature of $S$ is less than or equal to $1/C$ for all points in $S$. If we imagine two circles of radius $C$ tangent to $S$ at zero, we can see that the curve lies between them- see Figure 3-1 for illustration. Since $S$ lies between the circles, the distance from any point $\xi$ to $S$ (denoted by $d_1$ in Figure 3-1) does not exceed the distance from $\xi$ to the further of the two circles (denoted by $d_2$). This is the geometrical bound we use. Note that if the maximal curvature of $S$ goes to zero at all points (so that $C \to \infty$) then the two bounding circles converge to the tangent line to $S$ at zero on any bounded set. Further, note that the distribution of the distance $d_2$ from a normal random vector to the furthest of two circles depends only on $C$ and is easy to simulate.

![Figure 3-1: Bounding a line between two circles.](image)

The logic of this example is quite straightforward to generalize to the case of a $k-1$-dimensional manifold in $\mathbb{R}^k$, known as a hyper-surface or a manifold of co-dimension 1.\(^2\) If a regular $k-1$ dimensional manifold $S$ in $\mathbb{R}^k$ has curvature $\kappa_q(S) \leq 1/C$ at

\(^2\)The co-dimension of a manifold is the difference between the dimension of the space and the dimension of the manifold.
all $q \in S$, consider the two $k - 1$-dimensional spheres of radius $C$ that are tangent to the manifold at zero. One can show (see Theorem 10 below) that as in the one-dimensional case $S$ lies between these two spheres. Hence, we again have that the distance from any point $\xi$ to $S$ is bounded above by the distance from $\xi$ to the furthest of the two spheres. Likewise, if the maximal curvature of $S$ goes to zero (so that $C \to \infty$) we again have that on any bounded set the two spheres converge to the tangent space to $S$ at zero, which in this case is a $k - 1$-dimensional hyperplane.

Dealing with manifolds of co-dimension greater than 1 is much more challenging, but the basic principle of the approach can be illustrated using a curve in $\mathbb{R}^3$.

**Case 2 (k=3, p=1): A curve in $\mathbb{R}^3$.** Suppose now that we have a one-dimensional space curve $S$ in $\mathbb{R}^3$ which passes through zero and whose curvature at all points is bounded above by $1/C$. We construct our envelope by considering the collection of all one-dimensional circles of radius $C$ tangent to $S$ at zero. Equivalently, one can take a given circle tangent to $S$ at zero and rotate it around the tangent line. An example of the resulting surface is given on the left panel of Figure 3-2: as in the case of co-dimension 1, we can see that the curve $S$ lies inside the envelope. One can show that the distance from any point $\xi$ to the curve $S$ (denoted by $d_1$ in Figure 3-2) is bounded above by the distance from $\xi$ to the furthest circle in the collection used.
to construct the envelope (denoted \(d_2\)). Note that if the curvature of \(S\) goes to zero at all points (so that \(C \to \infty\)) then on any bounded set the envelope we consider converges to the tangent line to \(S\) at zero.

This geometric bound immediately implies a bound on the distribution of \(\rho^2(\xi, S)\). For \(\xi \sim N(0, I_3)\) the distribution of the distance \(d_2\) from \(\xi\) to the furthest circle is simple to simulate. One can show that it is distributed as the squared distance from a two-dimensional random vector \(\eta\) depicted on the right panel of Figure 3.2 to the circle of radius \(C\) with center \((0, -C)\) where the coordinates of \(\eta\) are distributed as independent \(\sqrt{\chi_1^2 + \chi_2^2}\) random variables.

**General case**  With the intuition provided by these examples, we now turn to the general case. Let \(S\) be a regular connected \(p\)-dimensional manifold in \(\mathbb{R}^k\) passing through zero. By the rotation invariance of standard normal vectors we can assume without loss of generality that the tangent space \(T_0(S)\) to manifold \(S\) at zero is spanned by the first \(p\) basis vectors. For each \(x \in \mathbb{R}^k\), let \(x = (x^{(1)}, x^{(2)})\) where \(x^{(1)} = (x_1, \ldots, x_p) \in \mathbb{R}^p\) contains the first \(p\) coordinates of \(x\) while \(x^{(2)} = (x_{p+1}, \ldots, x_k) \in \mathbb{R}^{k-p}\) contains the last \(k-p\). In what follows, we restrict attention to points on the manifold that lie inside of a (large) finite cylinder \(D_C = \{x = (x^{(1)}, x^{(2)}): \|x^{(1)}\| \leq C, \|x^{(2)}\| \leq C, x^{(1)} \in \mathbb{R}^p, x^{(2)} \in \mathbb{R}^{k-p}\} \subset \mathbb{R}^k\). Let \(S_C\) be the intersection \(S \cap D_C\) if it is connected or the connected part of \(S \cap D_C\) that passes through zero (that is, the part of \(S \cap D_C\) which can be reached by continuous paths lying in \(S \cap D_C\) which pass through zero) if \(S \cap D_C\) is not connected. Note that \(\rho(\xi, S) \leq \rho(\xi, S_C)\).

To obtain some of our bounding results, we need one further assumption:

**Assumption 7**  For any \(y^{(1)} \in \mathbb{R}^p\) with \(\|y^{(1)}\| \leq C\) there exists a point \(x \in S_C\) such that \(x^{(1)} = y^{(1)}\).

Assumption 7 requires that the projection of \(S_C\) on its tangent space at zero cover a \(p\)-dimensional ball of radius \(C\) centered at zero, and hence that \(S_C\) have dimension \(p\) in a global sense. By a local property we mean one that holds on an infinitesimal neighborhood of a point. In contrast, by a global property we mean one that holds on a fixed bounded set. We have already imposed a local dimensionality assumption on
S by restricting the rank of the tangent space at all points. The distribution of the minimum distance statistic, however, depends on global properties of the manifold S and so to bound the distribution we need a global dimensionality assumption. To illustrate why local dimensionality assumptions are insufficient, imagine a strip $S = \{(x, y, z) \in \mathbb{R}^3 : z = 0, -\varepsilon < y < \varepsilon, x \in \mathbb{R}\}$ in $\mathbb{R}^3$. At any point $q \in S$ the dimension of the tangent space is equal to 2, but if $\varepsilon > 0$ is small enough then S does not satisfy Assumption 1. For $\varepsilon$ sufficiently small, however, the distance from $\xi$ to S behaves like the distance from $\xi$ to the line $S^* = \{(x, y, z) : y = 0, z = 0\}$, which is one dimensional both locally and globally.

**Theorem 10** Let $S$ be a regular $p$-dimensional manifold in $\mathbb{R}^k$ passing through zero. Assume that the tangent space $T_0(S)$ is spanned by the first $p$ basis vectors. Assume that for some constant $C > 0$ we have that $\kappa_q(S) < \frac{1}{C}$ for all points $q \in S_C$. Then:

(a) Manifold $S_C$ lies inside the set $\mathcal{M} \cap D_C$, where

$$\mathcal{M} = \{\|x^{(1)}\|^2 + (C - \|x^{(2)}\|)^2 \geq C^2\}. \quad (3.9)$$

(b) If Assumption 1 is satisfied, then for any point $\xi \in \mathbb{R}^k$ we have

$$\rho(\xi, S) \leq \max_{u \in \mathbb{R}^{p-k}, \|u\| = 1} \rho(\xi, N_u),$$

where $N_u = \{x \in \mathbb{R}^k : x = (x^{(1)}, zu), x^{(1)} \in \mathbb{R}^p, z \in \mathbb{R}_+, \|x^{(1)}\|^2 + (C - z)^2 = C^2\}$.

(c) $\max_{u \in \mathbb{R}^p, \|u\| = 1} \rho(\xi, N_u) = \rho(\xi, N_{\tilde{u}})$, where $\tilde{u} = -\frac{1}{\|\xi^{(2)}\|} \xi^{(2)}$.

(d) If $\xi \sim N(0, I)$ we have for all $x, y$:

$$P \left\{ \max_{u \in \mathbb{R}^{p-k}, \|u\| = 1} \rho^2(\xi, N_u) \leq x, \|\xi\| \leq y \right\} = P \left\{ \rho^2(\eta, N^C_{x,y}) \leq x, \|\eta\| \leq y \right\},$$

where the coordinates of the 2-dimensional random vector $\eta = (\sqrt{\xi_1^2}, \sqrt{\xi_2^2}) \in \mathbb{R}^2$ are independently distributed, $N^C_{x,y} = \{(z_1, z_2) \in \mathbb{R}^2 : z_1^2 + (C + z_2)^2 = C^2\}$ is
a circle of radius $C$ with the center at $(0, -C)$, and $\rho_2$ is Euclidian distance in $\mathbb{R}^2$.

Figure 3-3: The stochastic bound described in Theorem 10 (d).

Theorem 10 (a) establishes that the manifold $S_C$ lies inside the set $\mathcal{M}$ bounded by an envelope we construct from a collection of $p$-dimensional spheres $N_u$. Statement (b) asserts that the distance from a point $\xi$ to the manifold $S$ is bounded by the distance from $\xi$ to the furthest sphere in this collection, while (c) picks out exactly which sphere $N_{\xi(\xi)}$ is the furthest away for a given $\xi$. Finally, (d) shows that the distribution of the distance from $\xi \sim N(0, I_k)$ to $N_{\xi(\xi)}$ is the same as the distribution of the distance from a random variable $\eta$ to a particular circle in $\mathbb{R}^2$ as depicted in Figure 3-3.

Theorem 10 (b) relies on Assumption 7. In some models, like our empirical example in Section 3.5, this assumption holds trivially. In other contexts, we may try to check Assumption 7 numerically. In particular, we can draw points at random from the manifold $S$, for example running MCMC on the restricted model. We can then project these draws on the tangent space. As the number of draws increases, we can check that the sampler draws points in all parts of the ball of radius $C$ in the tangent space. This bears a resemblance to how one would check that a Gibbs sampling procedure visits the whole parameter space.
3.2.3 Stochastic bound

Theorem 10 implies a bound on the distribution of the distance from $\xi \sim N(0, I_k)$ to a $p$-dimensional manifold $S$. Assume that for some $C > 0$, $S$ satisfies all the assumptions of Theorem 10 including Assumption 7. Then almost surely,

$$\rho^2(\xi, S) \leq \rho^2(\xi, N_2^C)$$

(3.10)

as follows from statements (b) and (c) of Theorem 10. By Theorem 10 (d), the distribution of the right hand side of (3.10) is the same as the distribution of the random variable $\psi_C$ defined as

$$\psi_C = \rho^2(\eta, N_2^C)$$

(3.11)

where the coordinates of the two-dimensional random vector $\eta = (\sqrt{\chi^2_{p}}, \sqrt{\chi^2_{k-p}}) \in \mathbb{R}^2$ are independently distributed, $N_2^C = \{(z_1, z_2) \in \mathbb{R}^2 : z_1^2 + (C + z_2)^2 = C^2\}$ is a circle of radius $C$ with the center at $(0, -C)$, and $\rho_2$ is Euclidean distance in $\mathbb{R}^2$. Combining these results, we establish the bound

$$P\{\rho^2(\xi, S) \geq x\} \leq P\{\psi_C \geq x\} \quad \text{for all } x > 0,$$

so the distribution of $\psi_C$ is an upper bound on the distribution of $\rho^2(\xi, S)$. We make the following observations:

1. The distribution of $\psi_C$ depends only on the dimension of the space $k$, the dimension of the manifold $p$ and the maximal value of the curvature, $\frac{1}{C}$.

2. The distribution of $\psi_C$ is stochastically increasing in the maximal curvature and hence stochastically decreasing in $C$, so if $C_1 < C_2$ then $\psi_{C_1}$ first-order stochastically dominates $\psi_{C_2}$.

3. $\psi_C \Rightarrow \chi^2_{k-p}$ as $C \to \infty$, so if the curvature converges to zero at all relevant points then our bounding distribution converges to the distribution of the distance from $\xi \sim N(0, I_k)$ to a $p$-dimensional linear subspace.
(4) At the other extreme, $\psi_C \Rightarrow \chi_k^2$ as $C \to 0$ so if the curvature of the manifold becomes arbitrarily large our bound coincides with the naive bound (3.3) that can be imposed without any assumptions on the manifold.

We want to emphasize that what we suggest is a stochastic bound that holds under quite general assumptions. If the model of interest has additional structure, this can potentially be exploited to obtain tighter bounds.

### 3.3 Statistical application of the stochastic bound and its modifications

In this section, we discuss how we can use the bounds obtained above to construct tests based on $MD$ which control size in finite samples under the assumption that $\hat{\theta} - \theta_0$ is normally distributed with known variance $\Sigma$. In this section we also propose three modifications of our baseline procedure that offer practical advantages in some contexts and which, as we discuss below, may be freely combined to suit the preferences of the researcher and the circumstances at hand.

As we note in equation (3.2) the minimum distance statistic is equal to $\rho^2(\xi, S)$, where $\xi \sim N(0, I_k)$ and the manifold $S = \{\Sigma^{-1/2}(\theta(\beta) - \theta_0), \beta \in U \subset \mathbb{R}^p\} \subset \mathbb{R}^k$ describes the restrictions imposed on the reduced-form parameters by the null, and passes through zero if the null is true. If the manifold $S$ satisfies the assumptions of Theorem 10 then by the results of Section 3.2.3 the $MD$ statistic is stochastically dominated by $\psi_C$ under the null. Thus if we use $F_{1-\alpha}(C, k, p)$, the $(1 - \alpha)$-quantile of $\psi_C$ (which is easy to simulate), as a critical value the resulting test has size not exceeding $\alpha$.

A practical question is what value of $C$ to use. According to Theorem 10, $C$ is tied to the maximal curvature of $S$ over the intersection of $S$ with a cylinder $D_C$ centered at zero. In practice, however, we do not observe the manifold $S$, which depends on the unknown $\theta_0$. Nonetheless, we can see that the desired curvature is the same as the maximal curvature of the observed manifold $S^* = \{\Sigma^{-1/2}\theta(\beta), \beta \in U \subset \mathbb{R}^p\} \subset \mathbb{R}^k$.
over all points in the intersection of $S^*$ with the cylinder $D_c^*(x_0) = \{ x \in \mathbb{R}^k : x - x_0 \in D_c \}$ centered at $x_0 = \Sigma^{-1/2} \theta_0$. This maximal curvature, in turn, is clearly bounded above by the maximal curvature over the whole manifold, so if we take $C^* = 1 / (\max_{\eta \in S^*} \kappa_{\eta^*}(S^*))$, using critical values based on $\psi_{C^*}$, provides a test that controls size. Moreover, since $C^*$ does not depend on any unobservables, a test based on these critical values is feasible.

If the null hypothesis has a global parametrization, as when $H_0 : \theta_0 = \theta(\beta)$, $\beta \in U$, let $\kappa(\beta) = \kappa_{\eta^*=\Sigma^{-1/2}\theta(\beta)}(S^*)$. The latter is a function on $U$ and depends only on the first two derivatives of $\theta(\beta)$. Hence, if we can evaluate these derivatives, finding $C^* = 1 / (\max_{\beta \in U} \kappa(\beta))$ is a standard non-stochastic optimization problem. If $\theta(\beta)$ is fairly tractable we may be able to solve for $C^*$ analytically, while if not we can use the usual menu of numerical optimization techniques, such as Newton’s method.

3.3.1 Modification 1: curvature over a smaller set

There are a variety of problems in which using $C^*$ may be unappealing. For example, it may be that searching numerically for the maximal curvature over the whole manifold is quite time-consuming, or that the manifold has irregularities or points of high curvature which are far away from $\hat{\theta}$. In such cases we may wish to restrict attention to the curvature of the manifold over some smaller set, which raises two issues. First, we do not know the true value $\theta_0$ and hence the center of the cylinder $D_c^*(x_0)$. Second, if the manifold is close to flat (so $C$ is large) to find the maximal curvature over $D_c^*(x_0)$ we might need to check the curvature over a huge set, which could again be very computationally demanding.

We suggest a test which overcomes both of these problems and is easy to implement in practice. For a fixed value $R$, let $C \wedge R = \min\{C, R\}$. Denote by $F_\alpha(C, R, k, p)$ the $\alpha-$quantile of the distribution of $\psi_C(R)$ defined as

$$
\psi_C(R) = \left\{ \begin{array}{ll} 
\rho^2_2(\eta, N^C_2) & \text{if } ||\eta|| \leq R; \\
||\eta||^2 & \text{if } ||\eta|| > R,
\end{array} \right.
$$

(3.12)
where \( \eta \) and and \( N_2^C \) are defined in statement (d) of Theorem 10. For any finite \( R \) the distribution of \( \psi_C(R) \) first order stochastically dominates the distribution of \( \psi_C \). In Lemma 6 below we show that one may calculate curvature only over that part of the manifold lying inside a ball of radius proportional to \( R \), but that one must compensate for this by using larger critical values, specifically quantiles \( \psi_C(R) \) rather than \( \psi_C \). This is the price paid for calculating curvature over a smaller set of points.

**Lemma 6** Assume that we have a single observation \( \hat{\theta} \) from a population \( \hat{\theta} \sim N(\theta_0, \Sigma) \) with unknown mean \( \theta_0 \). We wish to test the hypothesis \( H_0 : \theta_0 = \theta(\beta) \) for some \( \beta \in U \subset \mathbb{R}^p \). Let \( S^* = \{ \Sigma^{-1/2}\theta(\beta), \beta \in U \subset \mathbb{R}^p \} \subset \mathbb{R}^k \) be a regular \( p \)-dimensional manifold, and \( B^* = B_{R}(\bar{x}) \) a ball of radius \((1 + \sqrt{2})R\) around \( \bar{x} = \Sigma^{-1/2}\hat{\theta} \), where \( R \) is such that \( P\{X^2 \geq R^2\} < \alpha \). Let

\[
C_R^* = \begin{cases} 
\min_{\eta \in S^* \cap B^*} \{ 1/k_{\eta}(S^*) \} \wedge R, & \text{if } S^* \cap B^* \neq \emptyset; \\
0, & \text{if } S^* \cap B^* = \emptyset.
\end{cases}
\]

Assume that for any \( x \in S^* \) such that \( \|x - \bar{x}\| \leq R \) we have that the projection of \( S^* \cap B_R(x) \) onto \( T_x(S^*) \) contains a \( p \)-dimensional ball centered at \( x \) with radius \( C_R^* \wedge R \). Then the test which rejects the null if and only if \( MD > F_{1-\alpha}(C_R^*, R, k, p) \) has size not exceeding \( \alpha \).

**3.3.2 Modification 2: Working with a subset of parameters**

Suppose we wish to test a hypothesis of the form \( H_0 : \theta_0 = \theta(\beta) \) for some value of the \( p \)-dimensional structural parameter \( \beta \in \mathbb{R}^p \). The procedures discussed above treat the multi-dimensional vector \( \beta \) in such a way that only the direction of highest curvature affects the value of \( C \) and thus influences the critical values. Imagine instead that \( \beta \) can be divided into two sub-sets of parameters \( \beta = (\beta'_1, \beta'_2)' \) in such a way that the curvature corresponding to directions \( \beta_1 \) is high, but the null hypothesis exhibits only a low degree of curvature in the directions corresponding to \( \beta_2 \). Let \( p_1 \) be the dimension of \( \beta_1 \), \( p_2 \) the dimension of \( \beta_2 \), and \( p = p_1 + p_2 \). In this section we propose a modified procedure that treats \( \beta_1 \) and \( \beta_2 \) differently. In particular we may reduce the
critical value of the test due to the low curvature with respect to $\beta_2$ while projecting over $\beta_1$. The modified procedure may be more powerful if the difference in curvature with respect to parameters $\beta_1$ and $\beta_2$ is large.

For any value $\beta_1$ consider the $p_2$-dimensional manifold $S(\beta_1) = \{\Sigma^{-1/2} \theta(\beta_1, \beta_2), \beta_2 \in \mathbb{R}^{p_2}\}$. For any point $q = \Sigma^{-1/2} \theta(\beta_1, \beta_2) \in S(\beta_1)$ find the curvature $\kappa_q(S(\beta_1))$. Let

$$C_{\beta_2} = \min_{\beta_1} \min_{q \in S(\beta_1)} 1/\kappa_q(S(\beta_1)),$$

be the inverse of the maximal curvature with respect to $\beta_2$, where the maximum is taken over all $p_2$-dimensional sub-manifolds $S(\beta_1)$. One also needs that the analog of Assumption 7 is satisfied, specifically that for the true value $\beta_{1,0}$ the projection of $S(\beta_{1,0})$ onto its tangent space at zero covers a $p_2$-dimensional ball of radius $C_{\beta_2}$. The test which rejects the null if and only if $MD > F_{1-n}(C_{\beta_2}, k, p_2)$ controls size. Note that the new critical value may increase since $p_2 < p$, but it may also decrease if the curvature corresponding to directions $\beta_2$ is substantially smaller than the maximal curvature over all directions.

If we calculate curvature multiple times, projecting over any collection of different subsets of $\beta$ and using the smallest critical value, the resulting test will still control size. Thus, it is perfectly valid to search over all subsets of the parameter vector and choose the one which, when we project over it, yields the smallest critical value. Moreover, we can combine projection with the approach discussed in Section 3.3.1 and calculate the maximal curvature of $S(\beta_1)$ only on a ball of radius $(1 + \sqrt{2})R$ around the reduced-form parameter estimate, and the resulting tests will again control size.

### 3.3.3 Modification 3: Pre-test for weak identification

If for some reason a researcher prefers not to use our robust critical values, we propose a simple “pre-test of weak identification” which is of a similar flavor to procedures based on the first-stage $F$-statistic in linear IV. Specifically, imagine that a researcher wants to use a robust procedure (for example our robust critical values or the projection method) unless she knows that nonlinearity will not cause large size distortions,
in which case she prefers to use standard, non-robust critical values. Our bounds can be used to address this question, and in particular to determine whether nonlinearity constitutes a problem in a given setting. Below we suggest an approach which, used as the first step in a two-step test, ensures that the procedure as a whole controls size.

To proceed, let us introduce the notion of a “tolerance level”. Suppose we would like to have a test of size $\alpha$, but are uncertain whether the usual strong-identification asymptotics provide a reasonable approximation in our context; in the event that these approximations are imperfect, we are willing to accept a test with true size $\alpha + \alpha^*$ in exchange for the additional power and convenience of using conventional critical values. The potential size distortion $\alpha^*$ is called the tolerance level and was previously discussed by e.g. Stock and Yogo (2005).

The pretest we propose asks whether the curvature of the model is sufficiently small to ensure that tests based on classical $\chi^2_{k-p}$ critical values (that is, tests which treat the nuisance parameter $\beta$ as well-identified) with nominal size $\alpha$ have true size not exceeding $\alpha + \alpha^*$. To determine whether this is the case we calculate $C^*$, the smallest value $C$ such that the $(1-\alpha)$-quantile of a $\chi^2_{k-p}$ distribution does not exceed $F_{1-(\alpha+\alpha^*)}(C, k, p)$. The cut-off $C^*$ depends on the dimension $k$ of the reduced-form parameter vector and the dimension $p$ of the nuisance parameter. To implement the pre-test we can calculate the value $C$ over the whole manifold and compare $C$ to $C^*$. If $C > C^*$ the researcher can safely concentrate out $\beta$ and use $\chi^2_{k-p}$ critical values while if $C \leq C^*$ she should use a robust procedure. We can guarantee that the resulting two-step test will have size less than $\alpha + \alpha^*$. Table S1 in the Supplementary Appendix reports the cut-offs $C^*$ for nominal 5% tests with tolerance level 5% for different values of $p$ and $k$. We can see that for a fixed dimension $k$ of the reduced-form parameter, increasing the number of nuisance parameters $p$ tightens the restrictions imposed on curvature if one wants to concentrate out the nuisance parameters. Rather than calculating the maximal curvature over the whole manifold, or over all parameters, we can also create pre-tests based on the modified robust critical values discussed in Sections 3.3.1 and 3.3.2.
The main difference between the pre-testing procedures we suggest and the majority of existing tests for weak identification is that our approach guarantees that the true size does not exceed the pre-specified tolerance level. In contrast, many alternative approaches to detecting identification failure in nonlinear models, for example those discussed by Inoue and Rossi (2011), Iskrev (2010) and Wright (2003), do not control the size distortion of two-step testing procedures.

3.4 Asymptotic procedure

3.4.1 Uniformity result

The main procedure described above guarantees that finite-sample size is controlled when the reduced-form parameter estimates are normally distributed with a known covariance matrix. In this section we show that the procedure is asymptotically correct uniformly over a large set of models for which the reduced-form parameter estimator is asymptotically Gaussian.

We take a model to be a set consisting of the true value of the $k$-dimensional reduced-form parameter $\theta_0$, its estimator $\hat{\theta}$, the true value of uncertainty associated with this estimate (asymptotic covariance matrix $\Sigma$), an estimator $\hat{\Sigma}$, and a link function connecting the structural and reduced form parameters, or more generally a manifold $\tilde{S}$ describing the null hypothesis. The hypothesis of interest is $H_0: \theta \in \tilde{S}$. Here by estimators $\hat{\theta}$ and $\hat{\Sigma}$ we understand some procedures or algorithms for producing such estimators, where $\hat{\theta}_n$ and $\hat{\Sigma}_n$ are estimators for the sample of size $n$. We allow $\tilde{S}$ to change with the sample size $n$ as well, and will denote it $\tilde{S}_n$. This allows for sequences of link functions such as those which arise under drifting asymptotic embeddings like those of Andrews and Cheng (2012) and Stock and Wright (2000).

We consider a set of models $\mathcal{M} = \{M: M = (\theta_0, \hat{\theta}, \hat{\Sigma}, \tilde{S})\}$ satisfying the following assumptions.

**Assumption 2**

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, \Sigma)$ uniformly over $\mathcal{M}$;
(ii) $\hat{\Sigma}_n \rightarrow^p \Sigma$ uniformly over $\mathcal{M}$;

(iii) The highest and lowest eigenvalues of $\Sigma$ are bounded above and away from zero uniformly over $\mathcal{M}$;

(iv) For each $n$ the manifold $S_n = \{x = \sqrt{n}\Sigma^{1/2}(y - \theta_0), y \in \tilde{S}_n\}$ belongs to the set $\mathcal{S}$ described below.

**Assumption 3** Let $\mathcal{S}$ be a set of regular $p$-dimensional manifolds in $\mathbb{R}^k$ passing through zero where for each $S \in \mathcal{S}$ there is some $C(S)$ such that for $S \in \mathcal{S}$ as defined in Section 3.2.2, $C(S) = 1/(\sup_{q \in \tilde{S}_n} \kappa_q(S))$ and the orthogonal projection of $S_C$ onto the tangent space $T_0(S)$ covers the ball of radius $C$ centered at zero.

**Description of the procedure.** Let us introduce a manifold $\tilde{S}_n = \{\sqrt{n}\Sigma_n^{-1/2}(x - \theta_0) : x \in \tilde{S}_n\}$, which differs from $S_n$ in using an estimator $\hat{\Sigma}_n$ in place of $\Sigma$. Let $\tilde{C}_n = 1/(\sup_{q \in \tilde{S}_n} \kappa_q(\tilde{S}_n))$. Our main test uses the statistic $n \min_{q \in \tilde{S}_n} (\hat{\theta}_n - \theta)^T \hat{\Sigma}_n^{-1}(\hat{\theta}_n - \theta)$ along with critical value $F_{1-\alpha}(\tilde{C}_n, k, p)$, where we denote by $F_n(C, k, p)$ the $\alpha$-quantile of the random variable $\psi_C$ discussed in Section 3.2.3.

**Theorem 11** The testing procedure described above has uniform asymptotic size $\alpha$:

$$\lim_{n \to \infty} \sup_{M \in \mathcal{M}} P \left\{ n \min_{\theta \in \tilde{S}_n} (\hat{\theta}_n - \theta)^T \hat{\Sigma}_n^{-1}(\hat{\theta}_n - \theta) > F_{1-\alpha}(\tilde{C}_n, k, p) \right\} \leq \alpha$$

In the Supplementary Appendix we show that the test which maximizes curvature over the ball of radius $(1 + \sqrt{2})R$ around $\tilde{\Sigma}^{-1/2}\hat{\theta}$ (the modification suggested in Section 3.3.1) likewise has correct uniform asymptotic size over $\mathcal{M}$, and establishing analogous results for the other modifications discussed in the previous section is straightforward.

**3.4.2 Curvature of strongly identified parameters.**

In Section 3.1.1 we motivated our geometric perspective by showing that if a set of nuisance parameters is strongly identified in the sense of Stock and Wright (2000), the null hypothesis is asymptotically linear in these parameters. Here we show that the
curvature corresponding to strongly identified parameters is asymptotically of order $O(1/\sqrt{n})$, where $n$ is the sample size.

Consider a sample of size $n$ from a model parameterized by structural parameter $\beta$ that belongs to some bounded set $U \subseteq \mathbb{R}^p$ and assume that one has consistent and asymptotically normal estimates for the reduced-form parameter $\theta$:

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, \Sigma).$$

Assume that the relation between structural and reduced-form parameters $\theta = \theta(\beta)$ is fixed (not changing with $n$), twice continuously differentiable with respect to $\beta$, and that the matrix $\frac{\partial}{\partial \beta} \theta(\beta)$ has full rank in a neighborhood of $\beta_0$, which is the only point in the closure of $U$ that solves the equation $\theta_0 = \theta(\beta)$. The null hypothesis manifold $S_n$ for sample size $n$ is the graph of function $x_n(\beta) = \sqrt{n}\Sigma^{-1/2}(\theta(\beta) - \theta_0), \beta \in U$. The maximal curvature over all points of the manifold $S_n$ is equal to $1/\sqrt{n}$ times the maximal curvature of the manifold $S_1$ obtained for the sample of size 1, assuming the maximal curvature is finite. This can easily be seen by examining the role of the scale of parametrization $x_n$ in formula (3.8). As a result, the inverse curvature $C_n$ for the sample of size $n$ is proportional to $\sqrt{n}$ and so diverges to infinity. Consequently, the critical value $F_{1-\alpha}(C_n, k, p)$ converges to the $(1-\alpha)$-quantile of the $\chi^2_{k-p}$-distribution, which is the true asymptotic distribution under the assumption of the strong identification.

To summarize, our procedure makes no assumption about the strength of identification of the nuisance parameter but if the nuisance parameter happens to be strongly identified our robust critical values converge to those derived under the assumption of strong identification. Thus, our baseline approach to constructing robust critical values does not sacrifice power in cases with strongly identified nuisance parameters.
3.4.3 Comparison with other methods available for testing hypotheses with weak nuisance parameters

As previously discussed, there is a wide literature discussing weak-identification-robust tests for the full parameter vector and for hypotheses with strongly identified nuisance parameters, but much less is known about testing with weakly identified nuisance parameters.

**Projection method.** The projection method has recently been the standard approach to inference with weakly identified nuisance parameters. The projection method was popularized in econometrics by Dufour and Jasiak (2001) and Dufour and Taamouti (2005), and recent applications to non-standard testing problems in econometrics include Dufour, Khalaf, and Kichian (2006), Guerron-Quintana, Inoue and Kilian (2013), and Qu (2013).

The projection method is based on the observation that the hypothesis $H_0 : \alpha = \alpha_0$ with nuisance parameter $\beta$ is equivalent to the hypothesis

$$H_0 : \exists \, \beta_0 \text{ s.t. } \alpha = \alpha_0, \beta = \beta_0.$$

Let $MD(\alpha_0, \beta_0) = n(\hat{\theta} - \theta(\alpha_0, \beta_0))' \hat{\Sigma}^{-1}(\hat{\theta} - \theta(\alpha_0, \beta_0))$ be the test statistic for the hypothesis $H_0 : \alpha = \alpha_0, \beta = \beta_0$, and note that under the assumptions of Theorem 11 it is uniformly asymptotically $\chi_k^2$ under the null. Recall that the minimum distance statistic for testing a hypothesis on $\alpha$ alone is $MD(\alpha_0) = \inf_{\beta} MD(\alpha_0, \beta^*)$, so since

$$MD(\alpha_0) = \inf_{\beta^*} MD(\alpha_0, \beta^*) \leq MD(\alpha_0, \beta_0) \Rightarrow \chi_k^2,$$

we know that comparing $MD(\alpha_0)$ to $\chi_k^2$ critical values will yield a test which controls size. The name "projection method" stems from the fact that constructing confidence sets for $\alpha$ based on this procedure is equivalent to constructing a joint confidence set for $(\alpha_0, \beta_0)$ using the full-vector $MD$ statistic and then projecting this set on the parameter space for $\alpha$. 

187
The obvious advantage of the projection method is that it requires no assumption about the strength of identification of $\beta$, since it relies only on the validity of the test for the full parameter vector. Other advantages include its ease of implementation and broad applicability. The primary disadvantage of the projection method is its conservativeness. Our test, introduced in Section 3.3, is based on the same statistic as the projection method but uses smaller critical values while still maintaining size. Only in the limiting case of infinitely high curvature ($C = 0$) do our critical values equal those of the projection method. As a result, except for this limiting case our test is strictly more powerful than the projection method and produces smaller confidence sets in all realizations of the sample.

Concentrating out nuisance parameters. If one knows that the nuisance parameter $\beta$ in a given testing problem is strongly identified then he/she can simply “concentrate out” the nuisance parameter and pair the $MD(\alpha_0)$ statistic with $\chi^2_{k-p}$ critical values. As discussed in Section 3.1.1, this reduction in degrees of freedom (from $k$ to $k - p$) stems from the fact that any manifold corresponding to a hypothesis with a strongly identified nuisance parameter converges to a linear subspace asymptotically.

The obvious advantage of this approach is that it is strictly more powerful than the projection method. However, the assumption of strong identification of the nuisance parameter is essential, and the test may over-reject if this assumption fails. In many practical settings, including the DSGE and Phillips curve examples discussed below, the exact nature and source of weak identification is not clear, and we are unaware of any previous test of the null of weak identification which can be used to separate the weakly and strongly identified parameters. In contrast, the test we suggest in this paper does not employ any assumptions about the strength of identification of any parameter in point identified models. Indeed, since our approach is based on a finite-sample perspective we do not even require that there be a meaningful distinction between weakly and strongly identified structural parameters in the model.
Other methods. There are very few other papers that work directly with weakly identified nuisance parameters. One alternative approach, developed by D. Andrews and Cheng (2012), proceeds by assuming we know the structure of weak identification. In particular, they assume we know which parameters are weakly identified and that there is a known parameter that controls the strength of identification. Under these assumptions, they show that statistics for testing hypotheses with weakly identified nuisance parameters have non-standard asymptotic distributions which depend on the value of the nuisance parameter. They then propose creating robust tests by simulating the asymptotic distribution of the test statistic for different values of the nuisance parameter and taking the "least favorable" among those distributions over a set of relevant nuisance parameter values. Unfortunately, this approach can become quite computationally demanding in models with more than a few nuisance parameters. Moreover, the assumption that a known parameter controls the strength of identification rules out many models of economic interest, for example the new Keynesian Phillips curve and DSGE examples discussed below. As one might expect given the additional structure imposed by D. Andrews and Cheng's approach, in contexts where both their results and those developed in this paper can be applied, their approach will generally yield smaller critical values.

Another example of inference with weakly identified nuisance parameters is given in I. Andrews and Mikusheva (2013). That paper considers a case when concentrating out a weakly identified parameter leads to asymptotically correct inferences, but this result holds only for weakly identified parameters which enter the log-likelihood function linearly.

3.5 Empirical application: new Keynesian Phillips curve

To illustrate the methods developed in this paper we consider the application of our approach to an empirical example drawn from Magnusson and Mavroeidis (2010)
(henceforth MM), who study the problem of identification-robust inference on new Keynesian Phillips curve (NKPC) parameters. These parameters govern the behavior of inflation and play a central role in monetary policy-making. There is substantial evidence that NKPC parameters are weakly identified: see Mavroeidis, Plagborg-Moller and Stock (2013) for extensive discussion of this issue.

3.5.1 Model setting

MM consider identification-robust inference on new Keynesian Phillips curve parameters, building on the minimum-distance approach of Sbordone (2005). They study the simple NKPC model

\[
\pi_t = \frac{(1 - \nu)^2}{\nu(1 + \varrho)} x_t + \frac{1}{1 + \varrho} E[\pi_{t+1}|I_t] + \frac{\varrho}{1 + \varrho} \pi_{t-1} + \varepsilon_t
\]  

(3.14)

where \(x_t\) is a measure of marginal costs, \(\pi_t\) is inflation, \(E[\cdot|I_t]\) denotes an expectation conditional on information available at time \(t\), and \(\varepsilon_t\) is an exogenous shock with \(E[\varepsilon_{t+1}|I_t] = 0\). The structural parameters \(\nu\) and \(\varrho\) denote the degree of price stickiness and price indexation, respectively. MM further assume that \((\pi_t, x_t)\) follows a third order VAR process which can be written in companion form as

\[
z_t = A(\theta) z_{t-1} + \epsilon_t,
\]

where \(z_t = (\pi_t, x_t, \pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2})'\) is a \(6 \times 1\) vector, \(A(\theta)\) is a \(6 \times 6\) companion form matrix, the reduced-form parameter \(\theta\) is a vector of 12 unknown VAR coefficients, and \(\epsilon_t\) are VAR innovations with \(E[\epsilon_{t+1}|I_t] = 0\). MM note that the NKPC model (3.14) implies that the true parameter value \((\theta, \nu, \varrho)\) satisfies the 6-dimensional restriction

\[
f(\theta, \nu, \varrho) = A(\theta)' \left\{ \left[ I - \frac{1}{1 + \varrho} A(\theta)' \right] e_\pi - \frac{(1 - \nu)^2}{\nu(1 + \varrho)} e_x \right\} - \frac{\varrho}{1 + \varrho} e_\pi = 0,
\]

where \(e_\pi\) and \(e_x\) are unit vectors with \(e'_\pi z_t = \pi_t\) and \(e'_x z_t = x_t\).

Under mild assumptions the VAR estimate \(\hat{\theta}\) for the reduced-form parameter
θ satisfies $\sqrt{T} \left( \hat{\theta} - \theta \right) \to_d N(0, \Sigma)$, where the asymptotic covariance matrix Σ is consistently estimable, so from this perspective θ is strongly identified. MM propose testing the hypothesis $H_0 : (\nu, \varrho) = (\nu_0, \varrho_0)$ using the statistic

$$MM = T \cdot f \left( \hat{\theta}, \nu_0, \varrho_0 \right)^T \hat{\Sigma}_f^{-1} f \left( \hat{\theta}, \nu_0, \varrho_0 \right),$$

where $\hat{\Sigma}_f$ is a $\Delta$-method estimator for the asymptotic variance of $f \left( \hat{\theta}, \nu_0, \varrho_0 \right)$, together with $\chi_0^2$ critical values. MM argue that this test controls size regardless of whether the parameters $\nu$ and $\varrho$ are weakly identified.

The minimum-distance approach studied in this paper provides an alternative route to inference on $(\nu, \varrho)$. For any fixed $(\nu_0, \varrho_0)$ the restriction $f \left( \theta, \nu_0, \varrho_0 \right) = 0$ defines a 6-dimensional manifold in the 12-dimensional space of reduced-form parameter values $\theta$, $S_{\nu_0, \varrho_0} = \{ \theta \in \mathbb{R}^{12} : f \left( \theta, \nu_0, \varrho_0 \right) = 0 \}$. Hence, to test $H_0 : (\nu, \varrho) = (\nu_0, \varrho_0)$ we can consider the statistic

$$MD = \min_{\theta \in S_{\nu_0, \varrho_0}} T \cdot \left( \hat{\theta} - \theta \right)^T \hat{\Sigma}_g^{-1} \left( \hat{\theta} - \theta \right)$$

together with robust critical values based on the curvature of the manifold $S_{\nu_0, \varrho_0}$.

### 3.5.2 Empirical Results

We calculate 95% confidence sets for $\nu$ and $\varrho$ based on the MD statistic together with our robust critical values and compare these confidence sets to those obtained from inverting the MM test. We use the same data as MM, which consists of quarterly observations on US GDP deflator $(\pi_t)$ and labor share $(x_t)$ from 1984 to 2008- see MM for details.

To implement our robust MD test, we need to calculate an appropriate curvature-based critical value. We show in the Supplementary Appendix that for all $(\nu_0, \varrho_0)$ the manifold $S_{\nu_0, \varrho_0}$ can be parametrized by a 6-dimensional sub-vector of VAR parameters.

---

3See the Supplementary Appendix for details. MM propose other identification-robust tests as well, but since these tests produce similar confidence sets in their empirical application we will focus on this one, which is the most closely related to our MD test.
Figure 3-4: 95% confidence sets based on inverting our robust MD test and the MM test proposed by Magnusson and Mavroeidis (2010). Robust MD test compares the MD statistic $MD_{\nu, \varphi}$ to $\chi^2_5$ critical values, while the MM test compares the MM statistic to $\chi^2_6$ critical values.

We find that the maximal curvature of $S_{\nu_0, \varphi_0}$ is quite high for many values $(\nu_0, \varphi_0)$, but that by projecting over two elements of $\theta_1$ we can reduce curvature with respect to the remaining parameters to zero. Hence, for all values $(\nu_0, \varphi_0)$ we use robust critical values equal to the 95th percentile of a $\chi^2_5$ distribution, or equivalently $F_{0.95}(\infty, 12, 4)$ where $p = 4$ rather than 6 because we project over two parameters. To calculate our robust MD confidence set for the parameters $\nu$ and $\varphi$ jointly, we take a fine grid of values $(\nu_0, \varphi_0)$ and for each value in this grid test the hypothesis $H_0 : (\nu, \varphi) = (\nu_0, \varphi_0)$. The confidence sets consist of all those values not rejected, and are plotted in Figure 3-4. As we can see from this figure, in this application our robust MD confidence set is contained entirely within MM’s confidence set.

In addition to building joint confidence sets for $(\nu, \varphi)$ together we can construct confidence intervals for these parameters separately. To do so, we first check whether it is possible to obtain smaller critical values by calculating curvature with respect
Table 3.1: 95% confidence intervals for $\nu$ and $\varrho$ based on our robust MD test and the MM test. In both cases, confidence intervals correspond to projection-method confidence intervals obtained from joint confidence sets in Figure 3-4.

<table>
<thead>
<tr>
<th></th>
<th>$\nu$</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>[0.91, 0.99]</td>
<td>[0.39, 0.56]</td>
</tr>
<tr>
<td>MM</td>
<td>[0.81, 0.99]</td>
<td>[0.27, 0.63]</td>
</tr>
</tbody>
</table>

to $\nu$ or $\varrho$. We find that curvature with respect to these parameters is quite high and that the smallest robust critical value we can obtain again corresponds to a $\chi^2_8$ distribution. Thus, our robust MD confidence intervals for $\nu$ and $\varrho$ correspond to the projection of the joint MD confidence set on the coordinate axes, and are reported in Table 3.1. From these results, we again see that confidence intervals based on our robust MD test are significantly smaller than those based on the MM test in this context. One reason the confidence sets are small is that the NKPC model is barely compatible with the estimated VAR coefficients. Indeed, a test of overidentifying restrictions using our robust critical values rejects the model at the 10%, though not 5%, level.

### 3.5.3 Size Simulation

One might worry that, despite the asymptotic validity of our robust MD test, the small volume of robust MD confidence sets in this application might reflect poor size control for the model and sample size considered. To address this issue we calculate the size of our robust MD test and the MM test for $H_0 : (\nu, \varrho) = (\nu_0, \varrho_0)$ in a model calibrated to match the empirical application. For comparison, we also calculate the size of the non-robust MD test that incorrectly treats all parameters as strongly identified, which uses $\chi^2_6$ critical values, and the projection-method MD test, which uses $\chi^2_{12}$ critical values. The simulated size of these tests is reported in Table 3.2. As might be expected, the non-robust MD test substantially over-rejects, while the projection-method MD test is conservative. With simulated size 7% our robust MD test is only mildly over-sized, reflecting small-sample issues in the distribution of

\footnote{For details on the simulation design, see the Supplementary Appendix.}
<table>
<thead>
<tr>
<th></th>
<th>Robust MD</th>
<th>Non-robust MD</th>
<th>Projection MD</th>
<th>MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>7.0%</td>
<td>13.5%</td>
<td>1.9%</td>
<td>11.9%</td>
</tr>
</tbody>
</table>

Table 3.2: Size of nominal 5% tests of $H_0 : (\nu, \varrho) = (\nu_0, \varrho_0)$ based on 100 observations. Robust MD compares the $MD$ statistic to $\chi^2$ critical values (based on the curvature), while non-robust MD and projection MD use $\chi^2$ and $\chi^2_{12}$ critical values, respectively. MM is the AR-MD test proposed in Magnusson and Mavroeidis (2010). Size is calculated based on 10,000 simulations.

Finally, we see that the MM test substantially over-rejects. While this may seem surprising given that Magnusson and Mavroeidis show that this test is robust to identification issues for $\nu$ and $\varrho$, the $\chi^2_{12}$ distribution of $MM$ under the null relies on the $\Delta$-method approximation $f(\hat{\theta}, \nu_0, \varrho_0) \approx \frac{\partial}{\partial \theta} f(\theta, \nu_0, \varrho_0)(\hat{\theta} - \theta)$. Such an approximation, if accurate, would also suggest low curvature for the manifold $S_{\nu_0, \varrho_0}$ which is inconsistent with the data. As a result, in the sample-size considered the $\Delta$-method approximation underlying the MM test is unreliable, while our robust MD test successfully accounts for the curvature of $S_{\nu_0, \varrho_0}$. Hence, in addition to yielding smaller confidence sets in this application, robust MD confidence sets also have better finite-sample coverage than MM confidence sets for the parameter values considered.

3.6 Simulation Example: DSGE Model

In this section we further illustrate our results by considering tests of Dynamic Stochastic General Equilibrium (DSGE) model parameters in a small simulation example. DSGE models are highly non-linear, very multi-dimensional dynamic models that describe the evolution of the main macro indicators in the economy. These models are currently quite popular in applied Macroeconomics and are used by many central banks. A recent literature raises concerns about identification in these models (e.g. Ruge-Murcia (2007), Canova and Sala (2009), Iskrev (2010)), and many authors have noted that standard frequentist statistical procedures seem to be unreliable. The source and extent of weak identification in these models is not well understood, and it is typically impossible to distinguish which parameters are weakly identified using currently-available procedures. Motivated by these issues, several recent papers
(Dufour, Khalaf and Kichian (2009), Guerron-Quintana, Inoue and Kilian (2013), I. Andrews and Mikusheva (2013), and Qu (2013)) suggest tests for full parameter vector hypotheses which are robust to weak identification. With the exception of I. Andrews and Mikusheva (2013) and Guerron-Quintana, Inoue and Kilian (2013), these papers use the projection method for inference on subsets of parameters which, due to the high dimension of the parameter vector, tends to be quite conservative.

Most DSGE models can be cast into our framework, which is natural here as one suggestion for how to estimate DSGE models is through two-step matching procedures (Christiano and Eichenbaum (1992), Rotemberg and Woodford (1997), Ruge-Murcia (2012), Andreasen, Fernandez-Villaverde, and Rubio-Ramirez (2013)). Log-linearized DSGE models can be written in state-space form

\[ z_t = A(\tilde{\beta})z_{t-1} + B(\tilde{\beta})u_t, \]

where \( z_t \) is a vector of state-space variables at time \( t \), \( u_t \) are i.i.d. mean zero shocks with identity covariance matrix, the state-space parameter matrices \( A(\tilde{\beta}) \) and \( B(\tilde{\beta}) \) are non-linear functions of the structural parameter \( \tilde{\beta} \) which typically need to be evaluated numerically, and we observe \( x_t = Cz_t \). In this context, a natural choice of reduced-form parameters is the auto-covariances of the observed vector-series \( x_t \). In particular, let \( \Sigma_x(j) \) be \( j \)-th order auto-covariance of \( x_t \) (for details see Iskrev (2010)):

\[ \theta_j(\tilde{\beta}) = vec(\Sigma_x(j)) = (C \otimes CA') (I - (A \otimes A))^{-1} vec(BB'). \]

One may choose the reduced form parameter \( \theta \) to be some subset of \( vec(\Sigma_x(j)) \), so \( \theta = m(\tilde{\beta}) = W(\theta_0(\tilde{\beta})', \ldots, \theta_j(\tilde{\beta})')', \) where \( W \) is a selection matrix. In the absence of persistence (exact or near unit roots) the usual estimators

\[ \hat{\theta}_j = vec \left( \frac{1}{T-j-1} \sum_{t=1}^{T-j} (x_{t+j} - \bar{x})(x_t - \bar{x})' \right) \]

of \( \theta_j \) satisfy a central limit theorem and achieve normality quite quickly. As a result, normal approximations to the distribution of \( \hat{\theta} = (\hat{\theta}_0, \ldots, \hat{\theta}_j) \) are usually reliable for
realistic sample sizes. Hence, we can conduct inference on the structural parameters \( \beta \) using a minimum-distance approach focused on matching auto-covariances.

While our discussion focuses on log-linearized DSGE model, one could equally well apply our approach to nonlinear DSGE models, for example matching covariances calculated from higher-order approximations as in Andreasen, Fernandez-Villaverde, and Rubio-Ramirez (2013). Even in the log-linear case we could base inference on quantities other than auto-covariances, for example matching the parameters of the state-space representation. Indeed, by matching the parameters of the matrices \( A(\tilde{\beta}) \) and \( B(\tilde{\beta}) \) we may be able to obtain a more powerful test. However, one should be cautious with this choice of reduced-form parameters since it is known that state-space coefficients are sometimes poorly identified, casting doubt on the normal approximation to the distribution of state-space parameter estimates. In particular, if the dynamics of the data are described by a VARMA process such that some VAR and MA roots are close to each other (near-canceling roots), the corresponding state-space representation coefficients are weakly identified. This issue has been raised by Schorfheide (2013) in the DSGE context and is studied in an ARMA model by D. Andrews and Cheng (2012).

### 3.6.1 Simulation study: A Small-scale DSGE Model

We apply our approach to a small-scale DSGE model based on Clarida, Gali and Gertler (1999). The (log-linearized) equilibrium conditions for the model are

\[
\begin{align*}
\begin{cases}
    bE_t \pi_{t+1} + \kappa x_t - \pi_t + \varepsilon_t = 0, \\
    -[r_t - E_t \pi_{t+1} - rr_t^*] + E_t x_{t+1} - x_t = 0, \\
    \lambda r_{t-1} + (1 - \lambda) \phi_n \pi_t + (1 - \lambda) \phi_x x_t + u_t = r_t, \\
    rr_t^* = \rho \Delta a_t,
\end{cases}
\end{align*}
\]  

(3.15)

where the exogenous variables (\( \Delta a_t \) and \( u_t \)) evolve according to

\[
\Delta a_t = \rho \Delta a_{t-1} + \varepsilon_{a,t}; \quad u_t = \delta u_{t-1} + \varepsilon_{u,t};
\]

\[
(\varepsilon_t, \varepsilon_{a,t}, \varepsilon_{u,t})' \sim iid N(0, \Sigma); \quad \Sigma = \text{diag}(\sigma^2, \sigma^2_a, \sigma^2_u).
\]
Here we assume that a researcher observes data on inflation $\pi_t$, the interest rate $r_t$ and some measure of real activity $x_t$. This model has ten parameters: the discount rate $b$, the structural parameters $\kappa, \phi_x, \phi_r$, and $\lambda$, and the parameters describing the evolution of the exogenous variables. We set $b = 0.99$ and estimate the remaining parameters by maximum likelihood using de-meaned US macro data from Smets and Wouters (2007). Since the estimated values of $\rho$ and $\kappa$ are quite close to the boundaries of their respective parameter spaces (one and zero respectively), we set $\rho = .85$ and $\kappa = .1$ while calibrating the remaining parameters to their ML estimates, in particular taking $\phi_x = 2.278$, $\phi_r = 2.023$, $\lambda = .898$, $\delta = .103$, $\sigma_\alpha = .325$, $\sigma_\delta = .265$, and $\sigma = .556$. We generate samples of size 300 from this model and then discard the first 100 observations, using only the last 200 observations for the remainder of the analysis. Given well-documented problems with estimating $b$ in many models, from this point forward we treat this parameter as known and focus on the remaining nine parameters.

As discussed above, to test hypotheses on the structural parameters in DSGE models we can test the implied restriction on the auto-covariances $\theta(\tilde{\beta})$. In particular, we let $\theta$ consist of the unique entries of the covariance matrix of $(x_t, \pi_t, r_t)$ and their first auto-covariance, giving us 15 reduced-form parameters. To focus on the problem of weak identification and abstract from the issues which may arise from HAC covariance matrix estimation, we treat the true covariance matrix $\Sigma$ of our reduced-form parameter estimates as known.

We consider the problem of separately testing that each of the structural parameters is equal to its true value (as one needs to do to construct confidence sets for each parameter individually). For example, to test $H_0 : \kappa = \kappa_0$, we let $\beta$ contain all the parameters other than $\kappa$ (that is, $\tilde{\beta} = (\kappa, \beta)$) and consider the $MD$ statistic $MD(\kappa_0) = \min_{\beta} T(\tilde{\beta} - \theta(\kappa_0, \beta))^\Sigma^{-1}(\tilde{\beta} - \theta(\kappa_0, \beta))$. As before, the key issue is what critical values to use. The projection method uses the 95th percentile of a $\chi^{15}_{15}$, which is equal to 25. If we assume that a $p$-dimensional sub-vector of $\beta$ is strongly identified, we can use $\chi^2_{15-p}$ critical values instead, which are equal to 14.07 (for $p = 8$). Applying our robust critical values, in contrast, requires no assumption on the strength of iden-
tification. As we might expect in a poorly identified model, tests which concentrate out the nuisance parameters do not have correct size. We simulated tests for each parameter separately, and all of them over-reject, though the degree of over-rejection is limited. The largest size distortion occurs for tests of $\kappa = \kappa_0$, where nominal 5% tests have true size 11.4%.

For each of the nine parameters we calculate the curvature of the submanifold of $\{\Sigma^{-1/2}(\theta(\beta))\}$ obtained by holding that parameter equal to its null value, intersected with the ball of radius $\sqrt{2}R$ around $\theta_0$, for $R$ the .99 quantile of a $\chi^2_{15}$ distribution. We find quite substantial curvature: the manifold corresponding to $\kappa = \kappa_0$, for example, has a maximal curvature of 5.16, which if plugged into $F(C, R, 15, 8)$ implies a robust critical value nearly as large as that used by the projection method.

These large values of curvature result from calculating curvature with respect to all nuisance parameters simultaneously. As noted above, however, if some nuisance parameters correspond to directions of high curvature (we may loosely call them weakly identified) while others enter nearly linearly, we may be able to substantially reduce our critical values by projecting over the weakly identified parameters. Specifically, we find that projecting over $\phi_x$ and $\phi_\pi$ often substantially reduces curvature with respect to the remaining parameters. Motivated by this fact we calculate curvature projecting over the following subsets of parameters: $\phi_x, \phi_\pi, (\phi_x, \phi_\pi), (\phi_x, \phi_\pi, \lambda), (\phi_x, \phi_\pi, \rho)$, $(\phi_x, \phi_\pi, \kappa)$. To test hypotheses on each of the nine structural parameters we use the smallest robust critical value obtained by projecting over one of these six subsets.

For each structural parameter, Table 3.3 reports the robust critical value obtained from this exercise (column 2), together with the simulated size (based on 10,000 simulations) of nominal 5% tests based on our robust critical values (column 3), and projection-method tests (column 4). As we can see, projection-method based tests (using critical values of 25) are typically quite conservative, with simulated size less than or equal to 0.6% for all parameters but $\kappa$. In contrast, for most parameters our robust critical values fall about halfway between projection method critical values (25) and non-robust critical values (14.07). Tests based on our robust critical values control size and, while conservative, are for the most part substantially less so than
### Table 3.3: Nominal 5% Tests of one-dimensional hypotheses on structural parameters.

The first column lists the tested parameter for each row, while the other parameters are treated as nuisance parameters. The statistic is $AR$ minimized over nuisance parameters. Projection method critical values are obtained by considering projection over $\phi_x$, $\phi_\pi$, $(\phi_x, \phi_\pi)$, $(\phi_x, \phi_\pi, \lambda)$, $(\phi_x, \phi_\pi, \rho)$, $(\phi_x, \phi_\pi, \kappa)$ and using the smallest critical value projection-method tests.

<table>
<thead>
<tr>
<th>Parameter tested</th>
<th>Robust Critical Value</th>
<th>Robust Test Size</th>
<th>Projection Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_x$</td>
<td>17.3</td>
<td>2.5%</td>
<td>0.4%</td>
</tr>
<tr>
<td>$\phi_\pi$</td>
<td>17.3</td>
<td>2.8%</td>
<td>0.6%</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>20.4</td>
<td>1.0%</td>
<td>0.3%</td>
</tr>
<tr>
<td>$\rho$</td>
<td>20.2</td>
<td>1.4%</td>
<td>0.6%</td>
</tr>
<tr>
<td>$\delta$</td>
<td>20.2</td>
<td>0.9%</td>
<td>0.3%</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>22.5</td>
<td>2.4%</td>
<td>1.7%</td>
</tr>
<tr>
<td>$\sigma_a$</td>
<td>17.7</td>
<td>3.0%</td>
<td>0.4%</td>
</tr>
<tr>
<td>$\sigma_u$</td>
<td>19.0</td>
<td>2.4%</td>
<td>0.6%</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>19.5</td>
<td>1.5%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

3.7 Appendix

#### Proof of Theorem 10

The proof is based on the following lemma:

**Lemma 7** Assume the curve $\alpha(s) : [0, b] \rightarrow D_C \subset \mathbb{R}^k$ is parameterized by arc length and that its curvature $\kappa(s) = \|\dot{\alpha}(s)\| \leq \frac{1}{C}$ for all points $s$. Assume that $\alpha(0) = 0$ and $\dot{\alpha}(0) = v \in \text{span}\{e_1, \ldots, e_p\}$, where $e_1, \ldots, e_p$ are first $p$ basis vectors. Then the curve $\alpha(s)$ is contained in the set $M_v \cap D_C$, where

$$M_v = \{ x : (x, v)^2 + (C - \|x - (x, v)v\|)^2 \geq C^2 \}.$$  \hfill (3.16)

**Proof of Lemma 7.**

Consider the curve defined by $\beta(s) = \dot{\alpha}(s)$, the first derivative of $\alpha$. Since the curve $\alpha$ is parameterized by arc length $\|\beta(s)\| = \|\dot{\alpha}(s)\| = 1$ and the new curve $\beta$ lies on the unit sphere $Sp = \{ x \in \mathbb{R}^k : \|x\| = 1 \}$, with $\beta(0) = v$. Let $t \leq \frac{\pi}{2} C$ and $t \leq b$. 

199
Consider the arc length of the restriction of the curve $\beta$ to the interval $[0, t]$:

$$\text{length}(t) = \int_0^t \|\dot{\beta}(s)\|ds = \int_0^t \|\dot{\alpha}(s)\|ds = \int_0^t \kappa(s)ds \leq \frac{t}{C}.$$ 

This implies that the geodesic (a curve of a shortest length) on the sphere $Sph$ connecting $\beta(0)$ and $\beta(t)$ has length less than or equal to $\frac{t}{C}$ or, equivalently, that the angle between vectors $\beta(0) = v$ and $\beta(t)$ is less than or equal to $\frac{t}{C}$. Hence

$$\langle v, \beta(t) \rangle = \langle v, \dot{\alpha}(t) \rangle \geq \cos\left(\frac{t}{C}\right).$$

Since $\alpha(s)$ is parameterized by arc length, from inequality (3.17) we have:

$$\|\dot{\alpha}(t) - \langle v, \dot{\alpha}(t)\rangle v\| \leq |\sin\left(\frac{t}{C}\right)|.$$  \hspace{1cm} (3.18)

This, in turn, implies that

$$\|\alpha(t) - \langle v, \alpha(t)\rangle v\| = \| \int_0^t (\dot{\alpha}(s) - \langle v, \dot{\alpha}(s)\rangle vds\| \leq \int_0^t \|\dot{\alpha}(s) - \langle v, \dot{\alpha}(s)\rangle v\|ds \leq \int_0^t \sin\left(\frac{s}{C}\right)ds = C - C \cos\left(\frac{t}{C}\right).$$

Inequality (3.17) also implies that

$$\langle v, \alpha(t) \rangle \geq \int_0^t \cos\left(\frac{s}{C}\right)ds = C \sin\left(\frac{t}{C}\right).$$  \hspace{1cm} (3.19)

Combing these results yields

$$\langle v, \alpha(t) \rangle^2 + (C - \|\alpha(t) - \langle v, \alpha(t)\rangle v\|)^2 \geq C^2$$

for all $t \leq \frac{t}{2}C$. Notice that (3.19) implies that for $\tau = \frac{t}{2}C$ we have $\langle v, \alpha(\tau) \rangle \geq C$ and thus for the first $p$ coordinates of $\alpha(\tau)$, which we denote $\alpha^{(1)}(\tau)$, we have $\|\alpha^{(1)}(\tau)\| \geq C$ so the curve is leaving or has already left the cylinder $D_C$ and thus $b \leq \frac{t}{2}C$. This concludes the proof of the lemma. \(\Box\)
Proof of statement (a) of Theorem 10. First, let us show that

$$\bigcup_{v \in T_0(S), \|v\|=1} M_v = \{ \|x^{(1)}\|^2 + (C - \|x^{(2)}\|)^2 \geq C^2 \} = \mathcal{M},$$

(3.20)

where $M_v$ is defined in (3.16), $\mathcal{M}$ is defined in (3.9) and $T_0(S)$ is the tangent space to $S$ at zero and is spanned by first $p$ basis vectors. Indeed, the set on the left hand side consists of points $x$ for which there exists a vector $v \in span\{e_1, ..., e_p\}, \|v\| = 1,$ such that

$$\langle x, v \rangle^2 + (C - \|x - \langle x, v \rangle v\|)^2 \geq C^2.$$  

(3.21)

For each $x$ let us find the maximum of the expression on left-hand side of inequality (3.21) over $v \in T_0(S), \|v\| = 1$ :

$$\langle x, v \rangle^2 + (C - \|x - \langle x, v \rangle v\|)^2 =$$

$$= \langle x, v \rangle^2 + C^2 + \|x\|^2 - \langle x, v \rangle^2 - 2C\|x - \langle x, v \rangle v\| =$$

$$= C^2 + \|x\|^2 - 2C\|x - \langle x, v \rangle v\|$$

where we used that $\|x - \langle x, v \rangle v\|^2 = \|x\|^2 - \langle x, v \rangle^2$. We see that maximizing the left-hand side of (3.21) over $v \in span\{e_1, ..., e_p\}, \|v\| = 1$ is equivalent to minimizing $\|x - \langle x, v \rangle v\|$. The minimum is achieved at the projection of $x$ onto $T_0(S) = span\{e_1, ..., e_p\}$, that is, $v = \frac{1}{\|x^{(1)}\|}(x^{(1)}, 0, ..., 0)$, where $x^{(1)} \in \mathbb{R}^p$ consists of the first $p$ components of $x$. As a result, the maximum of the left-hand side of (3.21) equals

$$C^2 + \|x\|^2 - 2C\|x^{(2)}\| = \|x^{(1)}\|^2 + (C - \|x^{(2)}\|)^2.$$  

This proves statement (3.20).

Now assume that statement (a) of Theorem 10 is incorrect and there exists a point $q \in S_C$ with $q \notin \mathcal{M}$. Take a geodesic (a curve of the shortest distance lying in $S_C\) \alpha(s)$ connecting $q$ and 0 lying in $S_C$, where such a curve exists since $S_C$ is
a connected manifold. Parameterize this curve by the arc length. The curve \( \alpha(s) \) is a geodesic in \( S \) if and only if at any point \( q = \alpha(t) \) the second derivative \( \dot{\alpha}(t) \) is perpendicular to \( T_q(S) \) (see Spivak (1999) for discussion of geodesics, v.3, p.3). As a result, the curvature of the geodesic \( \alpha \) at each point \( q = \alpha(t) \) is equal to \( \kappa_q(X, S) \) (where \( X = \dot{\alpha}(t) \)), and thus it is less than \( \frac{1}{C} \). Denote the tangent to this curve at 0 by \( v \in T_0(S) \). Applying Lemma 7 we obtain that the curve belongs to \( M_v \cap D_C \) and thus belongs to \( \mathcal{M} \cap D_C \). We have arrived at a contradiction. \( \square \)

**Proof of statement (c) of Theorem 10.** Let

\[
 f(u) = \rho^2(\xi, N_u) = \min_{\xi(1) \in \mathbb{R}^p, z \in \mathbb{R}_+} \left\{ \| \xi(1) - x(1) \|^2 + \| \xi(2) - z u \|^2 \right\}. 
\]

We need to find the maximizer of \( f(u) \) subject to the constraint \( \| u \| = 1 \). To differentiate \( f(u) \) we use the “envelope theorem” that allows one to differentiate a function which is the optimum of a constrained optimization problem and yields

\[
 \frac{df(u)}{du} = -2(\xi(2) - zu). 
\]

Hence, the first-order condition for finding \( \bar{u} \) implies that \( u \) is proportional to \( \xi(2) \). The sign is a reflection of the fact that we search for a max rather than a min. \( \square \)

**Proof of statement (b) of Theorem 10.** For a given point \( \xi \in \mathbb{R}^k \) find the sphere \( N_{\bar{u}} \) furthest from \( \xi \), \( \bar{u} \) is described in Theorem 10 (c), and the point \( \tau \in N_{\bar{u}} \) such that \( \rho(\xi, N_{\bar{u}}) = \rho(\xi, \tau) \). Consider the \( k - p \) dimensional linear space \( R_\tau = \{ x \in \mathbb{R}^k : x(1) = \tau(1) \} \) that restricts the first \( p \) components of \( x \) to coincide with the first \( p \) components of \( \tau \). We will put forward two statements: first, that all points in the intersection of \( R_\tau \cap \mathcal{M} \cap D_C \) are not further from \( \xi \) than \( \tau \); and second, that this intersection \( R_\tau \cap \mathcal{M} \cap D_C \) contains at least one point from \( S \). Together, these two statements imply that \( \rho(\xi, S) \leq \rho(\xi, \tau) \).

The intersection of the three sets \( R_\tau \cap \mathcal{M} \cap D_C \) can be written as follows:

\[
 R_\tau \cap \mathcal{M} \cap D_C = \{ x = (\tau(1), x(2)) \in D_C : \| \tau(1) \|^2 + (C - \| x(2) \|)^2 \geq C^2 \} = \\
 = \left\{ x = (\tau(1), x(2)) : \| x(2) \| \leq C - \sqrt{C^2 - \| \tau(1) \|^2} \right\}. 
\]
Now let us show that for each \( x \in R_t \cap M \cap D_C \) we have \( \rho(\xi, x) \leq \rho(\xi, \tau) \). Indeed, one can solve the constrained maximization problem

\[
\rho(\xi, x)^2 = \|\xi^{(1)} - \tau^{(1)}\|^2 + \|\xi^{(2)} - x^{(2)}\|^2 \to \max \text{ s.t. } x \in R_t \cap M \cap D_C.
\]

From the first-order condition for this problem one can see that the maximum is achieved at \( x^{(2)} \) proportional to \( \xi^{(2)} \). We recall that \( \tau \in N_{\tilde{u}} \) and by statement (c) \( \tau^{(2)} \) is proportional to \( \xi^{(2)} \). Further inspection reveals that the maximum is achieved at \( x = \tau \). Hence, all points lying in the intersection \( R_t \cap M \cap D_C \) have distance to \( \xi \) less or equal than \( \rho(\xi, N_{\tilde{u}}) \).

To complete the proof we need only show that \( R_t \cap M \cap D_C \) contains at least one point from the manifold \( S \). Recall that from the definition of \( \tau \in N_{\tilde{u}} \) it follows that \( \|\tau^{(1)}\| \leq C \). Then Assumption 7 guarantees that the intersection of \( S_C \) with \( R_t \) is non-empty, while statement (a) of Theorem 10 implies that \( S_C \subseteq M \cap D_C \). □

**Proof of statement (d) of Theorem 10.** Note that since \( \tilde{u} \) is proportional to \( \xi^{(2)} \) by statement (c), both \( \xi \) and \( N_{\tilde{u}} \) belong to the same \( p + 1 \)-dimensional linear sub-space \( L_{\tilde{u}} = \{ x : x = (x^{(1)}, -z\tilde{u}), x^{(1)} \in \mathbb{R}^p, z \in \mathbb{R} \} \). Let us restrict our attention to this subspace only. Let \( (x^{(1)}, z) \) be the coordinate system in this sub-space, so \( \xi \) corresponds to \( \tilde{\xi} = (\xi^{(1)}, \|\xi^{(2)}\|) \), and \( N_{\tilde{u}} \) corresponds to the sphere \( N^C = \{ x = (x^{(1)}, z) \in \mathbb{R}^{p+1} : \|x^{(1)}\|^2 + (C + z)^2 = C^2 \} \). The distance implied by the distance in \( \mathbb{R}^k \) is the usual Euclidean metric, which we denote by \( \tilde{\rho} \). So far, we proved that \( \rho(\xi, N_{\tilde{u}}) = \tilde{\rho}(\tilde{\xi}, N^C) \). By invariance of the distance to orthonormal transformations of first \( p \) components we have \( \tilde{\rho}(\tilde{\xi}, N^C) = \tilde{\rho}(\xi^*, N^C) \), where \( \xi^* = (\|\xi^{(1)}\|, 0, ..., 0, \|\xi^{(2)}\|) \in \mathbb{R}^{p+1} \). From this it is easy to see that

\[
\rho(\xi, N_{\tilde{u}}) = \rho_2(\eta, N^C_{2}),
\]

where \( \eta = (\|\xi^{(1)}\|, \|\xi^{(2)}\|) \in \mathbb{R}^2 \), \( N^C_{2} = \{ (z_1, z_2) \in \mathbb{R}^2 : z_1^2 + (C + z_2)^2 = C^2 \} \), and \( \rho_2 \) is Euclidian distance in \( \mathbb{R}^2 \). It then follows that if \( \xi \sim N(0, I_k) \) then components of \( \eta \) have independent \( \sqrt{\chi_p^2} \) and \( \sqrt{\chi_{k-p}^2} \) distributions, respectively. □
Proof of Lemma 6

Proof of Lemma 6. Let \( \xi = \Sigma^{-1/2}(\hat{\theta} - \theta_0) \sim N(0, I_k) \) and \( S = \{ \Sigma^{-1/2}(\theta - \theta_0), \theta \in H_0 \} \subset \mathbb{R}^k \). Let \( C \) be the largest value for which all assumptions of Theorem 10 are satisfied. Let \( \psi_C(\xi, R) \) be defined as

\[
\psi_C(\xi, R) = \begin{cases} 
\rho^2(\xi, N_u), & \text{if } \|\xi\| \leq R; \\
\|\xi\|^2, & \text{if } \|\xi\| > R,
\end{cases}
\]

where \( N_u = \{ x \in \mathbb{R}^k : x = (x^{(1)}, z\bar{u}), x^{(1)} \in \mathbb{R}^p, z \in \mathbb{R}_+, \|x^{(1)}\|^2 + (C - z)^2 = C^2 \}, \bar{u} = -\frac{1}{\|\xi\|^2} \xi \). Random variable \( \psi_C(\xi, R) \) has the same distribution as \( \psi_C(R) \) defined in Section 3.3.1, but is defined on a different probability space, as \( \psi_C(R) \) is written in terms of random vector \( \eta \in \mathbb{R}^2 \) described in Theorem 10 (d). Consider the infeasible test \( \varphi \) which rejects (\( \varphi = 1 \)) if and only if \( \psi_C(\xi, R) \geq F_{1-\alpha}(C, R, k, p) \). The size \( E_\varphi(\xi) = \alpha \), so since \( P\{ \chi_k^2 \geq R \} < \alpha \) we know that \( \varphi \) rejects for all realizations of \( \xi \) where \( \|\xi\| > R \) as \( \|\xi\| \geq \rho(\xi, N_u) \). This test is infeasible, however, since we do not know the true value of \( \theta_0 \) and hence cannot calculate \( \xi \). The (feasible) test described in Lemma 6 is

\[
\tilde{\varphi} = \begin{cases}
1, & \text{if } MD \geq F_{1-\alpha}(C^*, R, k, p); \\
0, & \text{otherwise}.
\end{cases}
\]

We claim that \( \tilde{\varphi} \leq \varphi \) almost surely (realization-by-realization). To show that this is the case, assume that \( \tilde{\varphi} = 1 \). If at the same time \( \|\xi\| > R \) then \( \varphi = 1 \), so the claim holds. If, on the other hand, \( \|\xi\| \leq R \), then the cylinder \( D_R(x_0) \) around \( x_0 = \Sigma^{-1/2}\theta_0 \) lies inside of ball \( B^* \) of radius \( 1 + \sqrt{2}R \) around \( \hat{x} = x_0 + \xi \), and thus

\[
C^*_R = \left( \min_{q \in S^* \cap B^*} 1/\kappa_q(S^*) \right) \wedge R \leq \left( \min_{q \in S^* \cap D_R(x_0)} 1/\kappa_q(S^*) \right) \wedge R \leq C.
\]

Indeed, to justify the last inequality, consider two cases \( R \leq C \) and \( R > C \). In the first case \( C^*_R \leq R \leq C \), in the second case \( D_C^* \subset D_R^* \) and thus \( \min_{q \in S^* \cap D_C^*(x_0)} 1/\kappa_q(S^*) \leq \min_{q \in S^* \cap D_C^*(x_0)} 1/\kappa_q(S^*) \leq C \).
Note that the function $F_{1-\alpha}(c, R, k, p)$ is decreasing in $c$, and hence $F_{1-\alpha}(C, R, k, p) \leq F_{1-\alpha}(C^*_R, R, k, p)$. Further, all the assumptions of Theorem 10 are satisfied so $MD = \rho^2(\xi, S) \leq \rho^2(\xi, N_u) \leq \psi_C(\xi, R)$. Combining these results we obtain that

$$F_{1-\alpha}(C, R, k, p) \leq F_{1-\alpha}(C^*_R, R, k, p) \leq MD = \rho^2(\xi, S) \leq \psi_C(\xi, R),$$

and thus $\varphi = 1$. Hence whenever $\tilde{\varphi} = 1$, we get that $\varphi = 1$ as well, so $\tilde{\varphi} \leq \varphi$ as we wanted to show, and the size of the feasible test $\tilde{\varphi}$ is bounded above by $\alpha$, completing the proof. □

**Supplementary Appendix**

**Proof of Theorem 11.** Assume that $\xi \sim N(0, I_k)$. Our main theorem states:

$$P \left\{ \rho^2(\xi, S) > F_{1-\alpha}(C, k, p) \right\} \leq \alpha,$$

for all $S \in \mathcal{S}$ if $C = C(S)$ as described in Assumption 3. For the rest of the proof we drop the subscript $n$ in $S_n$, $\hat{S}_n$ and $\tilde{S}_n$ and the corresponding $C$’s, as all statements below hold for *all* sets $S \in \mathcal{S}$ (that is, uniformly over $\mathcal{S}$) and $S_n \in \mathcal{S}$ for all $n$. Technically for any given set $S \in \mathcal{S}$, the sets $\hat{S}$ and $\tilde{S}$ depend on $n$ since they use $\hat{S}_n$ and/or $\sqrt{n}$ in their definition, but in what follows we suppress the index $n$ for notational simplicity. Let $\xi_n = \sqrt{n} \tilde{\Sigma}_n^{-1/2}(\tilde{\theta}_n - \theta_0)$. We note that the statistic of interest can be written as $n \min_{\theta \in S}(\tilde{\theta}_n - \theta)^T \tilde{\Sigma}_n^{-1} (\tilde{\theta}_n - \theta) = \rho^2(\xi_n; \tilde{S})$. Assumption 2 (i) - (iii) imply that $\xi_n \Rightarrow N(0, I_k)$ uniformly over $\mathcal{M}$.

This weak convergence can be metrized by Prokhorov’s metric. Let $\beta_n \geq 0$ be Prokhorov’s distance between the distributions of random variables $\xi$ and $\xi_n$, where all terms are implicitly indexed by model $M$. According to Dudley’s (1968) result we can construct on a probability space two random variables $\tilde{\xi}_n$ and $\tilde{\xi}$ with the same marginal distributions as $\xi_n$ and $\xi$ such that $P \left\{ \| \tilde{\xi} - \tilde{\xi}_n \| > \beta_n \right\} \leq \beta_n$. From now on for simplicity of notation we will drop tildes and assume that $\xi_n$ and $\xi$ satisfy this
condition. So, we have:

\[
P \left\{ \rho(\xi, \hat{S}) > F_{1-a}^{\frac{1}{2}}(\hat{C}, k, p) \right\} \leq \]
\[
\leq P \left\{ \rho(\xi, \hat{S}) > F_{1-a}^{\frac{1}{2}}(\hat{C}, k, p) - \beta_n \right\} + P \left\{ \|\xi - \xi_n\| > \beta_n \right\},
\]

(3.23)

where we used that \(|\rho(\xi, \hat{S}) - \rho(\xi, \tilde{S})| \leq \|\xi - \xi\|\). The second term on the right hand side in (3.23) does not exceed \(\beta_n\), and \(\beta_n\) converges to zero uniformly over set of models \(M\).

Let \(C = 1/(\sup_{\eta \in \mathcal{S}} \kappa_\eta(S))\), and note that by Assumption 3, Assumption 1 from the paper holds for this value of \(C\) and the manifold \(S\). Fix some small \(\varepsilon > 0\). We can notice that:

\[
P \left\{ \rho(\xi, \hat{S}) > F_{1-a}^{\frac{1}{2}}(\hat{C}, k, p) - \beta_n \right\} \leq P \{ |\rho(\xi, \hat{S}) - \rho(\xi, S)| > \varepsilon \} +
\]
\[
+ P \left\{ \rho(\xi, S) > F_{1-a}^{\frac{1}{2}}(C, k, p) - 2\varepsilon - \beta_n \right\} +
\]
\[
+ P \left\{ |F_{1-a}^{\frac{1}{2}}(\hat{C}, k, p) - F_{1-a}^{\frac{1}{2}}(C, k, p)| > \varepsilon \right\},
\]

(3.24)

Below we show that the first and third terms on the right hand side of equation (3.24) are asymptotically negligible uniformly over \(M\), while by choosing small \(\varepsilon\) we can bound the second term from above by a number arbitrarily close to \(\alpha\).

For the first term, note that the manifold \(\hat{S} = \{Ax : x \in S\}\) for matrix \(A = \hat{\Sigma}_n^{-1/2}\Sigma_1^{1/2}\). Let \(\|X\|\) denote the matrix norm of a square matrix \(X\) (that is, the maximal eigenvalue in absolute value). Below we show that

\[
|\rho(\xi, \hat{S}) - \rho(\xi, S)| \leq 2\|\xi\| \max \{\|I - A\|, \|I - A^{-1}\|\}. \tag{3.25}
\]

Indeed, consider first the case when \(\rho(\xi, \hat{S}) \geq \rho(\xi, S)\), and assume that \(\rho(\xi, S) = \rho(\xi, x)\) for a point \(x \in S\). Then \(Ax \in \hat{S}\), and \(\rho(\xi, \hat{S}) \leq \rho(\xi, Ax)\). This implies that

\[
0 \leq \rho(\xi, \hat{S}) - \rho(\xi, S) \leq \rho(\xi, Ax) - \rho(\xi, x) \leq \rho(x, Ax) \leq \|I - A\| \cdot \|x\|
\]

206
Next, we notice that since $0 \in S$, $\|x\| \leq \|\xi\| + \rho(\xi, x) \leq 2\|\xi\|$. The case when $\rho(\xi, \tilde{S}) < \rho(\xi, S)$ can be considered analogously. This establishes the validity of inequality (3.25). Since $\|\xi\|^2$ is distributed as $\chi^2_k$, and according to assumptions 2 (ii)-(iii) the two maximal eigenvalues in equation (3.25) converge to zero uniformly, we can see that the first term in (3.24) is asymptotically small uniformly over $\mathcal{M}$.

For the second term, our main theorem guarantees that

$$P\left\{ \rho(\xi; S) > \psi_{1,\alpha}(C, k, p) - 2\varepsilon - \beta_n \right\} \leq P\left\{ \psi_{1,\alpha}(C, k, p) - 2\varepsilon - \beta_n \right\},$$

where $\psi_{1,\alpha}(C, k, p)$, for $\eta$ a random vector with two independent coordinates distributed as $(\sqrt{\chi^2_{k-p}}, \sqrt{\chi^2_{C}})$ and $N^C_2$ a circle of radius $C$ with center at point $(0, -C)$, while $\rho_2$ is a Euclidian distance in $\mathbb{R}^2$. First notice that function $f(x, C) = \rho_2(x, N^C_2)$ for $x \in \mathbb{R}^2$ is continuous in $x$ uniformly over all values of $C$, indeed, $|\rho_2(x, N^C_2) - \rho_2(y, N^C_2)| \leq \rho_2(x, y)$. Since $\eta$ is continuously distributed with bounded pdf and the variable $\psi_C$ has a pdf which is bounded above uniformly over $C$, this means that by choosing small enough $\varepsilon$ and $\beta_n$, we can make the right hand side in equation (3.26) arbitrarily close to $\alpha$.

It is easy to see that the function $\rho_2(x, N^C_2)$ is uniformly continuous in $C$. Thus $F_{1-\alpha}(C, k, p)$ the $\alpha$-quantile of random variable $\psi_C$, is continuous in $C$ uniformly over all values of $C$. As $C \to \infty$ the $\alpha$ quantile $F_{1-\alpha}(C, k, p)$ converges to the $\alpha$ quantile of a $\chi^2_{k-p}$ distribution, which can be called $\chi^2_{\alpha}(\infty, k, p)$. For any small $\varepsilon > 0$ there exists a constant $c$ such that for any $C > c$ we have $|F_{1-\alpha}(C, k, p) - F_{1-\alpha}(\infty, k, p)| < \varepsilon$.

What we are left to show is that (i) $\hat{C} \to C$ uniformly over the subset of models $\mathcal{M}$ for which $C < c(1 + \varepsilon)$ and (ii) for any probability arbitrarily close to one there exists a sample size such that for all models in $\mathcal{M}$ with $C > c(1 + \varepsilon)$ we have $\hat{C} > c$.

First we examine the asymptotic relationship between $C$ and $\hat{C}$. Let us consider a point $q \in S$ and curvature $\kappa_q(S) = \|\sum_{j=1}^{p} u_j V_{ij}^\perp\|$, where $(u_1, ..., u_p)$ is the optimizer from formula (3.8) with the condition $\|\sum_{j=1}^{p} u_j v_j\| = 1$. By Theorem 6.4 in chapter III of Kobayashi and Nomizu (1963), there exists a unique geodesic $\alpha(t) \in S$ defined for $t$ in an open neighborhood of zero with initial conditions: $\alpha(0) = q$ and $\alpha(0) = 207$
\[ \sum_{i=1}^{p} u_i v_i \in T_q(S). \] In particular, the fact that \( \alpha(t) \) is a geodesic curve on \( S \) means that \( \dot{\alpha}(0) = \sum_{i,j=1}^{p} u_i u_j v_{ij} \) is perpendicular to tangent space \( T_q(S) \) spanned by \( Z = (v_1, \ldots, v_p) \). This implies that \( \kappa_q(S) = \|\dot{\alpha}(0)\| = \|N_Z \dot{\alpha}(0)\| \) and \( \|\dot{\alpha}(0)\| = 1. \)

Let us consider a curve \( \hat{\alpha}(t) = A\alpha(t) \) and notice that this curve lies on manifold \( \hat{S} \) and passes through the point \( \hat{q} = Aq \in \hat{S} \). Let \( Z \) be the set of vectors spanning the tangent space \( T_q(S) \), then \( AZ \) spans \( T_{\hat{q}}(\hat{S}) \).

From formula (3.8) we can see that
\[
\kappa_{\hat{q}}(\hat{S}) \geq \frac{\|N_Z A \frac{d^2 \hat{\alpha}(0)}{dt^2}\|}{\|\dot{\hat{\alpha}}(0)\|^2} = \frac{\|N_Z A \dot{\alpha}(0)\|}{\|\dot{\alpha}(0)\|^2},
\]
as the left-hand-side expression is the maximum of the right-hand-side expression taken over all possible curves in \( \hat{S} \) passing through \( \hat{q} \).

Let us consider the following sequence of inequalities:
\[
\kappa_q(S) = \|\dot{\alpha}(0)\| = \|N_Z \dot{\alpha}(0)\| \leq \|(N_Z A - N_Z) \dot{\alpha}(0)\| + \|N_Z A \dot{\alpha}(0)\| \leq \|N_Z A - N_Z\| \|\dot{\alpha}(0)\| + \kappa_{\hat{q}}(\hat{S}) \|\dot{\alpha}(0)\|^2. \tag{3.27}
\]
We can notice that \( \|A \dot{\alpha}(0)\| \leq \|A\| \) since \( \|\dot{\alpha}(0)\| = 1 \). Finally, notice that
\[
N_Z A - N_Z = A - AZ(Z'AZ)^{-1}Z'A^2 - I + Z(Z'Z)^{-1}Z = (A - I) + (I - A)Z(Z'Z)^{-1}Z + AZ(Z'Z)^{-1}Z(I - A^2) + AZ(Z'Z)^{-1}Z'(A^2 - I)Z(Z'AZ^{-1}Z'A^2
\]
where we use \( A^2 \) to denote \( A' A \). Recall that \( A \rightarrow^\rho I \) uniformly over \( \mathcal{M} \), thus \( \|N_Z A - N_Z\| \leq C\|I - A\| \) with probability approaching one uniformly over \( \mathcal{M} \), where \( C \) is a constant that does not depend on \( M \). Putting this reasoning together with inequality (3.27) we obtain that with probability tending to one uniformly over \( \mathcal{M} \)
\[
\kappa_q(S) \leq C\|I - A\| \kappa_q(S) + \kappa_{\hat{q}}(\hat{S}) \|A\|^2, \tag{3.28}
\]

208
or
\[ \kappa_q(S) - \kappa_q(\hat{S}) \leq C\|I - A\|\kappa_q(S) + \kappa_q(\hat{S})(\|A\|^2 - 1). \]

Symmetric reasoning reversing the roles of the “hatted” and “non-hatted” variables yields
\[ \kappa_q(\hat{S}) \leq C\|I - A^{-1}\|\kappa_q(\hat{S}) + \kappa_q(S)\|A^{-1}\|^2, \]
which implies that
\[ \kappa_q(\hat{S}) - \kappa_q(S) \leq C\|I - A^{-1}\|\kappa_q(\hat{S}) + \kappa_q(S)(\|A^{-1}\|^2 - 1), \]
and
\[ \kappa_q(\hat{S}) \leq \frac{1}{1 - C\|I - A^{-1}\|}\kappa_q(S)\|A^{-1}\|^2. \]

Since \( A \rightarrow^{p} I \) uniformly over \( \mathcal{M} \), we get that for any finite constant \( K \), \( |\kappa_q(\hat{S}) - \kappa_q(S)| \rightarrow^{p} 0 \) uniformly over all points \( q \in S \) such that \( \kappa_q(S) \leq K \) and uniformly over the set of models \( \mathcal{M} \).

What we have just shown is that for any fixed constants \( K_1 \) and \( K_2 \) and \( \varepsilon > 0 \) we have that \( |\hat{C} - C| \rightarrow^{p} 0 \) uniformly over all models in \( \mathcal{M} \) with \( K_1 < C < K_2 \) and \( P\{\hat{C} > K_2(1 - \varepsilon)\} \rightarrow 1 \) uniformly over all models in \( \mathcal{M} \) with \( C > K_2 \). Inequality (3.28) also implies that if \( C < K_1 \), that is if there exists a point \( q \in S \) with \( \kappa_q(S) > 1/K_1 \), then \( \kappa_q(\hat{S}) \) is also large and for any \( \varepsilon \) there is a sample size that guarantees \( \hat{C} < K_1(1 + \varepsilon) \) with high probability for all such models uniformly over \( \mathcal{M} \). Thus we have that (i) \( \hat{C} \rightarrow^{p} C \) uniformly over the subset of models \( \mathcal{M} \) for which \( C < c(1 + \varepsilon) \) and (ii) for any probability arbitrarily close to one there exists a sample size such that for all models in \( \mathcal{M} \) with \( C > c(1 + \varepsilon) \) we have \( \hat{C} > c \) with at least this probability. This concludes the proof of Theorem 11.

**Curvature maximization over restricted sets**

In this section, we show that the modified procedure that calculates curvature on a finite ball around the reduced-form parameter estimate has correct uniform asymp-
totic size. In particular, combining modifications stated in Sections 3.4 and 3.3.1 of the paper produces uniformly asymptotically correct procedure.

Description of the procedure. Suppose we have a reduced-form parameter estimator \( \hat{\theta} \) and covariance estimator \( \hat{\Sigma} \) as described in Section 3.7. Again assume that the hypothesis of interest is \( H_0 : \theta \in \tilde{S} \). For any \( R \) such that \( R^2 > \chi^2_{k,1-\alpha} \), let 
\[
\tilde{B}_R = \{ x : \| x - \hat{\Sigma}^{-1/2}\hat{\theta} \| \leq (1 + \sqrt{2})R \}
\]
be the ball of radius \( (1 + \sqrt{2})R \) around the reduced-form parameter estimate. Let

\[
\tilde{C}_R = \begin{cases} 
R \wedge \left[ 1 / \left( \max_{\tilde{\theta} \in \tilde{S}} \tilde{\Sigma}^{-1/2} \tilde{\theta} \right) \right], & \text{if } \tilde{\Sigma}^{-1/2} \tilde{S} \cap \tilde{B}_R \neq \emptyset; \\
0, & \text{if } \tilde{\Sigma}^{-1/2} \tilde{S} \cap \tilde{B}_R = \emptyset.
\end{cases}
\]

The modified version of our test uses the statistic \( n \min_{\theta \in \tilde{S}} (\hat{\theta}_n - \theta) \tilde{\Sigma}_n^{-1}(\hat{\theta}_n - \theta) \) along with critical value \( F_{1-\alpha}(\tilde{C}_R, R, k, p) \), where we denote by \( F_{\alpha}(C_R, R, k, p) \) the \( \alpha \)-quantile of the random variable \( \psi_C(R) \) discussed in Section 3.3.1 of the paper.

**Theorem 12** Under Assumptions 2 and 3 the testing procedure described above has uniform asymptotic size \( \alpha \):

\[
\lim_{n \to \infty} \sup_{M \in M} P \left\{ n \min_{\theta \in \tilde{S}} (\hat{\theta}_n - \theta) \tilde{\Sigma}_n^{-1}(\hat{\theta}_n - \theta) > F_{\alpha}(\tilde{C}_R, R, k, p) \right\} \leq \alpha
\]

**Proof of Theorem 12.**

The proof of this Theorem combines the proofs of Theorem 2 from the Supplementary Appendix and Lemma 1 from the paper. In particular, in the paper we introduced

\[
C^*_R = \begin{cases} 
R \wedge \left[ 1 / \left( \max_{\theta \in S^*} \Sigma_n^{-1/2} \theta \right) \right], & \text{if } S^* \cap B^*_R \neq \emptyset; \\
0, & \text{if } S^* \cap B^*_R = \emptyset.
\end{cases}
\]

where \( B^*_R = \{ x : \| x - \xi \| \leq (1 + \sqrt{2})R \} \). The quantity \( C^*_R \) differs from \( \tilde{C}_R \) in two respects: first, it relates to the curvature of \( S^* = \Sigma^{-1/2} \tilde{S} \), while \( \tilde{C}_R \) is connected to the curvature of \( \tilde{\Sigma}^{-1/2} \tilde{S} \); second the maximal curvature is found over the ball \( B^*_R \) which is centered at \( \xi \), while \( \tilde{B}_R \) is a ball around \( \tilde{\Sigma}^{-1/2} \tilde{\theta} \). Lemma 1 in the paper states that
for any $r > \chi^2_{1-a,k}$ and for any $S \in \mathcal{S}$ we have

$$P\{\rho^2(\xi, S) > F_{1-a}(C^*_r, r, k, p)\} \leq \alpha. \quad (3.30)$$

We proceed analogous to the proof of Theorem 2 to obtain the following inequality that holds for any $r > \chi^2_{1-a,k}$:

$$P \left\{ \rho(\xi_n, \hat{S}) > F^{1/2}_{1-a}(\tilde{C}_R, R, k, p) \right\} \leq$$

$$\leq P \left\{ \rho(\xi, S) > F^{1/2}_{1-a}(C^*_r, r, k, p) - 2\varepsilon - \beta_n \right\} +$$

$$+ P \left\{ ||\xi - \xi_n|| > \beta_n \right\} +$$

$$+ P \left\{ F^{1/2}_{1-a}(\tilde{C}_R, R, k, p) < F^{1/2}_{1-a}(C^*_r, r, k, p) - \varepsilon \right\}. \quad (3.31)$$

As we argued in the proof of Theorem 2, the second and the third terms in (3.31) are uniformly asymptotically negligible. Due to statement (3.30) and the fact that $\psi_C(r)$ has uniformly bounded density, the first term in (3.31) can be made uniformly asymptotically bounded by any value larger than $\alpha$ by way of choosing small $\varepsilon > 0$, since $\beta_n \to 0$ uniformly over $\mathcal{M}$. We are left only to prove that for some choice of $r$ the last term in (3.31) is uniformly asymptotically negligible. We choose $r = (1 - \delta)R$ for small $\delta > 0$.

First, we notice that the distribution of random variable $\psi_C(R)$, which is defined in the proof of Lemma 1 in the paper, is uniformly continuous in $R$, thus we can always choose $\delta$ small enough that $\sup_{C^*} |F_{1-a}(C, R, k, p) - F_{1-a}(C, r, k, p)| < \varepsilon/2$. It is then enough to show that

$$\lim_{n \to \infty} \sup_{\mathcal{M}} P \left\{ F^{1/2}_{1-a}(\tilde{C}_R, R, k, p) < F^{1/2}_{1-a}(C^*_r, R, k, p) - \varepsilon/2 \right\} = 0.$$

Given the monotonicity of $F_{1-a}(C, R, k, p)$ it is enough to show that for any $\varepsilon_2 > 0$ (where we choose $\delta$ above so that $\delta < \varepsilon_2/R$) we have $\tilde{C}_R \leq C^*_r + \varepsilon_2$ with probability arbitrarily close to 1 uniformly over $\mathcal{M}$ in large samples.

Let $\tilde{q} = Aq^*$, where $A = \tilde{\Sigma}^{-1/2}\Sigma^{1/2}$. Note that $q^* \in S^*$ is equivalent to $\tilde{q} \in \mathcal{S}$.
Now let $\tilde{q} \in \tilde{B}_R$, again defined as $\|q - \tilde{\Sigma}^{-1/2}\hat{\theta}\| \leq (1 + \sqrt{2})R$. We have that $\tilde{q} - \tilde{\Sigma}^{-1/2}\hat{\theta} = Aq^* - \xi_n$. Given that $A$ uniformly converges to $I$ and $\xi - \xi_n$ uniformly converges to zero, we have that $AB^*_t \subset \tilde{B}_R$ with probability arbitrarily close to 1 in large samples, which in turn implies

$$\max_{\tilde{q} \in \tilde{\Sigma}^{-1/2}\tilde{S}} \kappa_{q^*}(\tilde{\Sigma}^{-1/2}\tilde{S}) \geq \max_{\tilde{q} \in \tilde{\Sigma}^{-1/2}\tilde{S}} \kappa_{q^*}(\tilde{\Sigma}^{-1/2}\tilde{S}).$$

(3.32)

In the proof of Theorem 2 we showed that for $q^* \in S^*$ and $\tilde{q} = Aq^* \in \tilde{\Sigma}^{-1/2}\tilde{S}$ we have that $|\kappa_{q^*}(S^*) - \kappa_{q^*}(\tilde{\Sigma}^{-1/2}\tilde{S})|$ converges to zero uniformly over points $q^*$ at which curvature is below a fixed constant and over $\mathcal{M}$. Hence, asymptotically (for any $\varepsilon_3 > 0$) with probability arbitrarily close to 1 we have.

$$\max_{\tilde{q} \in \tilde{\Sigma}^{-1/2}\tilde{S} \cap AB^*_t} \kappa_{q^*}(\tilde{\Sigma}^{-1/2}\tilde{S}) \geq \max_{q^* \in S^* \cap AB^*_t} \kappa_{q^*}(S^*) - \varepsilon_3$$

Joining this last inequality with (3.32) and the definitions of $C^*_R$ and $\tilde{C}_R$, we arrive to the conclusion that for any positive $\varepsilon_2$ and any probability arbitrarily close to 1, there exists a sample size such that $\tilde{C}_R \leq C^*_R + \varepsilon_2$ holds with at least this probability uniformly over $\mathcal{M}$. This concludes the proof of Theorem 12.

**Pretest curvature cutoffs**

Section 4.3 of the paper discusses a pre-test which can be used in a two-step testing procedure to guarantee that the resulting test has size distortion bounded above by a pre-specified tolerance level. The pretest we propose asks whether the curvature of the model is sufficiently small to ensure that tests based on classical $\chi^2_{k-\alpha}$ critical values (that is, tests which concentrate out the nuisance parameter $\beta$) with nominal size $\alpha$ have true size not exceeding $\alpha + \alpha^*$. To determine whether this is the case we calculate $C^*$, the smallest value $C$ such that the $(1 - \alpha)$-quantile of a $\chi^2_{k-p}$ distribution does not exceed $F_{1-(\alpha + \alpha^*)}(C, R, k, p)$. The cut-off $C^*$ depends on the dimension $k$ of the reduced-form parameter vector, the dimension $p$ of the nuisance parameter, and
Table 3.7 reports the cut-offs $C^*$ for nominal 5% tests and tolerance level 5% for different values of $p$ and $k$ for $R$ equal to the 0.99 quantile of a $\sqrt{\chi^2_k}$. Based on Table S1 we can see that for a fixed dimension $k$ of the reduced-form parameter, increasing the number of nuisance parameters $p$ tightens the restrictions imposed on curvature if one wants to concentrate out the nuisance parameters.

Details of NKPC application

This section contains additional details concerning empirical example discussed in Section 3.5 of the paper.

The MM Test

The statistic proposed by MM (which they call MD-AR) is

$$\text{MM} = T \cdot \hat{\mathbf{f}}^{\prime} \left(\hat{\theta}, \nu_0, \varrho_0\right) \hat{\Sigma}_f^{-1} \hat{\mathbf{f}} \left(\hat{\theta}, \nu_0, \varrho_0\right),$$

where $\hat{\Sigma}_f = \frac{\partial}{\partial \theta} \hat{\mathbf{f}} \left(\hat{\theta}, \nu_0, \varrho_0\right)^{\prime} \hat{\Sigma}_\theta \frac{\partial}{\partial \theta} \hat{\mathbf{f}} \left(\hat{\theta}, \nu_0, \varrho_0\right).$ Standard $\Delta$-method arguments imply that as $T \to \infty$, since $\sqrt{T} \left(\hat{\theta} - \theta\right) \to_d N(0, \Sigma_\theta)$ we have that under the null

$$\hat{\Sigma}_f \to_p \Sigma_f = \frac{\partial}{\partial \theta} \mathbf{f} \left(\theta, \nu_0, \varrho_0\right)^{\prime} \Sigma_\theta \frac{\partial}{\partial \theta} \mathbf{f} \left(\theta, \nu_0, \varrho_0\right)$$

and $\sqrt{T} \mathbf{f} \left(\hat{\theta}, \nu_0, \varrho_0\right) \to_p N(0, \Sigma_f).$ This distributional approximation relies on the accuracy of a first-order Taylor approximation of $f \left(\hat{\theta}, \nu_0, \varrho_0\right)$ around $\theta$, $f \left(\hat{\theta}, \nu_0, \varrho_0\right) \approx \frac{\partial}{\partial \theta} f \left(\theta, \nu_0, \varrho_0\right) \left(\hat{\theta} - \theta\right)$, which seems to be problematic for the sample size in MM's empirical application.
<table>
<thead>
<tr>
<th>Value of ( R )</th>
<th>0.73</th>
<th>0.28</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>23.58</td>
<td>21.78</td>
<td>21.88</td>
<td>19.97</td>
<td>20.02</td>
<td>19.88</td>
<td>18.38</td>
<td>18.16</td>
<td>18.00</td>
<td>17.80</td>
<td>16.38</td>
<td>16.12</td>
<td>15.95</td>
<td>15.72</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4.00</td>
<td>3.63</td>
<td>3.36</td>
<td>3.13</td>
<td>2.93</td>
<td>2.74</td>
<td>2.57</td>
<td>2.41</td>
<td>2.26</td>
<td>2.11</td>
<td>1.97</td>
<td>1.84</td>
<td>1.71</td>
<td>1.58</td>
</tr>
<tr>
<td>10</td>
<td>4.00</td>
<td>3.63</td>
<td>3.36</td>
<td>3.13</td>
<td>2.93</td>
<td>2.74</td>
<td>2.57</td>
<td>2.41</td>
<td>2.26</td>
<td>2.11</td>
<td>1.97</td>
<td>1.84</td>
<td>1.71</td>
<td>1.58</td>
</tr>
<tr>
<td>9</td>
<td>4.00</td>
<td>3.63</td>
<td>3.36</td>
<td>3.13</td>
<td>2.93</td>
<td>2.74</td>
<td>2.57</td>
<td>2.41</td>
<td>2.26</td>
<td>2.11</td>
<td>1.97</td>
<td>1.84</td>
<td>1.71</td>
<td>1.58</td>
</tr>
<tr>
<td>8</td>
<td>4.00</td>
<td>3.63</td>
<td>3.36</td>
<td>3.13</td>
<td>2.93</td>
<td>2.74</td>
<td>2.57</td>
<td>2.41</td>
<td>2.26</td>
<td>2.11</td>
<td>1.97</td>
<td>1.84</td>
<td>1.71</td>
<td>1.58</td>
</tr>
</tbody>
</table>

Table SI: CITE levels of \( C^* \) described in Section 3.7 for different values of \( k \) and \( p \). Value of \( R \) is equal to 0.99-quantile of distribution \( F \), based on 10 million simulations.
Our Robust MD Test

We begin by showing that we can explicitly parametrize the manifold

\[ S_{\nu, \varrho} = \{ \theta \in \mathbb{R}^{12} : f(\theta, \nu, \varrho) = 0 \} \]

in terms of six VAR parameters. Let

\[
A(\theta) = \begin{bmatrix}
\theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} \\
\theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} & \theta_{26} \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}
\]

and note that \( e_\pi = (1, 0, 0, 0, 0, 0)' \) and \( e_x = (0, 1, 0, 0, 0, 0)' \). The constraint function \( f(\theta, \nu, \varrho) \) can be re-written as

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

Note that

\[
A(\theta)' A(\theta)' e_\pi = A(\theta)' \theta_1 = \theta_{11} \cdot \theta_1 + \theta_{12} \theta_2 + B \cdot \theta_i
\]

where \( \theta_i = \begin{bmatrix} \theta_{i1} & \theta_{i2} & \theta_{i3} & \theta_{i4} & \theta_{i5} & \theta_{i6} \end{bmatrix}' \) and

\[
B = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.
\]
Together with the observation that $A(\theta)'e_\pi = \theta_1$, $A(\theta)'e_x = \theta_2$, this allows us to re-write

$$f(\theta, \nu, \varrho) = \theta_1 - \frac{(1 - \nu)^2}{\nu (1 + \varrho)} \theta_2 - \frac{1}{1 + \varrho} (\theta_{11} \cdot \theta_1 + \theta_{12} \cdot \theta_2 + B \cdot \theta_1) - \frac{\varrho}{1 + \varrho} e_\pi$$

$$= \left( I - \frac{1}{1 + \varrho} (\theta_{11} \cdot I + B) \right) \theta_1 - \left( \frac{(1 - \nu)^2}{\nu (1 + \varrho)} + \frac{1}{1 + \varrho} \theta_{12} \right) \theta_2 - \frac{\varrho}{1 + \varrho} e_\pi.$$

Thus, provided $\frac{(1 - \nu)^2}{\nu (1 + \varrho)} + \frac{1}{1 + \varrho} \theta_{12} \neq 0$, in order to impose the restriction that $f(\theta, \nu, \varrho) = 0$ we can set

$$\theta_2(\theta_1, \nu, \varrho) = \frac{1}{\frac{(1 - \nu)^2}{\nu (1 + \varrho)} + \frac{1}{1 + \varrho} \theta_{12}} \left( \left( I - \frac{1}{1 + \varrho} (\theta_{11} \cdot I + B) \right) \theta_1 - \frac{\varrho}{1 + \varrho} e_\pi \right),$$

$$\theta(\theta_1, \nu, \varrho) = \begin{bmatrix} \theta_1 \\ \theta_2(\theta_1, \nu, \varrho) \end{bmatrix}.$$  

Hence, for any fixed $(\nu, \varrho)$ we can explicitly parametrize the manifold $S_{\nu, \varrho}$ in terms of the 6-dimensional parameter $\theta_1$. In the expression for $\theta_2(\theta_1, \nu, \varrho)$, if we fix $\theta_{11}$ and $\theta_{12}$ the remaining four parameters $(\theta_{13}, \theta_{14}, \theta_{15}, \theta_{16})$ enter linearly. Hence, if we project over $\theta_{11}$ and $\theta_{12}$ it is easy to see that the curvature with respect to remaining four parameters is zero.

To calculate confidence sets, we follow MM and take a grid of values for $\nu$ ranging from 0.01 to 0.99, with grid points spaced 0.0025 apart, and a grid of values for $\varrho$ ranging from 0 to 1 with grid points spaced 0.0025 apart.

**Size Simulations: Simulation Design**

Section 6 of the paper compares the finite-sample size of our robust MD test and the MM test in simulation. Below we discuss how we calibrate these simulations to match the empirical setting.

Following MM, we assume that pair $(\pi_t, x_t)$ follows a 3rd order VAR process. For this process to be consistent with the NKPC model we need the parameters $(\theta, \nu, \varrho)$
to satisfy the restriction \( f(\theta, \nu, \varphi) = 0 \). Let \( \hat{\theta} \) be the VAR coefficients estimated from the data with estimated variance matrix \( \hat{\Sigma}_\theta \). We estimate the coefficients \((\hat{\theta}, \hat{\nu}, \hat{\varphi})\) for use in our simulations via minimum distance estimator

\[
(\hat{\theta}, \hat{\nu}, \hat{\varphi}) = \arg \min_{\theta, \nu, \varphi} \{ f(\theta, \nu, \varphi) = 0 \} \quad \hat{\Sigma}_\theta^{-1} (\hat{\theta} - \theta).
\]

This yields the combination of reduced form and structural coefficients consistent with the NKPC model closest to the VAR coefficients estimated from the data. Using the residuals \( \hat{\epsilon}_t \) from the unrestricted VAR regression, we estimate the covariance matrix of the driving shocks by

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

We simulate samples \((\pi^*_t, x^*_t)\) from VAR model with VAR coefficients \( \hat{\theta} \) and normally distributed shocks with a covariance matrix \( \hat{V}_e \). The sample size in our simulations is 100 and the number of performed simulations equals to 10,000.


