Testing Identification Strength*

Bertille Antoine† and Eric Renault‡

February 16, 2013

Abstract

We consider models defined by a set of moment restrictions that may be subject to weak identification. Following the recent literature, the identification of the structural parameters is characterized by the Jacobian of the moment conditions. We unify several definitions of identification that have been used in the literature, and show how they are linked to the consistency and asymptotic normality of GMM estimators. We then develop two tests to assess the identification strength of the structural parameters in models that are (i) either linear or separable; (ii) neither linear nor separable. Both tests are straightforward to apply and allow to test specific subvectors without assuming identification of the components not under test. In simulations, our tests are well-behaved when compared to contenders, both in terms of size and power. In addition, we show how pretesting specific subvectors can improve inference by delivering shorter confidence regions with comparable coverage probability. Finally, when applied to the estimation of the Elasticity of Intertemporal Substitution, our test indicates weak instruments in a larger number of countries than tests proposed by Stock and Yogo (2005) and Montiel Olea and Pflueger (2012).

Keywords: GMM; Weak IV; Test; Subvector.

JEL classification: C32; C12; C13; C51.

*This paper has been circulating under the title "Specification tests for strong identification".

†*Simon Fraser University. Email: bertille_antoine@sfu.ca.* *The author gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada and the Economics Department at Brown University where part of this work was completed.*

‡*Brown University. Email: eric_renault@brown.edu*

1 Introduction

Following Hahn and Hausman (2003), the weak instruments problem should be defined 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 some reinsurance against the fear of the second feature[1]. In our opinion, this goal imposes to test the null hypothesis that instruments are too weak for reliable application of "standard asymptotic theory". It is only when the sample provides some compelling evidence that the null can be rejected that the practitioner will conclude that she can safely rely on standard inference.

Such a pretest approach has sometimes been criticized for two main reasons. First, there is the common pretesting problem: fully correct inference should take the error of the pretesting step into account. Second, the pretesting step may be skipped by relying on inference procedures that are robust to weak identification. We do not formally address the first issue, but rely on our Monte-Carlo evidence to illustrate that this issue is only of second-order with respect to the severe distortions of inference due to the weak instruments problem. In response to the second criticism, we argue that if our pretest is sufficiently well-behaved, considerable gains in inference precision can be achieved without sacrificing coverage properties by avoiding the systematic use of the worst case scenario, as robust procedures do. Power gain in pretesting for weak instruments is arguably the main contribution of this paper. We achieve this power gain by testing weak identification on some specific subset of 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 pretest without assuming anything about

[1] A former draft of this paper, which has been circulated under the title "Specifications tests for strong identification", was also addressing the first feature. This study is now included in a companion paper Antoine and Renault (2013).

the identification of the other components. In our Monte-Carlo study, it is quite striking how pretesting specific components (rather than the whole vector) can lead to improved inference.

The main trick of this paper is germane to a common practice of macroeconomists in their empirical study of DSGE models. They often suspect what are the parameters 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 contamination of 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 detect consistently such a perturbation. Depending on several identification patterns, we develop similar tests for different identification strengths, all 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 overidentification 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 an infinite distortion of the true value may deliver a positive p-value.

On theoretical grounds, the main contribution of this paper is, for the purpose of testing, to characterize the relevant amount of infinite distortions that may deliver positive p-values. 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. Two streams of alternative asymptotics have been considered so far. On the one hand, Staiger and Stock's (1997) asymptotic approximation (see also Stock and Wright (2000) for its non-linear generalization) is such that IV estimators have non-standard distributions. On the other hand, a more recent literature still considers the IV estimators as approximately normal, but such that the standard asymptotic variance estimators may not be as reliable as in the strong

instrument approach. While several authors, including Hansen, Hausman and Newey (2008) in the linear case, and Newey and Windmeijer (2009) in the non-linear case, have justified such adjustments of Gaussian-based confidence intervals by the so-called many-instrument asymptotics, others are more agnostic and simply acknowledge that slower rates of convergence towards normality may occur: see e.g. Hahn and Kuersteiner (2002) in the linear case, Antoine and Renault (2009) and Caner (2010) in the non-linear case. In this respect, the fact that the number of instruments may be seen as going to infinity with the sample size is only one possible interpretation of these non-standard rates. See also the recent paper by Andrews and Cheng (2012) where many of the above identification cases are discussed.

More generally, 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 DGP) lead to various rates of convergence for estimators of structural parameters. Besides the extreme case of weak identification studied by Staiger and Stock (1997) and Stock and Wright (2000), a slightly less severe identification issue, the so-called near-weak identification, may ensure asymptotic normality (albeit at a slower rate than standard root-T), allowing almost standard GMM inference; see also Antoine and Renault (2009) and Andrews and Cheng (2012).

For the purpose of our testing issue, we first unify several definitions of identification that have been used in the literature, and show how they are linked to the asymptotic properties of GMM estimators. Since the goal is always to make sure that, when rejecting the null of weak identification, the empirical researcher is able to safely apply standard inference rules, we stress the need to distinguish several cases. In the linear case, the question of interest is whether some parameters are consistently estimated. In the setting of Hahn and Kuersteiner (2002), one would then be sure that confidence sets can be built from asymptotic normal distributions, while realizing that rates of convergence may be slower than the standard root-T.

In non-linear settings, the issue may be more involved. The problem is that 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. If the convergence of the GMM estimator is too slow, this will introduce some asymptotic distortions[2]. This may lead the empirical researcher to consider that (nearly) weak 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. This difficulty has led Antoine and Renault (2009) and Caner (2010) to dub "near-strong identification" the case where a sufficiently fast rate was insured.
As a result, our tests are designed in the following three cases of interest:

(i) The moment conditions are linear (or affine) with respect to the suspicious parameters, so that the null of weak identification is simply akin to say they are not consistently estimated[3].

(ii) A similar null hypothesis is worth considering even in some non-linear cases, when a well-suited separability[4] of the moment conditions is maintained between parameters whose identification is under test and other parameters.

(iii) By contrast, in the general non-linear and non-separable case, we must test a null hypothesis of no near-strong identification.

In all cases, we devise tests with a controlled asymptotic size that may be conservative but 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. We are also able to keep valid testing procedures based on distorted J-test statistics, even when lack of consistency under the null prevents us from estimating consistently the efficient weighting matrix for asymptotic chi-square distributions. We apply a projection technique in the spirit of Chaudhurhi and Zivot (2011). It is also important

[2]Such distortions typically arise because some second-order terms in Taylor expansions may not remain negligible in front of first-order terms.

[3]Linearity with respect to the suspicious parameters ensures that they cannot make first-order derivatives excessively noisy.

[4]Here again, the idea is to avoid the aforementioned contamination in Taylor expansions.

to recall that, when the empirical researcher is lucky enough to reject the relevant null hypothesis, she knows for sure that she can apply standard inference procedures, even if rates of convergence slower than root-T are possibly at stake. The key idea is that she would not need to know these rates of convergence, for instance for Wald test or overidentification test, because studentization protects her against possibly slower rates (see e.g. Antoine and Renault (2009) and Newey and Windmeijer (2009)).

The related literature, albeit too vast to be properly summarized here, can be classified as follows. As already mentioned, several authors rather recommend inference procedures that are robust to weak identification as an alternative to pretesting. While this robust approach includes several important contributions (see e.g. Moreira (2003), Kleibergen (2005), Guggenberger, Kleibergen, Mavroeidis and Chen (2013, references therein and the survey by Andrews and Stock (2007)), we consider here as only benchmark the robust procedure proposed by Stock and Wright (2000), that is indeed an extension of the seminal work of Anderson and Rubin (1949). The pretesting literature can be divided into two categories, depending whether the null under test is poor identification (as in this paper) or on the contrary strong identification. The testing procedures proposed by Staiger and Stock (1997) and Stock and Yogo (2005) belong to the first category while Hahn and Hausman (2002) as well as Inoue and Rossi (2011), have rather considered the second strategy.

In our Monte Carlo study, we consider the linear IV regression model (with or without conditional heteroskedasticity) as well as a (persistent) AR(1) model stemming from a continuous time Ornstein-Uhlenbeck process and calibrated to interest rate data. Infill asymptotic (with time interval between consecutive observations going to zero) then provides an interesting new example of nearly-weak identification. In the linear model, we compare the performance of our pretesting procedure to Staiger and Stock's (1997) and Stock and Yogo's (2005). We also revisit the empirical application of Yogo (2004) (estimation of the Elasticity of Intertemporal Substitution from Euler equations) and consider the test recently developed by Montiel Olea and Pflueger (2012) to account for heteroskedasticity, autocorrelation and clustering.

The paper is organized as follows. In section 2, we introduce our framework and characterize the identification strength of structural parameters through the asymptotic behavior of the Jacobian of the moment restrictions. We also show how it is linked to the asymptotic prop-

erties of GMM estimators. In section 3, we propose three tests to assess the identification strength of the structural parameters in linear, separable and general settings respectively. In section 4, we illustrate the finite sample performance of our tests through Monte-Carlo simulations. We consider the linear IV regression model and a (persistent) AR(1) model calibrated to interest rate data. In section 5, we apply our procedures to the estimation of the Elasticity of Intertemporal Substitution. Section 6 concludes. The proofs of the theoretical results and the tables of empirical results are gathered in the Appendix in the Supplemental Material.

The following notation is used throughout the paper. The symbols "$\xrightarrow{p}$" and "$\xrightarrow{d}$" denote convergence in probability and in distribution, while "`Plim`" denotes the probability limit of a random expression. $o_p(1)$ denotes a random variable that converges to 0 in probability, whereas $\mathcal{O}_p(1)$ denotes a random variable that is bounded in probability. For any (k, p)-matrix M, "M'" denotes the transpose matrix of M, `Rank`(M) denotes the rank of M, and $\|M\| \equiv \max\{\sqrt{\lambda}, \lambda \text{ is an eigenvalue of } M'M\}$. $\mathbf{I}_q$ denotes the identity matrix of size q. $\chi^2(k)$ denotes the central chi-square random variable with k degrees of freedom. "With respect to" is written "w.r.t".

2 General framework

2.1 Identification strength

We consider the true unknown value θ^0 of the parameter $\theta \in \Theta \subset \mathbb{R}^p$ defined as the solution of the moment conditions,

$$E[\phi_t(\theta)] = 0 \quad \text{for some known function } \phi_t(.) \text{ of size } K. \tag{2.1}$$

Since the seminal work of Stock and Wright (2000), the weakness of the moment conditions (or instrumental variables) is usually captured through a drifting DGP such that the informational content of the estimating equations shrinks towards zero (for all θ) while the sample size T grows to infinity. The strength of identification of the parameters is then reflected by the Jacobian of the moment equations with respect to the parameters. We maintain the assumptions that the moment function $\phi_t(.)$ is continuously differentiable with

respect to θ on the interior of the set of possible parameter values Θ, $\texttt{int}(\Theta)$, and that the true unknown value θ^0 belongs to $\texttt{int}(\Theta)$. We now unify several definitions of identification strength of θ that have been used in the literature.

Definition 2.1. *(Identification strength of θ)*
The identification strength of θ is characterized by a sequence M_T of deterministic nonsingular matrices of size p such that

$$\Gamma(\theta^0) \equiv \texttt{Plim}\left[\frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'} M_T\right] \qquad \textit{exists and is full-column rank.} \tag{2.2}$$

We borrow the terminology "identification strength" to Kleibergen and Mavroeidis (2009) who set the focus (see their Assumption 6) on the special case where

$$\frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'} = \Gamma(\theta^0) M_T^{-1} .$$

They stress the importance of characterizing the identification strength of θ to draw valid inference about some other parameters of the model of interest. The faster the sequence of matrices M_T diverges to infinity, the lesser θ is identified. It is actually strongly identified when M_T can be taken as the identity matrix. The concept of identification strength has been extensively studied in Antoine and Renault (2009, 2010). In the context of many instruments, it is revisited in Assumption 1(ii) in Newey and Windmeijer (2009). An important message of all these papers is that different linear combinations of θ may display different strengths (or degrees) of identification. More generally, the identification strength of the possible linear combinations of θ is tightly related to the rate of convergence of the eigenvalues of $(M_T M_T')$ to infinity, while the eigenvectors describe the linear combinations corresponding to different degrees of identification (see Antoine and Renault (2009, 2010) and assumption 4 below). Qu (2011) uses the same concept in the context of Maximum Likelihood inference.
The role of the sequence of matrices M_T in the asymptotic distributional theory of the GMM estimator of θ is well-understood, at least in the linear case, as first pointed out by Staiger and Stock (1997) and reminded in the example below.

Example 2.1. *(Linear IV regression)*

We consider the following structural linear equation,

$$y_t = x_t'\theta^0 + u_t \qquad \text{for } t = 1, \cdots, T,$$

where the p explanatory variables x_t may be endogenous. The true unknown value θ^0 of the structural parameter is identified through $K \geq p$ instrumental variables z_t uncorrelated with u_t. In other words, the estimating equations for standard IV estimation are

$$\overline{\phi}_T(\hat{\theta}_T) = \frac{1}{T} Z' \left(y - X\hat{\theta}_T \right) = 0\,, \tag{2.3}$$

where X (respectively Z) is the (T, p) (respectively (T, K)) matrix which contains the available observations of the p explanatory variables (respectively the K instrumental variables) and $\hat{\theta}_T$ denotes the standard IV estimator of θ. The reduced form equation for X can be written as

$$X = Z\Pi_T + V\,, \tag{2.4}$$

where the K columns of Z and the p columns of V are uncorrelated. Note that, at the price of more tedious notations, one could easily accommodate the general model considered in Staiger and Stock (1997) where additional exogenous variables W show up in both the structural and the reduced form equation. Actually, everything can be understood in the more general setting by considering orthogonal projections on the orthogonal space of the range of W. Then, we have

$$\frac{\partial \overline{\phi}_T(\theta)}{\partial \theta'} = -\frac{Z'X}{T} = -\frac{Z'Z}{T}\Pi_T - \frac{Z'V}{T}\,. \tag{2.5}$$

Under standard regularity conditions, $(Z'V/T)$ and $(Z'Z/T)$ converge respectively towards zero and a nonsingular matrix Σ_Z. Therefore, (2.5) can be used to reinterpret the above definition 2.1 in terms of an asymptotic specification of the matrix Π_T,

$$\Pi_T = \Pi M_T^{-1} \quad \text{with} \quad \texttt{Rank}(\Pi) = p\,. \tag{2.6}$$

In other words, instead of a fixed full-column rank matrix Π, drifting reduced form parameters Π_T are used to capture the fact that identification may be weaker than usual (e.g. when

coefficients of M_T go to infinity). Staiger and Stock (1997) actually define weak identification by considering $\Pi_T = \Pi/\sqrt{T}$, that is $M_T = \sqrt{T}\mathbf{I}_p$. The conformity between condition (2.6) and definition 2.1 follows from (2.5) by noting that we have for all $\theta \in \Theta$,

$$\frac{\partial \overline{\phi}_T(\theta)}{\partial \theta'} M_T = -\frac{Z'Z}{T}\Pi - \frac{Z'V}{\sqrt{T}}\frac{M_T}{\sqrt{T}}\,.$$

In the weak identification case ($M_T/\sqrt{T} = \mathbf{I}_p$), an extension of definition 2.1 would lead to consider a random matrix $\Gamma(\theta^0)$ since the Jacobian matrix rescaled by $\sqrt{T}$ is asymptotically normal as $Z'V/\sqrt{T}$. The effects of this randomness have been documented by Kleibergen (2005) for a score test on the whole parameter vector. By contrast, such randomness is implicitly precluded in Kleibergen and Mavroeidis (2009) (see their Assumption 6) for the Jacobian matrix of parameters not under test. The only way to ensure that the matrix $\Gamma(\theta^0)$ in definition 2.1 is not random is to assume, in addition, that

$$\lim_T \left(\frac{M_T}{\sqrt{T}}\right) = 0\,. \tag{2.7}$$

Condition (2.7) has been dubbed near-weak identification by Hahn and Kursteiner (2002) who typically consider

$$M_T = T^\lambda \mathbf{I}_p \quad \text{with} \quad 0 < \lambda < 1/2\,.$$

The extreme cases $\lambda = 0$ and $\lambda = 1/2$ correspond respectively to strong and weak identification. Precluding the extreme case of weak identification, or, in other words, maintaining the rank condition (2.6) with the upper bound (2.7) on the rate of weakness is key to get asymptotic normality of the IV estimator $\hat{\theta}_T$ with standard studentized statistics. To see this, simply rewrite (2.3) as

$$\begin{aligned} & \frac{Z'X}{T}(\hat{\theta}_T - \theta^0) = \frac{Z'u}{T} \\ \Leftrightarrow \quad & \frac{Z'Z}{T}\Pi\sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) + \frac{Z'V}{\sqrt{T}}\frac{M_T}{\sqrt{T}}\sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) = \frac{Z'u}{\sqrt{T}}\,. \end{aligned} \tag{2.8}$$

Under standard regularity conditions, (2.8) delivers asymptotic normality of

$$\sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0)\,,$$

after noting that, thanks to (2.7), we have

$$\Sigma_Z \Pi \sqrt{T} M_T^{-1} (\hat{\theta}_T - \theta^0) = \frac{Z'u}{\sqrt{T}} + o_P(1)\,, \tag{2.9}$$

with $\Sigma_Z \Pi$ full-column rank and $Z'u/\sqrt{T}$ asymptotically normal. Of course, since near-weak identification entails some coefficients of the matrix M_T diverging to infinity (albeit not as fast as $\sqrt{T}$), the rate of convergence of the IV estimator to normality may be slower than $\sqrt{T}$. The fact that the matrix M_T may not be proportional to the identity matrix (and not even diagonal) allows for linear combinations of θ to have different degrees of identification. Assuming M_T diagonal means that different identification strengths are assigned to different columns of $Z\Pi$, and not to different instruments (or columns of Z). The two are not equivalent in the overidentified case since Π is not a square matrix. Therefore, maintaining the diagonality of M_T, albeit commonly done, would be overly restrictive[5].

Assumption 1 in Newey and Windmeijer (2009) highlights the importance of the near-weak identification condition (see their condition $\mu_{jn}/\sqrt{n} \to 0$; see also Hansen, Hausman and Newey (2008) for the linear case). As already explained, this assumption allows, at least in the linear case, to get consistent asymptotically normal estimators whose rates of convergence are described by the sequence of matrices $\sqrt{T} M_T^{-1}$ (see (2.8)). The non-linear case for near-weak identification, as studied by Antoine and Renault (2009, 2010) and Caner (2010) works similarly. It can be shown under very general conditions (see Antoine and Renault (2009, 2010)) that any GMM estimator of θ will display a rate of convergence at least equal to $\sqrt{T}/\left\|M_T\right\|$. It is worth stressing that while we allow moments to display some singularities, the GMM estimators we consider are all defined in a standard way from a positive definite weighting matrix.

Definition 2.2. *(GMM estimator)*
For any sequence of possibly random symmetric matrices Ω_T of size K that converges in probability towards a positive definite matrix Ω, a GMM estimator $\hat{\theta}_T$ is defined as any solution of

$$\min_{\theta \in \Theta} \left[\overline{\phi}_T(\theta)' \Omega_T \overline{\phi}_T(\theta) \right] .$$

[5]See further assumption 4 for a convenient generalization of diagonality.

As already mentioned, based on Antoine and Renault (2009, 2012), we will admit throughout that the condition $\lim_T(M_T/\sqrt{T}) = 0$ is sufficient to ensure consistency of any GMM estimator. Regarding asymptotic normality of such GMM estimators, the proof requires a Taylor expansion of the first-order conditions to get a linear representation of the GMM estimator that generalizes the linear case (2.9). Non-linearity may then entail an additional technical difficulty due to the fact that the concept of identification strength may not be robust to plugging in the Jacobian matrix a consistent estimator of the true unknown value. This is why we consider the following high-level condition that strengthens definition 2.1.

Definition 2.3. *(Near-weak identification)*
In the context of definition 2.1, θ is said near-weakly identified if there exists a sequence M_T of deterministic nonsingular matrices of size p such that

$$\lim_T \left(\frac{M_T}{\sqrt{T}} \right) = 0 \,,$$

and, for any GMM estimator $\hat{\theta}_T$ as in definition 2.2 and any sequence θ_T between θ^0 and $\hat{\theta}_T$ component by component[6]*, we have*

$$\Gamma(\theta^0) = \texttt{Plim} \left[\frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'} M_T \right] ,$$

where $\Gamma(\theta^0)$ is the full-column rank matrix introduced in (2.2).

When definition 2.1 is fulfilled with $M_T/\sqrt{T}$ going to zero, near-weak identification of θ will be warranted in many cases. It is worth realizing that going from the former to the latter only amounts to assume that

$$\theta_T \in [\theta^0, \hat{\theta}_T] \quad \Rightarrow \quad \texttt{Plim} \left[\left(\frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'} - \frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'} \right) M_T \right] = 0 \,. \tag{2.10}$$

The required zero-limit in (2.10) is obviously ensured for the components of the moment vector $\overline{\phi}_T(.)$ that are linear with respect to the parameters θ. For those which are not linear with respect to some subset θ_1 of components of θ, the issue at stake is to know whether their

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

rate of convergence along the sequence θ_T is sufficient to supersede the possible convergence to infinity of the sequence M_T. As explained above, we expect for the GMM estimator $\hat{\theta}_T$ (and thus also for $\theta_T \in [\theta^0, \hat{\theta}_T]$) a rate of convergence at least equal to $\sqrt{T}/\|M_T\|$. More precisely, we expect, as in the linear case, that $\sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) = \mathcal{O}_P(1)$. Hence, the validity of (2.10) would mean that in relevant directions, the convergence to zero of $M_T/\sqrt{T}$ dominates the convergence to infinity of the sequence M_T. Roughly speaking, $\|M_T\|$ should not blow-up as fast as $T^{1/4}$. This threshold is the key concept of near-strong identification promoted by Antoine and Renault (2009, 2010) as a sufficient condition for near-weak identification. A less restrictive, albeit related, point of view will be warranted in section 3 when testing for the identification strength of subvectors θ in non-linear settings. We first present the asymptotic theory for GMM estimators under the high-level assumption of near-weak identification. While a careful study of rates of convergence of GMM estimators is provided in Antoine and Renault (2009, 2010), we simplify the exposition by directly maintaining the required high-level assumptions.

Assumption 1. *θ is near-weakly identified as in definition 2.3.*

Assumption 2. *In the context of assumption 1, any GMM estimator $\hat{\theta}_T$ as in definition 2.2 is such that $\sqrt{T}M_T^{-1}(\hat{\theta}_T - \theta^0) = \mathcal{O}_P(1)$.*

2.2 Asymptotic Theory

As usual, asymptotic normality of a GMM estimator results from a central limit theorem applied to the moment conditions evaluated at the true unknown value of the parameters.

Assumption 3. *$\sqrt{T}\bar{\phi}_T(\theta^0)$ converges in distribution towards a normal distribution with mean zero and variance $S(\theta^0)$.*

The following theorem extends the asymptotic normality result given in the linear case, as well as results previously given in Antoine and Renault (2009, 2010).

Theorem 2.1. *(Asymptotic normality)*
Let $\hat{\theta}_T$ denote any GMM estimator as in definition 2.2. Under assumptions 1 to 3,

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

is asymptotically normal with mean zero and variance

$$\Sigma(\theta^0) = \left[\Gamma'(\theta^0)\Omega\Gamma(\theta^0)\right]^{-1}\Gamma'(\theta^0)\Omega S(\theta^0)\Omega\Gamma(\theta^0)\left[\Gamma'(\theta^0)\Omega\Gamma(\theta^0)\right]^{-1}.$$

As already acknowledged, assumptions 1 and 2 are high-level assumptions, and we refer the interested reader to Antoine and Renault (2010) for more primitive conditions. In any case, Theorem 2.1 paves the way for a concept of efficient estimation. By a common argument, the unique limit weighting matrix Ω minimizing the above covariance matrix is clearly $\Omega = [S(\theta^0)]^{-1}$.

Theorem 2.2. *(Efficient GMM estimator)*
Under the assumptions of Theorem 2.1, any GMM estimator $\hat{\theta}_T$ as in definition 2.2 with a weighting matrix $\Omega_T = S_T^{-1}$, where S_T denotes a consistent estimator of $S(\theta^0)$, is such that

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

is asymptotically normal with mean zero and variance $[\Gamma'(\theta^0)S^{-1}(\theta^0)\Gamma(\theta^0)]^{-1}$.

In our framework, the terminology efficient GMM must be carefully qualified. For all practical purposes, Theorem 2.2 states that, 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

$$M_T^{-1}\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}M_T^{-1\prime}, \tag{2.11}$$

since $\Gamma(\theta^0) = \texttt{Plim}\left[\frac{\partial\overline{\phi}_T(\theta^0)}{\partial\theta'}M_T\right]$. However, it is incorrect to deduce from formula (2.11) that $\sqrt{T}\left(\hat{\theta}_T - \theta^0\right)$ can be seen (for T large enough) as a Gaussian vector with mean zero and variance consistently estimated by

$$\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}. \tag{2.12}$$

The above matrix (2.12) is actually the inverse of an asymptotically singular matrix. In this sense, a truly standard GMM theory does not apply and some components of $\sqrt{T}(\hat{\theta}_T - \theta^0)$ actually blow-up. Quite fortunately, standard inference procedures work, albeit for

non-standard reasons. For all practical purposes related to inference about the structural parameter θ, the knowledge of the matrix M_T is not required; see also the discussion in Antoine and Renault (2010). In particular, the J-test for overidentification can be performed as usual as stated in the following result.

Theorem 2.3. *(J-test)*
Under the assumptions of Theorem 2.2, for any GMM estimator as in definition 2.2 with a weighting matrix $\Omega_T = S_T^{-1}$, where S_T denotes a consistent estimator of $S(\theta^0)$, we have

$$T\overline{\phi}_T'(\hat{\theta}_T)S_T^{-1}\overline{\phi}_T(\hat{\theta}_T) \xrightarrow{d} \chi^2(K-p)\,.$$

Of course, the J-test is a "black box" that conceals the fact that the quality of identification is heterogenous. The asymptotic singularity of (2.12) means that the actual rate of convergence of the GMM estimator may vary depending on the linear combinations of the structural parameter vector θ. The following high-level assumption[7] helps to characterize the relevant directions in the parameter space.

Assumption 4. *The sequence of matrices M_T such that assumptions 1 and 2 are fulfilled can be chosen as*

$$M_T = R\Lambda_T\,,$$

for some fixed non-singular matrix R and a sequence Λ_T of diagonal matrices.

As shown in the appendix, the main intuition is that, from an initial sequence of matrices M_T that does not fulfill assumption 4, we can build the diagonal coefficients of Λ_T as the singular values of the matrix M_T (as the square-roots of eigenvalues of $(M_T M_T')$), while the matrix R is the limit of a sequence of orthogonal matrices of eigenvectors of $(M_T M_T')$. Then, Theorems 2.1 and 2.2 characterize the asymptotic normal distribution of $\sqrt{T}\Lambda_T^{-1}R^{-1}(\hat{\theta}_T - \theta^0)$. In other words, when considering the reparametrization $\eta \equiv R^{-1}\theta$, the j-th component of $\hat{\eta}_T \equiv R^{-1}\hat{\theta}_T$ is a consistent asymptotically normal estimator of η_j^0 (with $\eta^0 \equiv R^{-1}\theta^0$) with a rate of convergence $\sqrt{T}/\lambda_{jT}$ (with λ_{jT} the j-th diagonal coefficient of Λ_T). Moreover, Antoine and Renault (2010) have shown that this rate of convergence is not impacted by a preliminary consistent estimation of the matrix R, making this asymptotically normal

[7] More primitive justifications are provided in the appendix

estimation feasible. The characterization of the different rates of convergence through the sequence of diagonal coefficients of the matrix Λ_T may matter for interpretation, even though their knowledge is not necessary to run Wald inference from (2.11).

2.3 Identification of subvectors

When testing for identification strength, we will check whether our data set allows us to reject the null hypothesis that some components of the vector θ of structural parameters are (very) poorly identified. Throughout, θ_1 denotes a vector of p_1 components of θ, while $\theta_{\setminus 1}$ collects the p_2 $(= p - p_1)$ remaining components of θ that are not included in θ_1. For simplicity, θ_1 corresponds hereafter to the first p_1 components of θ, that is $\theta = (\theta_1' \ \theta_{\setminus 1}')'$. Typically, we consider cases where the econometrician's is concerned about the poor identification of the components of $\theta_{\setminus 1}$ while prior knowledge warrants "sufficiently strong" identification of θ_1.[8]

Note that it is only when a sequence of matrices M_T characterizing the identification strength of θ is block-diagonal,

$$M_T = \begin{bmatrix} M_{1T} & \mathbf{0} \\ \mathbf{0} & M_{\setminus 1T} \end{bmatrix}, \tag{2.13}$$

that we can deduce the identification strengths of θ_1 and $\theta_{\setminus 1}$ from the identification strength of θ. A well-known example is the setting put forward by Stock and Wright (2000) where the two subsets of components of θ are disentangled as follows[9],

$$\begin{aligned} &E\overline{\phi}_T(\theta) = E\overline{\phi}_{1T}(\theta_1) + \frac{1}{T^\lambda} E\overline{\phi}_{2T}(\theta), \quad \text{with} \ \ 0 < \lambda \leq 1/2\,, \\ &\texttt{Rank}\left(E\frac{\partial \overline{\phi}_{1T}(\theta_1^0)}{\partial \theta_1'} \right) = p_1 \quad \text{and} \quad \texttt{Rank}\left(E\frac{\partial \overline{\phi}_{2T}(\theta^0)}{\partial \theta_{\setminus 1}'} \right) = p_2\,. \end{aligned}$$

M_T can then be defined as in (2.13) with $M_{1T} = \mathbf{I}_{p_1}$ and $M_{\setminus 1T} = T^\lambda \mathbf{I}_{p_2}$.

It is worth pointing out that, up to a convenient reparameterization, the above block-diagonal structure of (2.13) is not really restrictive. From assumption 4, we define the new vector of parameters, $\eta \equiv R^{-1}\theta$ whose identification strength (in the sense of definition 2.1) is

[8] Note that our test does not exclude the case where $\theta_{\setminus 1} = \theta$.

[9] Strictly speaking, Stock and Wright (2000) only consider the limit case with $\lambda = 1/2$.

described through the sequence of diagonal matrices Λ_T. Hence, the maintained assumption (2.13) simply means that such a reparameterization is possible with a block-diagonal matrix R. It makes sense to question the identification strength of $\theta_{\backslash 1}$ while maintaining a (near-weak) identification assumption on θ_1, precisely because the two subvectors are disentangled in the classification of directions as regards identification strength.

Our tests for identification strength will provide some (partial) answers to the following question: taking for granted that θ_1 is near-weakly identified, do our data confirm the identification of a larger vector of unknown parameters? Our null hypothesis will be devised such that, failing to reject it means that we cannot rely upon standard inference based on the Gaussian asymptotic theory of section 2.2 when any of the parameters in $\theta_{\backslash 1}$ are considered as unknown. In front of such a negative evidence, only two strategies are available:

- either 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;

- or, following the common practice of calibration, one may fix the value of parameters in $\theta_{\backslash 1}$ 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 θ_1. 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 $\theta_{\backslash 1}$ is fixed at its true unknown value $\theta^0_{\backslash 1}$, the remaining moment problem is well-behaved.

Definition 2.4. *(Near-weak identification of a subvector)*
In the context of definition 2.1, with the block-diagonal structure (2.13) for the sequence of matrices M_T, θ_1 is near-weakly identified if the two following conditions are fulfilled.
(i) Assumptions 1 and 2 are fulfilled for the sequence of matrices M_{1T} in the context of the (infeasible) moment model

$$E\left[\phi_t(\theta_1, \theta^0_{\backslash 1})\right] \;=\; 0 \quad \textit{with} \quad \theta_1 \in \Theta(\theta^0_{\backslash 1}) = \left\{\theta_1 \in \mathbb{R}^{p_1}; (\theta_1, \theta^0_{\backslash 1}) \in \Theta\right\} .$$

(ii) For any GMM estimator $\hat{\theta}_T$ as in definition 2.2 and any sequence θ^*_T such that $\theta^*_{1T} = \hat{\theta}_{1T}$

and $\theta^*_{\backslash 1T} - \hat{\theta}_{\backslash 1T} = o_P(T^{-1/4})$, *we have*

$$\frac{\partial \overline{\phi}_T(\theta^*_T)}{\partial \theta'_{\backslash 1}} M_{\backslash 1T} = \mathcal{O}_p(1)\,.$$

When wondering whether a parameter vector that strictly nests θ_1 is near-weakly identified, let us recall that the key issue is to check that the convergence condition (2.10) holds for any sequence θ_T between the true value θ^0 and some GMM estimator $\hat{\theta}_T$, that is

$$\texttt{Plim}\left[\left(\frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'} - \frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'}\right) M_T\right] = 0\,. \tag{2.14}$$

The reason why the maintained assumption $\lim_T(M_T/\sqrt{T}) = 0$ may not be sufficient to get (2.14) is that a rate of convergence for θ_T strictly slower than $\sqrt{T}$ may be unable to protect against the asymptotic blow-up of the sequence M_T. This issue will obviously be easier to control for, when the weakest parameters $\theta_{\backslash 1}$ will not be multiplied by the most explosive part of the sequence M_T, namely $M_{\backslash 1T}$. This explains why we consider first the most favorable circumstances of linearity with respect to $\theta_{\backslash 1}$.

• Case i): Moment conditions affine w.r.t. $\theta_{\backslash 1}$, $\phi_t(\theta) = a_t(\theta_1) + B_t(\theta_1)\theta_{\backslash 1}$.

Regarding condition (ii) of definition 2.4, note that we have

$$\frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'} M_T \;=\; \left[\begin{array}{cc} \frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'_1} M_{1T} & \bar{B}_T(\theta_{1T}) M_{\backslash 1T} \end{array}\right]. \tag{2.15}$$

To get the second block of (2.15) consistent with (2.14) when the only assumption is that $\lim_T(M_{\backslash 1T}/\sqrt{T}) = 0$, θ_{1T} has to be $\sqrt{T}$-consistent, that is $M_{1T} = \mathbf{I}_{p_1}$. In this case, since we know $\theta_{\backslash 1T}$ consistently estimates $\theta_{\backslash 1}$, it is clear that a continuity assumption on $\frac{\partial \overline{a}_T(.)}{\partial \theta'_1}$ and $\frac{\partial \overline{B}_T(.)}{\partial \theta'_1}$ will be sufficient to deduce (2.14) from (2.15). Overall, up to some regularity conditions, we can conclude the following.

Golden rule for the test of identification strength in the linear case:
If the moment conditions are affine w.r.t. a subvector $\theta_{\backslash 1}$, strong identification of the other components θ_1 ($M_{1T} = \mathbf{I}_{p_1}$) jointly with ($\lim_T(M_{\backslash 1T}/\sqrt{T}) = 0$) is sufficient to warrant near-weak identification of the whole vector θ. Note that when the moment conditions are affine w.r.t. the whole vector θ, no strong identification condition is required.

• Case ii): Separable (non-linear) moment conditions *a la Stock and Wright* (see Assumption C p1061),

$$E\overline{\phi}_T(\theta) = E\overline{\phi}_{1T}(\theta_1) + \frac{1}{T^\lambda}E\overline{\phi}_{2T}(\theta)\,.$$

In this framework, θ_1 is strongly identified, while $\theta_{\setminus 1}$ may be weakly identified. Note that compared to Stock and Wright's original setup, we allow λ to take values between 0 and 1/2 since our interest lies in testing whether $\theta_{\setminus 1}$ is near-weakly identified or not. In such setup, the matrix M_T is block-diagonal with the first block $M_{1T} = \mathbf{I}_{p_1}$, and the second block $M_{\setminus 1T} = T^\lambda \mathbf{I}_{p_2}$. Regarding condition (ii) of definition 2.4, note that we have

$$E\frac{\partial\overline{\phi}_T(\theta_T)}{\partial\theta'}M_T \;=\; \left[\; E\frac{\partial\overline{\phi}_{1T}(\theta_{1T})}{\partial\theta_1'} + \frac{1}{T^\lambda}E\frac{\partial\overline{\phi}_{2T}(\theta_T)}{\partial\theta_1'} \quad E\frac{\partial\overline{\phi}_{2T}(\theta_T)}{\partial\theta_{\setminus 1}'} \;\right]. \tag{2.16}$$

Hence, it is clear that a continuity and a rank assumptions of $\frac{\partial\overline{\phi}_{1T}(\theta_{1T})}{\partial\theta_1'}$ (and $\frac{\partial\overline{\phi}_{2T}(\theta_{1T})}{\partial\theta_{\setminus 1}'}$) will be sufficient to get (2.14) when the only assumption is $\lim_T(M_{\setminus 1T}/\sqrt{T}) = 0$, that is $\lambda < 1/2$. Overall, up to some regularity conditions, we can conclude the following:

> **Golden rule for the test of identification strength in the separable case:**
> If the moment conditions are separable w.r.t. a subvector $\theta_{\setminus 1}$, strong identification of the other components θ_1 ($M_{1T} = \mathbf{I}_{p_1}$) jointly with $\lim_T(M_{\setminus 1T}/\sqrt{T}) = 0$, that is $\lambda < 1/2$, is sufficient to warrant near-weak identification of the whole vector θ.

• Case iii): General (non-separable and non-linear) moment conditions.
Losing linearity w.r.t. $\theta_{\setminus 1}$ implies that the second block of (2.15) is now $\left(\frac{\partial\overline{\phi}_T(\theta_T)}{\partial\theta_{\setminus 1}'}\right) M_{\setminus 1T}$, with $\left(\frac{\partial\overline{\phi}_T(\theta_T)}{\partial\theta_{\setminus 1}'}\right)$ that usually depends on the estimator of the weakest parameters $\theta_{\setminus 1T}$. Since these parameters are only consistent at a rate $\left\|M_{\setminus 1T}\right\|/\sqrt{T}$, a Taylor expansion of $\left(\frac{\partial\overline{\phi}_T(\theta_T)}{\partial\theta_{\setminus 1}'}\right)$ (assuming ϕ twice continuously differentiable) will allow us to prove (2.14) only if $\left\|M_{\setminus 1T}\right\|^2/\sqrt{T}$ goes to zero when T goes to infinity. In other words, the condition to ensure that the poor identification of $\theta_{\setminus 1}$ does not impair near-weak identification of the whole vector θ is that $\left\|M_{\setminus 1T}\right\| = o(T^{1/4})$, or equivalently that the rate of convergence of all parameters in $\theta_{\setminus 1}$ (rate defined by the sequence of matrices $M_{\setminus 1T}/\sqrt{T}$) is more than $T^{1/4}$. As already emphasized by

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. Even without an infinite dimensional issue, we intuitively want to make sure that second-order terms in Taylor expansions (see the discussion above regarding a Taylor expansion of $\left(\frac{\partial \overline{\phi}_T(\theta_T)}{\partial \theta'_{\setminus 1}}\right)$) remain negligible in front of first-order terms. The fact that this condition is a byproduct of second-order Taylor expansions explains why no threshold like $T^{1/4}$ pops up in the linear case. In the non-linear case, the rule will then be as follows.

> **Golden rule for the test of identification strength in the non-separable and non-linear case:**
> When the subvector θ_1 is near-weakly identified, but the moment conditions are neither affine nor separable w.r.t. the complementary subvector $\theta_{\setminus 1}$, considering any linear combination of $\theta_{\setminus 1}$ as unknown (in addition to θ_1) may impair global near-weak identification except if we assume that this linear combination is consistently estimated at a rate faster than $T^{1/4}$.

Following Antoine and Renault (2009), this latter property will be dubbed near-strong identification of the associated linear combination.

3 Testing identification strength

In this section, we are interested in assessing the identification strength of the structural parameter in order to detect weaker patterns of identification. Staiger and Stock (1997) propose a rule of thumb to detect weak instruments, whereas Stock and Yogo (2005) propose a formal characterization of the weakness of instruments based on the 2SLS bias as well as on the size of associated tests. Both Staiger and Stock (1997) and Stock and Yogo (2005) consider the null hypothesis that the instruments are weak, even though the parameters might be identified. Following these pioneer papers, we design two specifications tests which correspond respectively to the three golden rules stated in section 2. The null hypotheses will be designed in a way such that, failing to reject the null means that we have no sufficiently compelling evidence to trust an assumption of global near-weak identification. Accordingly,

no standard asymptotic theory based on asymptotic normality is available and the researcher may resort to identification-robust procedures. Note also that both tests set the focus on the identification of subvector which is in contrast with existing procedures.
Even though they apply to two different settings (one with linearity or separability w.r.t. the parameters under test, one without any linearity or separability assumption at the cost of contemplating faster rates of convergence), the two testing strategies share a common structure: both amount to a conservative J-test questioning the rate of convergence of a given GMM estimator. This is the reason why we first build a general theoretical framework before discussing the feasibility of our tests in two more practically oriented subsections.

3.1 Theoretical framework

Throughout section 3, null hypotheses under test are about the rate of convergence of a subset $\hat{\theta}_{\setminus 1T}$ of components of a given GMM estimator $\hat{\theta}_T$ defined according to definition 2.2. We maintain assumptions 2, 3, and 4, and in particular we know that

$$\sqrt{T} M_T^{-1} \left(\hat{\theta}_T - \theta^0 \right) = \mathcal{O}_P(1) \, .$$

Note however that we do not maintain assumption 1 of near-weak identification since it is precisely the focus of our interest. As done in section 2.3, we assume that only the subset θ_1 is near-weakly identified while the question is about the other parameters gathered in $\theta_{\setminus 1}$. In particular, the matrix M_T is block-diagonal and we do not know yet whether $\lim_T \left[M_{\setminus 1T} / \sqrt{T} \right] = 0$. We do not even know whether $\hat{\theta}_{\setminus 1T}$ is consistent.
To formulate a well-suited null hypothesis about the rate of convergence of $\hat{\theta}_{\setminus 1T}$, several remarks are in order.
(i) Following the practice that has been dominant since Staiger and Stock (1997), the null hypothesis reckons with the worse case scenario regarding the identification of $\theta_{\setminus 1}$, that is the rate of convergence of $\hat{\theta}_{\setminus 1T}$. Note however that our first extension w.r.t. the common practice is to set the focus on subvectors of θ. It should be clear that nothing prevents us from testing the identification of the whole parameter θ.
(ii) As stressed by the three golden rules of section 2, our worst case scenario regarding the rate of convergence a_T of $\hat{\theta}_{\setminus 1T}$ will be that, in the linear or separable case, it is not even infinite whereas in the non-linear case it is not greater than $T^{1/4}$.

(iii) As made explicit in the third golden rule, the worst case scenario of interest is actually that no linear combination of the parameters can be estimated at a satisfactory rate.
(iv) For a given GMM estimator $\hat{\theta}_T$ (and any given linear combination of $\hat{\theta}_{\backslash 1T}$), what really matters is not merely its rate of convergence but the rate of convergence of a well-suited subsequence. After all, a well-suited subsequence is able to properly identify the true unknown value of the linear combination of interest.
Therefore, for any real number $\nu \in [0, 1/2[$, we will generally consider null hypotheses of the following type:

> $H_0(\nu)$ *(No identification within* $\theta_{\backslash 1}$ *at rate faster than* ν*)*:
> For any subsequence of the estimator $\hat{\theta}_T$, for any deterministic sequence a_T such that $a_T/T^\nu \to \infty$, no non-zero linear combination of the subsequence $(\hat{\theta}_{\backslash 1T} - \theta^0_{\backslash 1})$ is $\mathcal{O}_P(1/a_T)$.

Note that, for sake of notational simplicity, we do not use an explicit notation like $\hat{\theta}_{m_T}$ for subsequences of $\hat{\theta}_T$. This abuse of notation will be maintained throughout. The key intuition for our proposed test of $H_0(\nu)$ comes from the following lemma, proved in the appendix.

Lemma 3.1. *(i) Under the null hypothesis* $H_0(\nu)$*, for any deterministic sequence* a_T *such that* $a_T/T^\nu \to \infty$*, we have*

$$\lim_T \left[\frac{\sqrt{T}}{a_T} M^{-1}_{\backslash 1T} \right] = 0 \, .$$

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

$$\lim_T \left\| \frac{\sqrt{T}}{a_T} M^{-1}_{\backslash 1T} \right\| = \infty \, .$$

As explained in the appendix, the required regularity condition amounts to the following application of Prohorov's theorem. By definition, under the alternative, we can find a deterministic sequence a_T with $a_T/T^\nu \to \infty$ such that, for some non-zero vector $\delta \in \mathbb{R}^{p_2}$, we have (for a well-suited subsequence)

$$a_T \delta' \left(\hat{\theta}_{\backslash 1T} - \theta^0_{\backslash 1} \right) = \mathcal{O}_P(1) \, ,$$

that is,

$$a_T \gamma'(\hat{\eta}_{\backslash 1T} - \eta^0_{\backslash 1}) = \mathcal{O}_P(1)\,,$$

for the non-zero vector $\gamma = R'\delta$. Then, since by our maintained assumption 2,

$$\sqrt{T}\Lambda^{-1}_{\backslash 1T}\left(\hat{\eta}_{\backslash 1T} - \eta^0_{\backslash 1}\right) = \mathcal{O}_P(1)\,,$$

Prohorov's theorem tells us that $(\hat{\eta}_{\backslash 1T} - \eta^0_{\backslash 1})$ (at least for a convenient subsequence) is endowed with an asymptotic distribution such that each component $\left(\hat{\eta}_{j,\backslash 1T} - \eta^0_{j,\backslash 1}\right)$ has a rate of convergence $\lambda_{j,\backslash 1T}/\sqrt{T}$ (with obvious notation for diagonal coefficients of $\Lambda_{\backslash 1T}$). Our regularity condition will amount to a non-degeneracy assumption about the joint limit distribution to ensure that $\gamma'(\hat{\eta}_{\backslash 1T} - \eta^0_{\backslash 1})$ does not go to zero at a rate faster than the minimal rate, $\min_j \left[\lambda_{j,\backslash 1T}/\sqrt{T}\right]$. Otherwise, the proposed test would have no power against the alternative defined by the linear combination $\delta = R'^{-1}\gamma$.

Lemma 3.1 allows us to characterize the behavior of moment conditions computed at a conveniently distorted value of the GMM estimator $\hat{\theta}_T$. This distortion will depend on a deterministic sequence a_T and on a direction $\delta \in \mathbb{R}^{p_2}$. More precisely, we define the distorted estimator $\hat{\theta}^{a,\delta}_T$ as

$$\hat{\theta}^{a,\delta}_{1T} = \hat{\theta}_{1T} \quad \text{and} \quad \hat{\theta}^{a,\delta}_{\backslash 1T} = \hat{\theta}_{\backslash 1T} + \frac{\delta}{a_T}\,.$$

Then, the proposed test will be based on the comparison of norms of moment conditions computed as

$$J_T(\Omega) = T\overline{\phi}'_T(\hat{\theta}_T)\Omega_T\overline{\phi}_T(\hat{\theta}_T) \quad \text{and} \quad J^{a,\delta}_T(\Omega) = T\overline{\phi}'_T(\hat{\theta}^{a,\delta}_T)\Omega_T\overline{\phi}_T(\hat{\theta}^{a,\delta}_T)\,.$$

where Ω_T is a sequence of symmetric matrices converging in probability towards a positive definite matrix Ω. Then, Lemma 3.1 allows us to show the following. As already explained through our three golden rules, the test of $H_0(\nu)$ will be especially relevant in the three following cases:

- Case i): Moment conditions affine w.r.t. θ_1 and $\nu = 0$;
- Case ii): Separable moment conditions and $\nu = 0$; this corresponds to $\lambda = 1/2$ as in Stock and Wright (2000);
- Case iii): General moment conditions and $\nu = 1/4$.

Corollary 3.2. *(i) Under the null hypothesis $H_0(\nu)$, in case i), ii) or iii) above, for any deterministic sequence a_T such that $a_T/T^\nu \to \infty$, we have for any $\delta \in \mathbb{R}^{p_2}$,*

$$\texttt{Plim}\left[J_T^{a,\delta}(\Omega) - J_T(\Omega)\right] = 0\,.$$

(ii) Assume that δ is drawn randomly according to some absolutely continuous probability distribution on $\mathbb{R}^{p_2}$. Then, under the alternative hypothesis to $H_0(\nu)$, with convenient regularity conditions, there exists a deterministic sequence a_T such that $a_T/T^\nu \to \infty$ such that, at least for a convenient subsequence,

$$\texttt{Plim}\left[J_T^{a,\delta}(\Omega)\right] = \infty\,. \tag{3.1}$$

The convenient regularity conditions, made explicit in the appendix, are not really restrictive. They are implied in particular by the assumption that the moment conditions are affine w.r.t. $\theta_{\setminus 1}$. The key intuition is that when $\lim_T \left\| \frac{\sqrt{T}}{a_T} M_{\setminus 1 T}^{-1} \right\| = \infty$ as in Lemma 3.1, we can be sure that $\lim_T \left\| \frac{\sqrt{T}}{a_T} M_{\setminus 1 T}^{-1} \delta \right\| = \infty$ for generically all directions δ. Then, the result (3.1) follows by standard Taylor expansions (up to unlikely singularities of the Jacobian matrix introduced by non-linearities w.r.t. $\theta_{\setminus 1}$), knowing that $\sqrt{T} M_T^{-1} \left(\hat{\theta}_T - \theta^0\right) = \mathcal{O}_p(1)$.

3.2 Detecting near-weak identification

In this subsection, we explain how to test weak identification in the three cases highlighted above: first, the cases where the moment conditions are either linear, or non-linear but separable; second, the general case with non-linear and non-separable moments. In all cases, we will discuss the detection of near-weak identification for the whole vector θ, as well as for the subvector $\theta_{\setminus 1}$ when the remaining subvector is either assumed near-weakly identified, or without assuming anything about it.

3.2.1 Linear or Separable cases

As explained by our first and second golden rules, the case where moment conditions are affine with respect to the parameter θ, or separable leads us to simply wonder whether some linear combinations of the parameters under test $\theta_{\setminus 1}$ can be consistently estimated. In other words, we want to test the null $\theta_{\setminus 1}$ is not identified, or only weakly identified (more precisely defined as $H_0(0)$ in section 3.1 above):

H_0: $\theta_{\setminus 1}$ is only weakly identified or not identified.

As explained in the former subsection, we consider a well-suited distortion of a GMM estimator $\hat{\theta}_T$. For sake of expositional simplicity, we will assume throughout this subsection that $\hat{\theta}_T$ has been computed with an "efficient" weighting matrix, that is:

$$\hat{\theta}_T = \arg\min_{\theta} \left[T\overline{\phi}_T'(\theta) S_T^{-1} \overline{\phi}_T(\theta) \right] ,$$

where S_T stands for a consistent estimator of $S(\theta^0)$. Note that the consistent estimation of $S(\theta^0)$ may not be feasible if we do not have at our disposal a first-step consistent estimator of θ^0. As shown in the subsection 3.3 below, this assumption can always be relaxed (even in the general case), at the price of a more involved approach.

For testing H_0, let us consider some deterministic sequence a_T such that $a_T \to \infty$. It will shortly become obvious that the slower the sequence a_T converges to infinity, the more powerful the resulting test will be; for instance, one may consider $a_T = \log(\log T)$, or an even slower sequence. As in the former subsection, this sequence a_T is used to build a distorted version $\hat{\theta}_T^{a,\delta}$ of the GMM estimator $\hat{\theta}_T$:

$$\hat{\theta}_{1T}^{a,\delta} = \hat{\theta}_{1T} \text{ and } \hat{\theta}_{\setminus 1T}^{a,\delta} = \hat{\theta}_{\setminus 1T} + \frac{\delta}{a_T} , \tag{3.2}$$

where δ is a given deterministic vector of size p_2. Our asymptotic conservative test for H_0 will be based on the corresponding distorted J-test statistic for overidentification

$$J_T^{a,\delta} = T\overline{\phi}_T'(\hat{\theta}_T^{a,\delta}) S_T^{-1} \overline{\phi}_T(\hat{\theta}_T^{a,\delta}) . \tag{3.3}$$

We can show the following result.

Theorem 3.3. *(Test of weak identification of $\theta_{\setminus 1}$ in the separable case)*
For an arbitrary choice of a deterministic sequence a_T such that $a_T \to \infty$ and of a vector $\delta \in \mathbb{R}^{p_2}$, we define two asymptotic tests with respective critical regions $W_T^{a,\delta}$ and $\tilde{W}_T^{a,\delta}$

$$W_T^{a,\delta} = \left\{ J_T^{a,\delta} > \chi^2_{1-\alpha}(K) \right\} \quad \textit{and} \quad \tilde{W}_T^{a,\delta} = \left\{ J_T^{a,\delta} > \chi^2_{1-\alpha}(K - p_1) \right\}$$

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 to 4, the test $W_T^{a,\delta}$ is asymptotically conservative at level α for the

null hypothesis H_0 of "weak identification within $\theta_{\setminus 1}$". The test $W_T^{a,\delta}$ is consistent against any alternative that makes the choice (a_T, δ) conformable to (3.1).
(ii) If we assume in addition that θ_1 is near-weakly identified, the test $\tilde{W}_T^{a,\delta}$ is also asymptotically conservative for the same null hypothesis H_0.

The test $\tilde{W}_T^{a,\delta}$ is obviously consistent whenever $W_T^{a,\delta}$ is. Note also that intermediate critical values $\chi^2_{1-\alpha}(K-q)$ with $0 < q < p_1$ can be considered after assuming that at least q components of θ_1 are near-weakly identified. Hence, in order to obtain a procedure that is less conservative, parameters known not to be at least weakly identified (e.g. the intercept) should not be included in $\theta_{\setminus 1}$ but rather in θ_1. It is also worth mentioning that nothing prevents us to test the whole vector θ.
Of course, the consistency claim above (based on equation (3.1)) may look somewhat tautological. The important point is to remember that Lemma 3.1 and Corollary 3.2 have shown that, under the alternative hypothesis, we are likely to be successful in our choice of the pair (a_T, δ). The key intuition is that under the alternative $\left\|\sqrt{T}M_{\setminus 1T}^{-1}\right\|$ goes to infinity. Our main task is to pin down a rate a_T strictly slower than the rate of divergence of $\left\|\sqrt{T}M_{1T}^{-1}\right\|$. Hence, the slower the sequence a_T converges to infinity, the more powerful the resulting test will be. In finite samples, the choice of this tuning parameter takes a data-based selection rule that will be described in subsection 3.3. Let us first explain why the above tests cannot be oversized asymptotically. We have shown in the former subsection that, under the null,

$$J_T^{a,\delta} - J_T = o_P(1)$$

where

$$J_T = T\overline{\phi}_T'(\hat{\theta}_T)S_T^{-1}\overline{\phi}_T(\hat{\theta}_T)$$

is the standard J-test statistic for overidentification. By definition, we have

$$J_T \leq T\overline{\phi}_T'(\overline{\theta}_T)S_T^{-1}\overline{\phi}_T(\overline{\theta}_T) \leq T\overline{\phi}_T'(\theta^0)S_T^{-1}\overline{\phi}_T(\theta^0)\,,$$

where $\overline{\theta}_T$ is the (infeasible) GMM estimator computed when the components of $\theta_{\setminus 1}$ are fixed at their true (unknown) value $\theta^0_{\setminus 1}$. Under the maintained assumption that θ_1 is near-weakly identified, we know by Theorem 2.3 that $\left[T\overline{\phi}_T'(\overline{\theta}_T)S_T^{-1}\overline{\phi}_T(\overline{\theta}_T)\right]$ converges in distribution

towards a chi-square distribution with $(K - p_1)$ degrees of freedom. Therefore, under the null,

$$\lim_T P\left(\tilde{W}_T^{a,\delta}\right) = \lim_T P\left(\left\{J_T > \chi^2_{1-\alpha}(K-p_1)\right\}\right) \leq \alpha\,,$$

and the second test is asymptotically conservative at level α as announced.
Without any assumption about the identification of θ_1, we know by Theorem 2.3 that $\left[T\overline{\phi}_T'(\theta^0)S_T^{-1}\overline{\phi}_T(\theta^0)\right]$ converges in distribution towards a chi-square distribution with K degrees of freedom. Therefore, under the null,

$$\lim_T P(W_T^{a,\delta}) = \lim_T P\left(\left\{J_T > \chi^2_{1-\alpha}(K)\right\}\right) \leq \alpha\,,$$

and at least the first test is asymptotically conservative at level α as announced.

3.2.2 General non-linear and non-separable case

As explained by our third golden rule, the general case of moment conditions that may not be separable with respect to the parameters under test $\theta_{\backslash 1}$ forces us to 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 (more precisely defined as $H_0(1/4)$ in section 3.1 above):

H_0 : No identification within $\theta_{\backslash 1}$ at rate faster than $T^{1/4}$.

Our testing procedure in the general case is quite similar to the procedure described in the previous section. However, for clarity, the following highlights the key elements of our general procedure. We consider a sequence a_T such that $a_T/T^{1/4} \to \infty$ and a deterministic vector δ to build a distorted version $\hat{\theta}_T^{a,\delta}$ of the GMM estimator $\hat{\theta}_T$ as in (3.2). Our asymptotic conservative test for H_0 is then based on the corresponding distorted J-test statistic $J_T^{a,\delta}$ as in (3.3). We refer the interested reader to section 3.2.1 and Appendix C. We can show the following result.

Theorem 3.4. *(Test of near-weak identification of $\theta_{\backslash 1}$ in the general case)*
For an arbitrary choice of a deterministic sequence a_T such that $a_T/T^{1/4} \to \infty$ and of a vector $\delta \in \mathbb{R}^{p_2}$, we define two asymptotic tests with respective critical region $W_T^{a,\delta}$ and $\tilde{W}_T^{a,\delta}$,

$$W_T^{a,\delta} = \left\{J_T^{a,\delta} > \chi^2_{1-\alpha}(K)\right\} \quad \textit{and} \quad \tilde{W}_T^{a,\delta} = \left\{J_T^{a,\delta} > \chi^2_{1-\alpha}(K-p_1)\right\}\,,$$

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 to 4, the test $W_T^{a,\delta}$ is asymptotically conservative at level α for the null hypothesis H_0 of "no identification within $\theta_{\setminus 1}$ at rate faster than $T^{1/4}$". The test $W_T^{a,\delta}$ is consistent against any alternative that makes the choice (a_T, δ) conformable to (3.1).
(ii) If we assume in addition that θ_1 is near-weakly identified, the test $\tilde{W}_T^{a,\delta}$ is also conservative for the same null hypothesis H_0.

It is worth realizing that the same comments apply to Theorems 3.3 and 3.4: intermediate critical values depending on the number of parameters for which near-weak identification is granted, wide range of alternatives against which consistency is warranted.

3.3 Additional considerations

3.3.1 Choosing the parameter δ through a data-dependent procedure

For a given well-suited sequence a_T, we know that the test $W_T^{a,\delta}$ (and possibly $\tilde{W}_T^{a,\delta}$) is asymptotically conservative and consistent with probability one when the vector δ is chosen randomly (with an absolutely continuous probability distribution). However, for the sake of finite sample performance, it clearly matters a lot to fine tune the length of the perturbation δ in due proportion of the accuracy of estimation of θ already allowed by the available sample of size T. We detail in Appendix A a procedure to choose δ appropriately. This search clearly sets the focus on the length of δ rather than on its direction. The user of the procedure may eventually want to add a random component, by picking the direction of δ randomly in the the unit sphere.

As detailed in Appendix A, the proposed procedure of elicitation of δ entails two steps of subsampling. First, we consider the empirical distribution of the GMM estimators of θ across all the subsamples of $\lfloor T^\kappa \rfloor$ consecutive observations[10] (with κ given between 0 and 1; e.g. $\kappa = 2$). The range of this distribution suggests a relevant grid of values of θ (and by solving equation (3.2) a grid of values of δ) for the purpose of Monte-Carlo study.

Second, we consider S subsamples of $\lfloor T^{\kappa^*} \rfloor$ consecutive observations (where κ^* may or may not coincide with the aforementioned κ). For each value of δ in the grid and for each

[10] $\lfloor T^\kappa \rfloor$ refers to the largest integer below T^κ. We consider consecutive observations to accommodate possible serial dependencies.

subsample s (with $s = 1, \cdots, S$), we compute the corresponding distorted J-test statistic for overidentification as in equation (3.3). As a result, for each grid point δ_m, we obtain a cross-sectional distribution of the test statistic (3.3), say $(J^{a,\delta_m}_{\lfloor T^{\kappa^*} \rfloor,s})_{s=1,\cdots,S}$. We can then extract the $(1-\alpha^*)$-quantile of the test statistic, for some user-chosen α^*. We select the perturbation vector δ_{m^*} associated with the $(1-\alpha^*)$-quantile the closest to the $(1-\alpha^*)$-quantile of the chi-square distribution with the appropriate degrees of freedom[11]. Note that $(1-\alpha^*)$ may, or may not correspond to the actual asymptotic size of the designed test. Regardless of the chosen $(1-\alpha^*)$ and associated perturbation vector δ_{m^*}, the asymptotic size of the test $(1-\alpha)$ is always controlled as shown above.
In our Monte-Carlo experiments, we choose $\alpha = \alpha^*$; our results (through unreported experiments) are not too sensitive to this choice.

3.3.2 Feasibility without a consistent estimator of $S(\theta^0)$

As highlighted in section 3.2, our testing procedures crucially depends on a consistent estimator S_T of $S(\theta^0)$ to define the efficient GMM estimator $\hat{\theta}_T$; the existence of S_T is guaranteed by standard two-step procedures whenever $\lim_T \left(M_T/\sqrt{T} \right) = 0$. In this section, we relax this assumption. As a result, under H_0, there is no obvious (consistent) estimator of $S(\theta^0)$ since there is no (first-step) consistent estimator of θ^0. We propose the following testing procedure which is robust to the absence of consistent estimators of $S(\theta^0)$:
(i) Build a confidence region C_T for θ^0. We rely on the test statistic proposed by Stock and Wright (2000). If C_T is empty, the null hypothesis is rejected.
(ii) If C_T is not empty, minimize the following test statistic,

$$T\overline{\phi}_T'(\hat{\theta}_T^{a,\delta})S_T^{-1}(\theta)\overline{\phi}_T(\hat{\theta}_T^{a,\delta})$$

with respect to $\theta \in C_T$, and compare it to the appropriate chi-square quantile as discussed in Theorems 3.3 and 3.4. The detailed procedure as well as additional discussions are provided in Appendix A.

[11] Recall that the appropriate degrees of freedom depends on the identification assumption imposed on θ_1. For instance, when θ_1 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 Monte-Carlo evidence

In this section, we use Monte-Carlo methods to illustrate the finite samples properties of the tests introduced in section 3. We consider a standard linear IV regression model with one intercept and one endogenous regressor, as well as a (non-linear) diffusion process with continuous record and increasing time span asymptotic.

4.1 Linear IV regression model

Consider the following standard linear IV regression model with one intercept and one endogenous regressor,

$$\begin{aligned} y_t &= \alpha_0 + Y_{1t}\beta_0 + h(X_t)\varepsilon_t \,, \\ Y_{1t} &= X_t'\Pi_x + U_t \,, \end{aligned} \tag{4.1}$$

where Y_{1t} is a univariate endogenous regressor, while X_t is a vector of L_x (exogenous) instrumental variables that follows a standard normal distribution. (ε_t, U_t) is normally distributed and independent of X_t. We set $\theta^0 = (\alpha_0 \ \beta_0)' = (0 \ 0)'$. We consider two versions of the model: a homoskedastic model with $h(x) = 1$ and a heteroskedastic model with $h(x) = \sqrt{(1 + (e'x)^2)/(L_x + 1)}$ where e is the vector of ones of size L_x. In both models, $(h(X_t)\varepsilon_t, U_t)$ has mean $\mathbf{0}$, unit unconditional variances, and unconditional correlation ρ. Π_x is proportional to the vector e and is related to the first stage R^2 by,

$$R_x^2 = \frac{\Pi_x'\Pi_x}{\Pi_x'\Pi_x + 1} \,.$$

It is worth pointing out that the intercept parameter is always strongly-identified, while the slope parameter is more or less weakly identified depending on the value of R_x^2.

In this experiment, we are interested in knowing whether pretesting for weak identification improves the inference about the structural parameter θ. More specifically, we compare the following inference strategies:

(a) The naive inference procedure (when one believes there is no identification issue): without pretesting, always construct a confidence region about θ using GMM;

(b) The conservative inference procedure (when one believes identification issues are always

at stake): without pretesting, always construct a confidence region about θ using Stock and Wright (2000)'s identification robust procedure;
(c) The agnostic inference procedure (when one does not know whether there are identification issues): pretest first for weak identification, and use either GMM (when the pretest rejects weak identification), or Stock and Wright (when the pretest cannot reject weak identification) to construct a confidence region. Three pretests are considered: (i) our pretest introduced in section 3; (ii) Staiger and Stock's (1997) rule of thumb; (iii) Stock and Yogo's (2005) pretest based on 10% bias of 2SLS[12]. As already mentioned, fully accounting for the error of the pretesting step is beyond the scope of this paper. We refer the interested reader to the recent paper by McCloskey (2012) who develops powerful size-correction methods while minimizing the degree of "conservativeness"; see also references therein.
For each inference procedure, we compute the Monte-Carlo coverage probability (or probability that the associated confidence region contains the true value), as well as the average length of the confidence interval for each component. Results for the homoskedastic model with $R_x^2 = 0.14$ (stronger identification), 0.01 (weaker identification), and 0 (no identification) are respectively collected in Tables 1 to 3. Corresponding results for the heteroskedastic model are collected in Tables 4 to 6.
Each table contains three panels. Panel A provides the first four moments of the Monte-Carlo distribution of the standardized GMM estimator. The reader can then assess how far the Monte-Carlo distribution is from its asymptotic approximation. Recall that for the standard normal distribution, we expect mean 0, variance 1, skewness 0 and kurtosis 3. Panel B collects results for inference on the whole parameter vector θ with and without pretesting. We consider four pretests: two versions of our test, robust and not robust to a consistent estimator of the covariance matrix of the moment restrictions, hereafter AR (robust) and AR (not robust); the rule of thumb of Staiger and Stock, hereafter SS; Staiger and Stock pretest based on 10% bias of 2SLS, hereafter SY. For each pretest, we provide the rejection probabilities. In addition, we consider six inference procedures, the four agnostic procedures associated with each pretest, as well as the naive procedure (based on GMM) and the conservative one (based on Stock and Wright, hereafter SW). For each inference

[12]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.

procedure, we provide the coverage probabilities and average lengths of associated confidence regions. Since the coverage probability of SW is always equal to 1, the average lengths of the confidence intervals associated with the other procedures are given in percentage of the average length of the interval associated with SW. Finally, Panel C collects results for inference on subvectors based on our pretest: specifically, on the intercept without assuming identification of the slope, and inference on the slope after assuming identification of the intercept. We provide rejections probabilities, coverage probabilities and average length of associated confidence intervals.

Several comments are in order.

(1) The distribution of the standardized GMM estimator displayed in Panel A confirms that the asymptotic approximation always works well for the intercept parameter (which is always well-identified), whereas the approximation worsens for the slope parameter as R_x^2 decreases (that is, when the identification issues are more acute).

(2) AR robust vs AR not robust. When comparing the performance of the two versions of our pretest and associated inference procedures, it is clear that a lot of power is lost by the robust version of our pretest in order to ensure perfect coverage. This is especially striking for the test on the whole vector (intercept and slope jointly). However, when considering our pretest on the intercept only, both versions of our test reject weak identification with probability 1 in all cases, as expected. In addition, it is worth mentioning that the non-robust version of the joint test has lower coverage probability (at least 0.90 for cases with strong and weaker identification, and at least 0.70 for cases with no identification). However, when considering the non-robust test on a subvector (either the intercept, or the slope), the coverage probabilities are very good (around 0.94 in all cases).

(3) Pretest vs no pretest. When considering the performance of the three types of inference procedures, naive, conservative and agnostic, and more generally whether pretesting improves the inference on the parameter θ, it is not clear that pretesting the whole vector θ leads to improved inference. However, when pretesting specific components, as done with our pretest (especially non-robust), it appears that inference indeed improves as shorter confidence intervals can be obtained with similar coverage probabilities.

To conclude, we recommend testing identification strength of specific components rather than the whole vector in order to have good coverage and power properties.

4.2 Diffusion process with continuous record and increasing time span asymptotic

Consider the following continuous time Ornstein-Uhlenbeck process

$$dy_t = (\theta_0 - \theta_1 y_t)dt + \theta_2 dW_t \qquad \text{with} \qquad dW_t \overset{iid}{\sim} \mathcal{N}(0, dt)\,,$$

where $\theta_0/\theta_1 > 0$ represents the long run (unconditional) mean, $\theta_1 > 0$ captures the speed of the mean reversion, and $\theta_2 > 0$ gives the constant volatility of the process. It is well-known that its exact solution is the following discrete time AR(1) process

$$\begin{aligned} & y_t = a + b y_{t-\Delta} + \sqrt{c}\epsilon_t\,, \qquad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)\,, \\ \text{with} \qquad & a = \frac{\theta_0}{\theta_1}(1 - e^{-\theta_1 \Delta})\,, \qquad b = e^{-\theta_1 \Delta}\,, \qquad c = \theta_2^2 \left(\frac{1 - e^{-2\theta_1 \Delta}}{2\theta_1}\right)\,. \end{aligned} \tag{4.2}$$

For simplicity, the parameters θ_0 and θ_2 are assumed to be known throughout, and are fixed at their true values in the structural model, while only the parameter θ_1 is estimated. It is well-known that only the estimation of θ_1 may be problematic in finite sample.

Suppose that n observations of (4.2) are available for $t = \Delta, \cdots, n\Delta$ with $T \equiv n\Delta$. Define the associated OLS estimators of the three parameters, a, b, and c, respectively $\hat{a}_{n,ols}$, $\hat{b}_{n,ols}$, and $\hat{c}_{n,ols}$. For fixed Δ, the usual asymptotic result for OLS estimators holds, and we have

$$\sqrt{n}\begin{bmatrix} \hat{a}_{n,ols} - a(\theta_1) \\ \hat{b}_{n,ols} - b(\theta_1) \\ \sqrt{\hat{c}_{n,ols}} - \sqrt{c(\theta_1)} \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma(\Delta)) \qquad \text{with} \qquad \Sigma(\Delta) = \begin{pmatrix} cE\left[(X_i X_i')\right]^{-1} & 0 \\ 0 & c/2 \end{pmatrix}\,,$$

where X_i' represents the i-th row of the matrix X. Our estimation procedure for θ_1 relies on the (overidentified) GMM estimation with three moment conditions,

$$\begin{aligned} \hat{\theta}_{1,n} &= \arg\min_{\theta_1} \left[\phi(\theta_1)' \Omega_n \phi(\theta_1)\right] \\ \text{with} \quad \phi(\theta_1) &= \frac{1}{\Delta}\begin{bmatrix} \hat{a}_{n,ols} - a(\theta_1) \\ \hat{b}_{n,ols} - b(\theta_1) \\ \sqrt{\hat{c}_{n,ols}} - \sqrt{c(\theta_1)} \end{bmatrix} = \frac{1}{\Delta}\begin{bmatrix} \hat{a}_{n,ols} - \frac{\theta_0}{\theta_1}(1 - e^{-\theta_1 \Delta}) \\ \hat{b}_{n,ols} - e^{-\theta_1 \Delta} \\ \sqrt{\hat{c}_{n,ols}} - \sqrt{\theta_2^2 \left(\frac{1 - e^{-2\theta_1 \Delta}}{2\theta_1}\right)} \end{bmatrix}\,, \end{aligned}$$

where Ω_n is a sequence of symmetric positive definite random matrices of size 3 converging towards a positive definite matrix Ω. In Appendix C.1, we show that each moment condition

has a different identification strength controlled by Δ. More precisely, if we consider the three (just-identified) estimators obtained from the GMM estimation based on each moment condition separately, we get that, when $\Delta \to 0$ and $T \to \infty$:
- the estimator based on condition 2 converges at rate $\sqrt{T}$;
- the estimator based on condition 3 converges at rate $\sqrt{\Delta}\sqrt{T}$, with $\sqrt{\Delta}\sqrt{T} = o(\sqrt{T})$;
- the estimator based on condition 1 converges at rate $\Delta\sqrt{T}$, with $\Delta\sqrt{T} = o(\sqrt{\Delta}\sqrt{T})$.

Throughout, the following notations are used to distinguish the different estimators of θ_1 we consider:
- $\hat{\theta}_{all}$ refers to the (overidentified) GMM estimator based on the three moment conditions;
- $\hat{\theta}_{\setminus j}$ refers to the (overidentified) GMM estimator based on two moment conditions only, after condition j has been removed.

In this experiment, we are interested in testing the strength of identification. In our simple framework, we know that, asymptotically, the strongest moment condition dictates the rate of convergence of the associated estimator of θ_1.[13] In our experiment, we fix the time span T and vary the strength of identification by decreasing Δ (accordingly, n increases). Smaller values of Δ correspond to cases where the identification strength is weaker. The nominal size of the tests is 5%. Our results are displayed in Table 7. For each test, we provide the estimator of θ_1 being considered, as well as the associated Monte-Carlo rejection probability. Recall that rejection of the null hypothesis means that the estimator is sufficiently strongly identified for standard asymptotic results to hold. Further details regarding the implementation of this experiment are provided in Appendix C.1.

First, we consider the estimator $\hat{\theta}_{all}$ based on all moment conditions. As discussed above, this estimator is always strongly identified due to the moment condition 2. As a result, we expect our test to often be rejected regardless of the value of the parameter Δ. This is exactly what happens: the associated rejection probabilities are equal to 1 irrespective of the identification strength.

Second, we consider the estimator $\hat{\theta}_{\setminus 3}$. The presence of the moment condition 2 guarantees

[13] In general, this is not the case as discussed in Section 2. However, in our simple framework, there is only one parameter to identify, so it necessarily inherits the identification strength of the strongest moment condition available.

that this estimator is always strongly identified. And we also expect our test to often be rejected regardless of the value of the parameter Δ. The associated rejection probabilities are actually equal to 1 in all cases.

Finally, we consider the estimator based on moment conditions 1 and 3 only, $\hat{\theta}_{\backslash 2}$. As discussed above, this estimator is identified at rate $\sqrt{\Delta}\sqrt{T}$ due to the moment condition 3. As a result, we expect our test not to be rejected for sufficiently small values of Δ. The associated rejection probabilities are quite small in all cases, but do become even smaller as Δ gets smaller. This suggests that the identifying power of (missing) condition 2 is stronger than the one of the two other (included) conditions. To conclude, the test results are conformable to our expectations.

5 Empirical study: estimation of the Elasticity of Intertemporal Substitution

In this section, we apply our testing procedures to a well-known empirical example, the instrumental variables estimation of the Elasticity of Intertemporal Substitution (EIS) (Campbell (2003)). Following the literature, we consider the estimation of the linearized Euler equation in two standard IV frameworks.

$$\begin{aligned} \Delta c_{t+1} &= \nu + \psi r_{t+1} + u_{t+1}\,, \\ r_{t+1} &= \xi + \frac{1}{\psi}\Delta c_{t+1} + \eta_{t+1}\,, \end{aligned}$$

where ψ is the EIS, Δc_{t+1} the consumption growth at time $(t+1)$, r_{t+1} a real asset return at time $(t+1)$, ν and ξ 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 empirical study replicates partially the studies by Yogo (2004) and Montiel and Pflueger (2013) who have found that the EIS point estimates are small and close to zero.

Table 8 compares pretests for weak instruments for 11 countries and compares to Table 1 in Yogo (2004) and Table 2 in Montiel and Pflueger (2013). Panel A pretests for weak

identification with the ex-post real interest rate as the endogenous variable, while Panel B pretests for weak identification with the consumption growth as the endogenous variable. We compare seven pretests: Stock and Yogo test based on 10% bias of 2SLS, three versions of the test proposed by Pflueger and Montiel, and three versions of the test proposed in this paper. Pflueger and Montiel 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 pretests with simplified critical values c_{simp}, and generalized critical values, c_{TSLS} and c_{LIML}. We also report the results for three versions of our pretest: pretesting the intercept and the slope jointly, pretesting the intercept only, and pretesting the slope only. As already mentioned, the intercept (either ν or ξ) is always strongly identified, while only the slope (either ψ or $1/\psi$) may be weakly identified. This distinction cannot be captured by joint tests such as Stock and Yogo's or Montiel and Pflueger's. Our pretest for the intercept always conclude that it is strongly identified. Our pretest 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 pretests of Stock and Yogo and Montiel and Pflueger also lead to mixed results. However, it is worth mentioning that none of these pretests can reject weak identification in Panel B. This corresponds to the results of our test for the slope only.

To conclude and similarly to the Monte-Carlo results, we recommend using our pretest of a subvector rather than the joint pretest, which always yields to weak identification in both Panels A and B, as suggested by previous empirical studies.

6 Conclusion

This paper provides a new testing strategy for identification strength. The test statistic under the null is asymptotically upper bounded by a chi-square distribution and converges to infinity under the alternative with probability one. Among the advantages of the proposed strategy, we highlight the following:

(i) The test is simple to implement. It basically amounts to compute a J-test statistic for overidentification (at a point in the parameter space obtained by a suitable perturbation of the GMM estimator) and to compare its sample value to standard critical

values. The only user-dependent part of the strategy is the choice of tuning parameters for the perturbation, which may be important for finite sample performance of the test.

(ii) Unlike some other tests available in the literature, our test is conformable to the philosophy of the Neyman's approach. It is when the sample is sufficiently informative to allow the test to reject the null that the researcher can draw safely a conclusion about applying standard asymptotic distribution of GMM estimators and inference byproducts.

(iii) Unlike most of the tests available in the extant literature, the test is able to set the focus on the null hypothesis of poor identification of a subvector of parameters of interest. We show that setting the empirical research on the proper subvector of suspicious parameters is key to deliver a powerful test, that is to provide the empirical researcher with a chance to know that standard inference procedures can be safely applied more often.

(iv) The Monte-Carlo experiments show compellingly that such a powerful pretesting procedure is quite important to ensure accurate inference. Giving up pretests or using pretest with less power will too often lead to an excess of prudence with the application of robust inference strategies at the cost of efficiency loss in inference.

The aforementioned achievements are obtained thanks to a general characterization of patterns of identification strength in different directions in the parameter space. It is emphasized that depending of the shape of the moment conditions - linear (at least w.r.t. suspicious parameters), suitably separable, or general - the degree of "poor identification" that must be under test is different. Our testing methodology is typically able to adjust for such differences. It is the reason why we call our tests "tests for identification strength" and not only "tests for identification". The ability to address weak identification in non-linear settings is important, as already known for asset pricing or macroeconomic models through Euler equations. We provide a new route for interesting non-linear patterns of heterogeneous degrees of identification, with the example of estimation of continuous time processes with time series asymptotics in two dimensions (time span and frequency of observations).

In other words, we revisit the celebrated Merton's puzzle (irrelevance of high frequency data to estimate the drift of the process) within the terminology of weak identification. Even more generally, as extensively discussed in Antoine and Renault (2012), there are many settings where heterogeneous rates of convergence may pop up in GMM inference. In all these examples, testing for identification strength is worth doing for valid and accurate inference.

References

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

[2] D.W.K. Andrews, *Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity*, Econometrica **62** (1994), 43–72.

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

[4] D.W.K. Andrews and J.H. Stock, *Inference with Weak Instruments*, Econometric Society Monograph Series, vol. 3, ch. 8 in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Cambridge University Press, Cambridge, 2007.

[5] B. Antoine and E. Renault, *Efficient GMM with Nearly-weak Instruments*, The Econometrics Journal **12** (2009), 135–171.

[6] ———, *Efficient Inference with Poor Instruments: a General Framework*, ch. 2 in Handbook of Empirical Economics and Finance, pp. 29–70, edited by A. Ullah and D.E. Giles, Taylor & Francis, 2010.

[7] ———, *Efficient Minimum Distance Estimation with Multiple Rates of Convergence*, Journal of Econometrics (2012), 350–367.

[8] ———, *On the Relevance of Weaker Instruments*, Working paper (2013).

[9] M. Caner, *Testing, Estimation in GMM and CUE with Nearly-Weak Identification*, Econometric Reviews **29** (2010), 330–363.

[10] S. Chaudhuri and E. Zivot, *A new method of projection-based inference in GMM with weakly identified nuisance parameters*, Journal of Econometrics **164** (2011), 239–251.

[11] R. Dridi, A. Guay, and E. Renault, *Indirect Inference and Calibration of Dynamic Stochastic General Equilibrium Models*, Journal of Econometrics **136** (2007), 397–430.

[12] J.-M. Dufour, *Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models*, Econometrica **65** (1997), 1365–1388.

[13] J.-M. Dufour and P. Valéry, *Wald-type tests when rank conditions fail: a smooth regularization approach*, Working paper (2011).

[14] P. Guggenberger, F. Kleibergen, S. Mavroeidis, and L. Chen, *"On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression"*, Econometrica **80** (2012), 2649–2666.

[15] J. Hahn and J. Hausman, *A new Specification Test fir the Validity of Instrumental Variables*, Econometrica **70** (2002), 163–190.

[16] ———, *Weak Instruments: Diagnosis and Cures in Empirical Econometrics*, American Economic Review **93** (2003), 118–125.

[17] J. Hahn and G. Kuersteiner, *Discontinuities of Weak Instruments limiting Distributions*, Economics Letters **75** (2002), 325–331.

[18] C. Hansen, J. Hausman, and W. Newey, *Estimation with Many Instrumental Variables*, Journal of Business and Economic Statistics **26** (2008), 398–422.

[19] L.P. Hansen, *Large Sample Properties of Generalized Method of Moments Estimators*, Econometrica **50** (1982), no. 4, 1029–1054.

[20] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[21] A. Inoue and B. Rossi, *"Testing for Weak Identification in Possibly Non-linear Models"*, Journal of Econometrics **161** (2011), 246–261.

[22] F. Kleibergen, *Testing Parameters in GMM without assuming that they are identified*, Econometrica **73** (2005), 1103–1123.

[23] F. Kleibergen and S. Mavroeidis, *Inference on subsets of parameters in GMM without assuming identification*, Working paper, Brown University (2009).

[24] A. McCloskey, *Bonferroni-Based Size-Correction for Nonstandard Testing Problems*, Working paper (2012).

[25] M.J. Moreira, *A Conditional Likelihood Ratio Test for Structural Models*, Econometrica **71** (2003), 1027–1048.

[26] W.K. Newey and K.D. West, *A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix*, Econometrica **55** (1987), 703–708.

[27] W.K. Newey and F. Windmeijer, *GMM with Many Weak Moment Conditions*, Econometrica **77** (2009), 687–719.

[28] J.L. Montiel Olea and C. Pflueger, *A Robust Test for Weak Instruments*, Working paper (2012).

[29] Z. Qu, *Inference and Specification Testing in DSGE Models with Possible Weak Identification*, Working paper (2011).

[30] J.D. Sargan, *The Estimation of Economic Relationships Using Instrumental Variables*, Econometrica **26** (1958), 393–415.

[31] D. Staiger and J. Stock, *Instrumental Variables Regression with Weak instruments*, Econometrica **65** (1997), 557–586.

[32] J.H. Stock and J.H. Wright, *GMM with Weak Identification*, Econometrica **68** (2000), no. 5, 1055–1096.

[33] J.H. Stock and M. Yogo, *Asymptotic Distributions of Instrumental Variables Statistics With Many Instruments*, ch. 8 in Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, pp. 109–120, edited by D.W.K. Andrews and J.H. Stock, Cambridge University Press, Cambridge, U.K., 2005.

[34] J. Wright, *Detecting Lack of Identification in GMM*, Econometric Theory **19** (2003), 322–330.