Robust Inference in Models Identified via Heteroskedasticity

Daniel J. Lewis
Harvard University*

October 31, 2017

Preliminary draft. All comments welcome via email. Most recent version available here.

Abstract

Simultaneous equations models identified via heteroskedasticity may be subject to weak identification concerns due to proportional changes in variance across series. Considering data from Nakamura & Steinsson (2017), I calibrate simulations to show the relevance of these concerns in practice. I propose conditions under which robust inference methods for a subset of the parameter vector are valid. In simulation, these methods avoid dramatic size-distortions present in strong-identification methods. While there is power loss relative to standard inference, they are less conservative than the previous alternative of projection methods. I propose two tests for weak identification. I offer a method for robust inference on IRFs for SVARs identified via heteroskedasticity. An empirical application to Nakamura & Steinsson (2017) shows that weak identification is a problem with daily policy shocks, but not higher-frequency shocks. This confirms the authors’ suspicions, and shows the value of focusing empirical work on carefully constructed shock series.

JEL Classification: C12, C32, E43

*I would like to thank Jim Stock, Isaiah Andrews, Adam McCloskey, Jose Montiel Olea, Emi Nakamura, Mikkel Plagborg-Møller, and Jón Steinsson for their helpful feedback.
1 Introduction

Latent variables, like the shocks in the structural vector auto-regressions (SVARs) of Sims (1980), are ubiquitous in economic models across fields, where observed innovations are related to unobserved (structural) shocks by a linear combination matrix. A variety of identification schemes have been applied to map such innovations to structural shocks. In recent years, the literature has turned away from imposing identifying assumptions on the linear combination matrix between shocks and innovations, realizing such assumptions are too closely related to the structural relationships of interest. Instead, focus has turned to how assumptions on the structural shocks can contribute to identification. First proposed by Rigobon (2003), identification via heteroskedasticity uses the presence of multiple variance regimes to identify the structural parameters of interest. Reduced-form second moments are products of those parameters and second moments of the structural shocks. Identification via heteroskedasticity supposes that the variance regimes can be discerned (either using prior information or estimation), and sometimes that only a subset of variances change. It is reasonable to worry that in small samples, or in data measured with noise, that the variance regimes may not be markedly distinct. Such failure of the identifying assumptions can result in weak identification, as noted by Magnusson & Mavroeidis (2014), Nakamura & Steinsson (2017), and Hébert & Schreger (2017). Unfortunately, there do not currently exist general methods for assessing the strength of identification or conducting robust inference in this setting.

This paper draws on the extensive weak identification literature to provide a comprehensive treatment of the weak identification problem potentially present in models identified via heteroskedasticity. First, I provide a theoretical exposition of how such a system of VAR residuals may exhibit weak identification properties, drawing an analogy to classical weak-IV to build intuition. I develop a fully general model, allowing for \( n \) variables where multiple (or all) variances may change. A Monte Carlo study displays the extent to which asymptotic distributions are affected for an empirically-motivated calibration. Limiting representations are established for \( S \)-statistics and \( K \)-statistics for both the full parameter and a subset of parameters, facilitating non-conservative identification-robust inference. Conducting such inference on subsets of the parameter vector is crucial for applied work, and is not straightforward in weak identification settings. The tests proposed have the advantage over previously available approaches of being neither prohibitively conservative, unlike projection methods (Dufour (1997), Robins (2004), Dufour & Taoumouti (2005), Kleibergen (2005), Chaudhuri (2008), Chaudhuri et al (2010), Chaudhuri & Zivot (2011), McCloskey (2016)), nor limited to a simplified two-variable setting (Nakamura & Steinsson, 2017). Monte Carlo
analysis confirms that standard Wald tests exhibit dramatic size distortions, while the robust methods perform well. Next, tests of identification are offered; bias-based critical values for a first-stage $F$-test are suggested from Montiel Olea & Pflueger (2013) for an empirically-common simple case, and a fully-general size-based method is proposed from Andrews (2017).

I also offer a new approach to inference on the final object of interest in many macroeconomic applications, impulse response functions (IRFs). IRFs are transformations of both reduced-form and structural parameters. Point-wise confidence sets can be constructed using a simple test-inversion approach, with relatively small critical values. This parallels the recent work of Montiel Olea, Stock, & Watson (2016), who offer such a method in the external instruments setting; however, their method exploits linearities not present in the identification via heteroskedasticity context except in the simplest case. As such, the method proposed here is more general, and, I believe, the first to apply in this context.

Throughout, the paper considers an empirical application to the data of Nakamura & Steinsson (2017). They study the impact of forward guidance on treasury futures yields, using a variety of bivariate simultaneous equations models. They use identification via heteroskedasticity as a comparison to their main identification strategy. Two policy series are considered to track the latent monetary policy shocks: 1-day changes in treasury yields and 30-minute changes in a “policy news” series constructed using principle components. They find suggestive evidence of weak identification for the 1-day changes when they construct robust confidence sets using a bootstrap approach to an AR-type statistic. In applying the techniques developed here, I confirm the authors’ suspicions, with the one-day changes in the yield not passing any of the proposed tests of identification. On the other hand, the 30-minute policy shocks satisfy the tests at virtually all levels. In addition, Monte Carlo work calibrated from the data shows dramatic size-distortions resulting from using test statistics employing strong-identification asymptotics.

In light of these findings, scrutiny must be applied to the proliferation of applied papers now using the approach. While many applications come from monetary economics and international finance, examples now exist in public finance (Jahn & Weber, 2016), growth (Islam et al, 2017), trade (Lin et al, 2016, Feenstra & Weinstein, 2017), political economy (Rigobon & Rodrik, 2005, Khalid, 2016), environmental economics (Millimet & Roy, 2016, Gong et al, 2017), agriculture and energy (Fernandez-Perez et al, 2016), education (Hogan & Rigobon, 2009, Klein & Vella, 2009), marketing (Zaefarian et al, 2017), and even fertility studies (Mönkedick & Bras, 2016). Within the macrofinance core of the literature, notable papers include Rigobon & Sack (2003, 2004), Craine & Martin (2008), Eichengreen & Panizza (2016), Ehrmann & Fratzscher (2016), and Hébert & Schreger (2017). An entirely separate
strand of literature focuses on the method of Lewbel (2012), which is based on Rigobon’s (2003) approach; such analysis is also subject to these weak identification issues, but further work remains to extend the present results to Lewbel’s setting.

The paper proceeds as follows. Section 2 formally presents the identification via heteroskedasticity approach, and provides intuition behind the weak identification problem and when it may arise, supported by simulation evidence. Section 3 proposes methods for weak identification robust inference, both for the full parameter vector and a subset of parameters, and compares them to alternative methods in a Monte Carlo study. Section 4 offers two tests of identification. Section 5 develops a method to construct robust confidence sets for IRFs in non-linear identification schemes. Section 6 presents the empirical application. Section 7 concludes.

\( M_{ij} \) denotes the \( ij^{th} \) element of matrix \( M \)

\( M^{(j)} \) denotes the \( j^{th} \) column of matrix \( M \)

\( M_{(i)} \) denotes the \( i^{th} \) row of matrix \( M \)

\( \text{vech}(M) \) denotes the unique vectorization of matrix \( M \)

\( A_i \) denotes the matrix of coefficients corresponding to lag \( i \) of the lag polynomial \( A(L) \)

2 A weak identification problem

2.1 The standard setting

Before considering issues of weak identification, it is necessary to discuss how identification via heteroskedasticity behaves in an ideal environment. Consider an \( n \)-dimensional system of equations of the form

\[ \eta_t = H \varepsilon_t, \]

where \( \varepsilon_t \) are structural shocks with a diagonal covariance matrix. Often, these innovations, \( \eta_t \), result from a reduced-form VAR of the form

\[ A(L) Y_t = \eta_t, \]

which has the structural representation

\[ C(L) Y_t = \varepsilon_t, \]
where \( C(L) = H^{-1}A(L) \). The lag polynomial, \( A(L) \), and thus the reduced form residuals \( \eta_t \), can be estimated consistently by OLS. The conditional variance of \( \varepsilon_t \) is central to the identification scheme. Identification via heteroskedasticity exploits the additional moment equations generated by assuming multiple variance regimes for the structural shocks and thus the reduced-form residuals. This requires at least two variance regimes, in which case the system just-identified. The literature makes some standard assumptions:

**Assumption 1.** For all \( t = 1, 2, \ldots, T \) and regimes \( R_j, j = 1, 2, \ldots, k \),

1. \( H \) is fixed over time, invertible, and has a unit-diagonal,
2. \( E[\varepsilon_t | t \in R_j] = 0, \quad E[\varepsilon_t \varepsilon_t' | t \in R_j] = \Sigma_{\varepsilon,j} \),
3. \( \Sigma_{\varepsilon,j}^{-1/2} \varepsilon_t \), for \( t \) in \( R_j \), are i.i.d. cross-sectionally and across time.

The unit diagonal normalization means that a unit structural shock to variable \( i \) is assumed to have a unit impact on the reduced form innovation to \( i \). Assumption 1.2 assumes stationarity within each regime and the existence of the expectation \( \Sigma_{\varepsilon,j} \). The independence of the shocks allows estimators to be well-behaved, but could be relaxed to a mixing condition. The identical distribution assumption is made for convenience.

For simplicity of exposition, I discuss the two variance regime case throughout this paper. This is without loss of generality – the results can be extended to settings with more regimes, like that considered in Rigobon & Sack (2003). In particular, note that if a system of equations demonstrates weak identification for two regimes, further sub-dividing the regimes will not necessarily strengthen the variation.\(^1\)

A typical framework for identification via heteroskedasticity contrasts “event” observations and “control” observations, arguing that on the event days, when, for example, a piece of news reaches a market, variables are likely to be more volatile than on a typical day.\(^2\) One of the most common applications is typified by Nakamura & Steinsson (2017), studying monetary policy, where the event observations are changes in returns on announcement days, and the control observations are changes in returns on other days. On days when monetary policy announcements are made, relevant securities prices or interest rates should exhibit more dramatic movements than on a typical day. In keeping with this dominant application,\(^1\)

---

\(^1\)The extension of the theoretical results of this paper to accommodate additional regimes, up to \( k \), is quite straightforward. Most extend trivially. When a rank condition or similar is required for a theorem to apply, it is then taken with respect to a matrix containing \( k \) columns for the variance regimes, instead of 2.

\(^2\)When regimes and breaks must instead be estimated, there can be substantial bias in the estimation, as discussed in Lewis (2017).
I replace the sets of dates, \( R_j \), with \( P, C \subset T \), (\( P \) for policy, high variance, \( C \) for control, low variance) and \( T_p = |P| \) and similarly for \( C \). Then, by Assumption 1.1-1.2, the reduced-form covariance is

\[
\Sigma_{\eta,j} = H \Sigma_{\varepsilon,j} H',
\]

(1)

Given two periods, (1) yields \( 2 \times \frac{n(n+1)}{2} = n^2 + n \) equations, with \( n^2 - n + 2 \times n = n^2 + n \) unknowns. The system is just-identified in an order-condition sense; for now, assume that a rank condition also holds, so it is in fact identified. The system lends itself to estimation via GMM. In the \( n = 2 \) case, for each regime \( j \), the reduced form covariance can be decomposed as

\[
\begin{align*}
\sigma_{\eta_1,j}^2 &= \sigma_{\varepsilon_1,j}^2 + H_{12}^2 \sigma_{\varepsilon_2,j}^2 \\
\sigma_{\eta_2,j}^2 &= H_{21}^2 \sigma_{\varepsilon_1,j}^2 + \sigma_{\varepsilon_2,j}^2 \\
\sigma_{\eta_1,\eta_2,j} &= H_{21} \sigma_{\varepsilon_1,j}^2 + H_{12} \sigma_{\varepsilon_2,j}^2.
\end{align*}
\]

Assumption 1.2 and (1) imply \( E \left[ \eta_t' \eta_t \mid t \in T_j \right] = \Sigma_{\eta,j} \), which suggests the natural moment equations

\[
m_t (\eta_t, H, \Sigma_{\varepsilon,j}) = \begin{bmatrix}
\eta_1^2 - \sigma_{\varepsilon_1,j}^2 - H_{12}^2 \sigma_{\varepsilon_2,j}^2 \\
\eta_2^2 - H_{21}^2 \sigma_{\varepsilon_1,j}^2 - \sigma_{\varepsilon_2,j}^2 \\
\eta_1 \eta_2 - H_{21} \sigma_{\varepsilon_1,j}^2 - H_{12} \sigma_{\varepsilon_2,j}^2
\end{bmatrix},
\]

where date \( t \) is in regime \( j \). Define the vector \( \theta \in \Theta \) as the unique elements of \( H, \Sigma_{\varepsilon,C}, \) and \( \Sigma_{\varepsilon,P}, \) where \( \Theta \) is compact. Then, stacking the moment equations from each regime, the resulting moment function as

\[
\phi_t (\theta) = \begin{bmatrix}
1 [t \in T_C] (\text{vech} (\eta_t' \eta_t) - \text{vech} (H \Sigma_{\varepsilon,C} H')) \\
1 [t \in T_P] (\text{vech} (\eta_t' \eta_t) - \text{vech} (H \Sigma_{\varepsilon,P} H'))
\end{bmatrix},
\]

(2)

Clearly, \( E [\phi_t (\theta_0)] = 0 \) at \( \theta_0 \), the true parameter value. The standard GMM objective function is defined as

\[
S_T (\theta; \hat{\theta}_T (\theta)) = \left[ T^{-1/2} \sum_{t=1}^T \phi_t (\theta) \right]' W_T (\hat{\theta}_T (\theta)) \left[ T^{-1/2} \sum_{t=1}^T \phi_t (\theta) \right].
\]

(3)

Expressing identification via heteroskedasticity as a GMM problem is not novel; Rigobon (2003), Rigobon & Sack (2003, 2004), and Craine & Martin (2008) all use a related approach. Focus is hence restricted to the continuous-updating estimator (CUE) form of (3) with the efficient weighting matrix. This implies the use of \( S_T (\theta) \) and \( W_T (\theta) \) in (3), where \( W_T (\theta) = \Omega_T (\theta)^{-1}, \Omega_T (\theta) = E [\phi_t (\theta)' \phi_t (\theta)] \).
2.2 Identification in the simple case

The discussion above, and much of the literature, assumes a rank condition to hold, but that need not be the case. Consider a \(2 \times 2\) variable system where only the variance of one structural shock (say the second) changes, a common model in the literature. This is henceforth referred to as “the simple case”. Economically, it corresponds to the intuition that the monetary policy may be more volatile on FOMC announcement days, but other shocks should not be. Then, writing out the equations for each regime under Assumption 1, 

\[
\Sigma_{\eta,C} = \begin{pmatrix}
\sigma_{\eta_1,C}^2 & \sigma_{\eta_1\eta_2,C} \\
\sigma_{\eta_1\eta_2,C} & \sigma_{\eta_2,C}^2
\end{pmatrix} = H \begin{pmatrix}
\sigma_{\xi_1}^2 & 0 \\
0 & \sigma_{\xi_2,C}^2
\end{pmatrix} H',
\]

\[
\Sigma_{\eta,P} = \begin{pmatrix}
\sigma_{\eta_1,P}^2 & \sigma_{\eta_1\eta_2,P} \\
\sigma_{\eta_1\eta_2,P} & \sigma_{\eta_2,P}^2
\end{pmatrix} = H \begin{pmatrix}
\sigma_{\xi_1}^2 & 0 \\
0 & \sigma_{\xi_2,P}^2
\end{pmatrix} H'.
\]

\(H_{12}\) is the parameter of interest. This off-diagonal element measures the impact of a unit structural shock (say a policy shock) on another variable in the system. For instance, in Nakamura & Steinsson (2017), this represents the impact of policy news on treasury futures, which they argue proxy for market expectations of monetary policy. This is frequently the case in empirical work, where the policy shock is the one exhibiting heteroskedasticity, and the researcher is interested in its impact on the other series (the “dependent” variable). In this setting, \(H_{12}\) is identified in closed form:

\[
\frac{\sigma_{\eta_1\eta_2,P} - \sigma_{\eta_1\eta_2,C}}{\sigma_{\eta_2,P}^2 - \sigma_{\eta_2,C}^2} = \frac{H_{12} \Delta \sigma_{\xi_2}^2}{\Delta \sigma_{\xi_2}^2} = H_{12},
\]

where \(\Delta\) is the difference operator. This is in fact just one of three closed-form expressions for \(H_{12}\) (as the problem is over-identified).

Due to an observation in Rigobon & Sack (2004), this can be characterized as an IV problem:

\[
\hat{H}_{12} = \frac{\Delta \hat{\sigma}_{\eta_{12}}}{\Delta \hat{\sigma}_{\eta_{2}}} = \frac{\frac{1}{T_P} \sum_{t \in P} \eta_{1t} \eta_{2t} - \frac{1}{T_C} \sum_{t \in C} \eta_{1t} \eta_{2t}}{\frac{1}{T_P} \sum_{t \in P} \eta_{2t}^2 - \frac{1}{T_C} \sum_{t \in C} \eta_{2t}^2} = \frac{\sum_{t=1}^{T} \eta_{1t} Z_t}{\sum_{t=1}^{T} \eta_{2t} Z_t},
\]

where

\[
Z_t = \left[ 1 \ (t \in T_P) \times \frac{T}{T_P} - 1 \ (t \in T_C) \times \frac{T}{T_C} \right] \eta_{1t}.
\]

\(3\)The following development of weak identification applies equally to either of the alternative identifying equations.
When this instrument is not strongly correlated with the residuals, weak identification arises. Why might this be the case? Suppose the change in the first shock is quite small, such that it can be modeled as local-to-zero. Then the variance changes of the reduced form residuals are also local to zero:

$$\Delta \sigma_{\eta_2}^2 = d_2 / T^{1/2}$$

$$\Delta \sigma_{\eta_1 \eta_2} = d_{12} / T^{1/2}. $$

In this case, the estimator will have an asymptotic distribution described by

$$\hat{H}_{12} \xrightarrow{d} \frac{z_{12}}{z_2}, \text{ where } z = \left( \begin{array}{c} z_{12} \\ z_2 \end{array} \right) \sim \mathcal{N}\left( \begin{pmatrix} d_{12} \\ d_2 \end{pmatrix}, V \right). $$

This is the canonical weak instruments setting - the estimator is not consistent for $H_{12}$, nor is it asymptotically normally distributed, rather the ratio of two (generally correlated) normals. As a result, standard inference methods break down.

### 2.3 General case

I now extend the intuition present in the simple case to a general $n$-variable framework. In doing so, I take serious the rank condition, previously alluded to (and violated in the weak-IV discussion above). Sentana & Fiorentini (2001) establish a condition for $H$ to be globally identified in the presence of time-varying volatility, which is simplified in Proposition 1.

**Proposition 1.** $H$ is globally identified from $\Sigma_{\eta_1}$ and $\Sigma_{\eta_2}$ up to column order provided the rows of

$$\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}$$

are not proportional.

**Proof.** See Appendix. \qed

Proposition 1 presents the conditions under which the equations of identification via heteroskedasticity satisfy the rank condition, and the system is in fact identified. Under an additional assumption, distinguishing the columns of $H$ or the shocks, identification is to a point, not just up to column order. An example, adopted hence, is Assumption 2:

**Assumption 2.** The shock of interest experiences the largest relative change in variance across regimes.

This is one of a large class of assumptions that allow the researcher to impose a partial order on the columns of $H$. This assumption was implicit in the simple case above. For a more detailed discussion of this labeling problem in a related setting, see Lewis (2017).
Proposition 1 implies that weak identification can arise if two (or more) variances change by a similar factor. This nests the scenario illustrated in the above simple case, where one variance change is local-to-zero and the other change is precisely zero. As an example of variances changing by similar factors, consider a case where, following a macroeconomic announcement, two variables, including the variable of interest, experience shocks with about $\gamma$ times the usual volatility. This could occur if there are shocks to more than one dimension of monetary policy news. During the recovery from the financial crisis, the Fed made simultaneous announcements about the path of the Fed Funds rate and Quantitative Easing; the Reserve Bank of New Zealand routinely announces changes to its Official Cash Rate simultaneously with the release of detailed projections for a variety of key financial variables. Depending on the coordination between policy measures, it is plausible that their variances change by comparable factors. This general pathology of proportional changes is modeled as a local-to-unity relationship in the ratio of relative changes of two variances,

$$
\frac{\sigma^2_{\varepsilon_1P}}{\sigma^2_{\varepsilon_1C}} = 1 + \frac{d}{\sqrt{T}},
$$

where $d$ is finite. Without loss of generality, let $j = 1, k = 2$ and denote $\sigma^2_{\varepsilon_1P}/\sigma^2_{\varepsilon_1C} = \gamma$, so

$$
\sigma^2_{\varepsilon_2P} = \gamma \sigma^2_{\varepsilon_2C} \left(1 + \frac{d}{\sqrt{T}}\right).
$$

This nests the near-zero issue of the simple case when $\gamma = 1$.

**Proposition 2.** Adopting the modeling device in (4) and Assumption 1.1-1.3, $H$ is asymptotically unidentified.

**Proof.** See Appendix.

Under the local-to-unity modeling device, the proportionality requirement of Proposition 1 fails, resulting in an unidentified system, as stated in Proposition 2.

Dufour (1997) (Section 4) demonstrates the impact that such deficiencies can have on testing problems. He shows that the size of Wald tests tends to unity as a system tends towards non-identification. In this setting, it is straightforward to show, fixing $d$, that the size of a Wald test on the full parameter vector is unity asymptotically. In fact, the results of the Monte Carlo simulations in Table 2 illustrate this well, establishing the severity of the problems caused by weak identification in this setting.

It is important to note that in the case where $H_{12}$ is close to zero, there may also be a weak identification problem when $H_{12}$ is estimated using the analogy to linear IV. Hébert
Table 1: Monte Carlo Specifications

<table>
<thead>
<tr>
<th></th>
<th>very weak identification</th>
<th>data</th>
<th>strong identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>small sample</td>
<td>$T = 375, \delta = \delta/10$</td>
<td>$T = 375, \delta = \delta$</td>
<td>$T = 375, \delta = 10 \times \delta$</td>
</tr>
<tr>
<td>data</td>
<td>$T = 750, \delta = \delta/10$</td>
<td>$T = 750, \delta = \delta$</td>
<td>$T = 750, \delta = 10 \times \delta$</td>
</tr>
<tr>
<td>large sample</td>
<td>$T = 1500, \delta = \delta/10$</td>
<td>$T = 1500, \delta = \delta$</td>
<td>$T = 1500, \delta = 10 \times \delta$</td>
</tr>
</tbody>
</table>

Each cell represents a calibration of the Monte Carlo exercise. The sample split allocates $\text{ceil}(\frac{676}{750}T)$ of the observations to the control period, in keeping with the data sample studied in the empirical application.

$\hat{\delta} = \frac{\sigma_{1P}^2}{\sigma_{2P}^2/\sigma_{2C}^2} - 1$ is the estimated measure of identification from the empirical data.

& Schreger (2017) argue for the use of one of the three possible closed form identifying equations as the others have a zero denominator under their null hypothesis ($H_{12} = 0$), since $H_{12}$ appears in the denominator. This is not a concern in the GMM method discussed here. However, if, like in Hébert & Schreger (2017), one is interested in the null hypothesis that $H_{12} = 0$ the identification problems resulting from negligible variance changes are potentially exacerbated under the null.

### 2.4 Simulation evidence

#### Distribution of estimators

Monte Carlo evidence illustrates the scope of the problem in empirically-calibrated data. Indeed, the distributions of estimates under weak identification are far from those assumed under standard asymptotics. The central Monte Carlo specification is calibrated from estimates from the nominal future, 1-day window nominal yield data of Nakamura & Steinsson (2017), discussed in more detail in the empirical application. In particular,

$$H = \begin{bmatrix} 1 & 0.38 \\ 0.65 & 1 \end{bmatrix}, \Sigma_{\epsilon,C} = \begin{bmatrix} 0.0042 & 0 \\ 0 & 0.0004 \end{bmatrix}, \Sigma_{\epsilon,P} = \begin{bmatrix} 0.0065 & 0 \\ 0 & 0.0011 \end{bmatrix}. \quad (5)$$

Define $\delta \equiv \frac{\sigma_{1P}^2}{\sigma_{2P}^2/\sigma_{2C}^2} - 1$. I calculate the implied $\hat{\delta} = \frac{\hat{\sigma}_{1P}^2}{\hat{\sigma}_{2P}^2/\hat{\sigma}_{2C}^2} - 1$ as a measure of identification, which can be varied in the simulations. The specifications considered are described in Table I. $T = 750$ is taken as the central sample-length, in keeping with the empirical application. All other parameters are held constant. Estimation proceeds by CUE
Figure 1 presents histograms of the $t$–statistic on the parameter of interest, $H_{12}$, for 10,000 draws. It is clear that the estimates are not normally distributed for low degrees of identification, heuristically proxied by $\delta \times \sqrt{T}$. However, for the “strong identification” specifications, the distribution is closer to a $t$–distribution.

To further illustrate the implications of this problem in a macroeconomic context, I impose a reduced form VAR(1) specification estimated in the empirical application:

$$A_1 = \begin{bmatrix} -0.078 & -0.086 \\ 0.079 & 0.064 \end{bmatrix}.$$ 

I consider how weak identification propagates to the IRFs. I apply the “true” value of the reduced form coefficients to the estimated $\hat{H}$ matrices, as, in standard estimating procedures, $A_1$ will be estimated consistently, and contribute relatively little to the variation in distribution of the IRFs. For clarity of illustration, I calculate and plot the IRFs for only 200 draws, so the paths are actually distinguishable. Figure 2 presents the results. These should be concerning to macroeconomists - under weak identification, the distribution of the IRFs is highly diffuse - there should be little confidence in an estimated path from any single observed draw when weak identification is a concern. Even the sign of the responses is ambiguous.

### 3 Weak identification robust inference

#### 3.1 Asymptotic distribution of test statistics

Existing results for the asymptotic properties of weakly identified GMM apply directly in this context. The asymptotic distribution of GMM estimators, robust to weak identification, was established in Stock, Wright, & Yogo (1997) and Stock & Wright (2000). Instead of providing an asymptotic distribution for the parameter estimates, as in strongly identified GMM problems, they show that $S_T(\theta_0)$ follows a Chi-square distribution. Kleibergen (2005) provides a more refined statistic, which is asymptotically efficient, making use of the estimated Jacobian.

The present paper proceeds using the Kleibergen “$K$–statistic”, as defined in Kleibergen (2005), while the simpler $S$–statistic is retained in numerical work for comparison. Note that while there now exist test statistics that may exhibit more desirable properties than
$t-$statistics on $H_{12}$ calculated from 10,000 Monte Carlo draws, using the sample length in the left margin and the degree of identification in the bottom margin. Extreme outliers are truncated to allow comparison on the same axes. Calibration details are given in equation (5). Point estimation proceeds via Sims’ (2014) eigenvector method, with inference using these values in CUE GMM.
Figure 2: Distribution of IRFs

IRF paths for a unit shock for 200 Monte Carlo draws of $\hat{H}_{12}$, taking $A_1$ as given. Sample length is given in the left margin and the degree of identification in the bottom margin. Calibration details are given in equation (5). The horizontal axis is calibrated to days. Point estimation of $H$ proceeds via Sims’ (2014) eigenvector method, with inference using these values in CUE GMM.
the $K-$statistic (for example Andrews’ (2016) Conditional Linear Combination tests), it is pursued here since most newer statistics make use of a convex combination of the $K-$ and $J-$statistics. In the leading just-identified case, the $J-$statistic, testing overidentifying restrictions, cannot be used. However, researchers interested in overidentified systems should consider such refinements. To proceed, additional assumptions are needed.

**Assumption 3. Assume**

1. $E \left[ |\varepsilon_t|^{4+\rho} \right] < \infty$, $\rho > 0$ in all regimes for all $t = 1, 2, \ldots, T$,

2. As $T \to \infty$, $T_j \to \infty$ for all $j$.

Define

$$\tilde{\phi}_t(\theta) = \phi_t(\theta) - E(\phi_t(\theta))$$

$$q_t(\theta) = \text{vec}\left( \frac{\partial \phi_t(\theta)}{\partial \theta'} \right)$$

$$\tilde{q}_t(\theta) = q_t(\theta) - E(q_t(\theta))$$

Lemma 1 provides asymptotic distributions for $\tilde{\phi}_t(\theta_0)$ and $\tilde{q}_t(\theta_0)$.

**Lemma 1.** Under Assumptions 1 & 3,

$$\psi_T(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \begin{array}{c} \tilde{\phi}_t(\theta_0) \\ \tilde{q}_t(\theta_0) \end{array} \right) \xrightarrow{d} \left( \begin{array}{c} \psi_{\phi} \\ \psi_{\theta_0} \end{array} \right)$$

where $\psi = \left( \begin{array}{c} \psi_{\phi} \\ \psi_{\theta_0} \end{array} \right)$ is a $2(n^2 + n)$-dimensional normally distributed random variable with mean zero and positive semi-definite $2(n^2 + n) \times 2(n^2 + n)$-dimensional covariance matrix

$$V(\theta) = \begin{pmatrix} V_{\phi\phi}(\theta) & V_{\phi\theta}(\theta) \\ V_{\theta\phi}(\theta) & V_{\theta\theta}(\theta) \end{pmatrix} = \lim_{T \to \infty} \text{var} \left[ \frac{1}{\sqrt{T}} \begin{pmatrix} \phi_T(\theta) \\ q_T(\theta) \end{pmatrix} \right]$$

for $\phi_T(\theta) = \sum \phi_t(\theta)$.

Lemma 2 provides additional properties needed for the use of the estimated Jacobian in the $K-$statistic.
Lemma 2. Under Assumptions 1.3 & 3.1, the moment covariance matrix estimator \( \hat{V}(\theta_0) \) satisfies

\[
\hat{V}(\theta_0) \xrightarrow{p} V(\theta_0)
\]

and

\[
\frac{\partial \text{vec} \left( \hat{V}_{\phi\phi}(\theta_0) \right)}{\partial \theta'} \xrightarrow{p} \frac{\partial \text{vec} (V_{\phi\phi}(\theta_0))}{\partial \theta'}
\]

These two lemmas, proven in the Appendix, mirror those of Kleibergen (2005), and similarly yield Theorem 1 of that paper:

Theorem 1. If Lemmas 1 and 2 hold,

\[
K(\theta_0) \xrightarrow{d} \chi^2_{n^2+n'}
\]

This offers an asymptotic distribution for the \( K \)–statistic as defined above under the enumerated assumptions.

Monte Carlo: full vector Monte Carlo evidence shows that tests based on these identification-robust asymptotics perform far better than Wald tests. The distributions of \( t \)–statistics in the previous section suggests that strong identification asymptotics may break down, exemplified by Dufour’s (1997) unit-size result for the Wald test. To investigate this, Monte Carlo simulations are run using the calibrations of equation (5) and Table 1. For each sample size and degree of identification, 10,000 draws were taken, and Wald statistics, \( S \)–statistics, and \( K \)–statistics calculated to test the null hypothesis of the true parameter vector (a six-restriction test). These are compared to the appropriate limiting distribution 5% critical values (\( \chi^2_6 (0.95) \)). The resulting rejection rates are presented in Table 2. The Wald tests exhibit extremely large size distortions, in alignment with Dufour’s asymptotic unit size result, which improve with the strength of identification. For a given \( \delta \), increasing \( T \) increases the strength of identification. The \( S \)–tests and \( K \)–tests, however, are not systematically affected by the degree of identification, as expected of robust tests. Their size distortion does decrease with sample size, which is indicative of the respective asymptotics “kicking in”, not a lack of robustness. In a macroeconomic sense, it appears that performance of Wald-based inference approaches an acceptable level only for variance changes orders of magnitude larger than those observed empirically.
Table 2: Size of tests on the full parameter vector

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\delta}/10$ Wald</th>
<th>$\hat{\delta}/10$ S/K</th>
<th>$\hat{\delta} \times 10$ Wald</th>
<th>$\hat{\delta} \times 10$ S/K</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 375$</td>
<td>91.6</td>
<td>13.3</td>
<td>70.8</td>
<td>13.9</td>
</tr>
<tr>
<td>$T = 750$</td>
<td>91.5</td>
<td>10.1</td>
<td>64.9</td>
<td>9.7</td>
</tr>
<tr>
<td>$T = 1500$</td>
<td>91.57</td>
<td>7.9</td>
<td>58.3</td>
<td>8.1</td>
</tr>
</tbody>
</table>

Rejection rate of the true parameter vector based on 10,000 Monte Carlo draws. While S and K results are not identical, they are the same to rounding error, and thus displayed together. Calibration details are given in equation (5). Estimation via CUE GMM.

Subset inference A key challenge in the weak identification literature is inference on a subset of the parameter vector. Kleibergen (2005) offers a further refinement over Theorem 1, providing a limiting distribution for estimates of a sub-vector of parameters, provided the concentrated Jacobian, based on the remaining parameters, has full rank. Previous work with identification via heteroskedasticity has skirted the problem of subset inference besides in the simple case, leaving interested practitioners, should they worry about weak identification, to rely on projection methods based on the S−statistic (see Dufour (1997), Dufour & Taoumouti (2005), Chaudhuri (2008), Chaudhuri & Zivot (2008)) and the result of Theorem 1. I characterize the circumstances in which the present setting will satisfy Kleibergen’s rank condition, allowing the direct application of the subset K−statistic, and less conservative critical values.

To apply Kleibergen’s (2005) results, I partition θ into the parameter(s) of interest, $\beta$, and the remainder, $\alpha$, and make the following assumption:

**Assumption 4.** Conditional on $\beta$, $\alpha$ is strongly identified.

This can be operationalized as a condition on the Jacobian,

$$
J_{\alpha} (\alpha, \beta) = \lim_{T \to \infty} E \left\{ \frac{1}{T} \sum_{t=1}^{T} \left[ \left( \frac{\partial \phi_t (\alpha, \beta)}{\partial \alpha'} \right) \bigg|_{\alpha, \beta} \right] \right\} = E \left[ \left( \frac{\partial \phi_t (\alpha, \beta)}{\partial \alpha'} \right) \bigg|_{\alpha, \beta} \right]
$$

Under Assumption 2, Kleibergen (2005) proves the following Theorem.

**Theorem 2.** If Lemmas 1 and 2 hold, then under Assumption 4,

$$
K (\beta_0) \overset{d}{\to} \zeta_{p_{int}},
$$

where $\zeta_{p_{int}}$ is a $\chi^2$ random variable with $p_{int}$ degrees of freedom, and $p_{int}$ is the dimension of $\beta$.
Theorem 2 establishes that the subset $K$-statistic on the parameters of interest is asymptotically distributed Chi-square with degrees of freedom equal to the number of parameters of interest instead of the length of the full parameter vector. The proof of Kleibergen (2005) suffices under Lemmas 1 and 2 and Assumption 4. I henceforth refer to this as a “plug-in” statistic due to the approach taken with the nuisance parameters in the calculation of the statistic. The subset result, due to Kleigbergen (2005), extends Stock & Wright’s (2000) result for the concentrated $S$-statistic, which holds only when $\beta$ is the full set of weakly identified parameters.

I now characterize systems of equations identified via heteroskedasticity that satisfy Assumption 4. Since the identification results presented here are global, they also imply local identification, and thus imply the Jacobian condition, Assumption 4. I restrict analysis to the cases where $\beta$ is a single element of $H$ or a single column of $H$. In policy analysis, the object of interest is generally either the immediate impact of one shock on one variable, or an IRF, for which, a full column of $H$ is required.

Sentana & Fiorentini (2001) generalize the result described in Proposition 1 to partial identification. I reformulate their result in Proposition 3. First, define $\tilde{r}(M)$ as the “proportional row space” of matrix $M$. That is, it is the maximal set of rows in $M$ such that none are proportional.

**Proposition 3.** For $H$ partitioned $H_A; H_B$, $H_A$ is identified from the covariance matrices provided no row in the upper block of

$$\begin{bmatrix}
\text{vec} (\Sigma_{\eta A1}) & \text{vec} (\Sigma_{\eta A2}) \\
\tilde{r} (\text{vec} (\Sigma_{\eta B1}), \text{vec} (\Sigma_{\eta B2}))
\end{bmatrix}
$$

is proportional to another row.

**Proof.** This is a restatement of Sentana & Fiorentini (2001) Proposition 4, with linear independence replaced with the weaker non-proportionality, as explained in the Appendix under the proof of Proposition 1.

This result makes some progress with respect to Assumption 4. It offers partial identification when variance pathologies exist. However, in terms of inference, it shows only that if all elements in $H_B$ were fixed, the remainder would be identified, and Assumption 4 would be satisfied. However, given that the point of subset inference is to be able to choose a limited collection of parameters of interest and perform inference on them, having to fix all weakly identified parameters is not necessarily helpful. This result is loosely analogous to Theorem 3 in Stock & Wright (2000). Armed with the fact that attention can be limited to cases where the subset of interest is a single element or a single column, I extend the identification result of Proposition 3 from $H_A$ to all of $H$ in Theorem 3.

---

5 Contrast to the typical definition of the row space, which is further limited to not include bases that are any type of linear combination of the others.
Theorem 3. For $H$ partitioned $H_A: H_B$, where $H_B$ contains at most two columns, $H$ is identified from the covariance matrices provided no row in the upper block of
\[
\begin{bmatrix}
\text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\
\tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2}))
\end{bmatrix}
\]
is proportional to another row if

1. A single element $H_{ij}$ is fixed and $H_{jk} \neq 1/H_{ij}$ for $H^{(j)}, H^{(k)} \in H_B$, or

2. The full column $H^{(j)} \in H_B$ is fixed.

Proof. See Appendix.

This result is more useful in a subset testing context than Proposition 3. The Sentana & Fiorentini result yields partial identification, for the columns of $H$ unaffected by variance deficiencies. When there is proportionality in the variances, this does not say anything about the affected columns of $H$. As noted above, this means Assumption 4 cannot be met unless all affected columns of $H$ are conditioned upon. By explicitly incorporating the information to be used in the null hypothesis of the subset test, I obtain conditional strong identification for the remainder of $H_B$ by Theorem 3. A system of equations satisfying the conditions of Theorem 3 meets Assumption 4, and the subset $K$-statistic is valid.

To ground this in the weak identification setting, recall the local-to-unity device used in Section 2.3. Using this approach, I characterize weak identification settings for which the subset $K$-statistic is valid.

Condition 1. If there are two variances that change by an $O\left(1/\sqrt{T}\right)$ factor, and one is that of the shock of interest, then Assumption 4 is satisfied asymptotically.

While this result constitutes an improvement on the previous possibilities for subset inference, it warrants some important remarks.

Remark 1. The shock of interest must be one of those affected by the variance pathology. Given how identification via heteroskedasticity is usually applied, where focus is on changes in the variance of the shock of interest, this is not out of line with common practice.

Remark 2. Given how few variances are permitted to be affected by these pathologies, a researcher should err towards minimizing the number of series in the system of equations subject to the constraint that the reduced form innovations span the structural shocks (invertibility).

Remark 3. Moreover, in settings with three or more variables, if only one is thought to exhibit substantial variance changes across the samples, as is often the case, then the subset $K$-statistic cannot be used. This is due to the fact that the series of interest is not one
Monte Carlo: subset Repeating the Monte Carlo assessment of size-distortion for subset tests demonstrates scope for improvement over standard procedures. Now I compare four testing approaches: the Wald test, the projected $S$–test based on Theorem 1, the new proposed plug-in $K$–test, based on Theorem 2, and a plug-in $S$–test. The results are displayed in Table 3. The projected $S$–test represents the best available option for inference in this general setting, absent the results of Theorem 3. First, as with tests on the full parameter vector, the standard Wald test is substantially oversized, though the problem is not as serious as for the full vector. As identification gets stronger, the distortion shrinks. The $S$–test based on projection methods is substantially undersized, rejecting extremely rarely. The rejection rate is effectively zero in simulation. However, the plug-in method using both the $K$– and $S$–statistics is consistently well-sized, regardless of the degree of identification, or sample size.

In contrast, the “Fieller’s Method” strategy employed in Nakamura & Steinsson (2017) faces the same downsides as the standard projection tests. Nakamura & Steinsson construct their weak identification robust confidence intervals using a bootstrap-based version of Fieller’s Method, which they draw from Staiger, Stock, & Watson (1997). Asymptotically, this coincides with an $S$–statistic. This method allows them to test estimates of $H_{21}$, under the assumption of a single variance change. Their approach consists of constructing the test statistic

$$\omega(H_{21}) = \Delta \sigma_{\eta_1, \eta_2} - H_{21} \Delta \sigma_{\eta_1}^2$$

for candidate values of $H_{21}$, for a series of bootstrap draws from the sample. The 95% confidence interval then consists of the range of values of $H_{21}$ between the 2.5% and 97.5%
quantile of $\omega(H_{21})$. This approach can only be immediately applied to the simple case where one structural variance changes; as such, it is not suited to exercise of Table 3 or the fully general model allowing multiple variance changes. In theory, this method could be applied to other parameters or combinations of parameters in a more relaxed model, but it is not immediately possible to test each parameter separately - extending the method to more complicated specifications of $\omega(\cdot)$ would require an argument of several parameters to be specified by the researcher, and returns the standard projection problem. In other words, the use of a bootstrap approach for constructing an $S-$type statistic does not alter its limitations for subset inference. The bootstrap procedure might offer finite-sample improvements over the projection $S-$test. However, even assuming strong identification, Brüggemann, Jentsch, & Trenkler (2016) show that standard bootstrapping schemes (pairwise and wild) may fail in the presence of conditional heteroskedasticity. They propose a block bootstrap that is asymptotically valid under strong identification, but offer no discussion of weak identification.

3.2 Power improvements in subset testing

The power improvements offered by subset tests using smaller critical values are demonstrated in Monte Carlo simulation. To do so, the previous simulation approach is modified. I fix $T = 750$, but still consider a range of strengths of identification. I now test the null hypothesis of $H_{12} = 0.3787$ against a sequence of local alternatives, which are used to generate the data. Figure 3 computes power curves based on these simulations. Note that these are not size-adjusted; this is necessary to demonstrate the relative advantage of using the plug-in approaches over the projection approach. Since both are based on an identical $S$ or $K-$statistic, it is necessary to compare the values to asymptotic critical values to differentiate the tests. The strength of the plug-in approach is that it justifies using less conservative critical values in testing, not a new test statistic. At hypotheses relatively close to the true parameter value, the Wald test rejects with a significantly higher probability than either $S$-test or the plug-in $K-$test. For the two weaker identification calibrations, power of the $S-$ and $K-$tests gradually increases before increasing dramatically to close the gap on the Wald test at distant alternative values. For the strong identification calibration, the power of the $S-$ and $K-$tests rises more steadily. For all degrees of identification, there is a large spike in power around $H_{12} = 1.5$. Figure 5 suggests one possible explanation; under this particular calibration, the asymptotic variance of the parameter estimates approaches zero around this value. As it gets closer to zero, it is natural that the rejection rate of tests on that parameter increases dramatically. Naturally, the spike occurs for the weaker identifications.
Figure 3: Power

Power curves formed from estimates of rejection rates of the null hypothesis ($H_{12} = 0.38$) against a sequence of local alternatives ($x$-axis) based on 1000 Monte Carlo draws. Cubic smoothing was employed to smooth curves. The far right spikes in the robust methods is discussed in the text and Figure 5. Calibration details are given in equation (5). Estimation via CUE GMM.

Earlier as the variance is near-zero for a wider range of alternatives due to noisier estimation. A smaller effect can be seen for the Wald statistic in each parametrization. Naturally, the Wald test has better power properties the stronger the identification. The most important point is that the plugin tests outperform the projection tests universally.

Figure 4 repeats the above exercise, but now plots size-adjusted power instead of power for the Wald test, the $K$-test and $S$-test (since size-adjusted power of both $S$-tests is identical mechanically). It is clear that there is power loss due to using robust inference, particularly for alternatives close to the truth, but this must be weighed against the substantial size distortion of the $F$-tests. The plugin tests do, however, offer a substantial improvement over projection tests. It should also be noted that, particularly in more specific applied contexts, there is significant scope for more powerful robust tests (for example, the weighted $J - K$ test or the conditional linear combination test of Andrews (2016)).
Size-adjusted power curves formed from estimates of rejection rates of the null hypothesis ($H_{12} = 0.38$) against a sequence of local alternatives ($x$-axis) based on 1000 Monte Carlo draws. The critical values are based on quantiles from 10,000 Monte Carlo draws. Cubic smoothing was employed to smooth curves. The far right spikes in the robust methods is discussed in the text and Figure 5. Calibration details are given in equation (5). Estimation via CUE GMM.
The asymptotic variance of the estimator of $\hat{H}_{12}$ computed using an arbitrarily long time series to estimate the moment covariance matrix at the true value; the asymptote to zero around a local alternative of 1.5 may explain the spike in rejection rate visible in the power curves, Figures 3 and 4.

4 Tests of weak Identification

Tests for weak identification pose a challenge in general GMM settings. It is straightforward to reparametrize the GMM framework to afford tests of identification (i.e. $\delta = 0$); however, the suitable null is rather the presence of weak identification ($\delta$ close enough to 0 to cause problems in estimation). The simple case coincides with just-identified linear IV with a single endogenous variable, and a rule of thumb based on bias is provided based on the results of Staiger & Stock (1997) and Stock & Yogo (2005), specialized to a setting with heteroskedasticity. For the general case, I recommend the approach of Andrews (2017). Lütkepohl & Milunovich (2016) do offer tests of identification in systems identified via heteroskedasticity, but their work mainly considers GARCH-type models and makes distributional assumptions to allow maximum likelihood methods.

4.1 Single-variance change rule of thumb

The analogy to the just-identified linear IV case with a single endogenous regressor is no doubt an appealing feature of simpler systems of equations, and has been employed repeatedly in the literature. With that simplicity comes a familiar test. However, it must be noted that using this model fails to utilize all of the available moment conditions, and may pose misspecification problems if the assumption that other variances do not change is faulty (as discussed in the empirical application). Given the apparent preference for the simplified model in applied work, I discuss it anyway. This setting suggests utilizing the approach of Staiger & Stock (1997) and Stock & Yogo (2005), which yields the oft-cited “$F > 10$” rule.
Table 4: Critical values for first-stage $F$-test based on TSLS bias (5% significance level)

<table>
<thead>
<tr>
<th>Bias</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Value</td>
<td>37.42</td>
<td>23.11</td>
<td>15.06</td>
<td>12.05</td>
</tr>
</tbody>
</table>

Critical values for the first-stage $F$-statistic from Montiel Olea & Pflueger (2013); estimates deliver bias less than or equal to that listed in 95% of samples.

of thumb. The first-stage $F$-statistic is based on the first stage regression,

$$
\eta_{2t} = \Pi Z_t + \nu_t
$$

$$
= \Pi \left[ \mathbf{1}(t \in T_P) \times \frac{T}{T_P} - \mathbf{1}(t \in T_C) \times \frac{T}{T_C} \right] \eta_{2t} + \nu_t
$$

Stock & Yogo (2005) derive expressions for the worst-case asymptotic bias (relative to OLS) and size for a given degree of identification governed by $\Pi$, but assume homoskedasticity. Fortunately, Montiel Olea & Pflueger (2013) develop a test that is valid under the presence of arbitrary heteroskedasticity, autocorrelation, or clustering. They use a "scaled $F$-statistic", which, in this simplest setting, reduces to the standard $F$-statistic. They compare the Nagar bias of a TSLS estimator to a "worst case" benchmark. The Nagar bias approximates the distribution of the TSLS estimator under weak identification using a second-order Taylor approximation and computes the bias of that distribution. The benchmark coincides with the OLS comparison of Staiger & Stock (1997) and Stock & Yogo (2005) under those papers’ distributional assumptions. Table 4 presents the critical values for the first-stage weak instruments test based on TSLS bias for the 5% significance level from Montiel Olea & Pflueger (2013), analogous to Table 5.1 in Stock & Yogo (2005). Additional simulations, not reported here, based on an identification via heteroskedasticity DGP, produced critical values within Monte Carlo error of those of Montiel Olea & Pflueger. The size exercise of Stock & Yogo (2005) is not pursued as this setting makes it impossible to perform their controlled comparison since the worst-case size depends on the covariance between first and second stage residuals. This covariance depends mechanically on the degree of identification, resulting in non-monotonicity in the relationship between the degree of identification and worst-case size, rendering any threshold rule inappropriate. The generally-adopted threshold of $F > 10$ corresponds to the critical value for relative bias of 10%, which here translates to a new rule of thumb of $F > 23$, which can easily be adopted in future work in this simple setting.

4.2 Method for the general case

In the more general case, more complex methods are required. Tests of the null of non-identification are straightforward, as noted in Wright (2001), but testing of weak iden-
tification in general GMM has long been a problem. However, Andrews (2017) offers a two-step strategy, which can be adopted here. The necessary assumptions for its application are discussed and verified in the Appendix. Briefly, given a user-specified maximum allowable size distortion, $\xi$, a preliminary robust confidence set is constructed to have asymptotic coverage $1 - \nu - \xi$, regardless of identification. Then, $1 - \nu$ non-robust and robust confidence sets are constructed. If the preliminary confidence set is contained by the non-robust confidence set, the non-robust confidence set is adopted (strong identification); otherwise, the robust set should be used (weak identification). Asymptotically, the probability of making the correct determination converges to unity. These confidence sets can be constructed using $K$-statistics and Wald tests as elsewhere in this paper; Andrews, however, considers a linear combination of the $K$- and $S$-statistics due to favourable power properties. The interested reader should review the original paper for a more thorough discussion of the test’s properties and the choice of statistics to be used in its construction. Determination of the strength of identification can be conducted with respect to the full parameter vector or a subset of parameters of interest. These tests are applied in the empirical application.

5 Robust inference on impulse responses

The ability to perform robust inference on $H$ also enables inference on another object of interest, the impulse response function (IRF). $H$ is not the ultimate object of interest in many applications of identification via heteroskedasticity. Indeed, much macroeconomic policy analysis relies on the IRFs generated by SVARs. These are formed via non-linear combinations of the lag coefficients, $A(L)$ and $H$. The literature on this topic is nascent. Montiel Olea, Stock & Watson (2016) develop a method that exploits the linearity of the external instruments problem to offer an elegant solution in that context. Chevillon, Mavroeidis, & Zhan (2016) consider the case of long-run restrictions and offer a projection-based method that also accounts for cointegration issues they face. Neither of those methods can be employed for identification via heteroskedasticity. What follows is a general method that can be used across SVAR identification schemes with simple modifications, although it cannot immediately address the cointegration issues faced by Chevillon, Mavroeidis, & Zhan (2016).

Using a test-inversion approach, it is possible to compute confidence sets for IRFs. A structural impulse response function at horizon $h$ is computed as

$$\Lambda^0 = H,$$

$$\Lambda^h = \left[ \sum_{v=1}^{h} \Lambda^{h-v} A_v \right] \quad h = 1, 2, \ldots,$$  \(6\)
where \( A_v \) denotes the lag coefficient matrix corresponding to the \( v \)th lag. The response of the \( j \)th to a unit shock to \( k \) can be read off as the \( jk \) element of this object. It is helpful to define the related object, \( B^h \) given by

\[
B^0 = I,
\]

\[
B^h = \sum_{v=1}^{h} B^{h-v} A_v, \quad h = 1, 2, \ldots,
\]

such that \( \Lambda^h = B^h H \). Thus, \( B^h = B^h (A(L)) \), entirely a function of the lag coefficients. An element of interest in \( \Lambda^h \), \( \Lambda^h_{ij} \), is the product of the \( i \)th row of \( B^h \) and the \( j \)th column of \( H \).

I now present conditions under which the remainder of \( \Lambda^h \) is strongly identified conditional on \( \Lambda^h_{ij} \). First consider an analog to the partial identification result of Proposition 3:

**Proposition 4.** If \( B^h \) is invertible, for \( \Lambda^h \) partitioned \( \Lambda^h_A : \Lambda^h_B \), \( \Lambda^h_A \) is identified up to scale from the covariance matrices provided no row in the upper block of

\[
\begin{bmatrix}
\text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\
\tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2}))
\end{bmatrix}
\]

is proportional to another row.

**Proof.** See Appendix.

This follows the same argument as Proposition 3, simply replacing \( H \) with \( B^h H \), with invertibility of the product guaranteed by invertibility of \( B^h \) and \( H \). It says that IRFs to shocks with non-proportional variance processes are strongly identified. This leads to a stronger result.

**Theorem 4.** For \( \Lambda^h \) partitioned \( \Lambda^h_A : \Lambda^h_B \), where \( \Lambda^h_B \) contains two columns, \( \Lambda^h_A \) is identified up to scale from the covariance matrices provided

1. \( B^h \) is invertible,
2. No row in the upper block of

\[
\begin{bmatrix}
\text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\
\tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2}))
\end{bmatrix}
\]

is proportional to another row,
3. A single element \( \Lambda^h_{ij} \) is fixed and \( \Lambda^h_{jk} \neq 1/\Lambda^h_{ij} \) for \( \Lambda^h(j), \Lambda^h(k) \in \Lambda^h_B \).

**Proof.** See Appendix.

This is an extension of Theorem 3, again simply replacing \( H \) with \( \Lambda^h = B^h H \). Identification up to scale may be unfamiliar for IRFs, but recall that IRFs to each shock are always implicitly scaled by the normalization of the corresponding column of \( H \). Theorem 4
implies that, fixing an element of interest in \( \Lambda^h \), subject to the same remarks made following Theorem 3, an analog to Assumption 4 is met, and inference can proceed using \( \chi^2_1 \) critical values, as shown in Theorem 2. This is standard test inversion problem.

However, it is worth noting how this result can be applied computationally efficiently in practice. The following discussion assumes Theorem 4 applies. Recall that, asymptotically, moments identifying reduced form VAR coefficients \( (A(L)) \) and those identifying \( H \) have block-diagonal covariance, see e.g. Lütkepohl (2006). This fact will be useful, but some work is required first. Define a moment equation \( f_t^A (A, \Lambda^h, \Sigma_C, \Sigma_p) \) such that it stacks the moments yielding the reduced form VAR coefficients, \( f_t^A (A) \), and those pertaining to \( \Lambda^h \),

\[
\begin{bmatrix}
  \Lambda^h \\
  \Lambda^h
\end{bmatrix}
\end{bmatrix} = \begin{bmatrix}
  \eta_1 \eta_1' - H \Sigma_C H' B^h \\
  \eta_1 \eta_1' - H \Sigma_P H' B^h
\end{bmatrix}.
\]

Conditional on some \( \Lambda^h_{ij} \), the remainder of the parameters can be strongly identified from these moments by Theorem 4. Robust inference for \( \Lambda^h_{ij} \) can then proceed using \( \chi^2_1 \) critical values and the efficiently-weighted GMM objective function based on \( f_t^A (\cdot) \), denoted \( S_T^A (\cdot) \).

Note that \( f_t^A \) can be rewritten as

\[
vec \begin{bmatrix}
  \Lambda^h \\
  \Lambda^h
\end{bmatrix}
\end{bmatrix} = \begin{bmatrix}
  \eta_1 \eta_1' - H \Sigma_C H' B^h \\
  \eta_1 \eta_1' - H \Sigma_P H' B^h
\end{bmatrix}.
\]

since all moments are just scaled by \( B^h \). Because the efficient weighting matrix will be used, \( B^h \) can be ignored, as it is applied as a constant matrix and will drop out once the moments are re-weighted by the inverse of the moment covariance. In particular, for

\[
f_t^H (H, \Sigma_C, \Sigma_P) = \begin{bmatrix}
  \eta_1 \eta_1' - H \Sigma_C H' \\
  \eta_1 \eta_1' - H \Sigma_P H'
\end{bmatrix},
\]

\[
f_t^H (H, \Sigma_C, \Sigma_P) \Omega_{fH}^{-1/2} = f_t^A (A, \Lambda^h, \Sigma_C, \Sigma_P) \Omega_{fA}^{-1/2},
\]

where \( \Omega_{fH}^{-1/2} \) and \( \Omega_{fA}^{-1/2} \) are the inverse square roots of the relevant moment covariances. Denote the GMM objective function corresponding to \( f_t^{\text{block}} (A, H, \Sigma_C, \Sigma_P) = \begin{bmatrix}
  f_t^A (A) \\
  f_t^H (H, \Sigma_C, \Sigma_P) \Omega_{fH}^{-1/2}
\end{bmatrix} \) as \( S_T^{\text{block}} (A, H, \Sigma_C, \Sigma_P) \). Then

\[
S_T^A (A, \Lambda^h, \Sigma_C, \Sigma_P) = S_T^{\text{block}} (A, H, \Sigma_C, \Sigma_P),
\]
where $\Lambda^h = B^h (A) H^h$. By transforming the moment functions to separate the reduced-form and structural parameters, as is the case in $f^\text{block}_t (\cdot)$, the asymptotic block diagonality of the covariance matrix of such moments can be exploited. Asymptotically, using the fact that the covariance of $f^A_t (\cdot)$ and $f^H_t (\cdot)$ is block diagonal, $S_T^{\text{block}}$ can be decomposed as

$$S_T^{\text{block}} (A, H, \Sigma_C, \Sigma_P) = S_T^A (A) + S_T^H (H, \Sigma_C, \Sigma_P).$$

Finally, combining results, this implies

$$S_T^A (A) + S_T^H (H, \Sigma_C, \Sigma_P) = S_T^\Lambda (A, \Lambda^h, \Sigma_C, \Sigma_P),$$

asymptotically.

Armed with this decomposition, it is relatively straightforward to construct a confidence set for $\Lambda^h_{ij}$. From (7), $\lambda = \Lambda^h_{ij}$ can be in a $1 - \nu$ confidence set from the right hand side if and only if there is a pair $A, H$ in the $1 - \nu$ joint confidence set from the left hand side such that $[B^h (A) H]_{ij} = \lambda$. Thus, computing the values of $\lambda$ implied by the values of $A$ and $H$ in the $1 - \nu$ confidence joint confidence set implies a $1 - \nu$ set for $\Lambda^h_{ij}$. As noted above, concentrated $S-$statistics for $\Lambda^h_{ij}$ based on $S_T^A$ can be compared to $\chi_1^2$ critical values; such values can thus also be used for the left hand side. A confidence set can thus be constructed as follows:

1. Construct the $1 - \nu$ Wald set for $A$ using the $\chi_1^2 (1 - \nu)$ critical value. Since $A$ is strongly identified, this is asymptotically equivalent to the set of values such that $S_T^A (A) \leq \chi_1^2 (1 - \nu)$.

2. Construct the $1 - \nu$ concentrated robust confidence set for $H^{(j)}$, using the $\chi_1^2 (1 - \nu)$ critical value using $S_T^H (H^{(j)})$, even if there is more than one free parameter in $H^{(j)}$.

3. For each element $\tilde{A}^{(n)}$ of the first set, and each element $\tilde{H}^{(m)}$ of the second, compute the implied value $\tilde{\lambda}^{(nm)}$ using $[B^h (A) H]_{ij} = \lambda$.

4. If $S_T^A (\tilde{A}^{(n)}) + S_T^H (\tilde{H}^{(m)}) \leq \chi_1^2 (1 - \nu)$, then $\tilde{\lambda}^{(nm)}$ is in the $1 - \nu$ confidence set for $\Lambda^h_{ij}$.

The resulting set need not be connected, a common feature in weak identification settings. The test will be dominated by the confidence set for $H$, over which greater uncertainty almost always rests. The critical values used have 1 degree of freedom, equivalent to the

---

6 This mapping is unique since both $B^h$ and $H$ have been assumed to be invertible.

7 Steps 1 and 2, constructing preliminary confidence sets, are not strictly necessary, but rather allow the researcher to construct relevant sub-grids to focus the search for combinations of $A, H$ likely to be within the joint confidence set, before working in a higher-dimensional space for possible values of $\Lambda^h_{ij}$.
power of the method of Montiel Olea, Stock, & Watson (2016). These values differ from those of Chevillon, Mavroeidis, & Zhan (2016) due to their need to accommodate cointegration issues.

6 Empirical application

To demonstrate the use of these results in practice and the extent to which weak identification can impact applied work, I re-examine the data of Nakamura & Steinsson (2017). They analyze the impact of policy shocks on nominal and real treasury forward yields of varying maturities. They perform Rigobon-style analysis as a robustness check on their main results. Their analysis uses two different policy shock series: one is the one-day nominal 2-year treasury yields, and the other is the 30-minute change in a “policy news”, which series they construct from several macro variables using Principle Components Analysis. They use weak identification robust techniques, and express concern about weak identification problems for the response to 2-year nominal treasury yields, using a 1-day window; they believe that the 30-min changes behave as a strong instrument. Their main approach “regresses” 1-day changes of the forward rates on announcement days on the 1-day change in 2-year nominal yields. In the Rigobon-style analysis, they use announcement days (Tuesdays or Wednesdays) as the high-variance regime, and a sample of other Tuesdays and Wednesdays as the control period, or “low variance” sample. For more detail on the dataset, consult the original paper. For comparison, I also examine a specification using the authors’ constructed 30-minute window “policy news shocks”.

The data considered here is restricted to a subset of that considered by Nakamura & Steinsson. The dataset was reconstructed using data on 2-year nominal yields and 2-year nominal and real forward rates; it spans 2004 to 2014, shorter than the original paper, since the forward data is only publicly available back to 2004. Note that Nakamura & Steinsson (2017) analyze 5- and 10-year forwards in addition, but only the shortest maturity is selected for the current exercise. Their analysis takes observed changes in the interest rates to be reduced form residuals. This means that for the purpose of applying an IRF methodology, modifications must be made, as discussed below.

6.1 Specification & identification

Nakamura & Steinsson (2017) restrict only the variance of policy shocks to change on

---

8I am very grateful to Emi Nakamura and Jón Steinsson for making their policy news series available to me.

9They actually consider an additional instrument, one-day changes in the “policy news” series, but that is omitted here to permit a simple comparison between the two included policy measures.
announcement days. This places the analysis in the simple two-variable, one shock setting, with analogy to just-identified linear IV with a single endogenous regressor. However, this paper proposes a generalized estimation framework, making use of all moments, and allowing for the possibility that the variances of both structural shocks might change. Economically, it might be the case that only the variance of the policy shock should change, but if that is the case, the restriction need not be imposed mechanically, as the estimation will bear it out.

Table 5 reports estimates for this unrestricted model. For the 30 min policy news shock, the results are extremely close to those reported by Nakamura & Steinsson for their restricted model (on a slightly longer dataset); compare 0.94 to 0.96 and 1.13 to 1.10. This suggests that for a policy shock series that incorporates several dimensions of policy news and is calculated over a short window, the assumption that the policy series captures almost all of the heteroskedasticity in the true policy shocks is valid, and thus the single variance change assumption is innocuous. Further evidence of this is the fact that the policy shock variances are estimated to be the same regardless of the dependent variable, and the estimated shocks themselves are similar. On the other hand, the results for the 1-day change in the nominal yield are starkly different. The values of the $H$ parameters make less economic sense; the sizable values of $H_{21}$ are highly surprising, as are their opposite signs for nominal and real series. There should be no instantaneous impact on policy rates from treasuries in daily data. The lower pass-through to nominal futures is also at odds with the other results. Rather than suggesting that there is something wrong with the unrestricted estimation procedure, this is indicative of weak identification.

To assess these fears more rigorously, I apply the proposed tests of identification; they bear out suspicions noted in the original paper. Because this model can be cast in the two variable, one variance change just-identified IV setting, as it is by Nakamura & Steinsson, the first-stage $F$—statistic critical values can be used, as if the researcher intended to estimate that restricted model. These results are reported in the first panel of Table 6. For the 1-day nominal yield policy series, the first-stage $F$—statistic is very low compared to all of the critical values computed, and thus there is evidence of weak identification based on this bias metric. For the 30 minute policy news shock, the first stage $F$—statistic is large and surpasses all critical values. Applying the general 2-step size-based test without making an assumption on variance changes, reported in the second panel, shows similar results. The 1-day nominal yield policy variable gives weak identification for each level of acceptable distortion, for both dependent variables. For the 30 min policy news shock, there is evidence of weak identification only at the 5% distortion threshold, for both dependent variables, suggesting fairly strong identification. These results comport well with intuition, Nakamura
Table 5: Estimates

<table>
<thead>
<tr>
<th></th>
<th>Real, yield</th>
<th>Nominal, yield</th>
<th>Real, policy news</th>
<th>Nominal, policy news</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{12}$</td>
<td>1.10</td>
<td>0.38</td>
<td>0.94</td>
<td>1.13</td>
</tr>
<tr>
<td>$H_{21}$</td>
<td>-0.39</td>
<td>0.65</td>
<td>-0.0015</td>
<td>0.0003</td>
</tr>
<tr>
<td>$\sigma^2_{s,C}$</td>
<td>0.0025</td>
<td>0.0042</td>
<td>0.0034</td>
<td>0.0042</td>
</tr>
<tr>
<td>$\sigma^2_{i,C}$</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>$\sigma^2_{s,P}$</td>
<td>0.0044</td>
<td>0.0065</td>
<td>0.0051</td>
<td>0.0056</td>
</tr>
<tr>
<td>$\sigma^2_{i,P}$</td>
<td>0.0002</td>
<td>0.0011</td>
<td>0.0008</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

GMM CUE estimates allowing for changes in all variances, using 1-day changes in 2-year real or nominal forward rates as the “dependent” variable and 1-day changes in 2-year nominal yield or 30-minute changes in Nakamura & Steinsson’s policy news series as the “independent” variable.

Table 6: Tests of Identification

<table>
<thead>
<tr>
<th></th>
<th>First-stage $F$ (bias)</th>
<th>Andrews 2-step (size)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$ 0.2 0.1 0.05</td>
<td>0.2 0.1 0.05</td>
</tr>
<tr>
<td>Real, 1-day shock</td>
<td>5.77 × × ×</td>
<td>× × ×</td>
</tr>
<tr>
<td>Nominal, 1-day shock</td>
<td>× × ×</td>
<td>× × ×</td>
</tr>
<tr>
<td>Real, 30-min shock</td>
<td>1073.50 ✓ ✓ ✓</td>
<td>✓ ✓ ×</td>
</tr>
<tr>
<td>Nominal, 30-min shock</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ×</td>
</tr>
</tbody>
</table>

The first panel tests each shock series using the first-stage $F$–statistic bias-based critical values in Table 4. The second panel conducts the Andrews 2-step size test for each specification.

& Steinsson’s findings, and the discussion of estimates above.

6.2 Performance of tests

The simulation evidence presented previously offers insight into the performance of tests in this empirical context. Nakamura & Steinsson’s AR-type test for $H_{12}$ yields very wide confidence sets for the Rigobon procedure. Although they do not compare inference techniques, this is indicative that standard methods likely perform poorly. The Monte Carlo results of Section 3 establish that this is the case, using a calibration based on the 1-day window nominal yield and nominal forward specification. As shown in Table 2, a Wald test of the full parameter vector is found to have size tending to unity, and distortions are prohibitive even for an order of magnitude larger variance change than is present in the data and empirical sample size. Even in the best-case calibration considered, the Wald test is unambiguously outperformed by the robust tests considered. Turning to a subset of parameters, Table 3 shows that a $t$–test on $H_{12}$ yields size distortions between 5 and 15% for all but the strongest identified calibrations, while the robust tests perform well. These results imply that, empirically, there may be substantial returns to adopting robust testing procedures in terms of size distortion. While there is power loss compared to non-robust methods, the tests proposed are more powerful than those resulting from any projection-type argument.
6.3 Application to IRFs

The application to monetary policy shocks provides an opportunity to assess the performance of the proposed method for computing IRF confidence sets and to demonstrate the difference in performance compared to conventional techniques in the presence of weak identification. However, the precise setting considered by Nakamura & Steinsson follows an event study approach, and thus does not directly permit the computation of an IRF, as it considers disjoint series of days. For the purpose of this exercise, I adopt a new framework in the spirit of that paper. Instead of considering a sample of control and event days, I consider the full time series for the daily yield on 2-year nominal treasuries and daily yield on 2-year nominal treasury futures of all trading days between January 2004 and November 2015. I then estimate a VAR(1) on this data.

\[ y_t = \begin{pmatrix} s_t \\ i_t \end{pmatrix} = A_1 y_{t-1} + \eta_t \]

Now, instead of being raw 1-day changes in the variables, the residuals are the components of the time series not predictable by prior data. Estimation yields

\[ \hat{A}_1 = \begin{pmatrix} -0.078 & -0.086 \\ 0.079 & 0.064 \end{pmatrix}. \]

The point estimates are computed using GMM. Note that, unsurprisingly, given the presence of weak identification, the estimated value of \( \hat{H} \) is dramatically different to that on the main dataset. Now, the value of \( \hat{H}_{12} \) is 2.37. The test-inversion procedure proposed is then applied. The block diagonal structure of the moment covariance is exploited to form separate preliminary confidence sets for \( A \) and \( H \). Combining the gridded values and computing the \( S \)-statistic yields a set of response points that cannot be rejected by the test at the 95% level.

For comparison, two other confidence sets are displayed. The first is a standard Wald 95% confidence set using GMM, the multivariate delta method, and standard \( \chi^2 \) critical values. The second is a 95% confidence set computed using a conventional block bootstrapping procedure with block length equal to sixty days.

Figure 6 presents the results. The confidence sets get exponentially narrower at further horizons, in keeping with standard behaviour of VAR IRFs, as the \( B^h \) component in \( \Lambda^h \) tends towards a very precisely-estimated zero. The bootstrap set captures the asymmetry of the problem, but is virtually contained within the much wider robust set. These results show that inference based on non-robust confidence sets is likely to be misleading and exhibit substantial size distortions. The proposed method for computing robust confidence sets
appears to be advisable when a researcher is concerned the data may suffer from weak identification. In this example, using the standard confidence sets would be highly misleading as to the precision of the result. As noted above, this method will carry over to weak identification issues present in other SVAR identification schemes by merely changing the moment equations for $H$ accordingly. In an economic sense, when identification may be weak, non-robust inference makes overstating the significance of a policy impact likely.

7 Conclusion

This paper provides a comprehensive framework for practitioners to estimate and conduct inference on models identified via heteroskedasticity, in particular SVARs. It presents a conception of how weak identification may arise in this context and models the deficiency through which that weak identification arises. Simulations demonstrate that such identification issues significantly impact the performance of standard inference techniques, leading to substantial size distortions. The problem is presented in a standard GMM framework, and asymptotic distributions for test statistics are derived.

I establish the validity of non-conservative subset inference based on $K$-type statistic, which does not suffer from the size distortions of $F$- or Wald inference and is more powerful than projection methods. Measures are proposed to test for the presence of weak identification. The final theoretical contribution is to present a general method to construct confidence sets for impulse response functions in the presence of weak identification. This technique has applications beyond the present paper, throughout identification schemes for SVARs that may be at risk for weak identification.

These contributions are put through their paces on a subset of the data analyzed using the heteroskedasticity approach by Nakamura & Steinsson (2017). This exercise exposes the extent to which the judicious choice of a policy shock series can impact results; choosing a longer window and series encompassing less information yields weaker identification and may lead to misspecification. It suggests that for many of the ad hoc macroeconomic shock series that applied papers may use, standard inference may be threatened by identification problems. Under this weak identification, the new methods provide superior size control in a variety of contexts. Comparison of the new technique for computing robust confidence sets for IRFs to Wald and bootstrap techniques highlights substantial size distortions present in standard techniques when identification is weak as well as absent non-linearities.

Weak identification is a concern that should be taken seriously by those using heteroskedasticity assumptions to identify simultaneous equations models. This approach has considerable traction across many fields of economics, and beyond. The empirical exercise demonstrates its presence in relevant data as well as the need for the general, more flexible
Figure 6: IRF Confidence Sets

Response (in b.p.) of 2-year nominal forwards on Treasuries to a 1 b.p. monetary policy shock. The IRF path is computed based on the horizon-by-horizon IRF estimator (computing based on $H_{12}$ and $A_1$ estimates from the $0^{th}$ horizons results in negligible differences). A Wald 95% confidence set is plotted, along with a bootstrap confidence interval following a block bootstrap. The collapse of the confidence sets towards zero with the point estimate is due to the precision with which the near-zero $A_{1}^{h}$ values are estimated.
estimation framework proposed. Fortunately, with the proposed testing procedures, the need for robust inference should not be a deterrent to practitioners.
8 Appendix

Proof of Proposition 1

**Proposition 1.** \(H\) is globally identified from \(\Sigma_{\eta_1}\) and \(\Sigma_{\eta_2}\) up to column order provided the rows of \(\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}\) are not proportional.

**Proof.** This is a restatement of Sentana & Fiorentini (2001) Proposition 3, with linear independence replaced by proportionality due to the presence of only two regimes. The underlying condition needed by Sentana & Fiorentini is that the matrix \(\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}\) is of full rank. In their setting, where they in effect assume the number of regimes is greater than \(n\), this translates into linear independence. Here, with only two regimes, the matrix of variances is generally “tall” and can be of full column rank even if some rows are linearly dependent. Thus, the “no proportionality” condition is required, as without it, only the column space of the corresponding columns of \(H\) would be identified, not the individual columns. To see this, consider Sims’ (2014) argument that columns of \(H\) are identified as the right eigenvectors of \(\Sigma_{\eta_1}\Sigma_{\eta_2}^{-1}\) provided they correspond to unique elements (eigenvalues) of the matrix \(\Sigma_{\varepsilon_1}\Sigma_{\varepsilon_2}^{-1}\). Repeated eigenvalues mean that the corresponding eigenvectors are not uniquely determined, instead spanning the columnspace of corresponding elements of \(H\). Repeated eigenvalues occur if and only if rows of the matrix \(\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}\) are proportional. Since \(n \geq 2\), non-proportional rows imply that the matrix \(\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}\) is of full rank, 2, the weaker condition of Sentana & Fiorentini. □

Proof of Proposition 2

**Proposition 2.** Adopting the modeling device in (4), \(H\) is asymptotically unidentified.

**Proof.** I model the variance deficiency as

\[
\frac{\sigma_{\varepsilon_1P}^2/\sigma_{\varepsilon_1C}^2}{\sigma_{\varepsilon_2P}^2/\sigma_{\varepsilon_2C}^2} = 1 + \frac{d}{\sqrt{T}}.
\]

Under this device, the row of \(\begin{bmatrix} \Sigma_{\varepsilon_1} & \Sigma_{\varepsilon_2} \end{bmatrix}\) corresponding to \(\varepsilon_P\) is equal to \(\begin{bmatrix} \sigma_{\varepsilon_1P}^2 & \sigma_{\varepsilon_1P}\sigma_{\varepsilon_2C}^2 (1 + d/T^{1/2}) \end{bmatrix}\). In the limit, for finite \(d\), this equals \(\begin{bmatrix} \sigma_{\varepsilon_1P}^2 & \sigma_{\varepsilon_1P}\sigma_{\varepsilon_2C}^2 \end{bmatrix}\). However, this is \(\sigma_{\varepsilon_1P}^2/\sigma_{\varepsilon_1C}^2\) times \(\begin{bmatrix} \sigma_{\varepsilon_1C}^2 & \sigma_{\varepsilon_2C}^2 \end{bmatrix}\), the row corresponding to \(\varepsilon_C\), so the condition of Proposition 1 is violated, and only the columnspace of \(H\) is identified. □
Proof of Lemma 1

Lemma 1. Under Assumptions 1 & 3,

\[ \psi_T(\theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \bar{\phi}_t(\theta_0) \right) \xrightarrow{d} \begin{pmatrix} \psi_{\phi} \\ \psi_{\theta_0} \end{pmatrix} \]

where \( \psi = \begin{pmatrix} \psi_{\phi} \\ \psi_{\theta_0} \end{pmatrix} \) is a \( 2(n^2 + n) \)-dimensional normally distributed random variable with mean zero and positive semi-definite \( 2(n^2 + n) \times 2(n^2 + n) \)-dimensional covariance matrix

\[ \Omega(\theta) = V \begin{pmatrix} \Omega_{\phi \phi}(\theta) & \Omega_{\phi \theta}(\theta) \\ \Omega_{\theta \phi}(\theta) & \Omega_{\theta \theta}(\theta) \end{pmatrix} = \lim_{T \to \infty} \text{var} \left[ \frac{1}{\sqrt{T}} \begin{pmatrix} \phi_T(\theta) \\ q_T(\theta) \end{pmatrix} \right] \]

with \( \phi_T(\theta) = \sum_{t} \phi_t(\theta) \).

Proof. First, I prove that \( \psi_{\phi T}(\theta) = T^{-1/2} \sum_{t=1}^{T} \bar{\phi}_t(\theta) \) obeys a central limit theorem. For simplicity the proof assumes \( T_P = T_C = T/2 \), but this can trivially be generalized, as long as Assumption 3.2 is met. By Assumptions 1.1-1.2,

\[ \psi_{\phi T}(\theta) = T^{-1/2} \sum_{t=1}^{T} \begin{bmatrix} 1 [t \leq T/2] (\text{vech} (\eta_t \eta_t') - \text{vech} (H \Sigma_{\varepsilon,1} H')) \\ 1 [t > T/2] (\text{vech} (\eta_t \eta_t') - \text{vech} (H \Sigma_{\varepsilon,2} H')) \end{bmatrix} - E \begin{bmatrix} 1 [t \leq T/2] (\text{vech} (E [\eta_t \eta_t'] | t \leq T/2) - \text{vech} (H \Sigma_{\varepsilon,1} H')) \\ 1 [t > T/2] (\text{vech} (E [\eta_t \eta_t'] | t > T/2) - \text{vech} (H \Sigma_{\varepsilon,2} H')) \end{bmatrix} \]

\[ = T^{-1/2} \sum_{t=1}^{T} \begin{bmatrix} 1 [t \leq T/2] (\text{vech} (\eta_t \eta_t') - \text{vech} (E [\eta_t \eta_t'] | t \leq T/2)) \\ 1 [t > T/2] (\text{vech} (\eta_t \eta_t') - \text{vech} (E [\eta_t \eta_t'] | t > T/2)) \end{bmatrix} \]

\[ = 2^{1/2} \left( (T/2)^{-1/2} \sum_{t=1}^{T/2} \text{vech} (\eta_t \eta_t') - \text{vech} (E [\eta_t \eta_t'] | t \leq T/2) \right) \]

Each of the summations obeys a standard multivariate CLT with the addition of Assumptions 1.3 and 3.1. Thus, the entire expression obeys a CLT as a linear combination of random variables each obeying a CLT, for any \( \theta \in \Theta \) (the expression does not depend on the particular
value of $\theta$). Thus

$$
\psi_{\theta T}(\theta_0) \xrightarrow{d} N(0, V_{\phi\phi}(\theta_0))
$$

By definition, $\bar{q}_t(\cdot) = 0$ deterministically; note that $\frac{\partial \phi(\theta)}{\partial \theta'} = -\frac{\partial (vech(\Sigma_{\eta C})'vech(\Sigma_{\eta P}))'}{\partial \theta'} = E\left[\frac{\partial \phi(\theta)}{\partial \theta'}\right]$ since it contains only population parameters and no data (the moment equations are separable in data and parameters). This holds for any $\theta \in \Theta$; $\theta$ need not equal $\theta_0$. Thus $\psi_{\theta}$ is a degenerate random variable. That $V(\theta)$ takes the form given follows by construction. It remains to show that $V(\theta)$ is positive semi-definite. Since all but the top left block will be zeros, it suffices to show that $V_{\phi\phi}(\theta)$ is positive semi-definite. This follows as $V_{\phi\phi}(\theta_0)$ has the form $E[M M']$.

\[\square\]

**Proof of Lemma 2**

**Lemma 2.** Under Assumptions 1.3 & 3.1, the covariance matrix estimator $\hat{V}(\theta_0)$ satisfies

$$
\hat{V}(\theta_0) \overset{p}{\rightarrow} V(\theta_0)
$$

and

$$
\frac{\partial vec\left(\hat{V}_{\phi\phi}(\theta_0)\right)}{\partial \theta'} \overset{p}{\rightarrow} \frac{\partial vec\left(V_{\phi\phi}(\theta_0)\right)}{\partial \theta'}.
$$

Proof. By a standard Law of Large Numbers and Assumptions 1.3 and 3.1, the natural covariance estimator, is consistent, $\frac{1}{T} \sum_{t=1}^{T} \phi_t(\theta_0) \phi_t(\theta_0)' \overset{p}{\rightarrow} E\left[\phi_t(\theta_0) \phi_t(\theta_0)\right]$. Then

$$
V(\theta_0) = \lim_{T \to \infty} var\left[\frac{1}{\sqrt{T}} \phi_T(\theta_0)\right] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} var(\phi_t(\theta_0)),
$$

again by Assumption 1.3, which simplifies to $E\left[\phi_t(\theta_0) \phi_t(\theta_0)\right]$. Since $q_t$ is deterministic, this establishes the first part of the Lemma.

For the second part, note that $\frac{\partial \hat{V}_{\phi\phi}(\theta_0)}{\partial \theta'} = \frac{\partial}{\partial \theta'} \left[\frac{1}{T} \sum_{t=1}^{T} \phi_t(\theta_0) \phi_t(\theta_0)\right] = E\left[\frac{\partial \phi_t(\theta_0)}{\partial \theta'} \phi_t(\theta_0)\right]$. $\frac{\partial \phi_t(\theta_0)}{\partial \theta'}$ is a matrix of zeros, ones, and continuous functions of elements of $\theta$; it is entirely deterministic. Similarly, $\frac{\partial \hat{V}_{\phi\phi}(\theta_0)}{\partial \theta'} = E\left[\frac{\partial \phi_t(\theta_0)}{\partial \theta'} \phi_t(\theta_0)\right] = \frac{\partial \phi_t(\theta_0)}{\partial \theta'} E\left[\phi_t(\theta_0)\right]$, and since $E\left[\phi_t(\theta_0)\right]$ is consistently estimated, so too is $\frac{\partial \hat{V}_{\phi\phi}(\theta_0)}{\partial \theta'}$ by Slutsky’s Theorem.

\[\square\]
Proof of Theorem 1

Theorem 1.

\[ K(\theta_0) \xrightarrow{d} \chi^2_{n^2+n} \]

Proof. The result follows directly from Kleibergen (2005). Lemmas 1 and 2 establish Assumptions 1 and 2 from that paper, which are used to prove Theorem 1 therein. Note that Lemma 1 and part of Lemma 2 establish the required conditions of Stock & Wright (2001) Theorem 2 (their Assumption A and the consistency of the covariance matrix for the weighting matrix) so

\[ S_T(\theta_0) \xrightarrow{d} \chi^2_{n^2+n} \]

as an immediate corollary.

Proof of Theorem 3

Theorem 3. For \( H \) partitioned \( H_A; H_B \), where \( H_B \) contains at most two columns, \( H \) is identified from the covariance matrices provided no row in the upper block of \( \begin{bmatrix} \text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\ \tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2})) \end{bmatrix} \) is proportional to another row if

1. A single element \( H_{ij} \) is fixed and \( H_{jk} \neq 1/H_{ij} \) for \( H^{(j)}, H^{(k)} \in H_B \), or

2. The full column \( H^{(j)} \in H_B \) is fixed.

Proof. The proof follows from extending the proof of Proposition 4 in Sentana & Fiorentini (2001). They show that for a similarly partitioned \( H \), the columns of \( H_A \) are identified. However, the columns of \( H_B \) are identified only up to an orthogonal rotation \( Q, QQ' = Q'Q = I \). \( H_B \) represents the portion of \( H \) pertaining to proportional variance processes, and as such cannot contain just a single column. If \( H_B \) contains two columns, then \( Q \) is \( 2 \times 2 \). Consider first a single fixed element of \( H^{(j)} \), the subject of the null hypothesis for the subset test. Without loss of generality, let it be \( H_{2j} = x \). This yields the system of equations

\[
\begin{bmatrix}
1 & H_{1k} \\
\vdots & \vdots \\
H_{nj} & H_{nk}
\end{bmatrix}
\begin{bmatrix}
Q_{11} & Q_{12} \\
Q_{21} & Q_{22}
\end{bmatrix}
= 
\begin{bmatrix}
1 & \tilde{H}_{1k} \\
\vdots & \vdots \\
\tilde{H}_{nj} & \tilde{H}_{nk}
\end{bmatrix}.
\]

(8)
The normalization of $H_{1k}$ and $H_{2k}$ to unity is without loss of generality, as identification is only up to scale of each column. Since $Q$ is orthogonal, fixing column order, $Q_{11}^2 + Q_{21}^2 = 1$. Given this and the equation

$$xQ_{11} + Q_{21} = x,$$

I can solve for $Q_{11}$ and $Q_{21}$ where the sign is pinned down by the unit normalization. This yields two possible solutions for $Q_{11}$ and $Q_{21}$: \{\begin{align*}
Q_{11} &= 1, \quad Q_{21} = 0 \\
Q_{11} &= \frac{x^2-1}{x^2+1}, \quad Q_{21} = \frac{2x}{x^2+1}
\end{align*}\}. However, using an additional equation implied by (8), $Q_{11} + H_{1k}Q_{21}$, rules out the second solution unless $H_{1k} = 1/x$. With $Q_{11}$ and $Q_{21}$ thus pinned down, the other column of $Q$ is unique, and thus the entirety of $H$ is identified.

This argument extends to the case where the entirety of $H^{(j)}$ is fixed. Now, however, the solution is unique unless $H_{1k} = H_{1k}/H_{2k} = H_{lj}/H_{2j}$ for all $l$, in which case column $k$ is a scalar multiple of column $j$, making $H$ non-invertible, which is false by Assumption 1.1. Thus, the solution when a full column of $H$ is specified is unique.

The restriction that $H_B$ contain at most two columns is necessary to yield conditional identification without any assumptions on the variances. If three columns pertaining to proportional variance processes were included, and column $j$’s value conditioned upon, the remaining columns could not be distinguished.

**Proof of Proposition 4**

**Proposition 4.** If $B^h$ is invertible, for $\Lambda^h$ partitioned $\Lambda_A^h: \Lambda_B^h$, $\Lambda_A^h$ is identified up to scale from the covariance matrices provided no row in the upper block of

$$\begin{bmatrix}
\text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\
\tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2}))
\end{bmatrix}$$

is proportional to another row.

*Proof.* Proposition 3 follows from Sentana & Fiorentini’s (2001) Proposition 4. This proposition similarly follows by replacing $H$ with $\Lambda^h = B^hH$, where both are full-rank $n \times n$ matrices.

**Proof of Theorem 4**

**Theorem 4.** For $\Lambda^h$ partitioned $\Lambda_A^h: \Lambda_B^h$, where $\Lambda_B^h$ contains two columns, $\Lambda_A^h$ is identified up to scale from the covariance matrices provided

1. $B^h$ is invertible,
2. No row in the upper block of

$$\begin{bmatrix}
\text{vec}(\Sigma_{\eta A1}) & \text{vec}(\Sigma_{\eta A2}) \\
\tilde{r}(\text{vec}(\Sigma_{\eta B1}), \text{vec}(\Sigma_{\eta B2}))
\end{bmatrix}$$

is proportional to another row,
3. A single element $\Lambda^h_{ij}$ is fixed and $\Lambda^h_{jk} \neq 1/\Lambda^h_{ij}$ for $\Lambda^h(j), \Lambda^h(k) \in \Lambda^h_B$.

Proof. This follows from the proof of Theorem 3. The condition that $B^h$ is full rank guarantees that, like $H$, $\Lambda^h$ is full-rank, which is needed to yield Proposition 4, embedded in the theorem. Then, $H$ can simply be replaced in the proof of Theorem 3 with $\Lambda^h$. Note that this requires a unit normalization of $\Lambda^h$, which is unfamiliar. However, $\Lambda^h$ is always normalized, at least implicitly, by any normalization applied to $H$. Thus, the unit normalization is without loss of generality. Any identification results must similarly hold under a different normalization, where $\Lambda^h$ is rescaled back to the conventional $B^hH$, where $H$ has a unit diagonal.

Verification of Andrews (2017) Assumptions

Andrews’ (2017) framework for weak identification relies on the necessary conditions for his Theorem 1 being satisfied. For GMM, this requires conditions on both the robust confidence sets constructed and the non-robust confidence sets, as laid out in his Assumptions 2-6.

Assumption 2 coincides with Lemma 1. Assumption 3 largely coincides with Lemma 2, requiring the consistent estimation of $V(\theta)$. It also requires consistent estimation of the weighting matrix, which, given the focus on CUE in the present paper, also follows from Lemma 2. Assumption 4 requires, for any draw of data, the existence of normalizing matrices for both the Jacobian and orthogonalized Jacobian evaluated at $\theta_0$ such that when normalized they converge to full-rank matrices. Since the Jacobian is deterministic, it is identical to the orthogonalized Jacobian. Under strong identification, the identity matrix serves as such a normalizing matrix; it is trivial to construct a matrix with judiciously placed $O(T^{1/2})$ elements under weak identification. Unsurprisingly, if the system is unidentified, no such matrices exist.

Assumption 5 states four conditions underlying the performance of Wald tests under strong identification. First, given the separability of $\phi_t(\theta)$ in data and parameters, $\phi_T(\theta)$ converges uniformly over $\Theta$ since the sample second moments converge uniformly. Uniform boundedness follows similarly. The third condition requires that the estimated weighting matrix converges uniformly over $\Theta$, which follows since $W(\theta)$ is continuous in $\theta$ and measurable, $\Theta$ is compact, and $E\left[\text{vech}(\eta_t^T)\text{vech}(\eta_t^T)\right] < \infty$ by Assumptions 1.1 and 3.1. Since CUE is considered here, positive definiteness of $W(\theta)$ follows from the positive definiteness of $V_{\phi\phi}(\theta)$, which is established in the final point of Assumption 6 below. Finally, its maximal eigenvalue is bounded by Assumptions 1.1 and 3.1 and the minimal eigenvalue is bounded away from zero since $W(\theta)$ is positive definite. The second and fourth conditions are identification requirements, ensuring the population objective is small if and only if it
is evaluated in neighbourhood of the true parameter value; these are satisfied if the system meets the requirements for strong identification.

Assumption 6 imposes 5 conditions guaranteeing asymptotic normality of \( \hat{\theta} \) under strong identification. First, it requires that \( \theta_0 \) be in the interior of \( \Theta \). Second, it requires \( \phi_T (\theta) \) and \( \hat{\Omega} (\theta) \) be continuously differentiable; the first part holds as the Jacobian is a deterministic matrix of ones, zeros, and simple continuous functions of elements of \( \theta \), and the second part holds since \( \frac{\partial \hat{\Omega}(\theta)}{\partial \theta} = 2 \frac{\partial \phi_t(\theta)}{\partial \theta} \frac{1}{T} \phi_T (\theta)' \) and \( \frac{1}{T} \phi_T (\theta) \) is continuous in \( \theta \). The third set of conditions, on the Jacobian, are satisfied trivially since it is deterministic and of full rank under strong identification. 

\[ \sup_{\theta \in B(\theta_0)} \left\| \frac{\partial \hat{\Omega}(\theta)}{\partial \theta} \right\| = O_p (1) \] (stochastic boundedness) holds since \( \frac{1}{T} \phi_T (\theta) \) consists of linear combinations of estimated second moments of the data (which have finite second moments by Assumption 3.1) and multiplicative functions of \( \theta \) (which is finite for \( \theta \in B(\theta_0) \)). For the fifth condition, since the moment function is continuous in \( \theta \) and the asymptotic variance is a continuous function of the moment function, the asymptotic variance is continuous in \( \theta \). Thus, the covariance estimator is uniformly consistent on a ball around \( \theta_0 \). Since, for all \( \theta \), \( V_{\phi\phi}(\theta) = E \left[ \phi_t (\theta) \phi_t (\theta)' \right] \), an outer product, \( V_{\phi\phi}(\theta) \) is positive semi-definite on \( B(\theta_0) \). Since no elements in \( \phi_t (\theta) \) have zero variance, the result can be extended to positive definiteness.
References


