Abstract

We consider models defined by a set of moment restrictions that may be subject to weak identification. We propose a testing procedure to assess whether instruments are "too weak" for standard (Gaussian) asymptotic theory to be reliable. Since the validity of standard asymptotics for GMM rests upon a Taylor expansion of the first order conditions, we distinguish two cases: (i) models that are either linear or separable in the parameters of interest; (ii) general models that are neither linear nor separable. Our testing procedure is similar in both cases, but our null hypothesis of weak identification for a nonlinear model is broader than the popular one. Our test is straightforward to apply and allows to test the null hypothesis of weak identification of specific subvectors without assuming identification of the components not under test. In the linear case, it can be seen as a generalization of the popular first-stage F-test but allows us to fix its shortcomings in case of heteroskedasticity. In simulations, our test is well behaved when compared to contenders, both in terms of size and power. In particular, the focus on subvectors allows us to have power to reject the null of weak identification on some components of interest. This observation may explain why, when applied to the estimation of the Elasticity of Intertemporal Substitution, our test is the only one to find matching results for every country under the two symmetric popular specifications: the intercept parameter is always found strongly identified, whereas the slope parameter is always found weakly identified.

Keywords: GMM; Weak IV; Test; Subvector.

JEL classification: C32; C12; C13; C51.

*Simon Fraser University. Email: bertille_antoine@sfu.ca
†Brown University. Email: eric_renault@brown.edu
1 Introduction

Following Hahn and Hausman (2003), the weak instruments problem may be characterized by two features. The first relates to the bias of two-stage least squares (2SLS) towards OLS, while the second involves the inaccurate inference framework associated with the standard asymptotic theory. The main goal of this paper is to provide the practitioner with a tool to assess the second feature. Specifically, we propose to test the null hypothesis that instruments are too weak for reliable application of "standard asymptotic theory". When the sample provides compelling evidence against the null, the practitioner will conclude that she can safely rely on standard inference.

Power gain in testing for weak instruments is arguably the main contribution of this paper. We achieve this power gain by testing weak identification on some specific components of the vector of parameters of interest. One can often suspect which components of the structural parameter are likely to be poorly identified. Two cases of interest can then be distinguished when testing the null of weak identification on these suspicious components. On one hand, it should be quite intuitive to understand that maintaining the assumption that the other components are properly identified lead to substantial power gain. On the other hand, we also show how to test without assuming anything about the identification of the other components. In our Monte-Carlo study, it is quite striking how testing specific components (rather than the whole vector) can lead to power gains.

The main "trick" of this paper is related to macroeconomists’ common practice in their empirical study of DSGE models. They often suspect which parameters are likely to be poorly identified (like, for instance, the subjective discount factor) and they may prefer to fix the value of these parameters rather than to run the risk of contaminating inference on other parameters by a vain attempt to identify all parameters together. This practice also suggests a simple way to test for weak identification of some specific components. If we add to the GMM estimator of these components some well-tuned perturbation term, under the null of weak identification, the perturbation should have an immaterial impact on the value of the J-test statistic for overidentification proposed by Hansen (1982). By contrast, a sufficiently accurate identification (our alternative hypothesis in the testing procedure) will consistently detect such a perturbation. Depending on several identification patterns, we develop a testing strategy for different identification strengths, which is based on the perturbation of GMM estimators and corresponding J-tests. These tests are not only practitioner-friendly, but also conformable to a theoretical view of weak identification as developed very generally by Dufour (1997). When the degree of identification can be arbitrarily small, valid confidence
sets should be infinite with a positive probability. In terms of tests, it is akin to consider that a null hypothesis written as such distortions of the true value per unit of standard error may deliver a positive p-value.

Our analysis also sheds some light on the popular rule of thumb used to assess identification strength (see Staiger and Stock (1997)) through the magnitude of the first-stage F-statistic. Andrews (2018) has recently pointed out that "(pre)test based on the heteroskedasticity robust first-stage F-statistic fails to control coverage distortions in heteroskedastic linear IV". Our GMM approach based on distorted J-test statistics provides information not only about the behavior of the test statistic under the null of weak identification (what is also known for the F-statistic thanks to its robustification), but also under the alternative of stronger identification. We show that the GMM theory appropriately assess the identification strength (beyond the null of weak identification) by contrast with the robustified F-statistic. In linear models, the relevant amount of distortion of the GMM estimator - to be able to possibly reject the null of weak identification - is proportional to $1/\sqrt{T}$, precisely because the null of interest, following Staiger and Stock (1997), is akin to relevance of instruments vanishing as fast $1/\sqrt{T}$. It is also the case for separable models in the spirit of Stock and Wright (2000). An additional contribution of this paper is to characterize the relevant amount of infinite distortions that may deliver positive p-values for models that are neither linear nor separable with respect to the parameters under test. By doing so, we bridge the gap between Dufour (1997) and the literature on alternative asymptotics that has been developed to capture more accurately finite sample distributions of GMM estimators in the presence of weak instruments. The weak instrument literature can be understood by considering the reduced rank setting as the limit of a sequence of Data Generating Processes (DGP) indexed by the sample size. Antoine and Renault (2009, 2012) have characterized how various degrees of identification weakness (as defined by the rate of convergence towards reduced rank along the sequence of drifting DGPs) lead to various rates of convergence for estimators of structural parameters. The validity of standard asymptotic normal distributions of GMM estimators rests upon Taylor expansions of first-order conditions. These expansions involve the computation of derivatives not only at the true value of unknown parameters but also at some intermediate points between the true value and the GMM estimator. The slow convergence of the GMM estimator may introduce some asymptotic distortions. As a result, poor identification may be harmful not only when no consistent estimators are available (the case studied by Stock and Wright (2000)), but also, more generally, when consistent estimators at hand may not converge faster than the square root of square-root-T. In all cases, linear or not, we devise tests with a controlled asymptotic size that are consistent against all relevant
alternatives. We show that the finite sample power of these tests is significantly increased when the researcher is able to set the focus on a specific subvector, while possibly (but not necessarily) assuming proper identification of other components.

Two null hypothesis have actually been considered in the testing literature: the null of poor identification as done in Staiger and Stock (1997), Stock and Yogo (2005), and in this paper; the null of strong identification as done in Hahn and Hausman (2002), Inoue and Rossi (2011), and Escanciano and Otsu (2012). In line with our null hypothesis of weak identification, our Monte-Carlo study compares the performance of our test to Staiger and Stock’s (1997) and Stock and Yogo’s (2005) within a linear IV regression model. We also revisit the empirical application of Yogo (2004) (estimation of the Elasticity of Intertemporal Substitution from Euler equations) and compare with the recent test developed by Montiel-Olea and Pflueger (2013). Our test is the only one to find matching results for every country under the two symmetric specifications.

The paper is organized as follows. Section 2 introduces our framework and the motivating example of identification of the Elasticity of Intertemporal Substitution (EIS). In section 3, we set the focus on linear (or separable) models, and show how our testing procedure generalizes the first-stage F-test. General nonlinear and non-separable frameworks are considered in section 4. In section 5, we illustrate the finite sample performance of our tests through Monte-Carlo simulations in a linear IV regression model, and we apply our testing procedure to the estimation of the EIS. Section 6 concludes. The proofs of the theoretical results and the tables of empirical results are gathered in the Appendix. In the supplementary Appendix, we implement our testing procedure with a two-step GMM estimator, we show how some recent examples (Andrews and Cheng (2012), Dovonon and Renault (2013)) are nested in our framework, and provide additional Monte-Carlo results.

## 2 General framework and a motivating example

In this section, we first introduce our general framework that remains valid throughout the paper. Second, we revisit a well-known empirical example, the estimation of the Elasticity of Intertemporal Substitution to motivate our framework.

### 2.1 General framework

We consider models where the structural parameter of interest \( \theta (\theta \in \Theta \subset \mathbb{R}^p) \) is defined as solution of moment conditions, \( E[\phi_t(\theta)] = 0 \), for some known function \( \phi_t(\theta) \) of size \( k \geq p \)
defined for observations $t = 1, \cdots, T$. The moment conditions are assumed to define a true unknown value $\theta^0$ of the parameters in the interior of $\Theta$,

$$E \left[ \overline{\phi}_T(\theta) \right] = 0 \iff \theta = \theta^0 \quad \text{with} \quad \overline{\phi}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_t(\theta).$$

Our framework accommodates the case of a drifting DGP (see e.g. Stock and Wright (2000) and more examples in section 3 below) where the above expectation may be arbitrarily close to zero for a (large) set of parameters when $T$ is infinitely large.

We focus on the possibly weak identification of a subvector $\pi$ of $\theta$. More precisely, we consider the following partition of the vector $\theta$ of structural parameters,

$$\theta = (\psi', \pi')', \quad \psi \in \mathbb{R}^{p_1}, \quad \pi \in \mathbb{R}^{p_2}, \quad p = p_1 + p_2,$$

and we want to test the identification strength of $\pi$ (e.g. whether $\pi$ is weakly identified or not). Note that we do not specify any identification pattern for the nuisance parameters $\psi$. Obviously, inference about the parameters of interest $\pi$ will be simpler (and sharper) when we maintain the assumption that the nuisance parameters $\psi$ are strongly identified - however, it is important to emphasize that all possible cases will be considered in this paper: e.g. testing whether $\pi$ is weakly identified regardless of the identification properties of the nuisance parameters $\psi$.

For the purpose of GMM inference, we will consider the continuously updated GMM (CU-GMM) estimator\(^1\) $\hat{\theta}_T$ of $\theta^0$ defined as,

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta} J_T \left[ \theta, S_T^{-1}(\theta) \right] \quad \text{where} \quad J_T \left[ \theta, S_T^{-1}(\theta) \right] = T \overline{\phi}_T(\theta)' S_T^{-1}(\theta) \overline{\phi}_T(\theta) \quad (2.1)$$

and $S_T(\theta)$ consistent estimator of $S(\theta)$ the long run variance matrix of $\overline{\phi}_T(\theta)$ for all $\theta \in \Theta$,

$$\sqrt{T} \left\{ \overline{\phi}_T(\theta) - E \left[ \overline{\phi}_T(\theta) \right] \right\} \xrightarrow{d} \mathcal{N} \left( 0, S(\theta) \right),$$

with the maintained assumption that

$$\text{Plim} \left[ S_T \left( \theta^0 \right) \right] = S(\theta^0) \quad \text{and} \quad \left\| S_T^{-1}(\hat{\theta}_T) \right\| = \mathcal{O}_P(1).$$

In most cases, $S_T(\hat{\theta}_T)$ will actually converge in probability towards a positive definite deterministic matrix $S(\theta^0) = S$. However, when $S_T(\theta)$ does depend on $\theta$ and $\hat{\theta}_T$ may not be a consistent estimator of $\theta^0$, we can only assume that $S_T^{-1}(\hat{\theta}_T) = \mathcal{O}_P(1)$.

\(^1\)Whenever a first step consistent estimator $\tilde{\theta}_T$ of $\theta^0$ is available, we may also consider an efficient two-step GMM estimator of $\theta^0$ after replacing $S_T(\theta)$ by $S_T(\tilde{\theta}_T)$. However, in cases where no consistent estimator of $\theta^0$ is available (e.g. due to weak identification) only the first approach via CU-GMM is feasible. For brevity, we will always refer to the estimator $\hat{\theta}_T$ as GMM estimator; see the supplementary appendix A for a robust version of our procedure based on 2S-GMM.
2.2 Motivating example: the estimation of the EIS

To motivate our general framework, we now revisit a well-known empirical challenge, the estimation of the Elasticity of Intertemporal Substitution (EIS), a key parameter in macroeconomics and finance: for example, Woodford (2003, chapter 4) discusses its interpretation in the context of a monetary policy model; see Campbell and Viceira (1999) for its interpretation in a consumption and portfolio choice model.

Let $\beta$ be the subjective discount factor, $\gamma$ the coefficient of relative risk aversion, $\pi$ the EIS, and define $\lambda = (1 - \gamma)/(1 - 1/\pi)$. The Epstein-Zin (1989, 1991) objective function is defined recursively by

$$U_t = \left[ (1 - \beta) C_t^{(1 - \gamma)/\lambda} + \beta \left( E_t U_{t+1}^{1-\gamma} \right)^{1/\lambda} \right]^{\lambda/(1 - \gamma)},$$

where $C_t$ is the consumption at time $t$. The maximization of the above utility function subject to the intertemporal budget constraint (see Epstein and Zin (1991)) yields the following Euler equation

$$E_t \left[ \left( \beta \left( \frac{C_{t+1}}{C_t} \right)^{1/\pi} \right)^{1-\lambda} \left( \frac{1}{1 + R_{w,t+1}} \right)^{1-\lambda} \left( 1 + R_{i,t+1} \right) \right] = 1 \quad (2.2)$$

where $(1 + R_{i,t+1})$ is the gross real return on asset $i$ and $(1 + R_{w,t+1})$ is the gross real return on the portfolio of all invested wealth at time $(t+1)$. Given a vector of $k$ (valid) instruments $Z_t$, the preference parameters $\theta = (\psi' \pi)'$ with $\psi' = (\beta \lambda)$ can be estimated by CU-GMM through the above nonlinear Euler equation after defining the moment functions

$$\phi_t(\theta) = Z_t \left[ \left( \beta \left( \frac{C_{t+1}}{C_t} \right)^{1/\pi} \right)^{1-\lambda} \left( \frac{1}{1 + R_{w,t+1}} \right)^{1-\lambda} \left( 1 + R_{i,t+1} \right) - 1 \right].$$

The difficulty with this approach is two-fold: first, as mentioned by Epstein and Zin (1991), it requires observations on the returns on the wealth portfolio which includes returns on human capital; second, it is unclear whether sufficiently strong instruments can be found to allow conventional asymptotics and associated inference procedures to be reliable: see e.g. Stock and Wright (2000) and Antoine and Renault (2009) for the power utility case.

The first practical issue can easily be avoided by estimating the preference parameters from a linearized Euler equation as it does not require the knowledge of returns on the wealth portfolio. For example, when also using the market return, a second-order log-linearization\(^3\) of the Euler equation (2.2) yields:


\(^3\)When asset returns and consumption are jointly log normally distributed and homoskedastic conditional on information at time $t$, (2.3) is the exact log-linearization of (2.2).
\[ E_t r_{i,t+1} = \mu_i + \frac{1}{\pi} E_t \Delta c_{t+1} \quad (2.3) \]

where lowercase letters denote the logarithms of the corresponding uppercase variables, \( \Delta c_{t+1} \) is the consumption growth at time \((t + 1)\), and \( \mu_i \) is a constant term that depends on (conditional) second moments of \( r_{i,t+1} \) and \( \Delta c_{t+1} \): see Campbell and Viceira (2002, chapter 2) for full details and the explicit definition of \( \mu_i \). By setting

\[
\eta_{i,t+1} = r_{i,t+1} - E_t r_{i,t+1} - \frac{1}{\pi} (\Delta c_{t+1} - E_t \Delta c_{t+1})
\]

the model can be rewritten as the following linear IV regression model

\[
r_{i,t+1} = \mu_i + \frac{1}{\pi} \Delta c_{t+1} + \eta_{i,t+1},
\]

with associated moment functions,

\[
\phi_t(\mu_i, \pi) = Z_t \eta_{i,t+1} = Z_t \left[ r_{i,t+1} - \mu_i - \frac{1}{\pi} \Delta c_{t+1} \right].
\]

The obvious disadvantage of the linearized Euler equation is that the discount factor \( \beta \) cannot be identified. Nevertheless, the study of the linear IV regression model remains interesting because much more is known on how to detect weak identification: see e.g. Staiger and Stock (1997) and Stock and Yogo (2005). This is the focus of section 3. The detection of weak identification in general nonlinear models is dealt with in section 4, as it requires a more pervasive context of poor identification. We start by sketching the testing implications of this more or less severe context of poor identification, by analogy with testing of equivalence hypotheses which is popular in biostatistics\(^4\).

### 2.3 Testing of equivalence hypotheses for identification strength

Our test for identification strength is based on the asymptotic behavior of the Hansen’s overidentification test statistic (or the J-statistic) defined as \( J_T = J_T \left[ \hat{\theta}_T, S_T^{-1} \left( \hat{\theta}_T \right) \right] \). In the standard case of strong identification of the whole vector \( \hat{\theta}_T \), it is well-known that \( J_T \) is asymptotically distributed as \( \chi^2(k - p) \). However, in the general case, we only have:

**Lemma 2.1.** If \( \theta = (\varphi', \zeta')' \) with accordingly \( \Theta = \Xi \times \Upsilon, \varphi \in \Xi \) and if

\[
\text{Plim} \left[ \frac{\partial \varphi_T}{\partial \varphi}(\theta^0) \right] = \Gamma_\varphi(\theta^0) \quad \text{is a matrix of full-column rank (equal to dim(\varphi)),}
\]

\(^4\)We thank a referee for drawing our attention to this illuminating example.

7
then there exists a sequence of random variables $\overline{J}_T$ such that

$$J_T \leq \overline{J}_T \xrightarrow{d} \chi^2 [k - \dim(\theta)].$$

To see that, note that by definition of $\hat{\theta}_T$,

$$J_T = T\overline{\phi}_T(\hat{\theta}_T)'S^{-1}_T(\hat{\theta}_T)\overline{\phi}_T(\hat{\theta}_T) \leq \overline{J}_T \equiv \min_{\vartheta \in \Xi} T\overline{\phi}_T(\vartheta, \zeta_0)'S^{-1}_T(\vartheta, \zeta_0)\overline{\phi}_T(\vartheta, \zeta_0)$$

while the limit distribution of $\overline{J}_T$ comes from standard (strong identification) asymptotic theory of GMM. This argument remains valid when $\dim(\vartheta) = 0$, that is when no parameter is assumed to be strongly identified.

To detect any departure from standard asymptotic theory due to identification weakness - at least regarding the subvector $\pi$ of $\theta$ - we assess the impact of any perturbation in the estimation of $\pi$ on the $J$-test statistic. More precisely, for some non-zero $p_2$-dimensional vector $\delta$, we consider the asymptotic behavior of

$$J^\delta_T = J_T \left[ \hat{\theta}_T', S^{-1}_T(\hat{\theta}_T) \right] \quad \text{with} \quad \hat{\theta}_T = \left( \hat{\psi}_T', \hat{\pi}_T' \right)' \quad \text{and} \quad \hat{\theta}_T^\delta = \hat{\theta}_T + (0', \delta')' \quad (2.4)$$

Intuitively, the perturbed $J$-statistic should display different asymptotic behaviors under strong identification and ”sufficiently” weak identification. More precisely, if the Jacobian matrix of the moment conditions features some overly severe rank deficiency in the direction of $\pi$, one may not be able to detect some Pitman alternative $\delta_T \rightarrow 0$ making the perturbation $\delta$ only locally different from zero. By contrast, when this drifting alternative $\delta_T$ converges to zero slowly enough, one may expect that a sufficient level of identification strength will lead to a violation (by the sequence $J^\delta_T$) of an upper bound in the spirit of Lemma 2.1.

This leads us to define a null hypothesis of interest through an upper bound about the rate with which the asymptotic rank deficiency kicks in. For some real number $\nu \in ]0, 1/2]$, we consider the following null hypothesis.

**Null hypothesis of identification of $\pi$ at a rate no faster than $(1/2 - \nu)$:**

$$H_{0,\pi}(\nu) : \left\| \frac{\partial \overline{\phi}_T(\theta^0)}{\partial \pi'} \right\| = \mathcal{O}_P \left( \frac{1}{T^\nu} \right)$$

By testing the null hypothesis $H_{0,\pi}(\nu)$ for a suitable choice of $\nu$, our main goal consists in providing underpinnings for the validity of standard asymptotic inference as soon as the null of ”poor identification” is rejected. As described for instance in Romano (2005), a similar setup arises when trying to demonstrate equivalence of some pharmacokinetic parameters of a new drug to the standard drug. Romano (2005) stresses that the null hypothesis must
be formulated in such a way that "rejection of the null hypothesis is the same as declaring equivalence". By formulating the null hypothesis in this way, "the risk of marketing an alternative drug that does not behave like the standard drug is controlled". Similarly, we want to control the risk of using an inference procedure that may not behave like in a standard context of strong identification. Typically, we want to be able to at least reject the null $H_{0,π}(1/2)$ to maintain the hope that we do not need to resort to weak identification asymptotics à la Staiger and Stock (1997) and Stock and Wright (2000). We will even argue in section 4 that in general (non-separable) nonlinear settings, it may be necessary to test a more pervasive null hypothesis of weak identification like $H_{0,π}(1/4)$. The intuition is that in general settings, asymptotic distributional theory of GMM involves higher order derivatives, so that near-weak identification - in the sense that the norm of the Jacobian matrix may vanish like $(1/T^ν)$ with $1/4 ≤ ν < 1/2$ - may not be sufficient to warrant the use of standard asymptotics.

The bottom line is that the null hypothesis of interest is defined through a rate of convergence $ν$. The value $ν$ elicited for this test is dictated by the wish that "rejection of the null hypothesis is the same as declaring equivalence" with standard strong identification asymptotics (see section 4). In other words, the rate that defines the null hypothesis of interest is objective-driven, similarly to the possible rates in the interval $]0,1/2[$ for bioequivalence testing as discussed in Romano (2005) (see his remark 3.1). Equivalence testing has also been applied in econometrics by Lavergne (2014, 2015) for the purpose of model comparison. Without any concern for weak identification, but questioning the existence of a true value $θ^0$ fulfilling the moment conditions, Lavergne (2015) sets the focus on the null hypothesis:

$$\inf_{θ∈Θ} \left( E\bar{ϕ}_T(θ) \right)' S^{-1}(θ) \left( E\bar{ϕ}_T(θ) \right) ≥ \frac{γ^2}{n} \quad (2.5)$$

At the frontier of the null hypothesis, the sample counterpart of (2.5) follows asymptotically a non-central chi-square, allowing Lavergne to devise a consistent test for the null (2.5). Rejecting the null allows him to safely consider that the moment model is "approximately true". Interestingly enough, our approach to test the null $H_{0,π}(ν)$ of overly weak identification is tightly related to Lavergne’s result. Since our null of (near-)weak identification is defined by the fact that the Jacobian matrix of the moments goes to zero with the sample size, the impact of a perturbation on the estimated moments results in an asymptotically small discrepancy to zero of the minimum (2.5). Albeit for completely different purposes (weak identification vs model equivalence) our statistical strategies are arguably mathematically equivalent.

Our equivalence testing approach is in sharp contrast with approaches for which the null hypothesis under test is strong identification. For instance, the test devised by Bravo, Escanciano and Otsu (2012), albeit similar in spirit to our test, is only able to test the null hypothesis of strong identification against an alternative of (near-)weak identification. It is also the case for all classical tests for weak identification that are based on a Hausman test (see also Guggenberger and Smith (2005), Inoue and Rossi (2011)).

3 Testing for weak identification in a separable model

In this section, we introduce our testing procedure in the simpler framework of linear models, or possibly nonlinear but separable models as defined precisely in subsection 3.2. In such cases, standard asymptotic theory of Wald test statistics is known to be valid as long as identification is not as weak as it is under the null $H_{0,\pi}(1/2)$. Correspondingly, $H_{0,\pi}(1/2)$ is the null we are interested in testing in this section.

3.1 The linear intuition

In this subsection, we focus on moment conditions that are linear with respect to the parameters of interest $\pi$, that is,

$$
\phi_T(\psi, \pi) = a_T(\psi) + A_T \pi,
$$

for some sequences $a_T(\cdot)$ and $A_T$ of matricial functions that are known from the observations of a sample of size $T$; the intercept $a_T(\psi)$ may depend on the nuisance parameters, but not the slope $A_T$. In this context, the null hypothesis of interest can be rewritten as,

$$
H_{0,\pi}(1/2) : \left\| \sqrt{T}A_T \right\| = O_P(1)
$$

Example: We consider the general linear IV regression model with $n$ endogenous regressors $Y$ and $k_1$ included exogenous regressors $X$, 

$$
y = Y\pi + X\psi + u,
$$

$$
Y = Z\beta + X\gamma + v,
$$

where $Y$ is a $(T \times n)$ matrix of endogenous variables, $X$ is a $(T \times k_1)$ matrix of included exogenous variables, $Z$ is a $(T \times k_2)$ matrix of excluded exogenous variables to be used as instruments. Up to a pre-orthogonalization, we can assume without loss of generality,

$$
P\lim \left[ \frac{X'Y}{T} \right] = 0.
$$

10
Then, the moment conditions are
\[
\phi_T(\psi, \pi) = \frac{1}{T} \begin{bmatrix} X' \\ Z' \end{bmatrix} \begin{bmatrix} y - Y \pi - X \psi \end{bmatrix} \quad \text{with} \quad \sqrt{T} A_T = -\frac{1}{\sqrt{T}} \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix}.
\]

In this context, the null hypothesis of interest can be rewritten:
\[
H_{0, \pi} (1/2) : \frac{Z'Y}{\sqrt{T}} = \mathcal{O}_P(1)
\]

Note also that
\[
a_T(\psi) = \frac{1}{T} \begin{bmatrix} X' \\ Z' \end{bmatrix} \begin{bmatrix} y - X \psi \end{bmatrix} \quad \text{with} \quad \frac{\partial a_T}{\partial \psi'}(\psi) = -\frac{1}{T} \begin{bmatrix} X'X \\ Z'X \end{bmatrix},
\]
a matrix of full-column rank \(k_1\) insofar as the matrix \(X\) is full column rank.

More generally, the linear framework allows us to evaluate (in closed form) the impact of the perturbation on the GMM estimator \(\hat{\pi}_T\) and its tight connection with the identification strength of these instruments for inference on \(\pi\). To see that, write \(J^\delta_T\) (defined in (2.4)) as
\[
J^\delta_T = \left\| K_T + L^\delta_T \right\|_{S_T^{-1}(\hat{\theta}_T)}^2 \quad \text{with} \quad K_T = \sqrt{T} \phi_T(\hat{\theta}_T), \quad L^\delta_T = \sqrt{T} A_T \delta_T,
\]
and \(\left\| K \right\|_W^2 = K'WK\) for any conformable vector \(K\) and matrix \(W\).

We deduce that
\[
\left[ \left\| L^\delta_T \right\|_{S_T^{-1}(\hat{\theta}_T)} - \left\| K_T \right\|_{S_T^{-1}(\hat{\theta}_T)} \right]^2 \leq J^\delta_T \quad (3.1)
\]

The following proposition is a straightforward consequence of (3.1).

**Proposition 3.1.** (i) Under the conditions of Lemma 2.1, if \(\left\| \sqrt{T} A_T \right\| = \mathcal{O}_p(1)\) and \(\upsilon\) is a subvector of \(\psi\) conformable to Lemma 2.1, we have for any sequence \(\delta_T \in \mathbb{R}^{p_2}\):

\[
J^\delta_T \leq J_T + O_P(\delta_T) \quad \text{with} \quad J_T \xrightarrow{d} \chi^2(k - \dim(\upsilon)).
\]

(ii) If \(\operatorname{Plim} \left| \sqrt{T} A_T \delta_T \right| = +\infty\), then \(\operatorname{Plim} \left[ J^\delta_T \right] = +\infty\).

In the next subsection 3.2, we rely on Proposition 3.1 to build a test of \(H_{0, \pi} (1/2)\) of size \(\alpha\) using the statistic \(J^\delta_T\) when considering any sequence \(\delta_T\) converging to zero. Note that the size control will be uniform over the whole parameter set insofar as it is possible to elicit a vector \(\upsilon\) of minimum dimension fulfilling the conditions of Lemma 2.1.
Proposition 3.1(ii) shows that by selecting the sequence $\delta_T$ converging to zero sufficiently slowly, we should get a consistent test of the null hypothesis $H_{0,\pi}(1/2)$. More precisely, since by definition, under the alternative to $H_{0,\pi}(1/2)$, $\sqrt{T}A_T \not= O_P(1)$, there exists at least a direction $\delta$ such that the sequence $\sqrt{T}A_T\delta$ is unbounded. This direction may be multiplied by a real sequence $\varepsilon_T$ slowly converging to zero without ending up with the sequence $\sqrt{T}A_T\delta = \sqrt{T}A_T\delta\varepsilon_T$ being bounded.

3.2 The general test based on a separable perturbation

Proposition 3.1 suggests that a consistent test of the null of weak identification might be based on the distorted $J$-test statistic $J_T^{\delta_T}$ with a sequence $\delta_T$ slowly converging to zero: under the alternative, we expect $\text{Plim}[J_T^{\delta_T}] = +\infty$, while, under the null, we may hope that the distortion does not matter in the sense that $J_T^{\delta_T}$ still behaves asymptotically as

$$J_T = T\bar{\phi}_T(\hat{\theta}_T)'S_T^{-1}(\hat{\theta}_T)\bar{\phi}_T(\hat{\theta}_T) \leq T\bar{\phi}_T(\theta^0)'S_T^{-1}(\theta^0)\bar{\phi}_T(\theta^0) \overset{d}{\to} \chi^2(k).$$

As shown by (3.1) in the linear case, the fact that $J_T^{\delta_T}$ may behave under the null asymptotically as $J_T$ for any sequence $\delta_T$ going to zero (even slowly) comes from the singularity of the Jacobian matrix. The following first-order Taylor expansion may allow us to generalize the above to nonlinear settings:

$$\bar{\phi}_T\left(\hat{\theta}_T + \begin{bmatrix} 0 \\ \delta_T \end{bmatrix}\right) = \bar{\phi}_T(\hat{\theta}_T) + \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \pi'}\delta_T$$

with, up to the common abuse of notation, $\hat{\theta}_*^T$ in the interval $[\hat{\theta}_T, \hat{\theta}_T + \begin{bmatrix} 0 \\ \delta_T \end{bmatrix}]$. We can then write similarly to the previous section:

$$J_T^{\delta_T} \leq \|K_T\|_{S_T^{-1}(\hat{\theta}_T)}^2 + \|L_T^{\delta_T}\|_{S_T^{-1}(\hat{\theta}_T)}^2 + 2\|K_T\|_{S_T^{-1}(\hat{\theta}_T)}\|L_T^{\delta_T}\|_{S_T^{-1}(\hat{\theta}_T)}$$

where $K_T = \sqrt{T}\bar{\phi}_T(\hat{\theta}_T)$ and $L_T^{\delta_T} = \sqrt{T}\frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \pi'}\delta_T$.

Thus, for a nonlinear extension of Proposition 3.1, one needs to ensure that

$$\sqrt{T}\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \pi'} = O_P(1) \Rightarrow \sqrt{T}\frac{\partial \bar{\phi}_T(\hat{\theta}_*^T)}{\partial \pi'} = O_P(1).$$

A higher order Taylor expansion does not provide the answer, since, in general,

$$\sqrt{T}\frac{\partial \bar{\phi}_T(\hat{\theta}_*^T)}{\partial \pi'} = \sqrt{T}\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \pi'} + \frac{\partial^2 \bar{\phi}_T(\hat{\theta}_*^T)}{\partial \pi \partial \theta'} \sqrt{T}\left[\hat{\theta}_*^T - \theta^0\right]$$

12
with $\hat{\theta}_T^*$ in the interval $[\theta^0, \hat{\theta}_T^*]$. The difficulty comes from the fact that, under the null of weak identification, there is no reason to believe that $\sqrt{T}(\hat{\theta}_T^* - \theta^0) = \mathcal{O}_P(1)$, since we do not even know whether the GMM estimator is consistent. Therefore, the only way to ensure (3.3) is to assume that

$$\sqrt{T} \frac{\partial^2 \phi_T(\hat{\theta}_T^*)}{\partial \pi \partial \theta^0} = \mathcal{O}_P(1).$$  \hfill (3.4)

A sufficient condition for the validity (3.4) under the null of weak identification is the following separability assumption.

Assumption 1. (Separability wrt $\pi$) $\phi_t(\psi, \pi) = a_t(\psi) + A_T b_t(\psi, \pi)$.

Assumption 1 extends the linear intuition of the previous subsection by allowing the function $b_t(\psi, \pi)$ to be different from $\pi$. The null hypothesis of interest may then be characterized once again, up to singularities of the Jacobian matrix $E \left[ \frac{\partial b_t(\psi, \pi)}{\partial \theta^0} \right]$, by the condition

$$\left\| T^{1/2} A_T \right\| = \mathcal{O}_P(1).$$  \hfill (3.5)

Equation (3.5) is the condition maintained by Stock and Wright (2000) to characterize weak identification of $\pi$; it ensures in particular (3.4) and (3.3). We show the following Theorem.

Theorem 3.2. (Test of weak identification of $\pi$ in the separable case)

If $\theta = (\psi, \pi)$ and $\delta_T$ is a sequence of vectors with the same dimension as $\pi$ and converging towards zero, $d_v$ stands for the dimension of a subvector $v$ of $\psi$, we define a test of $H_{0,\pi}(1/2)$ by the sequence of critical regions $W_T^{\delta_T} = \left\{ J_T^{\delta_T} > \chi^2_{1 - \alpha} (k - d_v) \right\}$.

(i) If $\phi_t(\psi, \pi) = a_t(\psi) + A_T b_t(\psi, \pi)$ such that

(a) $\text{Plim} \left[ \frac{\partial^2 \phi_T(\theta^0)}{\partial \theta^0} \right] = \Gamma_v(\theta^0)$ is a matrix of rank $d_v$;

(b) Under the null $H_{0,\pi}(1/2)$, $\left\| T^{1/2} A_T \right\| = \mathcal{O}_P(1)$;

Then, the test $W_T^{\delta_T}$ is asymptotically at level $\alpha$ for the null hypothesis $H_{0,\pi}(1/2)$.

(ii) The test $W_T^{\delta_T}$ is consistent against any alternative that makes the choice $\delta_T$ conformable to $\text{Plim}[J_T^{\delta_T}] = +\infty$.

Up to the linearization achieved by the Taylor expansion (3.2), the proof of Theorem 3.2 is similar to the one written for Proposition 3.1. The only additional layer of complexity is the role played by the Jacobian matrix $E \left[ \frac{\partial b_t(\psi, \pi)}{\partial \theta^0} \right]$. When this matrix is not full rank: on the one hand, the null hypothesis $H_{0,\pi}(1/2)$ may be fulfilled even though $\left\| T^{1/2} A_T \right\| \neq \mathcal{O}_P(1)$;
on the other hand, the test $W_T^{\delta_T}$ may not be consistent even though the sequence $\delta_T$ has been chosen such that $\text{Plim} \left\| T^{1/2} A_T \delta_T \right\| = +\infty$.

Obviously, we obtain a test that is more powerful if we can increase the number $d_\psi$ of parameters for which we maintain the assumption of strong identification (in the infeasible case where all the other parameters would be known). Hence, parameters known to be strongly identified should not be included in $\pi$ but rather in $\psi$ and thus in $\psi$. The consistency claim (ii) in Theorem 3.2 may look somewhat tautological. One may interpret it as stating that the choice of a sequence $\delta_T$ fulfilling the consistency condition in (ii) is likely feasible in practice since by definition, under the alternative,

$$\left\| \sqrt{T} \frac{\partial \phi_T(\theta^0)}{\partial \pi'} \right\| \neq O_P(1).$$

While the choice of the tuning parameter $\delta_T$ is crucial as far as power is concerned, the simple fact that it goes to zero with the sample size ensures that, for any given DGP in the null hypothesis, the asymptotic probability of rejection is not larger than $\alpha$. The reason for that is the inequality stated in the first part of Proposition 3.1, itself due to the rank condition stated in Lemma 2.1. If necessary, this rank condition may actually be tested, since we should have at our disposal a root-$T$ consistent estimator of the Jacobian matrix, allowing us to apply for instance the rank test of Kleibergen and Paap (2006).

### 3.3 Revisiting the rule-of-thumb

We consider the standard linear IV regression model to revisit the first-stage $F$-statistic as a convenient and popular tool to detect weak identification. For sake of expositional simplicity and following common practice, we focus on the linear model with a single endogenous regressor, where all exogenous regressors have been partialled out:

$$y = Y \pi + u,$$

$$Y = Z \beta + v,$$

where the dependent variable $y$ and the endogenous regressor $Y$ are $(T, 1)$-vectors, $Z$ is the $(T, k)$-matrix of instruments whose validity is ensured by the maintained assumption, $E(u_t Z_t) = 0$. The second equation is a reduced form equation, so that by definition we also assume $E(v_t Z_t) = 0$. The variables $(u_t, v_t, Y_t, Z_t), t = 1, ..., T$ are assumed to be serially

---

6A data-based procedure to select the tuning parameter $\delta_T$ is discussed in subsection 3.4 below.
independent. Hence, with the notations of subsection 3.1, the moment conditions write

\[ \overline{\phi_T}(\pi) = \frac{1}{T} Z'(y - Y\pi), \]

and we have

\[ \sqrt{T} A_T = - \frac{1}{\sqrt{T}} Z'Y = - \frac{1}{T} Z' Z \sqrt{T} \beta - \frac{1}{\sqrt{T}} Z' \nu. \]

Therefore, the null hypothesis, \( H_{0,\pi}(1/2) : \|\sqrt{T} A_T\| = \mathcal{O}_P(1) \), can be written

\[ H_{0,\pi}(1/2) : \lambda = \frac{1}{2}, \]

within a general sequence of drifting models

\[ \beta = \beta_T = \frac{\gamma}{T^{\lambda}}, \lambda \in \left[ 0, \frac{1}{2} \right], \gamma \in \mathbb{R}^k, \gamma \neq 0. \]

Under the alternative hypothesis (\( \lambda < 1/2 \)), standard inference is possible because the GMM estimator \( \hat{\pi}_T \) is asymptotically normal.

**Proposition 3.3.** Under the alternative hypothesis, \( \lambda < 1/2 \), we have:

\[ T^{\frac{1}{2} - \lambda} \left[ \hat{\pi}_T - \pi^0 \right] = \mathcal{O}_P(1). \]

This allows us to characterize the asymptotic behavior of the \( J \)-test statistic under the alternative hypothesis of "not weak identification".

**Proposition 3.4.** Under the alternative hypothesis,

\[ \beta = \beta_T = \frac{\gamma}{T^{\lambda}}, \lambda \in \left[ 0, \frac{1}{2} \right], \gamma \in \mathbb{R}^k, \gamma \neq 0, \]

we have

\[ \frac{J_T^\delta}{T} \sim \beta_T E \left[ z_t z_t' \right] \left[ Q_z(u) \right]^{-1} E \left[ z_t z_t' \right] \beta_T \delta_T^2, \]

that is \( J_T^\delta \sim T^{1 - 2\lambda} \text{Slo}_J(\gamma) \delta_T^2 \) with

\[ \text{Slo}_J(\gamma) = \gamma' E \left[ z_t z_t' \right] \left[ Q_z(u) \right]^{-1} E \left[ z_t z_t' \right] \gamma, \quad Q_z(u) = E \left[ z_t z_t' \sigma_u^2(z_t) \right], \quad \sigma_u^2(z_t) = E[u_t^2 | z_t]. \]

Proposition 3.4 displays the rate of consistency of our test procedure as defined in Theorem 3.2. The consistency of our test is due to the fact that, outside of the null of weak identification (when \( \lambda < 1/2 \)), the distorted \( J \)-test statistic \( J_T^\delta \) goes to infinity with \( T \), insofar as the sequence of perturbations \( \delta_T \) does not go to zero too fast. However, \( J_T^\delta \) may go to
infinity at a (very) slow rate when \( \lambda \) is (very) close to 1/2; as a result, the test may have low power in such cases. For "local alternatives" (\( \lambda \) arbitrarily close to 1/2), the power of the test for moderate sample sizes highly depends on the magnitude of the slope \( \text{Slo}_J(\gamma) \). Since this magnitude is a squared norm of \( \gamma \), the power of the test may be characterized by the magnitude of the vector \( \beta_T \) of reduced form coefficients; see also related discussions in Moraes (2009). This puts on the table the natural and popular idea of checking the Fisher test statistic for the coefficients of the reduced form equation. The (heteroskedasticity corrected) Fisher test statistic can be written

\[
F_T = \frac{T - k}{k} \frac{\hat{\beta}_T \hat{\Sigma}^{-1}_T \hat{\beta}_T}{T}
\]

where \( \hat{\beta}_T \) is the OLS estimator of \( \beta \), \( \hat{\Sigma}_T = (Z'Z)^{-1}Z'Y \), and \( \hat{\Sigma}_T \) is a consistent estimator of the asymptotic variance \( \Sigma \) of \( \sqrt{T}(\hat{\beta}_T - \beta^0) \).

\[
\hat{\Sigma}_T = T\hat{\sigma}^2_v(Z'Z)^{-1} \quad \text{with} \quad \hat{\sigma}^2_v = \frac{1}{T - k} \sum_{t=1}^{T} (\hat{v}_t - \bar{v}_T)^2, \quad \bar{v}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{v}_t,
\]

and \( F_T \) is the standard F-test statistic. In case of heteroskedasticity, one may use an Eicker-White estimator of \( Q_z(v) \) to get a robustified version of the Fisher test statistic. In any case, we immediately get the following result.

**Proposition 3.5.** Under the alternative hypothesis,

\[
\beta = \beta_T = \frac{\gamma}{T\lambda}, \lambda \in \left[0, \frac{1}{2}\right], \gamma \in \mathbb{R}^k, \gamma \neq 0,
\]

we have \( kF_T \sim T^{1 - 2\lambda} \text{Slo}_F(\gamma) \) with

\[
\text{Slo}_F(\gamma) = \gamma' E\left[z_t z_t' | Q_z(u)\right]^{-1} E\left[z_t z_t' \gamma \right], \quad Q_z(v) = E\left[z_t z_t' \sigma_v^2(z_t) \right], \quad \sigma_v^2(z_t) = E[v_t^2 | z_t].
\]

The comparison of Propositions 3.4 and 3.5 shows that \( J_T^{\delta_F} \) and \( kF_T \) diverge at the same rate under the alternative, but that the slopes of the divergence, \( \delta^2 \text{Slo}_J(\gamma) \) vs \( \text{Slo}_F(\gamma) \), are different. Of course, such difference may be made immaterial by a proper calibration (see Stock and Yogo (2005) for calibration of the rule of thumb). In fact, under conditional homoskedasticity (in both structural and reduced forms), we have with obvious notations:

\[
\text{Slo}_J(\gamma) = \gamma' E\left[z_t z_t' | Q_z(u)\right]^{-1} E\left[z_t z_t' \gamma \right] = \gamma' E\left[z_t z_t' \gamma \right] \frac{\gamma}{\sigma_v^2},
\]

\[
\text{Slo}_F(\gamma) = \gamma' E\left[z_t z_t' | Q_z(v)\right]^{-1} E\left[z_t z_t' \gamma \right] = \gamma' E\left[z_t z_t' \gamma \right] \frac{\gamma}{\sigma_v^2}.
\]
Up to a proportionality coefficient that will typically be adjusted through a proper calibration, both approaches are based on the same norm of the vector $\gamma$ of reduced form parameters. It is no longer true in case of heteroskedasticity. This difference may matter (as far as power is concerned) when the rate of divergence $T^{1-2\lambda}$ is very close to zero ($\lambda$ close to the weak identification value $1/2$). By using a metric based on the variance matrix of the reduced form equation, $F_T$ may set the focus on irrelevant components of the reduced form parameter while the correct test statistic $J_T^\delta$ rightly sets the focus on the variance matrix of the structural equation. Note that this is also conformable to the CLR test proposed by Moreira (2003) for the null hypothesis $H_0 : \pi = \pi^0$. Like our J-test statistic, it sets the focus on the norm of $Z'(y - Y'\pi^0) = Z'u$.

However, and even more importantly, the crucial difference between our approach and the F-based rule of thumb is that our distorted J-statistic $J_T^\delta$ provides a formal test with a rigorous control of size, while the rule of thumb cannot, due to a nuisance parameter. We suspect that this explains Andrews’ (2018) finding that “(pre)test based on the heteroskedasticity-robust first stage F-statistic fails to control coverage distortions in heteroskedastic linear IV.” The presence of heteroskedasticity may complicate even more the calibration of the nuisance parameters. The formal result follows.

**Proposition 3.6.** Under the null hypothesis, $\beta = \beta_T = \hat{\gamma}_T$:

(i) $kF_T$ converges in distribution towards $\chi^2(k, \text{Slo}_F(\gamma))$, the non-central chi-square distribution with non-centrality parameter $\text{Slo}_F(\gamma)$;

(ii) $J_T^\delta \leq \left[ \sqrt{J_T} + \left\| L_T^\delta \right\|_{S_T^{-1}(\hat{\beta}_T)} \right]^2$ with $J_T^\delta \overset{d}{\to} \chi^2(k)$ and $\frac{1}{\delta_T} \left\| L_T^\delta \right\|^2_{S_T^{-1}(\hat{\beta}_T)} \overset{d}{\to} \chi^2(k, \text{Slo}_J(\gamma))$.

Proposition 3.6 does not fully characterize an upper bound for the asymptotic null distribution of $J_T^\delta$ (for fixed $\delta_T$) since $J_T$ and $L_T^\delta / \delta_T$ are not asymptotically independent in general. A critical value for a consistent test at level $\alpha$ must be based on the asymptotic null distribution $\chi^2(k)$ which is obtained with $\delta_T$ converging to zero. The fact that the perturbation $\delta_T$ goes to zero will obviously harm the power of the test based on $J_T^\delta$ since it will dampen the growth to infinity of $T^{1-2\lambda}\text{Slo}_J(\gamma)$; see Proposition 3.4. It is the price to pay to get rid of the nuisance parameter $\gamma$ and to deliver a tight size control with a critical value based on $\chi^2(k)$. The impact of nuisance parameter $\gamma$, through the non-centrality parameter $\text{Slo}_F(\gamma)$, is indeed the reason why the rule of thumb based on $F_T$ does not provide a formal test at level $\alpha$ and in particular does not shield from possible size distortions.
3.4 Practical implementation of our testing procedure

We now explain how to implement our testing procedure, that is how to choose the tuning parameter $\delta_T$. Recall that for a well-suited sequence $\delta_T$ with $\delta_T \xrightarrow{T} 0$ (e.g. $\delta_T = \delta / \log(\log(T)$) with $\delta$ deterministic vector), the test defined in Theorem 3.2 controls the size asymptotically and is consistent with probability one. However, for the sake of finite sample performance such as power, fine tuning the choice of the perturbation $\delta$ matters a lot. Our procedure of elicitation of $\delta$ relies on subsampling and is decomposed into two steps: first, the design of the grid of relevant candidate points for the perturbation parameter; second, the selection of the actual perturbation parameter.

- **Step 1**: design of the grid of candidate points for the perturbation vector. We consider the empirical distribution of the GMM estimators of $\theta$ across all the subsamples of $[T_\kappa]$ consecutive observations (with $\kappa$ user-chosen between 0 and 1). The range of this distribution suggests a relevant grid of values of $\theta$.

- **Step 2**: selection of the perturbation vector. We now consider $H$ subsamples of $[T_\kappa]$ consecutive observations (where $0 < \kappa^* < 1$ may or may not coincide with the aforementioned $\kappa$). For each perturbation vector $\delta_m$ in the grid and for each subsample $s$ (with $s = 1, \cdots, H$), we compute the associated GMM estimator $\hat{\theta}_{[T_\kappa^*],s}$, its local-to-zero version and the corresponding distorted J-test statistic,

$$\hat{\theta}_{[T_\kappa^*],s}^m = \hat{\theta}_{[T_\kappa^*],s} + \left( \frac{0_{p_1}}{\delta_m / \log(\log([T_\kappa^*]))} \right) \quad \text{and} \quad J_{[T_\kappa^*],s}^m = J_{[T_\kappa^*]}\left[ \hat{\theta}_{[T_\kappa^*],s}^m, S_T(\hat{\theta}_T) \right].$$

We obtain, for each perturbation vector $\delta_m$ in the grid, a cross-sectional distribution of the test statistic $(J_{[T_\kappa^*],s}^m)_{s=1,\cdots,H}$. Then, we compute $q(1 - \alpha^*, \delta_m)$, the $(1 - \alpha^*)$-quantile associated with $(J_{[T_\kappa^*],s}^m)_{s=1,\cdots,H}$ (for some user-chosen $\alpha^*$), and select the perturbation vector $\delta_m^*$ associated with $q(1 - \alpha^*, \delta_m)$ the closest to the $(1 - \alpha^*)$-quantile of the chi-square distribution with $(k - p_1)$ degrees of freedom; $(1 - \alpha^*)$ may, or may not correspond to the actual asymptotic size of the designed test. In our Monte-Carlo experiments, we use 15 grid points for each component of $\theta$, $\kappa = \kappa^* = 0.95$, and $\alpha = \alpha^*$; additional simulation results reported in the supplementary appendix are not too sensitive to these choices.

---

7There are other ways to obtain a meaningful grid of candidate points for the perturbation vector that may be less computer-intensive, or may rely on additional information about the values of the parameters.

8$[T_\kappa]$ refers to the largest integer below $T_\kappa$. We consider consecutive observations to accommodate possible serial dependencies.

9Recall that the appropriate degrees of freedom depend on the identification assumption imposed on $\theta_1$. For instance, when $\psi$ is assumed near-weakly identified, we use $(k - p_1)$ degrees of freedom, and when no such identification assumption is maintained, we rather use $k$ degrees of freedom.
4 Detecting poor identification in the general model

In this section, we consider general models defined by nonlinear and non-separable moment conditions. We are interested in assessing the identification strength of some components of the vector of structural parameters $\theta$. When testing for identification strength, we will specifically check whether our dataset allows us to reject the null hypothesis that some components of $\theta$ are too poorly identified - so poorly that standard (Gaussian) asymptotic theory is not reliable. Our testing strategy, similar to the one designed for linear or separable models, amounts to a J-test questioning the rate of convergence of a given GMM estimator. However, the generality of the model calls for a more extensive concept of "poor (weak) identification". More precisely, the null hypothesis under test (see definition in section 4.2 below) covers a broader variety of poor identification patterns than the strict definition of weak identification considered previously. We start with some preliminary theoretical results before defining our testing procedure.

4.1 Preliminary theoretical results

4.1.1 Asymptotic theory

To simplify the exposition, we present the asymptotic theory for CU-GMM estimators under the following high-level assumptions\(^\text{10}\).

**Assumption 2.** (High-level regularity assumptions)

Let $\hat{\theta}_T$ be the CU-GMM estimator defined in (2.1) and $M_T$ be a sequence of deterministic nonsingular matrices of size $p$.

(a) The identification strength of $\theta$ is characterized by $M_T$ with $\lim_T (M_T / \sqrt{T}) = 0$ and

$$\Gamma(\theta^0) \equiv \text{Plim} \left[ \frac{\partial \phi_T(\theta^0)}{\partial \theta} M_T \right] \text{ exists and is full-column rank.} \quad (4.1)$$

(b) For any sequence $\theta_T$ between $\theta^0$ and $\hat{\theta}_T$ component by component\(^\text{11}\), we have

$$\text{Plim} \left[ \frac{\partial \phi_T(\theta_T)}{\partial \theta} M_T \right] = \Gamma(\theta^0) \quad \text{the full-column rank matrix introduced in (4.1)}$$

(c) $\sqrt{T} \phi_T(\theta^0)$ converges in distribution towards a normal distribution with mean zero and variance $S(\theta^0)$.

(d) $\sqrt{T} M_T^{-1}(\hat{\theta}_T - \theta^0) = O_P(1)$.

\(^{10}\)More primitive conditions in the case of efficient 2S-GMM can be found in Antoine and Renault (2010).

\(^{11}\)Hereafter, we use the notation $\theta_T \in [\theta^0, \hat{\theta}_T]$. 

19
Going from assumption 2(a) to 2(b) only amounts to assuming that
\[ \theta_T \in [\theta^0, \hat{\theta}_T] \Rightarrow \text{Plim} \left[ \left( \frac{\partial \bar{\phi}_T(\theta_T)}{\partial \theta'} - \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} \right) M_T \right] = 0. \] (4.2)

The zero-limit in (4.2) is obviously ensured for the components of the moment vector \( \bar{\phi}_T(.) \) that are linear with respect to the parameters \( \theta \). For those which are not linear with respect to some components of \( \theta \), the issue at stake is to know whether their rate of convergence along the sequence \( \theta_T \) is sufficient to supersede the possible convergence to infinity of the sequence \( M_T \). Antoine and Renault (2009, 2010) show that it is a sufficient condition for asymptotic normality of the GMM estimator.

**Theorem 4.1. (Asymptotic normality and efficient estimation)**

Let \( \hat{\theta}_T \) denote the GMM estimator defined in (2.1). Under assumption 2, \( \sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) \) is asymptotically normal with mean zero and variance \( [\Gamma'(\theta^0)S^{-1}(\theta^0)\Gamma(\theta^0)]^{-1} \).

Theorem 4.1 extends the asymptotic normality result given in the linear case: for \( T \) large enough, \( \sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) \) can be seen as a Gaussian vector with mean zero and variance consistently estimated by
\[
M_T^{-1} \left[ \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta} S_T^{-1}(\hat{\theta}_T) \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} M_T^{-1} \quad \text{since} \quad \Gamma(\theta^0) = \text{Plim} \left[ \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} M_T \right].
\] (4.3)

However, it is incorrect to deduce from formula (4.3) that \( \sqrt{T}(\hat{\theta}_T - \theta^0) \) can be seen (for \( T \) large enough) as a Gaussian vector with mean zero and variance consistently estimated by
\[
\left[ \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta} S_T^{-1}(\hat{\theta}_T) \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1}.
\] (4.4)

In case of lack of strong identification, the above matrix (4.4) is actually the inverse of an asymptotically singular matrix. In this sense, the standard GMM theory does not apply and some components of \( \sqrt{T}(\hat{\theta}_T - \theta^0) \) actually blow-up. That being said, for all practical purposes related to inference about the structural parameter \( \theta \), the knowledge of the matrix \( M_T \) is not required and the J-test for overidentification can be performed as usual as stated in the following result12.

**Theorem 4.2. (J-test)**

Let \( \hat{\theta}_T \) be the CU-GMM estimator defined in (2.1). Under assumption 2,
\[
J_T \left[ \hat{\theta}_T, S_T^{-1}(\hat{\theta}_T) \right] \equiv T \bar{\phi}_T(\hat{\theta}_T)S_T^{-1}(\hat{\theta}_T)\bar{\phi}_T(\hat{\theta}_T) \overset{d}{\rightarrow} \chi^2(k - p).
\]

12See also the discussions in Antoine and Renault (2010, 2012) for the efficient 2S-GMM estimator.
The J-test is a "black box" that conceals the fact that the quality of identification is heterogeneous. The asymptotic singularity of (4.4) means that the actual rate of convergence of the GMM estimator may vary depending on the linear combinations of the structural parameter vector \( \theta \).

### 4.1.2 Identification of subvectors

When testing for identification strength, we will check whether our dataset allows us to reject the null hypothesis that some components of \( \theta \) are (too) poorly identified. Throughout, \( \psi \) denotes a vector of \( p_1 \) components of \( \theta \), while \( \pi \) collects the \( p_2 \) (\( = p - p_1 \)) remaining components of \( \theta \) that are not included in \( \psi \). For simplicity, \( \psi \) corresponds hereafter to the first \( p_1 \) components of \( \theta \), that is \( \theta = (\psi, \pi) \). We consider cases where the econometrician is concerned about the poor identification of \( \pi \) while prior knowledge warrants "sufficiently strong" identification of \( \psi \).\(^{13}\) It is only when a sequence of matrices \( M_T \) characterizing the identification strength of \( \theta \) is block-diagonal, that we can deduce the identification strengths of \( \psi \) and \( \pi \) from the identification strength of \( \theta \).

**Assumption 3.** The sequence of matrices \( M_T \) such that assumption 2 is fulfilled can be chosen as

\[
M_T = \begin{bmatrix} M_{\psi,T} & 0 \\ 0 & M_{\pi,T} \end{bmatrix}.
\]  

(4.5)

A well-known example is the separable case (see Assumption 1 in section 3.2). It has been put forward by Stock and Wright (2000) who consider\(^ {14}\) (see supplementary Appendix B),

\[
E\bar{\phi}_T(\theta) = E\bar{\phi}_{1T}(\psi) + \frac{1}{T^\nu}E\bar{\phi}_{2T}(\theta), \quad \text{with } 0 < \nu \leq 1/2,
\]

\[
\text{Rank} \left( E \frac{\partial \bar{\phi}_{1T}(\psi^0)}{\partial \psi'} \right) = p_1 \quad \text{and} \quad \text{Rank} \left( E \frac{\partial \bar{\phi}_{2T}(\theta^0)}{\partial \pi'} \right) = p_2.
\]

\( M_T \) can then be defined as in (4.5) with \( M_{\psi,T} = I_{p_1} \) and \( M_{\pi,T} = T^\nu I_{p_2} \). Our setting here is more general since the matrix \( M_{\psi,T} \) may also go to infinity.

Up to a convenient reparameterization, the above block-diagonal structure of (4.5) is not really restrictive; see Antoine and Renault (2010) for a more extensive discussion. It makes sense to question the identification strength of \( \pi \) while maintaining sufficiently strong identification on \( \psi \), precisely because the two subvectors are disentangled in the classification of directions as regards to identification strength.

\(^{13}\)Note that our procedure also accommodates testing the whole vector \( \theta \), and testing subvector \( \pi \) without assuming that \( \psi \) is sufficiently strongly identified.

\(^{14}\)Strictly speaking, Stock and Wright (2000) only consider the limit case with \( \nu = 1/2 \).
Our null hypothesis is devised such that failing to reject it implies that standard inference (using the Gaussian asymptotic theory of section 4.1.1) is not reliable when parameter $\pi$ is considered as unknown. In front of such a negative evidence, two strategies are available:

(i) one resorts to inference procedures that are robust to weak identification. Of course, robustness has a cost in terms of efficiency of estimators, power of tests, and maintained assumptions regarding nuisance parameters;

(ii) following the common practice of calibration, one may fix the value of parameters in $\pi$ at pre-specified levels provided by other studies hoping that these calibrated values are not too far from the unknown ones and will not contaminate inference on $\psi$. The validity of this practice has been extensively studied by Dridi, Guay and Renault (2007) who propose some encompassing tests for backtesting it.

In any case, both strategies will always maintain the assumption that, when $\pi$ is fixed at its true unknown value $\pi^0$, the remaining moment problem is well-behaved.

**Definition 4.1. (Sufficiently strong identification of a subvector)**
With the block-diagonal structure (4.5) for the sequence of matrices $M_T$ and a true unknown value $\pi^0$ for $\pi$, $\psi$ is sufficiently strongly identified if the two following conditions are fulfilled.

(i) Assumption 2 is fulfilled for the sequence of matrices $M_{\psi,T}$ in the context of the (infeasible) moment model

$$E \left[ \phi_t(\psi, \pi^0) \right] = 0 \quad \text{with} \quad \psi \in \Theta(\pi^0) = \{ \psi \in \mathbb{R}^{p_1}; (\psi, \pi^0) \in \Theta \}.$$  

(ii) For any GMM estimator $\hat{\theta}_T$ and any sequence $\theta^*_T$ such that $\psi^*_T = \hat{\psi}_T$ and $\pi^*_T - \hat{\pi}_T = o_p(T^{-1/4})$, we have

$$\frac{\partial \phi_T(\theta^*_T)}{\partial \pi^*_{\ell}} M_{\pi,T} = O_p(1).$$

When wondering whether a parameter vector $\theta = (\psi, \pi)$ that strictly nests $\psi$ is sufficiently strongly identified, the key issue is to check that the convergence condition (4.2) holds for any sequence $\theta_T$ between the true value $\theta^0$ and some GMM estimator $\hat{\theta}_T$, that is

$$\text{Plim} \left[ \left( \frac{\partial \phi_T(\theta_T)}{\partial \theta'} - \frac{\partial \phi_T(\theta^0)}{\partial \theta'} \right) M_T \right] = 0.$$  

(4.6)

The maintained assumption $\lim_T (M_T / \sqrt{T}) = 0$ may not be sufficient to get (4.6) in a general model, because a rate of convergence for $\theta_T$ strictly slower than $\sqrt{T}$ may be unable to protect against the asymptotic blow-up of the sequence $M_T$:

$$\frac{\partial \phi_T(\theta_T)}{\partial \theta'} M_T = \left[ \frac{\partial \phi_T(\theta_T)}{\partial \psi'} M_{\psi,T} : \frac{\partial \phi_T(\theta_T)}{\partial \pi'} M_{\pi,T} \right].$$  

(4.7)
In the second block of (4.7), \( \frac{\partial \phi_T(\theta_T)}{\partial \pi'} \) usually depends on the estimator of the weakest parameters \( \pi \). Since these parameters are consistent at a rate \( \|M_{\pi,T}\| / \sqrt{T} \), a Taylor expansion of \( \frac{\partial \phi_T(\theta_T)}{\partial \pi'} \) (assuming \( \phi \) twice continuously differentiable) will only allow us to prove (4.6) if \( \|M_{\pi,T}\|^2 / \sqrt{T} \) goes to zero when \( T \) goes to infinity. In other words, the condition to ensure that the poor identification of \( \pi \) does not impair the identification of the whole vector \( \theta \) is that \( \|M_{\pi,T}\| = o(T^{1/4}) \), or equivalently that the rate of convergence of all parameters in \( \pi \) (rate defined by the sequence of matrices \( M_{\pi,T}/\sqrt{T} \)) is faster than \( T^{1/4} \). As already emphasized in Antoine and Renault (2012), this condition is quite similar in spirit to Andrews’ (1994) study of MINPIN estimators, or estimators defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator. Intuitively, second-order terms in Taylor expansions have to remain negligible in front of first-order terms. Since this condition comes as a byproduct of second-order Taylor expansions, it explains why no threshold like \( T^{1/4} \) pops up in the linear case. Following Antoine and Renault (2009), this latter property will be dubbed near-strong identification of the associated linear combination.

The above issue is obviously easier to control for when the weakest parameters \( \pi \) are not multiplied by the most explosive part of the sequence \( M_T \), namely \( M_{\pi,T} \), as studied in section 3.

### 4.2 Detecting poor identification in the general case

In this section, null hypotheses under test are about the rate of convergence of \( \hat{\pi}_T \), a subset of components of a given GMM estimator \( \hat{\theta}_T \). Unlike existing literature, we do not maintain any identification assumption on the remaining components of \( \hat{\theta}_T \), namely \( \hat{\psi}_T \). To formulate a well-suited null hypothesis about the rate of convergence of \( \hat{\pi}_T \), we follow the practice that has been dominant since Staiger and Stock (1997) and setup the null hypothesis with the worst-case scenario regarding the identification of \( \pi \), that is the rate of convergence of \( \hat{\pi}_T \): no linear combination of \( \pi \) can be estimated at a satisfactory rate to guarantee standard asymptotic theory.

For some real number \( \nu \in [0, 1/2] \), we consider the following null hypothesis.

**Null hypothesis of identification of \( \pi \) at a rate no faster than \((1/2 - \nu)\):**

\[
H_{0,\pi}(\nu) : \left\| \frac{\partial \hat{\phi}_T(\theta)}{\partial \pi'} \right\| = O_P \left( \frac{1}{T^{\nu}} \right).
\]

The above null hypothesis encompasses the null hypothesis of weak identification of \( \pi \) (with \( \nu = 1/2 \)) and the strong identification of \( \psi \) (with \( \nu = 0 \)) considered in section 3. In a
general (nonlinear and non-separable) model, we will only be able to control the level of the test when $\nu \leq 1/4$; see Corollary 4.3 below. The case $\nu \leq 1/4$ is precisely the case of interest for general models, since, when the data allow us to reject the null hypothesis, we can conclude that the rate of convergence is faster than $T^{(1/2-\nu)} = T^{1/4}$, meaning that we have ”near-strong identification”.

The key intuition for our proposed test of $H_{0,\pi}(\nu)$ comes from a technical result (Lemma A.1 in the appendix) that allows us to characterize the behavior of moment conditions computed at a conveniently distorted value of the GMM estimator $\hat{\theta}_T$. Specifically, the distorted estimator $\hat{\theta}_T^{\delta_T}$ is defined as

$$\hat{\theta}_T^{\delta_T} = (\hat{\psi}_T, \hat{\pi}_T + \delta_T) \quad \text{with} \quad \delta_T = \delta/a_T,$$

where this distortion $\delta_T$ depends on a deterministic sequence $a_T$ and a direction $\delta \in \mathbb{R}^{p_2}$.

Our test computes the J-test at the distorted estimator. We can show the following result.

**Corollary 4.3.** Under Assumptions 2 and 3:

(i) Under the null hypothesis $H_{0,\pi}(\nu), \nu \leq 1/4$, for any deterministic sequence $a_T$ such that $a_T/T^{1/2-\nu} \rightarrow \infty$, we have for any $\delta \in \mathbb{R}^{p_2}$,

$$\text{Plim} \left[ J_T(\delta_T) - J_T(0) \right] = 0.$$

(ii) Assume that $\delta$ is drawn randomly according to some absolutely continuous probability distribution on $\mathbb{R}^{p_2}$. Then, under the alternative hypothesis to $H_{0,\pi}(\nu)$, there exists a deterministic sequence $a_T$ such that $a_T/T^{1/2-\nu} \rightarrow \infty$ and, at least for a convenient subsequence,

$$\text{Plim} \left[ J_T(\delta_T) \right] = \infty \quad \text{(4.9)}$$

**Comments:**

(i) When testing $H_{0,\pi}(\nu)$, the level of our test is controlled whenever $\nu \leq 1/4$: the equivalence result above between the usual J-test and the J-test evaluated at the distorted GMM estimator allows us to bound the asymptotic distribution of our test statistic under the null. This bound should hold uniformly (size control) as discussed in section 3.2. In this respect, the test designed for a general model actually works similarly to the one for a linear or separable model. Unfortunately, without the linearity or separability, the above equivalence only holds for $\nu \leq 1/4$.

(ii) The key intuition for the above result is that when $\lim_T \left\| \frac{\sqrt{T}}{a_T} M_{\pi}^{-1} \right\| = \infty$ as in Lemma A.1, we can be sure that $\lim_T \left\| \frac{\sqrt{T}}{a_T} M_{\pi}^{-1} \delta \right\| = \infty$ for generically all directions $\delta$. Then, the result
(4.9) follows by standard Taylor expansions, knowing that \( \sqrt{T} M_T^{-1} (\hat{\theta}_T - \theta^0) = O_p(1) \).

- **Testing procedure:**
  With general moment conditions that may not be separable with respect to the parameters under test \( \pi \), we wonder whether some linear combinations of these parameters can be consistently estimated at a rate faster than \( T^{1/4} \). In other words, we want to test the following null hypothesis (see section 4.1 above for a precise definition \( H_{0, \pi}(1/4) \)):

  \[ H_{0, \pi}(1/4) : \text{No identification within } \pi \text{ at rate faster than } T^{1/4}. \]

Our testing procedure in the general case is quite similar to the one described in the previous section. However, for clarity, the following highlights the key elements of our general procedure based on a J-test statistic computed at a distorted estimator as in (4.8).

**Theorem 4.4.** (Test of poor identification of \( \pi \) in the general case)

For an arbitrary choice of a deterministic sequence \( a_T \) such that \( a_T / T^{1/4} \to \infty \) and a vector \( \delta \in \mathbb{R}^{p_2} \), we define two asymptotic tests with respective critical region \( W_{T}^{a, \delta} \) and \( \tilde{W}_{T}^{a, \delta} \),

\[
W_{T}^{a, \delta} = \{ J_T(\delta_T) > \chi^2_{1-\alpha}(k) \} \quad \text{and} \quad \tilde{W}_{T}^{a, \delta} = \{ J_T(\delta_T) > \chi^2_{1-\alpha}(k - p_1) \},
\]

where \( \chi^2_{1-\alpha}(d) \) is the \((1-\alpha)\)-quantile of the chi-square distribution with \( d \) degrees of freedom.

(i) Under assumptions 2 and 3, the test \( W_{T}^{a, \delta} \) is asymptotically at level \( \alpha \) for the null hypothesis \( H_{0, \pi}(1/4) \). The test \( W_{T}^{a, \delta} \) is consistent against any alternative that makes the choice \( (a_T, \delta) \) conformable to (4.9).

(ii) If we also assume that \( \psi \) is sufficiently strongly identified, the test \( \tilde{W}_{T}^{a, \delta} \) is asymptotically at level \( \alpha \) for the null hypothesis \( H_{0, \pi}(1/4) \), albeit less conservative than \( W_{T}^{a, \delta} \). The test \( \tilde{W}_{T}^{a, \delta} \) is consistent against any alternative that makes the choice \( (a_T, \delta) \) conformable to (4.9).

The same comments apply to Theorems 3.2 and 4.4 regarding the use of intermediate critical values depending on the number of parameters for which sufficiently strong identification is granted, and the wide range of alternatives against which consistency is warranted.

## 5 Monte-Carlo and Empirical illustrations

### 5.1 Linear IV regression model

Consider the following linear IV regression model (see also section 3.3),

\[
y_t = \alpha_0 + Y_t \pi_0 + h(Z_t) \varepsilon_t, \quad Y_t = Z_t' \beta + U_t,
\]

(5.1)
where \( Y_t \) is a univariate endogenous regressor, while \( Z_t \) is a vector of \( L_z \) (exogenous) instrumental variables that follows a standard normal distribution. \((\varepsilon_t, U_t)\) is normally distributed and independent of \( Z_t \). We set \( \theta^0 \equiv (\alpha_0, \pi_0)' = (0 0)' \). Our benchmark model is heteroskedastic with \( h(z) = \sqrt{1 + (e'z)^2/(L_z + 1)} \) where \( e \) is the vector of ones of size \( L_z = 3 \). In addition, \((h(Z_t)\varepsilon_t, U_t)\) has mean 0, unit unconditional variances, and unconditional correlation \( \rho = 0.4 \); the sample size is \( T = 200 \). The intercept parameter \( \alpha \) is always strongly-identified, while the slope parameter \( \pi \) is more or less weakly identified depending on the value of \( \beta'\beta \) that ranges from 0 (no identification) to 0.30 (stronger identification) with \( \beta \) proportional to the vector \( e \).

In this experiment, we compare the properties of three tests: (i) our test introduced in section 3 (Antoine-Renault hereafter); (ii) Staiger and Stock’s (1997) rule of thumb based on the first-stage F test (SS hereafter); (iii) Stock and Yogo’s (2005) test based on 10% bias of 2SLS\(^{15} \) (SY hereafter). We consider a size of 0.95 throughout. To implement our test, we follow section 3.4 with a matching probability equal to 0.95, a subsample of size 133, and a perturbation of the form \( \delta_k/\log(\log(T)) \) with \( \delta_k \) chosen in a grid that contains 15 points per dimension of the parameter under test: in the present case, two when testing \((\alpha_0, \pi_0)\) jointly, and one when only testing \( \alpha_0 \) or \( \pi_0 \).

Monte-Carlo results with 5,000 replications are collected in Table 1 where we consider three cases, \( \beta'\beta = 0.30 \) (stronger identification), 0.02 (weaker identification), and 0 (no identification). For each case, we report in panel A the first four moments of the Monte-Carlo distribution of the standardized GMM estimator. The reader can then assess how far this Monte-Carlo distribution is from its standard asymptotic approximation: recall that for the standard normal distribution, we expect mean 0, variance 1, skewness 0 and kurtosis 3. In panel B, we report the results for testing weak identification of the whole parameter vector \( \theta \) jointly and of a subvector. When testing the whole vector, we consider the three above-mentioned tests, Antoine-Renault, SS, and SY; when testing a subvector, we consider two versions of Antoine-Renault: testing for weak identification of the intercept \( \alpha \) without assuming that the slope \( \pi \) is sufficiently strongly identified; testing for weak identification of the slope \( \pi \) under the assumption that the intercept \( \alpha \) is sufficiently strongly identified. For each test, we report the number of rejections and rejection probabilities. Several comments are in order.

1. The distribution of the standardized GMM estimator displayed in Panel A confirms that the asymptotic approximation works well for the intercept parameter in all cases,

\(^{15}\)This is the version of the test proposed by Stock and Yogo which is commonly used. Results for alternate versions of their test based on 5% bias, as well as 10% and 15% size distortion are available upon request.
whereas the approximation worsens for the slope parameter as the identification issues become more acute for small values of $\beta'\beta$.

2. When comparing the performances of the three tests on $\theta$, Antoine-Renault’s test has more power.

3. When considering Antoine-Renault’s tests on subvectors, the test on the intercept parameter $\alpha$ that does not assume that the slope $\pi$ is sufficiently strongly identified correctly concludes that the intercept is strongly identified with rejection probabilities very close to 1 in all cases. As expected, rejection probabilities of the test on the slope $\pi$ decrease as the identification strength worsens.

Moreover, we notice that Antoine-Renault’s test is more reliable when testing the identification of slope or intercept separately. In practice, it is rare that one is interested in testing all parameter vectors jointly, and our test allows to set the focus on subvectors of interest without having to assume that the components not under test are sufficiently strongly identified.

We conduct robustness checks by changing the DGP as well as the tuning parameters of our test: we consider a homoskedastic model; we increase the level of endogeneity, as well as the sample size and the number of instruments; we change the matching probability, and the size of the subsample. The results available in the supplementary appendix are qualitatively similar to the benchmark case.

5.2 Estimating the EIS

In this section, we apply our testing procedures to the motivating example discussed in section 2.2, the instrumental variables estimation of the Elasticity of Intertemporal Substitution (EIS). Recall the linearized Euler equation in two standard IV frameworks:

$$
\Delta c_{t+1} = \nu + \psi r_{t+1} + u_{t+1} \quad \text{and} \quad r_{t+1} = \xi + \frac{1}{\psi} \Delta c_{t+1} + \eta_{t+1},
$$

where $\psi$ is the EIS, $\Delta c_{t+1}$ the consumption growth at time $(t + 1)$, $r_{t+1}$ a real asset return at time $(t + 1)$, $\nu$ and $\xi$ two constants. The vector of (valid) instruments is denoted by $Z_t$.

Following Yogo (2004), we use the real return on the short-term interest rate for $r_t$, and two lags of nominal interest rate, inflation, consumption growth, and log dividend price-ratio for instruments. The quarterly data for 11 countries can be found on Yogo’s webpage at https://sites.google.com/site/motohiroyogo/. Our study replicates partially Yogo
(2004) and Montiel-Olea and Pflueger (2013) who found that the EIS point estimates are small and close to zero.

Table 2 compares tests for weak instruments for 11 countries and compares to Table 1 in Yogo (2004) and Table 2 in Montiel-Olea and Pflueger (2013). Panel A tests for weak identification with the ex-post real interest rate as the endogenous variable, while Panel B tests for weak identification with the consumption growth as the endogenous variable. We compare seven tests: Stock and Yogo test based on 10% bias of 2SLS, three versions of the test proposed by Montiel-Olea and Pflueger, and three versions of the test proposed in this paper. Montiel-Olea and Pflueger propose two testing procedures, simplified and generalized, that rely on the same test statistic and adjust the critical values. Accordingly, we consider the three associated tests with simplified critical values \( c_{\text{simp}} \), and generalized critical values, \( c_{\text{TSLS}} \) and \( c_{\text{LIML}} \). We also consider three versions of our test: testing the intercept and the slope jointly, testing only the intercept, and testing only the slope.

Our motivation for disentangling identification issues for the two parameters is the following. Obviously, the constant is not a weak instrument and thus the linear combination of \( \nu \) and \( \psi \) given by \( [E(\Delta c) - \nu - \psi E(\Delta r)] \) is strongly identified. In the two-dimensional parameter space, there should be at least one direction that is strongly identified. Strictly speaking, this direction is not given by the intercept \( \nu \) that might be contaminated by the possibly weak identification of the slope \( \psi \). However, when \( \nu \) is recovered as a linear combination of two directions in the parameter space, one strongly identified, one possibly weakly identified, most of the weight comes from the former and the latter has a negligible impact (see Antoine and Renault (2009) for a more extensive discussion). This is the reason why it is worth considering identification of \( \nu \) and \( \psi \) separately, the strong identification of the former giving more power to test the identification strength of the latter.

This distinction cannot be captured by joint tests such as Stock and Yogo’s or Montiel-Olea and Pflueger’s. Our test for the intercept always conclude that it is strongly identified. Our test for the slope always conclude that it is weakly identified. Our joint test is associated with mixed results: it rejects weak identification for 5 countries in Panel A (out of 11) and 2 countries in Panel B. The tests of Stock and Yogo and Montiel-Olea and Pflueger also lead to mixed results. However, it is worth mentioning that none of these tests can reject weak identification in Panel B. This corresponds to the results of our test for the slope only. Finally, our test is the only one to find matching results for every country under both specifications: the intercept parameter is always found strongly identified, while the slope parameter is always found weakly identified.

To conclude and similarly to the Monte-Carlo results, we recommend using our test of a
subvector rather than the joint test, which always yields consistent results of identification strength regardless of the formulation of the model as shown in Panels A and B.

6 Conclusion

This paper introduces a new test for identification strength based on sensitivity analysis of a GMM estimator. In linear homoskedastic cases, our test is asymptotically equivalent to the properly tuned rule of thumb based on a F-test performed on the reduced form. Moreover, by contrast with the naive F-test, even robustified for heteroskedasticity, our GMM approach provides the right metric for assessing the discrepancy from the null hypothesis. It is a formal test with properly controlled asymptotic level.

Our null hypothesis is defined by an identification strength that would be too weak for valid asymptotic inference. This concern for valid inference implies that we define the null hypothesis by a rate convergence, that may differ in linear and nonlinear contexts. The curvature of the model with respect to the possibly weakly identified parameters may make their slow convergence even more detrimental. The relevant rates govern the rate of distortion of GMM estimators introduced to detect weak identification. This careful definition is the price to pay to control the asymptotic behavior of tests and estimators through the setup of drifting DGPs. In addition, we provide a consistent cross-validation procedure to choose the relevant tuning parameters for the distortion of the GMM estimators. The important advantage of defining the null hypothesis by an "excessive level of weak identification" is that, when compelling evidence has been gathered by the practitioner to reject the null, she can safely perform standard GMM inference.

Our paper can also be understood as a general rule of thumb, independently of the asymptotic theory based on drifting DGPs. After all, it is natural to conclude that our GMM problem is well behaved when the overidentification test statistics is sufficiently sensitive to a well-tuned distortion of the GMM estimator. Moreover, other contexts of weak identification, like first-order underidentification studied by Dovonon and Renault (2013), are also characterized by slower rates of convergence, albeit not based on the arguably less conventional argument of drifting DGPs. Even more generally, the impossibility theorems devised by Dufour (1997) perfectly characterize the accuracy delusion that can be produced by overlooked identification issues. Our strategy to distort the GMM estimator appears as a relevant answer to this risk. And, at the end of the day, if distorting the GMM estimator does not have a significant impact on the overidentification test statistic, the practitioner should be worried that inference may not be as accurate as it was supposed to be.
References


Appendix

A Proofs of the main results

Proof of Proposition 3.1:
(i) Let us assume that the true unknown value $\psi^0$ of $\psi$ is such that, for some subvector $v,$

$$\text{Rank} \left\{ \text{Plim} \left[ \frac{\partial a_T}{\partial \psi} (\psi^0) \right] \right\} = \text{dim}(v)$$

Then, by using the notations of (3.1), we see that by lemma 2.1:

$$\|K_T\|_{S_{T^{-1}(\hat{\psi}_T)}}^2 \leq J_T \quad \text{with} \quad J_T \overset{d}{\to} \chi^2 (k - \text{dim}(v)),$$
while since \(\|S_T^{-1}(\hat{\theta}_T)\| = O_P(1)\), we have: 
\[
\sqrt{T} A_T = O_P(1) \implies L_T^{\delta_T} \|S_T^{-1}(\hat{\theta}_T) = O_P(\delta_T).
\]

The result then follows from the second inequality in (3.1).

(ii) The result follows from the first inequality in (3.1) and the condition:

\[
\text{Plim} \left\| \sqrt{T} A_T \delta_T \right\| = \text{Plim} \left\| L_T^{\delta_T} \|S_T^{-1}(\hat{\theta}_T) = +\infty .
\]

since the minimum eigenvalue of \(S_T^{-1}(\hat{\theta}_T)\) is bounded away from zero. ■

**Proof of Proposition 3.3:**

By definition: \(\hat{\pi}_T = \arg\min_{\pi} \frac{1}{T} (y - Y \pi)' Z S_T^{-1}(\pi) Z' (y - Y \pi)\). An application of Theorem 2.1 in Antoine and Renault (2012) reveals that, under the alternative hypothesis \((\lambda < 1/2)\), \(\hat{\pi}_T\) is consistent. Hence, as far as first-order asymptotics are concerned, this estimator is asymptotically equivalent to the following infeasible estimator

\[
\hat{\pi}_T = \arg\min_{\pi} \frac{1}{T} (y - Y \pi)' Z S^{-1}(\pi^0) Z' (y - Y \pi), \tag{A.1}
\]

where \(S(\pi^0) = \text{Plim} \left[ S_T(\pi^0) \right] = E \left[ z_t z_t' \sigma^2_u(z_t) \right] = Q_z(u) \text{ and } \sigma^2_u(z_t) = E[u_t^2 | z_t].\)

Note that we use the same notation \(\hat{\pi}_T\) for both the continuously updated GMM estimator and the infeasible Efficient GMM estimator since our focus of interest is only the asymptotic probability distribution. From (A.1), \(\hat{\pi}_T\) is defined as solution of the first-order conditions:

\[
Y' Z S^{-1}(\pi^0) Z' (y - Y \hat{\pi}_T) = 0 \iff \hat{\pi}_T = \left[ Y' Z S^{-1}(\pi^0) Z' Y \right]^{-1} Y' Z S^{-1}(\pi^0) Z' y
\]

\[
\iff \hat{\pi}_T = \pi^0 + \left[ Y' Z S^{-1}(\pi^0) Z' Y \right]^{-1} Y' Z S^{-1}(\pi^0) Z' u
\]

Therefore,

\[
T^{\lambda\gamma} \left[ \hat{\pi}_T - \pi^0 \right] = \left[ \left( T^{\lambda} \frac{Y' Z}{T} \right) S^{-1}(\pi^0) \left( T^{\lambda} \frac{Z' Y}{T} \right) \right]^{-1} \left[ \left( T^{\lambda} \frac{Y' Z}{T} \right) \right] S^{-1}(\pi^0) T^{1/2} \frac{Z'u}{T}.
\]

Since \(\frac{Z'u}{\sqrt{T}} = O_P(1)\) and \(\lambda < \frac{1}{2}\), we have:

\[
T^{\lambda\gamma} \frac{Z' Y}{T} = \frac{Z' Z}{T} T^{\lambda\beta} + T^{\lambda\gamma} \frac{Z'u}{T} = E[z_t z_t'] \gamma + o_P(1) \tag{A.2}
\]

Hence:

\[
T^{\lambda\gamma} \left[ \hat{\pi}_T - \pi^0 \right] = \left[ (E[z_t z_t'] \gamma) S^{-1}(\pi^0) (E[z_t z_t'] \gamma) \right]^{-1} (E[z_t z_t'] \gamma) \left[ S^{-1}(\pi^0) \frac{Z'u}{\sqrt{T}} + o_P(1) = O_P(1) \right.
\]

**Proof of Proposition 3.4:**

33
With the notations of subsection 3.1, we have:

\[ J_T^{\beta_T} = \left\| K_T + L_T^{\delta_T} \right\|_{S_T^{-1}(\pi_T)}^2 = K_T' S_T^{-1}(\hat{\pi}_T) K_T + 2K_T' S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T} + L_T^{\delta_T'} S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T}. \]

Then, \( K_T' S_T^{-1}(\hat{\pi}_T) K_T = J_T \leq T \phi_T(\pi^0) S_T^{-1}(\pi^0) \phi_T(\pi^0) \xrightarrow{d} \chi^2(k) \) while, as long as \( \delta_T = O_P(1) \)

\[ L_T^{\delta_T} = \sqrt{T} A_T \delta_T = -\left[ \frac{1}{T} Z' Z \sqrt{T} \beta + \frac{1}{T} Z' v \right] \delta_T = -E [z_t z_t'] \gamma T^{1/2 - \lambda} \delta_T + O_P(\delta_T) \]

\[ \xrightarrow{a} -E [z_t z_t'] \gamma T^{1/2 - \lambda} \delta_T, \]

and \( K_T = \sqrt{T} \phi_T(\hat{\pi}_T) = \frac{1}{\sqrt{T}} Z'(y - \hat{Y}_T) = \frac{1}{\sqrt{T}} Z'u + \frac{1}{\sqrt{T}} Z'Y(\pi^0 - \hat{\pi}_T) \)

\[ \xrightarrow{O_P(1)} T^{\lambda} Z'Y T^{-\frac{1}{2} - \lambda} [\hat{\pi}_T - \pi^0] \]

which gives, by using both Proposition 3.3 and (A.2):

\[ K_T = O_P(1) + \{E [z_t z_t'] \gamma + o_P(1)\} O_P(1) = O_P(1). \]

As a result:

\[ J_T^{\beta_T} = K_T' S_T^{-1}(\hat{\pi}_T) K_T + 2K_T' S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T} + L_T^{\delta_T'} S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T}, \]

\[ = O_P(1) \left\{ 1 + \left\| L_T^{\delta_T} \right\| + L_T^{\delta_T'} S_T^{-1}(\hat{\pi}_T) \right\} \]

\[ \xrightarrow{a} \gamma' E [z_t z_t'] \gamma T^{1/2 - 2\lambda} \delta_T^2, \]

\[ \Rightarrow J_T^{\beta_T} \xrightarrow{a} \beta_T'E [z_t z_t'] \gamma T^{1/2} E [z_t z_t'] \beta_T \delta_T^2 \]

**Proof of Proposition 3.6:**

\[ kF_T = \left( \sqrt{T} \hat{\beta}_T \right)' \Sigma^{-1} \left( \sqrt{T} \hat{\beta}_T \right) + o_P(1) \]

\[ \sqrt{T} \hat{\beta}_T = \left( \frac{Z'Z}{T} \right)^{-1} \left( \frac{Z'Y}{\sqrt{T}} \right) = (E [z_t z_t'])^{-1} \left( \frac{Z'Z}{T} \sqrt{T} \beta + \frac{Z'v}{\sqrt{T}} \right) + o_P(1) \]

\[ = \gamma + (E [z_t z_t'])^{-1} \frac{Z'v}{\sqrt{T}} + o_P(1) \]

\[ \Rightarrow \sqrt{T} \hat{\beta}_T \xrightarrow{d} \mathcal{N}(\gamma, \Sigma) \Rightarrow kF_T \xrightarrow{d} \chi^2(k, \gamma' \Sigma^{-1} \gamma), \gamma' \Sigma^{-1} \gamma = \text{Slo}_I(\gamma) = \gamma' E [z_t z_t'] \gamma T^{1/2} E [z_t z_t'] \gamma \]

Moreover, we have seen in the proof of Proposition 3.4 that:

\[ J_T^{\beta_T} = K_T' S_T^{-1}(\hat{\pi}_T) K_T + 2K_T' S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T} + L_T^{\delta_T'} S_T^{-1}(\hat{\pi}_T) L_T^{\delta_T}, \]

\[ \leq \left\| K_T \right\|_{S_T^{-1}(\hat{\pi}_T)}^2 + \left\| L_T^{\delta_T} \right\|_{S_T^{-1}(\hat{\pi}_T)}^2 + 2 \left\| K_T \right\|_{S_T^{-1}(\hat{\pi}_T)} \left\| L_T^{\delta_T} \right\|_{S_T^{-1}(\hat{\pi}_T)} \]

\[ \leq \mathcal{J}_T + \left\| L_T^{\delta_T} \right\|_{S_T^{-1}(\hat{\pi}_T)}^2 + 2 \sqrt{\mathcal{J}_T} \left\| L_T^{\delta_T} \right\|_{S_T^{-1}(\hat{\pi}_T)} = \left( \sqrt{\mathcal{J}_T} + \left\| L_T^{\delta_T} \right\|_{S_T^{-1}(\hat{\pi}_T)} \right)^2, \]

34
with \( J_T = T \tilde{\phi}_T(\pi^0)'S_T^{-1}(\pi^0)\tilde{\phi}_T(\pi^0) \overset{d}{\to} \chi^2(k) \), while
\[
L^{\delta^T} = - \left[ \frac{1}{T} Z' Z \sqrt{T} \beta + \frac{1}{\sqrt{T}} Z' v \right] \delta_T \Rightarrow \frac{1}{\delta_T} L^{\delta^T}_T = \mathbb{E} [z_t z_t'] \gamma + \frac{1}{\sqrt{T}} Z' v + o_p(1)
\]
\[
\Rightarrow \frac{1}{\delta_T} L^{\delta^T}_T \overset{d}{\to} \mathbb{N} (E [z_t z_t'] \gamma, Q_Z(u))
\]
\[
\Rightarrow \frac{1}{\delta_T} \left\| L^{\delta^T}_T \right\|_{S^{-1}_T(\delta_T)}^2 \overset{d}{\to} \chi^2(k, SloJ(\gamma))
\]
with \( SloJ(\gamma) = \gamma' E [z_t z_t'] [Q_Z(u)]^{-1} E [z_t z_t'] \gamma \)

**Proof of Theorem 4.1 (Asymptotic normality and efficient GMM estimator):**
A mean-value expansion of the moment conditions around \( \theta^0 \) for \( \hat{\theta}_T \) between \( \hat{\theta}_T \) and \( \theta^0 \) yields
\[
\tilde{\phi}_T(\hat{\theta}_T) = \tilde{\phi}_T(\theta^0) + \frac{\partial \tilde{\phi}_T(\theta^0)}{\partial \theta'} (\hat{\theta}_T - \theta^0) \tag{A.3}
\]
The first-order conditions,
\[
2 \frac{\partial \tilde{\phi}_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T(\hat{\theta}_T) \tilde{\phi}_T(\hat{\theta}_T) + \tilde{\phi}_T(\hat{\theta}_T) \frac{\partial S^{-1}_T(\hat{\theta}_T)}{\partial \theta} \tilde{\phi}_T(\hat{\theta}_T) = 0
\]
can be written more explicitly using Antoine, Bonnal and Renault (2007) or Newey and Windmeijer (2005): for example, with \( \theta \) univariate,
\[
\left[ \frac{\partial \tilde{\phi}_T(\hat{\theta}_T)}{\partial \theta} - \tilde{\phi}_T(\hat{\theta}_T) S^{-1}_T(\hat{\theta}_T) \text{Cov} \left( \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta}, \phi_T(\hat{\theta}_T) - \tilde{\phi}_T(\hat{\theta}_T) \right) \right] S^{-1}_T(\hat{\theta}_T) \tilde{\phi}_T(\hat{\theta}_T) = 0
\]
Under assumption 2a), we have
\[
\frac{\partial \tilde{\phi}_T(\hat{\theta}_T)}{\partial \theta'} M_T \overset{p}{\to} \Gamma(\theta^0) \quad \text{and} \quad \frac{\partial \tilde{\phi}_T(\hat{\theta}_T)}{\partial \theta'} M_T \overset{p}{\to} \Gamma(\theta^0)
\]
We combine (A.3), (A.4) and the above to get
\[
M^{-1}_T \sqrt{T}(\hat{\theta}_T - \theta^0) = - \left[ M'_T \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T(\hat{\theta}_T) \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta'} M_T \right]^{-1} M'_T \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T(\hat{\theta}_T) \sqrt{T} \tilde{\phi}_T(\theta^0) + o_p(1)
\]
which gives the desired result. ■

**Proof of Theorem 4.2 (J-test):**
We use intermediate results from the proof of Theorem 4.1 to get:
\[
\sqrt{T} \tilde{\phi}_T(\hat{\theta}_T) = \sqrt{T} \tilde{\phi}_T(\theta^0) - \frac{\partial \tilde{\phi}_T(\hat{\theta}_T)}{\partial \theta'} M_T \left[ M'_T \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T(\hat{\theta}_T) \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta'} M_T \right]^{-1} M'_T \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T(\hat{\theta}_T) \sqrt{T} \tilde{\phi}_T(\theta^0)
\]
\[
\Rightarrow TQ_T(\hat{\theta}_T) = \left[ \sqrt{T} \tilde{\phi}_T(\theta^0) \right]' S^{-1/2}_T(\hat{\theta}_T) \left[ I_k - P_X \right] S^{-1/2}_T(\hat{\theta}_T) \left[ \sqrt{T} \tilde{\phi}_T(\theta^0) \right] + o_p(1)
\]

35
with \( S_T^{-1}(\hat{\theta}_T) = S_T^{-1/2}(\hat{\theta}_T)S_T^{-1/2}(\hat{\theta}_T) \) and \( P_X = X(X'X)^{-1}X' \) for \( X = S_T^{-1/2}(\hat{\theta}_T)\frac{\partial \pi_T(\theta)}{\partial \theta} M_T \). And we get the expected result.

**Lemma A.1.** (i) Under the null hypothesis \( H_0(\nu) \), for any deterministic sequence \( a_T \) such that \( a_T/T^\nu \to \infty \), we have

\[
\lim_{T} \left[ \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \right] = 0.
\]

(ii) Under the alternative hypothesis to \( H_0(\nu) \), there exists a deterministic sequence \( a_T \) such that \( a_T/T^\nu \to \infty \) and, at least for a convenient subsequence,

\[
\lim_{T} \left\| \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \right\| = \infty.
\]

**Proof of Lemma A.1:** Our proof uses the notations introduced in section 4.

(i) Assume that we can find a deterministic sequence \( a_T \) with \( a_T/T^\nu \to \infty \) such that the sequence of matrices \( \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \) does not converge to zero. Then, there exists a vector \( \delta \in \mathbb{R}^{p_2} \) such that \( \sqrt{T} M_{\pi,T}^{-1} \frac{\delta}{a_T} \) does not converge to zero. Assume, for expositional simplicity, that the first coefficient \( b_T \) of this sequence does not converge to zero. Then, up to eliciting a well-chosen subsequence, we can claim that for some \( \varepsilon > 0 \), we have for all \( T, |b_T| > \varepsilon \).

However, we know that \( \sqrt{T} M_{\pi,T}^{-1} (\hat{\theta}_T - \theta^0) = \mathcal{O}_p(1) \); hence, \( \sqrt{T} M_{\pi,T}^{-1} (\hat{\pi}_T - \pi^0) = \mathcal{O}_p(1) \).

Note that \( M_T \) can be chosen as \( M_T = R\Lambda_T \) for some fixed non-singular matrix \( R \) and a sequence \( \Lambda_T \) of diagonal matrices\(^{16}\). Hence, \( R^{-1} \pi = \eta \), where \( \eta = R^{-1} \theta \) is the new vector of parameters. Thus we have, \( \sqrt{T} \Lambda_{\pi,T}^{-1} (\hat{\eta}_{\pi,T} - \eta^0) = \mathcal{O}_p(1) \), and, in particular, focusing on first (diagonal) coefficient \( \lambda_{1,\pi,T} \) of \( \Lambda_{\pi,T} \) and first coefficient \( \hat{\eta}_{1,\pi,T} \) of \( \eta_{\pi,T} \), we have:

\[
\frac{\sqrt{T}}{\lambda_{1,\pi,T}} (\hat{\eta}_{1,\pi,T} - \eta^0_{1,\pi}) = \mathcal{O}_p(1).
\]

However, since \( b_T \) has been defined as the first coefficient of \( \sqrt{T} M_{\pi,T}^{-1} \frac{\delta}{a_T} = \sqrt{T} \Lambda_{\pi,T}^{-1} R_{\pi}^{-1} \frac{\delta}{a_T} \), it can be written

\[
b_T = \frac{\sqrt{T}}{\lambda_{1,\pi,T}} \frac{\delta_1}{a_T},
\]

where \( \delta_1 \) stands for the first coefficient of \( R_{\pi}^{-1} \delta \). Note that \( \delta_1 \neq 0 \) (since \( |b_T| > \varepsilon \)) and we deduce from a comparison of the two above formulas that

\[
\frac{b_T}{\delta_1} a_T (\hat{\eta}_{1,\pi,T} - \eta^0_{1,\pi}) = \mathcal{O}_p(1).
\]

\(^{16}\)Antoine and Renault (2010) show how to write \( M_T \) as \( R\Lambda_T \) by defining the coefficients of \( \Lambda_T \) as the singular values of \( M_T \) (as the square-roots of eigenvalues of \( (M_T M_T') \)) and \( R \) as the limit of a sequence of orthogonal matrices of eigenvectors of \( (M_T M_T') \).
Since \(|b_T| > \varepsilon\) for all \(T\) (or at least a subsequence), this implies that, along a subsequence,
\[
a_T(\hat{\eta}_{1,T} - \eta^0_{1,T}) = O_P(1).
\]
Therefore, the null hypothesis \(H_0(\nu)\) must be violated since \(a_T/T^\nu \to \infty\) and \((\hat{\eta}_{1,T} - \eta^0_{1,T})\) has been built as a linear combination of \((\hat{\pi}_T - \pi^0)\).

(ii) Under the alternative, we can find a deterministic sequence \(b_T\) with \(b_T/T^\nu \to \infty\) such that for some non-zero vector \(\delta \in \mathbb{R}^{p_2}\) we have (for a well-suited subsequence):
\[
b_T\delta'(\hat{\pi}_T - \pi_0^0) = O_P(1).
\]
Then:
\[
\delta'(\hat{\pi}_T - \pi_0^0) = \gamma'(\hat{\eta}_{\pi,T} - \eta^0_{\pi}) = O_P(1/b_T),
\]
for some non zero vector \(\gamma = R'_\pi \delta\). Let us consider another deterministic sequence \(a_T\) with \(a_T/T^\nu \to \infty\) but \(a_T/b_T \to 0\). Then:
\[
\gamma'a_T(\hat{\eta}_{\pi,T} - \eta^0_{\pi}) = o_P(1).
\]
Since we maintain the assumption that, at least for a convenient subsequence, \(\gamma'(\hat{\eta}_{\pi,T} - \eta^0_{\pi})\) does not go to zero at a rate faster than \(\min_j \left[ \lambda_{j,\pi,T} / \sqrt{T} \right] \), we are able to conclude that at least one diagonal coefficient of \(A_T \Lambda_{\pi,T}^{-1} \sqrt{T}\) goes to zero. Therefore, since
\[
\frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} = \frac{\sqrt{T}}{a_T} \Lambda_{\pi,T}^{-1} R_\pi^{-1},
\]
at least one line of this matrix is such that the sum of the absolute value of its coefficients goes to infinity. In other words, the norm \(\| \cdot \|_{\infty}\) of this matrix (maximum row sum norm, see Horn and Johnson (1985) p295) goes to infinity. Since, for the spectral matrix norm \(\| \cdot \|\) we are using in this paper, we have \(\|M_{\pi,T}\| \geq \sqrt{p_2} \|M_{\pi,T}\|_{\infty}\) (see Horn and Johnson (1985) p314), we can conclude that, at least for a convenient subsequence,
\[
\lim_T \left\| \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \right\| = \infty
\]

**Proof of Corollary 4.3:** Our proof uses the notations introduced in section 4.

(i) From Lemma A.1(i), we get:
\[
\lim_T \left[ \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \delta \right] = 0.
\]
Moreover, a mean-value expansion of the moment conditions gives:
\[
\sqrt{T} \phi_T(\hat{\delta}_T) = \sqrt{T} \phi_T(\hat{\theta}_T) + \sqrt{T} \frac{\partial \phi_T}{\partial \delta_T}(\hat{\theta}_T)(\hat{\delta}_T - \hat{\theta}_T),
\]

37
where, with the standard abuse of notation, we have defined component by component some \( \tilde{\theta}_T \) between \( \hat{\theta}_T^{\delta^T} \) and \( \hat{\theta}_T \). Then, by definition of \( \hat{\theta}_T^{\delta^T} \),

\[
\sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) = \sqrt{T} \phi_T(\hat{\theta}_T) + \sqrt{T} \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \pi^T}(\hat{\theta}_T) \frac{\delta}{a_T} = \sqrt{T} \phi_T(\hat{\theta}_T) + \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \delta \quad (A.5)
\]

Definition 4.1 implies that

\[
\frac{\partial \phi_T(\hat{\theta}_T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} = \mathcal{O}_P(1) \quad \text{since} \quad \tilde{\psi}_T = \hat{\psi}_T \quad \text{and} \quad \tilde{\pi}_T - \hat{\pi}_T = \mathcal{O}(\delta/a_T) = o(T^{-1/4}) \quad (A.6)
\]

Then, from (A.4), (A.5) and (A.6), we deduce: \( \sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) = \sqrt{T} \phi_T(\hat{\theta}_T) + o_P(1) \), and the required result immediately follows.

(ii) Since from Lemma A.1, \( \lim_T \left\| \sqrt{T} a_T^{-1} M_{\pi,T}^{-1} \delta \right\| = \infty \), we have, for most vectors \( \delta \in \mathbb{R}^{p_2} \),

\[
\lim_T \left\| \sqrt{T} a_T^{-1} M_{\pi,T}^{-1} \delta \right\| = \infty \quad (A.7)
\]

Only vectors \( \delta \) in the orthogonal space of the relevant eigenspace would not fulfill this condition. In particular, condition (A.7) is fulfilled with probability 1 when \( \delta \) is drawn randomly according to some absolutely continuous probability distribution. Then, using (A.5), we expect \( \sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) \) to ”blow-up” like

\[
z_T \equiv \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} \frac{\sqrt{T}}{a_T} M_{\pi,T}^{-1} \delta.
\]

\( z_T \) must blow-up since, by definition 4.1, \( \left[ \frac{\partial \phi_T(\theta_0^T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} \right] \) is asymptotically full-column rank. If \( \left[ \frac{\partial \phi_T(\theta_0^T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} \right] \) is different from \( \left[ \frac{\partial \phi_T(\theta_0^T)}{\partial \pi^T}(\hat{\theta}_T) M_{\pi,T} \right] \) (due to some non-linearity w.r.t. \( \pi \)), it would take some perverse asymptotic singularity to erase the blow-up in (A.7). Note that insofar as \( \sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) \) blows up, we have \( \text{Plim} [J_T(\delta_T)] = \infty \) since, for \( T \) large,

\[
J_T(\delta_T) \geq \text{Mineg}(\Omega_T) \left\| \sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) \right\|^2 \geq \text{Mineg}(\Omega_T) \frac{1}{2} \left\| \sqrt{T} \phi_T(\hat{\theta}_T^{\delta^T}) \right\|^2,
\]

with probability one asymptotically, where \( \text{Mineg}(A) \) is the smallest eigenvalue of a matrix \( A \) and \( \text{Mineg}(\Omega) > 0 \) by positive definiteness. □

\textbf{Proof of Theorem 4.4:} It follows directly from Corollary 4.3. □

\section*{B\hspace{1em}Tables of results}

38
Case 1: strong identification with $\beta'\beta = 0.3$

1.A: Distribution of the standardized GMM estimator

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.04</td>
<td>1.04</td>
<td>0.01</td>
<td>2.97</td>
</tr>
<tr>
<td>Slope</td>
<td>0.11</td>
<td>1.11</td>
<td>0.52</td>
<td>3.82</td>
</tr>
</tbody>
</table>

1.B: Results of the tests

| Antoine-Renault | SS | SY | Antoine-Renault | Antoine-Renault |
| whole           | whole | whole | intercept, slope |
| Number of Rej.  | 5,000 | 4.366 | 4.591 | 5,000 | 4.988 |
| Rejection Prob. | 1.00 | 0.87 | 0.92 | 1.00 |

Case 2: weak identification with $\beta'\beta = 0.02$

2.A: Distribution of the standardized GMM estimator

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.04</td>
<td>0.95</td>
<td>0.02</td>
<td>2.79</td>
</tr>
<tr>
<td>Slope</td>
<td>0.28</td>
<td>1.17</td>
<td>1.00</td>
<td>4.73</td>
</tr>
</tbody>
</table>

2.B: Results of the tests

| Antoine-Renault | SS | SY | Antoine-Renault | Antoine-Renault |
| whole           | whole | whole | intercept, slope |
| Number of Rej.  | 4,978 | 6.00 | 12.00 | 5,000 | 3,015 |
| Rejection Prob. | 0.99 | 0.00 | 0.00 | 1.00 |

Case 3: no identification with $\beta'\beta = 0$

3.A: Distribution of the standardized GMM estimator

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.02</td>
<td>0.55</td>
<td>-0.10</td>
<td>4.04</td>
</tr>
<tr>
<td>Slope</td>
<td>1.30</td>
<td>1.29</td>
<td>0.74</td>
<td>3.37</td>
</tr>
</tbody>
</table>

3.B: Results of the tests

| Antoine-Renault | SS | SY | Antoine-Renault | Antoine-Renault |
| whole           | whole | whole | intercept, slope |
| Number of Rej.  | 4,935 | 0.00 | 0.00 | 4,999 | 1,422 |
| Rejection Prob. | 0.98 | 0.00 | 0.00 | 1.00 |

Table 1: Inference in the linear heteroskedastic model with $T = 200$, $M = 5,000$, $\rho = 0.4$, $\theta^0 = (0\ 0)'$. We consider 3 cases: $\beta'\beta = 0.3$ (strong identification), 0.02 (weak identification) and 0 (no identification). In each case, Panel A provides the first four moments of the Monte-Carlo distribution of the standardized GMM estimator, while Panel B collects the number of rejections and rejection probabilities when testing for weak identification on the whole parameter vector $\theta$ (with Antoine-Renault, SS, and SY), and on a subvector with Antoine-Renault: testing the intercept without assuming strong identification of the slope; testing the slope when assuming strong identification of the intercept.
Table 2: Test for weak instruments in the estimation of EIS. "0" indicates that the test cannot reject at 5%, whereas "1" indicates that the test rejects at 5%.