Efficient GMM with multiple rates of convergence and applications to nearly-weak identification*

Bertille Antoine† and Eric Renault‡

July, 16 2007

Abstract:
This paper considers an extension of asymptotic theory of GMM inference in a new setting where sample counterparts of estimating equations have multiple rates of convergence. GMM estimators are still consistent and asymptotically normally distributed, but with possibly different rates of convergence potentially slower than the usual square root of T. We then propose a convenient reparameterization which permits to identify directions in the parameter space associated with a specific rate of convergence.
Direct applications of our results allow us to address some issues of (nearly)-weak identification. In particular, in contrast to existing literature, we do not characterize weak identification directly in terms of drifting population moments but rather in terms of the content of the statistical information available about those moments. Drifting DGP is only a possible interpretation of weak identification and we actually have in mind a couple of alternative interpretations, especially one based on kernel smoothing.

JEL Classification: C13; C14; C32.

Keywords: GMM; Instrumental variables; Weak identification.

*We would like to thank Marine Carrasco, Jean-Marie Dufour, Alain Guay, Pascal Lavergne and Lonnie Magee for helpful discussions. We also thank Jonathan Wright for kindly providing us with computer code for generating artificial data.

†*Simon Fraser University. Email: bertille_antoine@sfu.ca.*

‡*University of North Carolina at Chapel Hill, CIRANO and CIREQ. Email: renault@email.unc.edu*

Contents

List of Tables

List of Figures

1 Introduction

The cornerstone of GMM estimation is a set of population moment conditions, often deduced from a structural econometric model. The limit distributions of GMM estimators are derived under a central limit theorem for the moment conditions and a full rank assumption of the expected Jacobian. The latter assumption is not implied by economic theory and many circumstances where it is rather unjustified have been documented in the literature (see Andrews and Stock (2005) for a recent survey).

Earlier work on the properties of GMM-based estimation and inference in the context of rank condition failures includes Phillips (1989) and Sargan (1983). In the context of a classical linear simultaneous equations model, Phillips (1989) considers the case of a *partially identified* structural equation. He notes that, in case of rank condition failure, it is always possible to rotate coordinates in order to isolate estimable linear combinations of the structural parameters while the remaining directions are completely unidentified. Asymptotic theory of standard IV estimators in this context is then developed through the general framework of limited mixed Gaussian family. This approach of *partially identified* models differs from Sargan (1983) *first order under-identification*. While for the former there is nothing between estimable parameters with standard root-T consistent estimators and completely unidentified parameters, the latter considers that asymptotic identification is still guaranteed but it only comes from higher order terms in the Taylor expansion of first order optimality conditions of GMM: higher order terms become crucial when first order terms vanish. They are responsible for slower rates of convergence of GMM estimators like $T^{1/4}$ and may lead to non-normal asymptotic distributions like a Cauchy distribution or a mixture of normal distributions.

Our contribution in this essay is to revisit an approach of partial identification *à la* Phillips (1989), while maintaining, like Sargan (1983), the complete identification of all parameters, but at possibly slower rates. Moreover, we remain true to asymptotic normality of GMM estimators deduced from first order identification but with an expected Jacobian that may vanish when the sample size increases. In this respect, we are in the line of the recent literature on weak instruments, which, following the seminal approach of Staiger and Stock (1997) and Stock and Wright (2000), captures weak identification by drifting population moment conditions. With respect to the existing literature, the contribution of this essay is as follows.
First, in sharp contrast with most of the recent literature on weak instruments, we do not

specify a priori which parameters are strongly or weakly identified. Conforming to the common wisdom that weakness should rather be assigned to specific instruments or more generally to some specific moment conditions, we follow Phillips (1989) to consider that the relevant partition of the set of structural parameters between strongly and weakly identified ones can only be achieved after a well-suited rotation in the parameter space. In nonlinear settings, this change of basis depends on unknown structural parameters and must itself be consistently estimated.

Second, like Caner (2005) (see also Hahn and Kuersteiner (2002) for the special case of linear 2SLS), we focus on the case dubbed *nearly-weak identification*, where the drifting DGP introduces a limit rank deficiency reached at a rate slower than root-T: this allows consistent estimation of all parameters, but at rates possibly slower than usual. It is then all the more important to identify the different directions in the parameter space endowed with the different rates. We consistently estimate these directions without assuming that the rates slower than root-T are known. We only maintain the assumption that the moment conditions responsible for approximate rank deficiency have been detected. Practically, this either may be thanks to prior economic knowledge (like market efficiency responsible for the weakness of instruments built from past returns in asset pricing models) or suggested by a preliminary study of the lack of steepness of the GMM objective function around plausible values of the structural parameters. Note that we only consider asymptotic rank deficiency such that all the rates of convergence of GMM estimators, possibly slower than root-T, are at least larger than $T^{1/4}$. The first order under-identification case of Sargan (1983), producing GMM estimators converging at rates $T^{1/4}$, can then be seen as a limit case of our approach. This is in sharp contrast with the weak instrument case *à la Stock and Wright* (2000) where the asymptotic rank deficiency is reached at a rate as fast as root-T: GMM estimators are not even consistent. The fact that all the GMM estimators are consistent with well-defined rates of convergence, albeit possibly unknown and slower than root-T, allows us to validate standard asymptotic testing approaches like Wald test or GMM-LM test of Newey and West (1987). In contrast with Kleibergen (2005), we do not need to modify the standard formulas for the LM test. Moreover, our approach is more general than Kleibergen (2005) since we explicitly take into account the possible simultaneous occurrence, in a given set of moment conditions, of heterogeneous rates of convergence.

As far as technical tools for asymptotic theory are concerned, we borrow to three recent developments in econometric theory.

First, as stressed by Stock and Wright (2000), (nearly)-weak identification in nonlinear settings makes asymptotic theory more involved than in the linear case because the occurrence of unknown parameters and observations in the moment conditions are not additively separable. Lee's (2004) minimum distance estimation with heterogeneous rates of convergence, albeit nonlinear, is also kept simple by this kind of additive separability. By contrast, this non-separability makes, in general, necessary resorting to a functional central limit theorem applied to the GMM objective function viewed as an empirical process indexed by unknown parameters.

Second, our approach to Wald testing with heterogeneous rates of convergence must be related to the former contribution of Lee (2005). The key issue is the following: when several directions (to be tested) in the parameter space are estimated at slow rates, while some linear combinations of them may be estimated at faster rates, a perverse asymptotic singularity is introduced and invalidates the common delta theorem. This situation, rather similar in spirit to cointegration, leads Lee (2005) to maintain an additional assumption for Wald testing. We are able to relax Lee's (2005) condition and to confirm that the common Wald test methodology always work, albeit with possibly nonstandard rates of convergence against sequences of local alternatives. The trick is again to consider a convenient rotation in the parameter space. Note that this issue makes even more important our extension of Kleibergen's (2005) setting to allow for different rates of convergence simultaneously.

A third debt to acknowledge is with respect to Andrews (1994, 1995) MINPIN estimators[1] and to Gagliardini, Gouriéroux and Renault (2005) XMM (Extended Method of Moments) estimators as well. Like them, we observe that GMM-like asymptotic variance formulas remain valid for strongly identified directions when slowly identified directions are estimated at rates faster than $T^{1/4}$. Rates even slower than that would imply a perverse contamination of the estimators of the standard directions by poorly identified nuisance parameters. In this respect, our approach should rather be dubbed *nearly-strong identification*. Of course, by doing so, we may renounce to capture severe weak identification cases arising even when the sample size is very large (see e.g Angrist and Krueger (1991)). However, our approach provides the empirical economist with estimation and inference procedures that are valid with the standard formulas, while warning her about rates of convergence in some specific directions that may be slower than the standard root-T. Moreover, these directions (strong and weak) can be disentangled

[1]MINPIN estimators are defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator.

and consistently estimated without modifying the overall rates of convergence of the implied linear combinations of structural parameters.

The chapter is organized as follows. Section 2 precisely defines our nearly-weak identification setting and proves consistency of both point GMM estimators of structural parameters θ and set estimators, that are equivalent to LM-tests of null hypotheses $\theta = \theta^0$. With nearly-weak global identification, consistency of point estimation rests upon an empirical process approach for the moment conditions, whereas set estimation rests upon nearly-weak local identification, characterized in terms of the expected Jacobian of the moment conditions. Our integrated framework restores the coherency between the two possible points of view about weak identification, global and local. In section 3, we show how to disentangle and to estimate the directions with different rates of convergence. We also prove the asymptotic normality of well-suited linear combinations of the structural parameters. The issue of Wald testing is addressed in section 4 while section 5 explicitly relates our setting to examples of weak identification well-studied in the literature. Section 6 is devoted to a couple of Monte-Carlo illustrations for two toys models: single-equation linear IV model and CCAPM.
All the proofs and figures are gathered in the appendix[2].

2 Consistent point and set GMM estimators

This section shows that a standard GMM approach works both for consistent point and set estimation, the latter through a score type test statistic. Typically, all the components of the parameters of interest are simultaneously estimated and tested without *a priori* knowledge of their heterogenous patters of identification.

2.1 Nearly-weak global identification

Let θ be a p-dimensional parameter vector with true (unknown) value θ^0, assumed in the interior of the compact parameter space Θ. The true parameter value satisfies the K equations,

$$E\left[\phi_t(\theta^0)\right] = 0 \tag{2.1}$$

[2]Most of the theoretical results are obtained in a more general context in a technical companion paper Antoine and Renault (2007).

with $\phi(.)$ some known functions. We have at our disposal a sample of size T, and we can calculate $\phi_t(\theta)$ for any value of the parameter in Θ and for every $t = 1, \cdots, T$.

Standard GMM estimation defines its estimator $\hat{\theta}_T$ as follows:

Definition 2.1 *Let Ω_T be a sequence of symmetric positive definite random matrices of size K which converges in probability towards a positive definite matrix Ω. A GMM estimator $\hat{\theta}_T$ of θ^0 is then defined as:*

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta) \quad \textit{where} \quad Q_T(\theta) \equiv \overline{\phi}'_T(\theta) \Omega_T \overline{\phi}_T(\theta) \tag{2.2}$$

with $\overline{\phi}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_t(\theta)$, the empirical mean of the moment restrictions.

Standard GMM asymptotic theory assumes that, for $\theta \neq \theta^0$, $\overline{\phi}_T(\theta)$ converges in probability towards its nonzero expected value because of some uniform law of large numbers. We consider here a more general situation where $\overline{\phi}_T(\theta)$ may converge towards zero even for $\theta \neq \theta^0$. And we show how this can be interpreted as identification issues.

More precisely, we imagine that we have here two groups of moment restrictions: one standard for which the empirical counterpart converges at the standard (usual) rate of convergence $\sqrt{T}$ and a weaker one for which the empirical counterpart still converges but potentially at a slower rate λ_T. At this stage, it is essential to stress that identification is going to be maintained (but through higher order asymptotic developments). More formally, we have k_1 standard moment restrictions such that

$$\sqrt{T} \left[\overline{\phi}_{1T}(\theta) - \rho_1(\theta) \right] = \mathcal{O}_P(1) \tag{2.3}$$

and k_2 $(= K - k_1)$ weaker moment restrictions such that

$$\sqrt{T} \left[\overline{\phi}_{2T}(\theta) - \frac{\lambda_T}{\sqrt{T}} \rho_2(\theta) \right] = \mathcal{O}_P(1) \ \text{ where } \ \lambda_T = o(\sqrt{T}) \ \text{ and } \ \lambda_T \overset{T}{\to} \infty \tag{2.4}$$

with $[\rho'_1(\theta) \ \rho'_2(\theta)] = 0 \ \Leftrightarrow \ \theta = \theta^0$.

λ_T measures the degree of weakness of the second group of moment restrictions. The corresponding component $\rho_2(.)$ is squeezed to zero and $Plim \left[\overline{\phi}_{2T}(\theta) \right] = 0$ for all $\theta \in \Theta$. Thus, the probability limit of $\overline{\phi}_T(\theta)$ does not allow to discriminate between θ^0 and any $\theta \in \Theta$. In such a context, identification is a combined property of the functions $\phi_t(\theta)$ and $\rho(\theta)$ and the asymptotic behavior of λ_T. The maintained identification assumption is the following:

Assumption 1 *(Identification)*
(i) $\rho(.)$ is a continuous function from a compact parameter space $\Theta \subset \mathbb{R}^p$ into $\mathbb{R}^K$ such that

$$\rho(\theta) = 0 \iff \theta = \theta^0$$

(ii) The empirical process $(\Psi_T(\theta))_{\theta\in\Theta}$ obeys a functional central limit theorem:

$$\Psi_T(\theta) \equiv \sqrt{T} \begin{bmatrix} \overline{\phi}_{1T}(\theta) - \rho_1(\theta) \\ \overline{\phi}_{2T}(\theta) - \frac{\lambda_T(\theta)}{\sqrt{T}}\rho_2(\theta) \end{bmatrix} \Rightarrow \Psi(\theta)$$

where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero.
(iii) λ_T is a deterministic sequence of positive real numbers such that

$$\lim_{T\to\infty} \lambda_T = \infty \quad and \quad \lim_{T\to\infty} \frac{\lambda_T}{\sqrt{T}} = 0$$

Following Stock and Wright (2000), assumption 1 reinforces the standard central limit theorem written for moment conditions at the true value ($\theta = \theta^0$) by maintaining a functional central limit theorem on the whole parameter set Θ. Stock and Wright (2000) use this framework to address the weak identification case corresponding to $\lambda_T = 1$. By contrast, as Hahn and Kuersteiner (2002) and Caner (2005), we focus here on nearly-weak identification where λ_T goes to infinity albeit slower than $\sqrt{T}$. Note that the standard strong identification case corresponds to $\lambda_T = \sqrt{T}$. The above functional central limit theorem[3] allows us to get a consistent GMM estimator, even in case of nearly-weak identification[4].

Theorem 2.1 *(Consistency of $\hat{\theta}_T$)*
Under assumption 1, any GMM estimator $\hat{\theta}_T$ like (2.2) is weakly consistent.

Besides the fact that all the components of the parameter of interest θ are consistently estimated, it is worth stressing another difference with Stock and Wright (2000). We do not

[3]Note that the asymptotic normality assumption is not necessary at this stage. In general, it might be replaced by some tightness assumption on $\Psi_T(.)$. See Antoine and Renault (2007).

[4]As stressed by Stock and Wright (2000) the uniformity in θ provided by the functional central limit theorem is crucial in case of nonlinear nonseparable moment conditions, that is when the occurrences of θ and the observations in the moment conditions are not additively separable. By contrast, Hahn and Kuersteiner (2002) (linear case) and Lee (2004) (separable case) do not need to resort to a functional central limit theorem.

assume the *a priori* knowledge of a partition $\theta = (\alpha' \ \beta')'$, where α is strongly identified and β (nearly)-weakly identified. By contrast, nearly-weak identification is produced by the rates of convergence of the moment conditions. More precisely, assumption 1 implies that, for the first set of moment conditions, we have (as for standard GMM),

$$\rho_1(\theta) = \underset{T\to\infty}{Plim}\ \overline{\phi}_{1T}(\theta)$$

whereas we only have for the second set of moment conditions

$$\rho_2(\theta) = \underset{T\to\infty}{Plim}\ \frac{\sqrt{T}}{\lambda_T}\overline{\phi}_{2T}(\theta)$$

It will be shown that this framework nests Stock and Wright (2000), Hahn and Kuersteiner (2002) and Caner (2005). More precisely, a rotation in the parameter space will allow us to identify some strongly identified directions and some others, only (nearly)-weakly identified. Subsection 2.2 below shows that the above rates of convergence naturally induce rates of convergence for the Jacobian matrices. This enables us to encompass the framework of Kleibergen (2005).

2.2 Nearly-weak local identification

As already explained, we simultaneously consider two rates of convergence to characterize the asymptotic behavior of the sample moments $\overline{\phi}_T(\theta)$ and the induced singularity issues in the sample counterparts of the estimating functions $\rho(\theta)$. In this respect, we differ from Sargan (1983) since we maintain the first-order identification assumption:

Assumption 2 *(First-order identification)*
(i) $\rho(.)$ *is continuously differentiable on the interior of* Θ *denoted as* $int(\Theta)$.
(ii) $\theta^0 \in \ int(\Theta)$.
(iii) The $(K \times p)$*-matrix* $[\partial\rho(\theta)/\partial\theta']$ *has full column rank* p *for all* $\theta \in \Theta$.
(iv) $\left[\begin{array}{cc} \frac{\partial\rho_1'(\theta^0)}{\partial\theta} & \frac{\partial\rho_2'(\theta^0)}{\partial\theta} \end{array}\right] = \underset{T\to\infty}{Plim}\left[\begin{array}{cc} \frac{\partial\overline{\phi}_{1T}'(\theta^0)}{\partial\theta} & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}'(\theta^0)}{\partial\theta} \end{array}\right]$
(v) $\sqrt{T}\left[\frac{\partial\overline{\phi}_{1T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_1(\theta^0)}{\partial\theta'}\right] = \mathcal{O}_P(1)$

The identification issue is not raised by rank deficiency of the moment conditions but by the rates of convergence. In other words, the implicit assumption in Kleibergen (2005) (see the

proof of his theorem 1 page 1122) that Jacobian matrices may have non-standard rates of convergence is made explicit in our framework. Assumptions 2(iv) and (v) are the natural extensions of assumption 1 on Jacobian matrices. Typically, Kleibergen (2005) maintains assumption 2(v) through a joint asymptotic normality assumption on $\overline{\phi}_T(\theta^0)$ and $\left[\partial\overline{\phi}_T(\theta^0)/\partial\theta'\right]$ (see his assumption 1).

While global identification (assumption 1) provides a consistent estimator of θ, local identification (assumption 2) provides an asymptotically consistent confidence set at level $(1-\alpha)$ or, equivalently, an asymptotically consistent test at level α for any simple hypothesis $H_0 : \theta = \theta_0$[5]. A score test approach, as defined in Newey and West (1987), does not resort to the asymptotic distributions of the estimators:

Theorem 2.2 *(Score test)*
The score statistic for testing $H_0 : \theta = \theta_0$ is defined as

$$LM_T(\theta_0) = \frac{T}{4}\frac{\partial Q_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial Q'_T(\theta_0)}{\partial\theta}$$

where S_T is a standard consistent estimator of the long-term covariance matrix[6].
Under H_0 and assumptions 1 and 2, $LM_T(\theta_0)$ has a $\chi^2(p)$ limit distribution.

In sharp contrast with Kleibergen (2005), we do not need to modify the standard score test statistic to replace the Jacobian of the moment conditions by their projections on the orthogonal space of the moment conditions. The reason for this maintained simplicity is that, in our nearly-weakly identified case,

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'}$$

has a deterministic limit which does not introduce any perverse correlations. By contrast, in the weakly identified case considered by Kleibergen (2005) (or $\lambda_T = 1$), the relevant limit of the sequence of Jacobian matrices is Gaussian. In this latter case, the limiting behavior of $\left[\partial\overline{\phi}_T(\theta^0)/\partial\theta'\right]$ is not independent of the limiting behavior of $\left[\overline{\phi}_T(\theta^0)\right]$ so the limiting distribution of the GMM score test statistic depends on nuisance parameters (see Stock and Wright

[5]Note that in general θ_0 might be different from the true (unknown) value of the parameter θ^0.

[6]Note that a consistent estimator S_T of the long-term covariance matrix $S(\theta_0)$ of $\Psi(\theta_0)$ can be built in the standard way (see in general Hall (2005)) from a preliminary inefficient GMM estimator $\tilde{\theta}_T$ of θ. However, under the null, one may simply choose $\tilde{\theta}_T = \theta_0$.

(2000)). Of course, the advantage of the K-statistic proposed by Kleibergen (2005) is to be robust in the limit case $\lambda_T = 1$ while, for us, λ_T must always converge towards infinity albeit possibly very slowly.

It is essential to realize that although the standard score test statistic has the common $\chi^2(p)$ distribution under the null, it works rather differently. Basically,

$$\left[\frac{\partial\overline{\phi}_T'(\theta^0)}{\partial\theta} S_T^{-1} \frac{\partial\overline{\phi}_T(\theta^0)}{\partial\theta'}\right] \tag{2.5}$$

is an asymptotically singular matrix since

$$\underset{T\to\infty}{Plim} \left[\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'}\right] = \lim_{T\to\infty} \left[\frac{\lambda_T}{\sqrt{T}} \frac{\partial\rho_2(\theta^0)}{\partial\theta'}\right] = 0$$

The proof of theorem 2.2 shows that the standard formula is actually recovered by well-suited matricial scalings of $\left[\partial Q_T(\theta^0)/\partial\theta'\right]$ and (2.5). The ultimate cancelation of these scalings must not conceal that testing parameter in GMM without assuming they are strongly identified requires a specific theory. It is in particular important to realize that both strong and (nearly)-weak identification may show up together in a given set of moment conditions. Note that this is immaterial as far as practical formulas for score testing are concerned. However, we show below that it has a dramatic impact on the power against local alternatives[7].

Another difference with Kleibergen (2005) is that our score test is consistent in all directions. Actually, ignoring the limit case ($\lambda_T = 1$) of weak identification allows us to write down consistent confidence sets and score tests. In terms of local alternatives, we get consistency at least at rate λ_T thanks to the following result:

Theorem 2.3 *(Rate of convergence)*
Under assumptions 1 and 2(i) to (iii), we have:

$$\left\|\hat{\theta}_T - \theta^0\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

In the remaining of the essay, we precisely focus on the identification of directions of local alternatives where consistency is kept at the standard rate $\sqrt{T}$.

[7]Kleibergen (2005) considers a simpler setting since he does not allow for two different kinds of identification (strong and weak) to be considered simultaneously (see the proof of his theorem 1). In addition, a full rank condition seems to be implicitly maintained in Kleibergen's proof.

3 Rates of convergence and asymptotic normality

In this section, we start with a kind of *rotation* in the parameter space which allows us to disentangle the rates of convergence. More precisely, some special linear combinations of θ are actually estimated at the standard rate of convergence $\sqrt{T}$, while some others are still estimated at the slower rate λ_T. This is formalized by a central limit theorem which allows the practitioner to apply the common GMM formula without knowing *a priori* the identification pattern.

3.1 Separation of the rates of convergence

We face the following situation:

(i) Only k_1 equations (defined by $\rho_1(.)$) have a sample counterpart which converges at the standard rate $\sqrt{T}$. These can be used in a standard way. Unfortunately, we have in general a reduced rank problem: $\left[\partial\rho_1(\theta^0)/\partial\theta'\right]$ is not full column rank. Its rank s_1 is strictly smaller than p and the first set of equations cannot identify θ. Intuitively, it can only identify s_1 directions in the p-dimensional space of parameters.

(ii) The k_2 remaining equations (defined by $\rho_2(.)$) should be used to identify the remaining $s_2(=p-s_1)$ directions[8]. However this additional identification comes at the slower rate λ_T.

We already have the intuition that the parameter space is going to be separated into two subspaces: the first one (defined through $\rho_1(.)$) collects s_1 standard directions and the second one (defined through $\rho_2(.)$) gathers the remaining (slow) directions. We now make this separation much more precise by defining a reparametrization. Each of the above subspaces is actually characterized as the range of a full column rank matrix: respectively the (p, s_1)-matrix R_1^0 and the $(p, p-s_1)$-matrix R_2^0.

Since R_2^0 characterizes the set of slow directions, it is natural to define it via the null space of $\left[\partial\rho_1'(\theta^0)/\partial\theta\right]$, or, in other words, everything that is not identified in a standard way (through $\rho_1(.)$):

$$\frac{\partial\rho_1(\theta^0)}{\partial\theta'}R_2^0 = 0 \tag{3.1}$$

[8]Recall that, by assumption, our set of moment conditions enables the identification of the entire vector of parameters θ.

Then these $(p - s_1)$ (slow) directions are completed with the definition of the remaining s_1 directions as follows:

$$R^0 = \left[R_1^0 \; R_2^0\right] \quad \text{and} \quad Rank\left[R^0\right] = p$$

Then R^0 is a nonsingular (p, p)-matrix that can be used as a matrix of a change of basis in $\mathbb{R}^p$. More precisely, we define the new parameter as $\eta = [R^0]^{-1}\theta$, that is

$$\theta = [R_1^0 \; R_2^0] \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \begin{matrix} \updownarrow s_1 \\ \updownarrow p - s_1 \end{matrix}$$

We will see in the next subsection that this reparametrization precisely isolates the two rates of convergence by defining two subsets of directions, each of them associated with a rate of convergence. The reparametrization also shows that, in general, there is no hope to get standard asymptotic normality of some components of the estimator $\hat{\theta}_T$ of θ^0. The reason is simple: after a standard expansion of the first-order conditions, $\hat{\theta}_T$ now appears as asymptotically equivalent to some linear transformations of $\overline{\phi}_T(\theta)$ which are likely to mix up the two rates. In other words, all components of $\hat{\theta}_T$ might be contaminated by the slow rate of convergence. Hence the main advantage of the reparametrization is precisely to separate these two rates. In section 5.1 where we carefully compare our theory with Stock and Wright (2000), we provide conditions under which some components of $\hat{\theta}_T$ are (by chance) converging at the standard rate. And this is exactly what is assumed *a priori* by Stock and Wright (2000) when they separate the structural parameters into one *standard-converging* group and one *slower-converging* one.

The reparametrization may not be feasible in practice since the matrix R^0 depends on the true unknown value of the parameter θ^0. However, we can still deduce a feasible inference strategy.

3.2 Efficient estimation

To be able to get an asymptotic normality result on the new set of parameters, we need some technical assumptions and preliminary results. More details can be found in the technical companion paper by Antoine and Renault (2007).

It is worth noting that, albeit with a mixture of different rates, the Jacobian matrix of moment conditions has a consistent sample counterpart. Let us first define the following (p, p) block

diagonal scaling matrix $\tilde{\Lambda}_T$, where Id_r denotes the identity matrix of size r:

$$\tilde{\Lambda}_T = \begin{pmatrix} \sqrt{T} Id_{s_1} & 0 \\ 0 & \lambda_T Id_{s_2} \end{pmatrix}$$

As it can be seen in the proof of theorem 2.2, assumption 2 ensures that:

$$\sqrt{T} \frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \xrightarrow{P} J^0 \quad \text{with} \quad J^0 \equiv \frac{\partial \rho(\theta^0)}{\partial \theta'} R^0 \tag{3.2}$$

where J^0 is the (K, p) block diagonal matrix with its two blocks respectively defined as the (k_i, s_i) matrices $\left[\partial \rho_i(\theta^0)/\partial \theta' \; R_i^0\right]$ for $i = 1, 2$. Note that the coexistence of two rates of convergence (λ_T and $\sqrt{T}$) implies zero north-east and south-west blocks for J^0.

Moreover to derive the asymptotic distribution of the GMM estimator $\hat{\theta}_T$ (through well-suited Taylor expansions of the first order conditions), the above convergence towards J^0 needs to be fulfilled even when the true value θ^0 is replaced by some preliminary consistent estimator θ_T^*. Hence, Taylor expansions must be robust to a λ_T-consistent estimator, the only rate guaranteed by theorem 2.3. This situation is rather similar to the one studied in Andrews (1994) for the so-called MINPIN estimator[9]. We do not want the slow convergence of some directions to contaminate the standard convergence of the others (see theorem 3.1 below): more precisely, we need to ensure that the slow rate λ_T does not modify the relative orders of magnitude of the different terms of the Taylor expansions. As Andrews (1995 p563) does for nonparametric estimators, we basically need to assume that our nearly-weakly identified directions are estimated at a rate faster than $(T^{1/4})$.[10] In addition, we want as usual uniform convergence of sample Hessian matrices. This leads us to maintain the following assumption:

Assumption 3 *(Taylor expansions)*
(i) $\lim_{T \to \infty} \left[\frac{\lambda_T^2}{\sqrt{T}} \right] = \infty$
(ii) $\overline{\phi}_T(\theta)$ *is twice continuously differentiable on the interior of* Θ *and is such that:*

$$\forall\ 1 \leq k \leq k_1 \ \frac{\partial^2 \overline{\phi}_{1T,k}(\theta)}{\partial \theta \partial \theta'} \xrightarrow{P} H_{1,k}(\theta) \quad \textit{and} \quad \forall\ 1 \leq k \leq k_2 \ \frac{\sqrt{T}}{\lambda_T} \frac{\partial^2 \overline{\phi}_{2T,k}(\theta)}{\partial \theta \partial \theta'} \xrightarrow{P} H_{2,k}(\theta)$$

[9] MINPIN estimators are estimators defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator. These nuisance parameters are estimated at slower rates and one wants to prevent their distributions to contaminate the asymptotic distribution of the parameters of interest.

[10] More details on the link between Andrews (1994, 1995) and this setting might also be found in Antoine and Renault (2007).

uniformly on θ in some neighborhood of θ^0, for some (p,p) matricial function $H_{i,k}(\theta)$ for $i = 1, 2$ and $1 \leq k \leq k_i$.

While common weak identification corresponds to $\lambda_T = 1$ and strong identification to $\lambda_T = \sqrt{T}$, our approach in the rest of the essay is actually a rather *nearly-strong* one since we assume λ_T strictly between $T^{1/4}$ and $\sqrt{T}$.[11]

Up to unusual rates of convergence, we get a standard asymptotic normality result for the new parameter η:

Theorem 3.1 *(Asymptotic Normality)*
(i) Under assumptions 1 to 3, the GMM estimator $\hat{\theta}_T$ defined by (2.2) is such that:

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[J^{0\prime}\Omega J^0\right]^{-1} J^{0\prime}\Omega S(\theta^0)\Omega J^0 \left[J^{0\prime}\Omega J^0\right]^{-1}\right)$$

(ii) Under assumptions 1 to 3, the asymptotic variance displayed in (i) is minimal[12] *when the GMM estimator $\hat{\theta}_T$ is defined with a weighting matrix Ω_T being a consistent estimator of $\Omega = [S(\theta^0)]^{-1}$:*

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[J^{0\prime}[S(\theta^0)]^{-1}J^0\right]^{-1}\right)$$

Note that $\hat{\eta}_T = [R^0]^{-1}\hat{\theta}_T$ can be interpreted as a consistent estimator of $\eta^0 = [R^0]^{-1}\theta^0$. Of course it is not feasible since R^0 is unknown. The issue of plugging in a consistent estimator of R^0 will be addressed in section 3.4. For the moment, our focus of interest are the implied rates of convergence for inference about θ. Since

$$\hat{\theta}_T = R_1^0\hat{\eta}_{1,T} + R_2^0\hat{\eta}_{2,T}$$

a linear combination $a'\hat{\theta}_T$ of the estimated parameters of interest will be endowed with a $\sqrt{T}$ rate of convergence of $\hat{\eta}_{1,T}$ if and only if $a'R_2^0 = 0$, that is a belongs to the orthogonal space of the range of R_2^0. By virtue of equation (3.1) the latter property means that a is spanned by the columns of the matrix $\left[\partial\rho_1'(\theta^0)/\partial\theta\right]$. In other words, $a'\theta$ is strongly identified if and only if it is identified by the first set of moment conditions $\rho_1(\theta) = 0$.

[11] It is worth reminding that the score test derived in section 2 is valid for λ_T arbitrarily close to the weak identification case.

[12] Note that efficiency is implicitly considered here for the given set of moment restrictions $\overline{\phi}_T(.)$. In section 3.3, we study the consequences of rewriting the moment conditions.

As far as inference about θ is concerned, several practical implications of theorem 3.1 are worth mentioning. Up to the unknown matrix R^0 and the unknown rate of convergence λ_T (which appears in $\tilde{\Lambda}_T$), a consistent estimator of the asymptotic covariance matrix $\left(J^{0\prime}\left[S(\theta^0)\right]^{-1}J^0\right)^{-1}$ is[13]

$$T^{-1}\tilde{\Lambda}_T[R^0]^{-1}\left[\frac{\partial\overline{\phi}'_T(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}[R^{0\prime}]^{-1}\tilde{\Lambda}_T \tag{3.3}$$

where S_T is a standard consistent estimator of the long-term covariance matrix[14]. From theorem 3.1, for large T, $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T-\theta^0)$ behaves like a gaussian random variable with mean zero and variance (3.3). One may be tempted to deduce that $\sqrt{T}(\hat{\theta}_T-\theta^0)$ behaves like a gaussian random variable with mean 0 and variance

$$\left[\frac{\partial\overline{\phi}'_T(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \tag{3.4}$$

And this would give the feeling that we are back to standard GMM formulas of Hansen (1982). As far as practical purposes are concerned, this intuition is correct: note in particular that the knowledge of R^0 is not necessary to perform inference. However, from a theoretical point of view, this is a bit misleading. First since in general all components of $\hat{\theta}_T$ converge at the slow rate, $\sqrt{T}(\hat{\theta}_T-\theta^0)$ has no limit distribution! In other words, considering the asymptotic variance (3.4) is akin to refer to the inverse of an asymptotically singular matrix. Second, for the same reason, (3.4) is not an estimator of the standard population matrix

$$\left[\frac{\partial\rho'(\theta^0)}{\partial\theta}[S(\theta^0)]^{-1}\frac{\partial\rho(\theta^0)}{\partial\theta'}\right]^{-1} \tag{3.5}$$

To conclude, if inference about θ is technically more involved than one may believe at first sight, it is actually very similar to standard GMM formulas from a pure practical point of view. In other words, if a practitioner is not aware of the specific framework with moment conditions associated with several rates of convergence (coming, say, from the use of instruments of different qualities) then she can still provide reliable inference by using standard GMM formulas. In this respect, we generalize Kleibergen's (2005) result that inference can be performed without *a priori* knowledge of the identification setting. However as already

[13] This directly follows from lemma B in the appendix.

[14] Note that a consistent estimator of S_T of the long-term covariance matrix $S(\theta^0)$ can be built in the standard way (see in general Hall(2005)) from a preliminary inefficient GMM estimator $\tilde{\theta}_T$ of θ.

mentioned in section 2, we are more general that Kleibergen (2005) since we allow moment conditions to display simultaneously different identification patterns[15].

Finally, the standard score test defined in theorem 2.2 may be completed by a classical overidentification test:

Theorem 3.2 *(J-test)*
Under assumptions 1 to 3, if Ω_T *is a consistent estimator of* $\left[S(\theta^0)\right]^{-1}$*, then*

$$TQ_T(\hat{\theta}_T) \xrightarrow{d} \chi^2_{K-p}$$

3.3 Orthogonalization of the moment restrictions

In this section, we investigate the consequences of transforming the moment restrictions to estimate the standard and slow directions. Since we deal simultaneously with standard and weaker moment conditions, we cannot consider any linear combination of the restrictions. In particular, we can only consider transformations preserving the central limit theorem in Assumption 1, and the fragile information of the weaker moment restrictions. Any valid transformation of the moment conditions, or transformation that does not affect the true moment conditions $\rho_1(.)$ and $\rho_2(.)$, can be written as follows:

$$\begin{bmatrix} \overline{\phi}^H_{1T}(\theta^0) \\ \overline{\phi}^H_{2T}(\theta^0) \end{bmatrix} = \begin{bmatrix} \overline{\phi}_{1T}(\theta^0) + H\overline{\phi}_{2T}(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) \end{bmatrix} \tag{3.6}$$

for some (k_1, k_2)- matrix H that may depend on the sample size T and the true unknown parameter θ^0.

A linear transformation of interest in the literature is the orthogonalization: the standard moment conditions are replaced by the residuals of their regression on the set of weaker moment conditions. The set of the empirical mean of the moment conditions $[\overline{\phi}'_{1T}\ \overline{\phi}'_{2T}]'$ is replaced by $[\tilde{\phi}'_{1T}\ \tilde{\phi}'_{2T}]'$ defined as follows:

$$\begin{bmatrix} \overline{\phi}_{1T}(\theta^0) - Cov\left(\sqrt{T}\overline{\phi}_{1T}(\theta^0), \sqrt{T}\overline{\phi}_{2T}(\theta^0)\right)\left[Var\left(\sqrt{T}\overline{\phi}_{2T}(\theta^0)\right)\right]^{-1}\overline{\phi}_{2T}(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) \end{bmatrix} \tag{3.7}$$

[15]For sake of notational simplicity, we only consider in this essay one speed of nearly-weak identification λ_T. The reader interested in working with an arbitrary number of different speeds might use the general framework of Antoine and Renault (2007).

Note that we still have the same central limit theorem, only the asymptotic variance is modified:

$$\sqrt{T}\left[\begin{array}{c}\tilde{\phi}_{1T}(\theta^0)-\rho_1(\theta^0)\\ \tilde{\phi}_{2T}(\theta^0)-\frac{\lambda_T}{\sqrt{T}}\rho_2(\theta^0)\end{array}\right]\Rightarrow\tilde{\Psi}(\theta^0)$$

where $\tilde{\Psi}(\theta)$ is a Gaussian random variable with mean zero and block diagonal matrix Σ^0 with respective blocks $\Sigma_1^0=S_1^0-S_{12}^0[S_2^0]^{-1}S_{21}^0$ and $\Sigma_2^0=S_2^0$.

The following theorem compares the asymptotic variances of the estimators associated to the original set of moment conditions $\hat{\eta}_T$ and to the orthogonalized one denoted as $\tilde{\eta}_T$:

Theorem 3.3 *(Orthogonalization)*
Consider the new parameter $\eta=[R^0]^{-1}\theta$. The two estimators obtained respectively from the GMM estimator associated with the original set of moment conditions $\hat{\theta}_T$ and from the GMM estimator associated with the orthogonalized set of moment conditions (3.7) $\tilde{\theta}_T$ are such that:
i) The orthogonalization improves the estimation of the standard directions (in terms of asymptotic variance matrix) ie

$$AVar\left[\hat{\eta}_{1T}\right]\geq\geq AVar\left[\tilde{\eta}_{1T}\right]$$

ii) The orthogonalization deteriorates the estimation of the slow directions (in terms of asymptotic variance matrix) ie

$$AVar\left[\hat{\eta}_{2T}\right]\leq\leq AVar\left[\tilde{\eta}_{2T}\right]$$

where $\leq\leq$ and $\geq\geq$ denote the comparisons between matrixes.

We show that the orthogalization of any valid set of moment restrictions (3.6) leads to the same set (3.7):

Proposition 3.4 *Any set of valid moment conditions like (3.6) leads to the same orthogonalized set of moment conditions (3.7).*

Denote by η_T^H the (transformed) estimator associated to the above moment conditions (3.6). We now show that among all the valid transformations, the orthogonalization is the best for the standard directions and the worse the slow ones.

Corollary 3.5 *Consider the new parameter $\eta = [R^0]^{-1}\theta$. The two estimators obtained respectively from the GMM estimator associated with the transformed set of moment conditions (3.6) $\hat{\theta}_T^H$ and from the GMM estimator associated with the orthogonalized set of moment conditions (3.7) $\tilde{\theta}_T$ are such that:*
i) The orthogonalization is the best (valid) transformation of the moment conditions in terms of the efficiency of the standard directions ie

$$AVar\left[\eta_{1T}^H\right] \geq\geq AVar\left[\tilde{\eta}_{1T}\right]$$

ii) The orthogonalization is the worst (valid) transformation of the moment conditions in terms of the efficiency of the slow directions ie

$$AVar\left[\eta_{2T}^H\right] \leq\leq AVar\left[\tilde{\eta}_{2T}\right]$$

3.4 Estimating the strongly-identified directions

In this subsection, we provide a feasible way to estimate the strongly-identified directions in the parameter space. Recall that these directions have been identified through the following reparametrization,

$$\left[R^0\right]^{-1}\theta \equiv \begin{pmatrix} A^0\theta \\ B^0\theta \end{pmatrix}$$

where $(A^0\theta)$ represent the s_1 standard directions while $(B^0\theta)$ are the weaker ones. In general, this reparametrization is unfeasible since it depends on the unknown value of the parameter θ^0. To make this approach feasible, the key lemma which allows us to replace the above directions by their estimated counterparts is the following:

Lemma 3.6 *(Estimating the rotation in the parameter space)*
Under assumptions 1 to 3, if the vector

$$\begin{bmatrix} \sqrt{T}A\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T B\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

is asymptotically gaussian and if $\hat{A}$ and $\hat{B}$ are consistent estimators of A and B such that

$$\left\|\hat{A} - A\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right) \quad and \quad \left\|\hat{B} - B\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

then the vector

$$\begin{bmatrix} \sqrt{T}\hat{A}\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T \hat{B}\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

is asymptotically gaussian.

In the proof of lemma 3.6, our *nearly-strongly* point of view is the essential key to keep the $\sqrt{T}$ convergence while relevant directions are only estimated at the slower rate λ_T: that is λ_T is small in front of $\sqrt{T}$ but large in front of $\left[T^{1/4}\right]$.[16]

4 Wald testing

In this section, we focus on testing a system of q restrictions about θ, say the null hypothesis $H_0 : g(\theta) = 0$, where $g(.)$ is a function from Θ to $\mathbb{R}^q$ continuously differentiable on the interior of Θ.

First, working under the null may conduct us to dramatically revisit the reparametrization $\eta = [R^0]^{-1}\theta$ defined in section 3. Typically additional information may lead us to define differently the linear combinations of θ estimated respectively with standard and slow rates of convergence. To circumvent this difficulty, we do not consider any constrained estimator and we focus on Wald testing. Caner (2005) overlooks this complication and derives the standard asymptotic equivalence results for the trinity of tests. This is because he only treats asymptotic testing when all the parameters converge at the same speed.

Second, as already explained, the main originality of this essay is to allow for the simultaneous treatment of different identification patterns. This more general point of view comes at a price when one wants to test. More precisely, we may face singularity issues when some tested restrictions estimated at the slow rate λ_T can be linearly combined so as to be estimated at the standard rate $\sqrt{T}$. Lee (2005) puts forward some high level assumptions (see his assumptions (R) and (G)) to deal with the asymptotic singularity problem. We show that our setting allows us to perform a standard Wald test even without maintaining Lee's (2005) high-level assumptions.

[16] As already mentioned, this is very similar in spirit to MINPIN estimators of Andrews (1994, 1995).

From our discussion in sections 2 and 3, we can guess that a Wald test statistic for H_0 can actually be written with a standard formula:

$$\zeta_T^W = Tg'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} g(\hat{\theta}_T)$$

Recall the standard rank assumption ensuring that the Wald test statistic is asymptotically chi-square with q degrees of freedom:

$$Rank \left[\frac{\partial g(\theta)}{\partial \theta'} \right] = q \tag{4.1}$$

for all θ in the interior of Θ, or at least in a neighborhood of θ^0. As well known, this condition is not really restrictive since it is akin to say that, at least locally, the q restrictions under test are linearly independent. Unfortunately, the existence of different rates of convergence may introduce (asymptotically) some perverse multicolinearity between the q estimated constraints.

The counterexample below points out the key issue.

Example 4.1 *(Counterexample)*
Assume that we want to test $H_0 : g(\theta) = 0$ with $g(\theta) = [g_j(\theta)]_{1 \leq j \leq q}$ and none of the q vectors $\partial g_j(\theta^0)/\partial \theta$, $j = 1, \cdots, q$ belongs to $Im[\partial \rho_1'(\theta^0)/\partial \theta]$[17]. Then the extension of the standard argument for Wald test would be to say that, under the null, $\lambda_T g(\hat{\theta}_T)$ behaves asymptotically like $\partial g(\theta^0)/\partial \theta' \lambda_T(\hat{\theta}_T - \theta^0)$, that is for large T, $\lambda_T g(\hat{\theta}_T)$ behaves like a gaussian

$$\mathcal{N}\left(0, \frac{\partial g(\theta^0)}{\partial \theta'} \left[\frac{\partial \overline{\phi}_T'(\theta^0)}{\partial \theta} S(\theta^0)^{-1} \frac{\partial \overline{\phi}_T(\theta^0)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\theta^0)}{\partial \theta}\right)$$

Imagine however that for some nonzero vector α,

$$\alpha' \frac{\partial g(\theta^0)}{\partial \theta'} = \sum_{j=1}^{q} \alpha_j \frac{\partial g_j(\theta^0)}{\partial \theta'}$$

belongs to $Im[\partial \rho_1'(\theta^0)/\partial \theta]$. Then (see comments after theorem 3.1) under the null $\sqrt{T}\alpha' g(\hat{\theta}_T)$ is asymptotically gaussian and thus

$$\lambda_T \alpha' g(\hat{\theta}_T) = \frac{\lambda_T}{\sqrt{T}} \sqrt{T} \alpha' g(\hat{\theta}_T) \xrightarrow{P} 0$$

[17]For any $(n \times m)$-matrix, $Im[M]$ represents the subspace of $\mathbb{R}^n$ generated by the column vectors of M. It is also referred to as $Col[M]$ and $Range[M]$.

In other words, even if the q constraints are locally linearly independent (ie $Rank[\partial g(\theta^0)/\partial\theta'] = q$*)* $\left[\lambda_T g(\hat{\theta}_T)\right]$ *does not behave asymptotically like a gaussian with a non-singular variance matrix. This is the reason why deriving an asymptotically* $\chi^2(q)$ *distribution for the Wald test statistic is more involved than usual.*

Lee (2005) avoids this kind of perverse asymptotic singularity by maintaining the following assumption:

Lee's (2005) assumption:
There exists a sequence of (q,q) *invertible matrices* D_T *such that for any* $\theta \in \Theta$

$$\underset{T\to\infty}{Plim}\left[D_T \frac{\partial g(\theta^0)}{\partial\theta'} R^0[\tilde{\Lambda}_T]^{-1}\right] = B_0$$

where B_0 *is a* (q,p) *deterministic finite matrix of full row rank.*

Lee's (2005) assumption clearly implies the standard rank condition (4.1). However, the converse is not true as it can be shown from the counterexample above[18]. And this is actually what is needed to justify a Wald test, through a delta-theorem approach as usual. Note that the above assumption implies that, under the null, $D_T g(\hat{\theta}_T)$ behaves like $D_T \partial g(\theta^0)/\partial\theta'(\hat{\theta}_T - \theta^0)$ that is like $B^0\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)$. From theorem 3.1, we know that $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)$ nicely behaves as an asymptotic gaussian distribution. In other words, the matrix D_T provides us with the right scaling to get asymptotic normality of $\partial g(\theta^0)/\partial\theta'(\hat{\theta}_T - \theta^0)$. However, we can prove that the standard practice of Wald testing is valid even without Lee's assumption:

Theorem 4.1 *(Wald test)*
Under the assumptions 1 to 3 and if $g(.)$ *is twice continuously differentiable, the Wald test statistic* ζ_T^W *for testing* $H_0 : g(\theta) = 0$ *is asymptotically* $\chi^2(q)$ *under the null.*

While a detailed proof of theorem 4.1 is provided in the appendix, it is worth explaining why it works in spite of the aforementioned singularity problem. The key intuition is somewhat related to the well-known phenomenon that the finite sample performance of the Wald test depends on the way the null hypothesis is formulated[19].

[18] By contrast, in the case of only $q = 1$ constraint, Lee's assumption is trivially fulfilled.

[19] In some respect, our approach of nearly-weak identification complements the higher order expansions of Phillips and Park (1988).

Let us first imagine a fictitious situation where the range of $\left[\partial\rho_1'(\theta^0)/\partial\theta\right]$ is known. Then it is always possible to define a (q,q) nonsingular matrix H and a q dimensional function $h(\theta) = Hg(\theta)$ to ensure a *genuine disentangling* of the strongly identified and nearly-weakly identified directions to be tested. By genuine disentangling, we mean that for some q_1 such that $1 \leq q_1 \leq q$:

- for $j = 1, \cdots, q_1$: $\left[\partial h_j(\theta^0)/\partial\theta\right]$ belongs to $Im\left[\partial\rho_1'(\theta^0)/\partial\theta\right]$
- for $j = q_1+1, \cdots, q$: $\left[\partial h_j(\theta^0)/\partial\theta\right]$ does not belong to $Im\left[\partial\rho_1'(\theta^0)/\partial\theta\right]$ and no linear combinations of them do.

Then the perverse asymptotic singularity of example 4.1 is clearly avoided. Of course, at a deeper level, the new restrictions $h(\theta) = 0$ to be tested should be interpreted as a nonlinear transformation of the initial ones $g(\theta) = 0$ (since the matrix H depends on θ). It turns out that, for all practical purposes, by fictitiously seeing H as known, the Wald test statistics written with $h(.)$ or $g(.)$ are numerically equal. The proof of theorem 4.1 shows that this is the key reason why standard Wald test always works (despite appearing invalid at first sight).

As far as the size of the test is concerned, the existence of the two rates of convergence does not modify the standard Wald result. Of course, the power of the test heavily depends on the strength of identification of the various constraints to test. More precisely, if, for the sake of notational simplicity, we consider only $q = 1$ restriction to test, we get:

Theorem 4.2 *(Local alternatives)*
Under assumptions 1 to 3, the Wald test of $H_0 : g(\theta) = 0$ (with $g(.)$ one dimensional continuously differentiable) is consistent under the sequence of local alternatives $H_{1T} : g(\theta) = 1/\delta_T$ if and only if either

$$\frac{\partial g(\theta^0)}{\partial\theta} \in Im\left[\frac{\partial\rho_1'(\theta^0)}{\partial\theta}\right] \quad and \quad \delta_T = o(\sqrt{T})$$

or

$$\frac{\partial g(\theta^0)}{\partial\theta} \notin Im\left[\frac{\partial\rho_1'(\theta^0)}{\partial\theta}\right] \quad and \quad \delta_T = o(\lambda_T)$$

The proof of theorem 4.2 is rather straightforward. In the line of the comments following theorem 3.1, a nonlinear function $g(.)$ of θ, interpreted as $\left[g(\theta^0) + \frac{\partial g(\theta^0)}{\partial\theta'}(\theta - \theta^0)\right]$, is identified at the standard rate $\sqrt{T}$ if and only if

$$\frac{\partial g(\theta^0)}{\partial\theta'} \in Im\left[\frac{\partial\rho_1'(\theta^0)}{\partial\theta}\right]$$

5 Examples

We now work out several examples to illustrate the general theory of the previous sections as well as to shed some light on the link between our approach and Stock and Wright (2000).

5.1 Single-equation linear IV model

As already mentioned, the major difference between Stock and Wright's (2000) framework and ours lies in considering the subvector of strongly identified parameters as known a priori. The context of the linear IV regression model sheds some light on the relationships linking the two procedures. Consider the following single-equation linear IV model with two structural parameters, two orthogonal instruments and no exogenous variables for convenience:

$$\left\{ \begin{array}{ccccccc} y & = & Y & \theta & + & u \\ \text{(T,1)} & & \text{(T,2)} & \text{(2,1)} & & \text{(T,1)} \\ Y & = & [X_1 \ X_2] & C & + & [V_1 \ V_2] \\ \text{(T,2)} & & \text{(T,2)} & \text{(2,2)} & & \text{(T,2)} \end{array} \right. \tag{5.1}$$

As commonly done in the literature the matrix of coefficients C is artificially linked to the sample size T in order to introduce some (nearly)-weak identification issues. However, to accommodate both interpretations of the identification issues, the matrices C_T are different. For our characterization (directly through the moment conditions) and for Stock and Wright's characterization (through the parameters) we have respectively:

$$C_T^{AR} = \left[\begin{array}{cc} \pi_{11} & \pi_{12} \\ \pi_{21}/T^{\lambda} & \pi_{22}/T^{\lambda} \end{array} \right] \quad \text{and} \quad C_T^{SW} = \left[\begin{array}{cc} \pi_{11} & \pi_{12}/T^{\lambda} \\ \pi_{21} & \pi_{22}/T^{\lambda} \end{array} \right] \tag{5.2}$$

Choosing C_T^{AR} modifies the explanatory power of the second instrument X_2 only. As a result, one standard moment condition naturally emerges (associated with X_1) and one less informative (associated with X_2). Intuitively, the standard restriction should identify one standard direction in the parameter space, which is so far unknown. On the other hand, choosing C_T^{SW} is equivalent to modeling θ_2 as weakly identified. The price to pay for such an early knowledge is the alteration of the explanatory powers of both instruments. Typically the moment

conditions,

$$E\left[(y_t - Y_t'\theta)X_t\right]$$

are respectively written as:

$$\left\{\begin{array}{cccccc} & E(X_{1t}^2)\pi_{11}(\theta_1^0-\theta_1) & + & & E(X_{1t}^2)\pi_{12}(\theta_2^0-\theta_2) \\ \frac{1}{T^\lambda} & E(X_{2t}^2)\pi_{21}(\theta_1^0-\theta_1) & + & \frac{1}{T^\lambda} & E(X_{2t}^2)\pi_{22}(\theta_2^0-\theta_2) \end{array}\right.$$

and

$$\left\{\begin{array}{ccccc} E(X_{1t}^2)\pi_{11}(\theta_1^0-\theta_1) & + & \frac{1}{T^\lambda} & E(X_{1t}^2)\pi_{12}(\theta_2^0-\theta_2) \\ E(X_{2t}^2)\pi_{21}(\theta_1^0-\theta_1) & + & \frac{1}{T^\lambda} & E(X_{2t}^2)\pi_{22}(\theta_2^0-\theta_2) \end{array}\right.$$

or, in a more compact way,

$$\left\{\begin{array}{l} \rho_1^{AR}(\theta_1,\theta_2) \\ \frac{1}{T^\lambda}\rho_2^{AR}(\theta_1,\theta_2) \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{l} m_{1s}^{SW}(\theta_1) + \frac{1}{T^\lambda} m_{1w}^{SW}(\theta_2) \\ m_{2s}^{SW}(\theta_1) + \frac{1}{T^\lambda} m_{2w}^{SW}(\theta_2) \end{array}\right. \tag{5.3}$$

for some real functions $\rho_1^{AR}(.)$, $\rho_2^{AR}(.)$, $m_{1s}^{SW}(.)$, $m_{1w}^{SW}(.)$, $m_{2s}^{SW}(.)$ and $m_{2w}^{SW}(.)$.

We now introduce our reparametrization of section 2 to identify the standard direction in the parameter space. The derivative of the standard moment restriction is

$$J_1^0 = \frac{\partial \rho_1(\theta^0)}{\partial \theta'} = \left[-E(Y_{1t}X_{1t}) \vdots - E(Y_{2t}X_{1t})\right] = \left[-E(X_{1t}^2)\pi_{11} \vdots - E(X_{1t}^2)\pi_{12}\right]$$

Hence the null space of J_1^0 is characterized by the following vector:

$$R = \left[\begin{array}{c} -\pi_{12} \\ \pi_{11} \end{array}\right]\mu \qquad \text{where } \mu \in \mathbb{R}^*$$

It is then completed into a legitimate matrix of change of basis R^0 in the parameter space $\mathbb{R}^2$:

$$R^0 = \left[\begin{array}{cc} a & -\pi_{12}\mu \\ b & \pi_{11}\mu \end{array}\right] \qquad \text{with } (a,b) \in \mathbb{R}^2 \ / \ a\pi_{11} \neq -b\pi_{12}$$

The new parameter η can now be defined as follows: $\eta = [R^0]^{-1}\theta$ that is

$$\left\{\begin{array}{l} \eta_1 = \frac{1}{a\pi_{11}+b\pi_{12}}\left[\pi_{11}\theta_1 + \pi_{12}\theta_2\right] \\ \eta_2 = -\frac{b}{\mu(a\pi_{11}+b\pi_{12})}\theta_1 + \frac{a}{\mu(a\pi_{11}+b\pi_{12})}\theta_2 \end{array}\right.$$

The standard direction is completely determined and not the weaker one. The main reason comes from the fact that everything that is not standard is weaker: in fact, the weaker directions contaminate the standard ones, when considering a linear combination of the two.

The above calculation shows that, strictly speaking, Stock and Wright (2000) and their linear reinterpretation of Staiger and Stock (1997) are not nested in our setting because each of their moment condition contains a strong part (that only depends on a subvector of parameter) and a weak part. Note that this setting (through the definition of the matrix C_T^{SW}) is conveniently built so as to know *a priori* which subset of the parameters is strongly identified. Now, if we pretend that we did not realize that the set of strongly identified parameter was known and we still perform the change of variables, we get:

$$J_1^0 \simeq [-\pi_{11}\ 0] \quad \text{hence} \quad R = [0\ \mu]' \quad \text{with} \quad \mu \in \mathbb{R}^*$$

and the change of basis is defined as:

$$R^0 = \begin{bmatrix} a & 0 \\ b & \mu \end{bmatrix} \quad \text{with } a \neq 0 \implies \eta = \begin{bmatrix} \mu & 0 \\ -b & a \end{bmatrix}$$

As expected, we identify the strongly identified direction as being parallel to θ_1. This is a nice internal consistency result of our procedure.

5.2 Non-linear IV model

As already mentioned in the linear case, our general setting does not strictly nest Stock and Wright (2000). However, we can show that the two procedures are relatively close to each other. Recall first the underlying assumptions on the moment restrictions:

$$\begin{aligned} Nearly - Weak \qquad & E\left[\overline{\phi}_T(\theta)\right] = \frac{\Lambda_T}{\sqrt{T}}\rho(\theta) \\ Staiger - Stock \qquad & E\left[\overline{\phi}_T(\theta)\right] = m_1(\theta_1) + \frac{1}{\sqrt{T}}m_2(\theta) \end{aligned}$$

where θ_1 is the *a priori assumed* strongly identified parameter.

Let us now derive the first-order conditions associated with our minimization problem when the weighting matrix is chosen to be block diagonal such that $\Omega_D = diag[\Omega_{D1}\ \Omega_{D2}]$ with Ω_{Di}

symmetric full rank (k_i, k_i)-matrix, i=1,2:

$$\min_{\theta} \left[\overline{\phi}'_{1T}(\theta)\Omega_{D1}\overline{\phi}_{1T}(\theta) + \overline{\phi}'_{2T}(\theta)\Omega_{D2}\overline{\phi}_{2T}(\theta) \right]$$

The associated first order conditions are

$$\frac{\partial \overline{\phi}'_{1T}(\hat{\theta}_T)}{\partial \theta}\Omega_{D1}\overline{\phi}_{1T}(\hat{\theta}_T) + \frac{\partial \overline{\phi}'_{2T}(\hat{\theta}_T)}{\partial \theta}\Omega_{D2}\overline{\phi}_{2T}(\hat{\theta}_T) = 0$$

The above first order condition can be seen as the selection of linear combinations of $\overline{\phi}_T$. If $\overline{\phi}_{1T}$ only depends on θ_1 then, after imposing $\lambda_T = 1$, our resulting linear combinations of the moment restrictions correspond to the ones of Stock and Wright (2000).

Note also that the null space used to reparametrize the problem can be defined directly from the above first order conditions after realizing that:

$$\left[\frac{\partial \overline{\phi}'_{1T}(\hat{\theta}_T)}{\partial \theta} \right] \Omega_{D1} \xrightarrow{P} \frac{\partial \rho'_1(\theta^0)}{\partial \theta}\Omega_{D1}$$

where $\Omega_{D1}\left[\partial \rho_1(\theta^0)/\partial \theta'\right]$ defines the same null space as $\left[\partial \rho_1(\theta^0)/\partial \theta'\right]$ since Ω_{D1} is a full rank squared matrix.

5.3 Estimation of the rates of convergence

In some special convenient cases (as a Monte Carlo study), it is possible to estimate the rate of convergence of our estimators via a linear regression. The idea is to simulate the model for several sample sizes: for each sample size, the simulation is replicated M times to get M draws of the estimator. The Monte-Carlo distribution of the estimate can then be deduced and its variance calculated. Finally, the regression of logarithm of the variance on the constant regressor and the logarithm of the sample size is performed:

$$\log(Var(\hat{\theta}_T)) = \alpha + \beta \log T + u_T \tag{5.4}$$

where u_T is some error term. β can be estimated by OLS and it gives an estimate of the square of the convergence rate.

Section 6 below provides some illustrations of the estimation of the rates of convergence.

6 Monte-Carlo Study

6.1 Single-Equation linear IV model

In our first Monte-Carlo study, our goal is to verify the finite sample relevance of our asymptotic theory. In particular, we use the linear regression technique developed in section 5.3 to estimate the rates of convergence of the transformed parameters as well as the ones of the original parameters. Recall first the linear model of example 1 in section 3.2:

$$\left\{ \begin{array}{ccccccc} y & = & Y & \theta & + & u \\ \text{(T,1)} & & \text{(T,2)} & \text{(2,1)} & & \text{(T,1)} \\ Y & = & [X_1 \ X_2] & C_T & + & [V_1 \ V_2] \\ \text{(T,2)} & & \text{(T,2)} & \text{(2,2)} & & \text{(T,2)} \end{array} \right. \tag{6.1}$$

$$\text{with} \quad C_T = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21}/T^{\mu} & \pi_{22}/T^{\mu} \end{bmatrix} \quad \text{and} \quad 0 < \mu < 1/2$$

The above model is estimated for several sample sizes as well as several degrees of weakness. We provide the results for $\mu = 1/5$: it corresponds to a *slow* convergence rate equal to $\lambda_T = T^{0.3}$, as introduced in section 2. This is a *strong nearly-weak* identification case and λ_T satisfies assumption 3.

Generally speaking the results are pretty good and conform to the theory. The main findings are listed here: i) The variance decreases with the sample size for the four estimators $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$, $\hat{\eta}_{1T}$ and $\hat{\eta}_{2T}$. Moreover figure 1 plots the log-variance as a function of the log-sample size: for the above estimators, it is a fairly straight decreasing line. This gives support to the fact that the variance is proportional to the sample size raised at some power;
ii) We now compare the rates of convergence among the sets of parameters by studying ratios of parameters, or specifically $\hat{\eta}_{2T}/\hat{\eta}_{1T}$ and $\hat{\theta}_{1T}/\hat{\theta}_{2T}$. From figure 2, the ratio of the new set of parameters increase with the sample size whereas the ratio of the original set of parameters is fairly flat. This supports the fact that $\hat{\eta}_{1T}$ converges faster than $\hat{\eta}_{2T}$, whereas $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ converge at a similar rate;
iii) Finally, we present the results of the estimation of the rates of convergence with the linear regression technique described in section 5.3. See tables 2 and 3. According to our asymptotic theory, we expect to find a standard rate of $T^{1/2}$ for $\hat{\eta}_{1T}$ and a slow rate equal to $T^{0.3}$ for the three remaining parameters. Over the entire sample, the standard rate is relatively well

estimated. On the other hand, the slow rate is less precise and we cannot conclude to the equality of the rates of convergence for $\hat{\eta}_{2T}$, $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$. However, when we consider only larger sample sizes (>5000), we get closer to the expected result. Since the convergence is slower, more data are expected to be needed to conclude.

6.2 CCAPM

In this section, we report some Monte-Carlo evidence about the intertemporally separable consumption capital asset pricing model (CCAPM) with constant relative risk-aversion (CRRA) preferences. The artificial data are generated in order to mimic the dynamic properties of the historical data. Hence, we can assess the empirical relevance of our general setting in such a context.

Moment conditions

The Euler equations lead to the following moment conditions:

$$E\left[h_{t+1}(\theta)|\mathcal{I}_t\right] = 0 \quad \text{with} \quad h_t(\theta) = \delta r_t c_t^{-\gamma} - 1$$

Our parameter of interest is then $\theta = [\delta \ \gamma]'$, with δ the discount factor and γ the preference parameter; (r_t, c_t) denote respectively a vector of asset returns and the consumption growth at time t. To be able to estimate this model, our K instruments $Z_t \in \mathcal{I}_t$ include the constant as well as some lagged variables. We then rewrite the above moment conditions as

$$E_0\left[\phi_{t,T}(\theta)\right] = E_0\left[h_{t+1}(\theta) \otimes Z_{t,T}\right]$$

Note that to stress the potential weakness of the instruments (and as a result of the moment function, see section 2.1), we add the subscript T.

Data Generation:

Our Monte-Carlo design follows Tauchen (1986), Kocherlakota (1990), Hansen, Heaton and Yaron (1996) and more recently Stock and Wright (2000). More precisely, the artificial data are generated by the method discussed in Tauchen and Hussey (1991). This method fits a 16 state Markov chain to the law of motion of the consumption and the dividend growths, so as to approximate a beforehand calibrated gaussian VAR(1) model (see Kocherlakota (1990)). The CCAPM-CRRA model is then used to price the stocks and the riskfree bond in each time period, yielding a time series of asset returns.

It is important to stress that since the data are generated from a general equilibrium model, even the econometrician does not know whether (δ, γ) are (nearly)-weakly identified or not. In a similar study, Stock and Wright (2000) impose a different treatment for the parameters δ and γ: typically, δ is taken as strongly identified whereas γ is not. We do not make such an assumption. We are actually able, through a convenient reparametrization, to identify some directions of the parameter space that are strongly identified and some others that are not.

Strong and weak moment conditions:

We consider here three instruments: the constant, the centered lagged asset return and the centered lagged consumption growth. To be able to apply our nearly-weak GMM estimation, we need to separate the instruments (and the associated moment conditions) according to their *strength*. Typically, a moment restriction $E[\phi_t(\theta)]$ is (nearly)-weak when $E[\phi_t(\theta)]$ is *close* to zero for all θ. This means that the restriction does not permit to (partially) identify θ. Hence, we decide to evaluate each moment restriction for a grid of parameter values. If the moment is uniformly *close* to 0 then we conclude to its weakness. Note that this study can always be performed and is not specifically related to the Monte-Carlo setting; the Monte-Carlo setting is simply convenient to get rid of the simulation noise by averaging over the many simulated samples.

Figure 3 has been built with a sample size of 100 and 2500 Monte-Carlo replications. Note that the conclusions are not affected when larger sample sizes are considered.

The above study clearly reveals two groups of moment restrictions: i) with the constant instrument, the associated restriction varies quite substantially with the parameter θ; ii) with the lagged instruments, both associated restrictions remain fairly small when θ vary over the grid. The set of instruments, and accordingly of moment conditions, is then separated as follows:

$$\phi_{t,T}(\theta) \quad = \quad \left(\begin{array}{c} (\delta r_t c_t^{-\gamma} - 1) \\ (\delta r_t c_t^{-\gamma} - 1) \otimes \left[\begin{array}{c} r_{t-1} - \bar{r} \\ c_{t-1} - \bar{c} \end{array} \right] \end{array} \right)$$

Accordingly,

$$\overline{\phi}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{t,T}(\theta) \quad \text{with} \quad \sqrt{T} E[\overline{\phi}_T(\theta)] = \left(\begin{array}{cc} \lambda_{1T} & 0_{1,2} \\ 0_{2,1} & \lambda_{2T} Id_2 \end{array} \right) \left(\begin{array}{c} \rho_1(\theta) \\ \rho_2(\theta) \end{array} \right)$$

As emphasized earlier, our Monte-Carlo study simulates a general equilibrium model. So, even the econometrician does not know in advance which moment conditions are weak and the level of this weakness. Hence, λ_{1T} and λ_{2T} must be chosen so as to fulfill the following conditions, $\lambda_{1T} = o(\sqrt{T})$, $\lambda_{2T} = o(\lambda_{1T})$ and $\lambda_{1T} = o(\lambda_{2T}^2)$.

In their theoretical considerations (section 4.1), Stock and Wright (2000) also treat differently the covariances of the moment conditions. The strength of the constant instrument is actually used to provide some intuition on their identification assumptions (δ strongly identified and γ weakly identified). However, we maintain that if γ is weakly identified, then it affects the covariance between r_t and $c_t^{-\gamma}$, and hence the identification of δ is altered too. This actually matches some asymptotic results of Stock and Wright (2000) where the weak parameter affects the strong one, by preventing it to converge to a standard gaussian random variable.

We now identify the strong directions in the parameter space via the reparametrization introduced in section 3.

Reparametrization:

First, we define the matrix of the change of basis (or reparametrization), that enables us to identify the standard directions in the parameter space. Recall that it is defined through the null space of the following matrix,

$$J_1^0 = \frac{\partial \rho_1(\theta^0)}{\partial \theta'}$$

Straightforward calculations lead to:

$$\left[\frac{\partial \phi_{1,t}(\theta)}{\partial \theta'}\right] = \left[\frac{\partial \phi_{1,t}(\theta)}{\partial \delta} \quad \frac{\partial \phi_{1,t}(\theta)}{\partial \gamma}\right] = \left[r_t c_t^{-\gamma} \;\vdots\; -\gamma\delta r_t c_t^{-\gamma-1}\right]$$

J_1^0 is then approximated as follows:

$$\hat{J} = \frac{\partial \hat{\rho}_1(\theta^0)}{\partial \theta'} = \left[\frac{1}{T}\sum_{t=1}^{T} r_t c_t^{-\gamma^0} \;\vdots\; -\frac{\gamma^0\delta^0}{T}\sum_{t=1}^{T} r_t c_t^{-\gamma^0-1}\right]$$

The null space of J_1^0 is defined via the (2,1)-matrix R_2 such that,

$$J_1^0 R_2 = 0 \Leftrightarrow R_2 = \nu \begin{bmatrix} -J_{12} \\ J_{11} \end{bmatrix} \quad \text{for any nonzero real number } \nu$$

R_2 is then completed with R_1 into the matrix R^0 so as to define a legitimate reparametrization.

In other words, R^0 is of full rank. So practically the only constraint is the following,

$$R_1 = \begin{bmatrix} a \\ b \end{bmatrix} \quad \text{with} \quad \frac{a}{b} \neq -\frac{J_{12}}{J_{11}}$$

We then get,

$$R^0 = \begin{bmatrix} a & -\nu J_2 \\ b & \nu J_1 \end{bmatrix} \quad \text{and} \quad [R^0]^{-1} = \frac{1}{\mu(aJ_{11} + bJ_{12})} \begin{bmatrix} \mu J_{11} & \mu J_{12} \\ -b & a \end{bmatrix}$$

And the new set of parameter is then obtained as,

$$\eta = [R^0]^{-1}\theta \Leftrightarrow \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{aJ_{11}+bJ_{12}} (J_{11}\delta + J_{12}\gamma) \\ \frac{1}{\mu(aJ_{11}+bJ_{12})}(-b\delta + a\gamma) \end{pmatrix}$$

We can see that the standard direction η_1 is completely determined: that is the relative weights on δ and γ are known. As a convention, we normalize all vectors to unity and we also ensure that the subspaces defined respectively by the columns of R_2 and of R_1 are orthogonal.

Asymptotic result:

Recall first the adapted asymptotic convergence result:

$$\begin{bmatrix} \lambda_{1T}(\hat{\eta}_{1T} - \eta_1^0) \\ \lambda_{2T}(\hat{\eta}_{2T} - \eta_2^0) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, (J^{0\prime} S(\theta^0)^{-1} J^0)^{-1})$$

We now provide some details on the calculation of the above asymptotic variance. J^0 is defined as:

$$J^0 = \begin{bmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_1 & 0 \\ 0 & \frac{\rho_2(\theta^0)}{\partial \theta'} R_2 \end{bmatrix}$$

The approximation of J^0 is easily deduced from what has been done above.

By assumption $S(\theta^0)$ is block diagonal and defined as,

$$S(\theta^0) = \begin{bmatrix} Var(\sqrt{T}\phi_{1T}(\theta^0)) & 0 \\ 0 & Var(\sqrt{T}\phi_{2T}(\theta^0)) \end{bmatrix}$$

and a usual sample estimator is used.

Results:

We now provide the results of our Monte-Carlo study. Again, we consider three instruments, the constant, the lagged asset return and the lagged consumption growth, and two sets of parameter: set 1 (or model M1a as in Stock and Wright (2000)) where $\theta^{0\prime} = [.97\ 1.3]$; set 2 (or model M1b) where $\theta^{0\prime} = [1.139\ 13.7]$. Model M1b has previously been found to produce non-normal estimator distributions.

First, as we have seen in the previous section, the matrix of reparametrization is not known (even in our Monte-Carlo setting) and it is actually data dependent. We then investigate the variability of the true new parameter η^0. We found that even with small sample size ($T = 100$), the (estimated) true new parameter is really stable and does not depend much on the realization of the sample. For our two models, we find the following true new parameter:

$$\text{Set 1: } \eta^0 = [-0.4015\ 1.5715];\quad [R^0]^{-1} = \begin{bmatrix} .6281 & -.7782 \\ .7782 & .6281 \end{bmatrix};\quad J_1^0 = [1.0321\ \ -1.2788]$$

$$\text{Set 2: } \eta^0 = [-13.5984\ 2.0176];\quad [R^0]^{-1} = \begin{bmatrix} .0642 & -.9979 \\ .9979 & .0642 \end{bmatrix};\quad J_1^0 = [.8986\ \ -13.9780]$$

To estimate our models, we use the 2-step nearly-weak GMM and we produce the estimation results also for the intermediate 1-step estimator.

Note also that the optimization resolution is not affected by the rates of convergence (λ_{1T} and λ_{2T}).

Our findings are: i) All the estimators are consistent; ii) The variances of the estimators (for both $\hat{\eta}_T$ and $\hat{\theta}_T$) decrease to 0 with the sample size. The direct comparison between the variances of the parameter is not much of interest, this is rather the ratio that carries some information; iii) According to our asymptotic results, in case of nearly-weak identification, the asymptotic variance of the new parameter $\hat{\eta}_{1T}$ should decrease (a lot) faster with the sample size than the one of $\hat{\eta}_{2T}$. Figure 4 investigates this feature by plotting the evolution of the ratio of the Monte-Carlo variance of $\hat{\eta}_{2T}$ and the Monte-Carlo variance of $\hat{\eta}_{1T}$ with the sample size.

For set 1, the ratio of variances is fairly constant: this suggests that the variances of both parameter $\hat{\eta}_{1T}$ and $\hat{\eta}_{2T}$ decrease at the same speed towards 0. This actually supports previous findings that this model presents less severe case of nonstandard behaviors. However, this

does not support our study of the strength of the moments (see figure 3) and the presence of plateaus for two of them. For set 2, the ratio of η slightly decreases with the sample size, while nothing like this can be observed for initial parameters. This provides some support to our asymptotic approach even though the difference between the identification issues in the two sets 1 and 2 is not very compelling from figure 4 or from the estimation of the rates of convergence (tables 4 and 5). When studying the ratios, the slope is not significantly different from 0 for the new parameters and slightly positive for the original parameters. Similarly, for set 2, all rates are also close to each other (0.48), slightly slower than for set 1 and significantly different from the usual rate $T^{1/2}$. The slope of the ratio of new parameters is significantly positive whereas this is not the case for the original parameters.

7 Conclusion

In a GMM context, this essay proposes a general framework to account for potentially weak instruments. In contrast with existing literature, the weakness is directly related to the moment conditions (through the instruments) and not to the parameters. More precisely, we consider two groups of moment conditions: the standard one associated with the standard rate of convergence $\sqrt{T}$ and the nearly-weak one associated with the slower rate λ_T. This framework ensures that GMM estimators of all parameters are consistent, but at rates possibly slower than usual. We also characterize the validity of the standard testing approaches like Wald and GMM-LM tests. Moreover, we identify and estimate some $\sqrt{T}$-consistent directions in the parameter space. Such results are practically relevant since the knowledge of the slower rate of convergence is not required.

For notational and expositional simplicity, we have chosen here to focus on two groups of moment conditions only. The extension to considering several degrees of weakness (think of a practitioner using several instruments of different informational qualities) is quite natural. Antoine and Renault (2007) specifically consider multiple groups of moment conditions associated with specific rates of convergence (which may actually be faster and/or slower than the standard rate $\sqrt{T}$). Note however that they do not explicitly consider any applications to identification issues, but rather applications in kernel, unit-root, extreme values or continuous time environments.

References

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

[2] ———, *Nonparametric Kernel Estimation for Semiparametric Econometric Models*, Econometric Theory **11** (1995), 560–596.

[3] D.W.K. Andrews and J. Stock, *Inference with Weak Instruments*, invited survey paper for the 2005 World congress of the Econometric Society (2005).

[4] J.D. Angrist and A. Krueger, *How compulsory School Attendance affect Schooling and Earnings*, Quarterly Journal of Economics **91** (1991).

[5] B. Antoine and E. Renault, *Efficient Minimum Distance Estimation with Multiple Rates of Convergence*, Working Paper (2007).

[6] M. Caner, *Testing, Estimation and Higher Order Expansions in GMM with Nearly-Weak Instruments*, Working Paper (2005).

[7] P. Dovonon and E. Renault, *GMM Overidentification Test with First-Order Unidentification*, Working Paper (2006).

[8] P. Gagliardini, C. Gouriéroux, and E. Renault, *Efficient Derivative Pricing by Extended Method of Moments*, Working Paper (2005).

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

[10] A.R. Hall, *Generalized Method of Moments*, Advanced Texts in Econometrics, Oxford University Press, 2005.

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

[12] L.P. Hansen, J. Heaton, and A. Yaron, *Finite Sample Properties of some Alternative GMM Estimators*, Journal of Business and Economic Statistics **14** (1996), 262–280.

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

[14] N. Kocherlakota, *On Tests of Representative Consumer Asset Pricing Models*, Journal of Monetary Economics **26** (1990), 285–304.

[15] L. Lee, *Pooling Estimators with Different Rates of Convergence - A minimum χ^2 Approach with an emphasis on a Social Interaction Model*, Working Paper (2004).

[16] ———, *Classical Inference with ML and GMM Estimates with Various Rates of Convergence*, Working Paper (2005).

[17] W.K. Newey and K.D. West, *Hypothesis Testing with Efficient Method of Moments Estimation*, International Economic Review **28** (1987), 777–787.

[18] P.C.B. Phillips, *Partially Identified Econometric Models*, Econometric Theory **5** (1989), 181–240.

[19] P.C.B. Phillips and J.Y. Park, *On the Formulation of Wald Tests of Nonlinear Restrictions*, Econometrica **56** (1988), 1065–1083.

[20] J.D. Sargan, *Identification and Lack of Identification*, Econometrica **51** (1983), 1605–1634.

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

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

[23] G. Tauchen, *Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data*, Journal of Business and Economic Statistics **4** (1986), 397–425.

[24] G. Tauchen and R. Hussey, *Quadrature-based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models*, Econometrica **59** (1991), 371–396.

Appendix

Proofs of the main results

Proof of Theorem 2.1 *(Consistency)*:
The consistency of the minimum distance estimator $\hat{\theta}_T$ is a direct implication of the identification assumption 1 jointly with the following lemma:

Lemma A

$$\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

Proof of lemma A: From (2.2), the objective function is written as follows

$$Q_T(\theta) = \left[\frac{\Psi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)\right]' \Omega_T \left[\frac{\Psi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)\right] \quad \text{where} \quad \Lambda_T = \begin{bmatrix} Id_{k_1} & 0 \\ 0 & \frac{\lambda_T}{\sqrt{T}} Id_{k_2} \end{bmatrix}$$

Since $\hat{\theta}_T$ is the minimizer of $Q(.)$ we have in particular:

$$\begin{aligned} & Q_T(\hat{\theta}_T) \leq Q(\theta^0) \\ \Longrightarrow \quad & \left[\frac{\Psi_T(\hat{\theta}_T)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\hat{\theta}_T)\right]' \Omega_T \left[\frac{\Psi_T(\hat{\theta}_T)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\hat{\theta}_T)\right] \leq \frac{\Psi'_T(\theta^0)}{\sqrt{T}}\Omega_T\frac{\Psi_T(\theta^0)}{\sqrt{T}} \end{aligned}$$

Denoting $d_T = \Psi'_T(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) - \Psi'_T(\theta^0)\Omega_T\Psi_T(\theta^0)$, we get:

$$\left[\Lambda_T\rho(\hat{\theta}_T)\right]' \Omega_T \left[\Lambda_T\rho(\hat{\theta}_T)\right] + 2\left[\Lambda_T\rho(\hat{\theta}_T)\right]' \Omega_T\Psi_T(\hat{\theta}_T) + d_T \leq 0$$

Let μ_T be the smallest eigenvalue of Ω_T. The former inequality implies:

$$\mu_T\|\Lambda_T\rho(\hat{\theta}_T)\|^2 - 2\|\Lambda_T\rho(\hat{\theta}_T)\| \times \|\Omega_T\Psi_T(\hat{\theta}_T)\| + d_T \leq 0$$

In other words, $x_T = \|\Lambda_T\rho(\hat{\theta}_T)\|$ solves the inequality:

$$x_T^2 - \frac{2\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T}x_T + \frac{d_T}{\mu_T} \leq 0$$

and thus with

$$\Delta_T = \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|^2}{\mu_T^2} - \frac{d_T}{\mu_T}$$

we have:

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} - \sqrt{\Delta_T} \leq x_T \leq \frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} + \sqrt{\Delta_T}$$

Since $x_T \geq \lambda_T \|\rho_T(\hat{\theta}_T)\|$ we want to show that $x_T = \mathcal{O}_P(1)$ that is

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \quad \text{and} \quad \Delta_T = \mathcal{O}_P(1)$$

which amounts to show that:

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \quad \text{and} \quad \frac{d_T}{\mu_T} = \mathcal{O}_P(1)$$

Note that since $det(\Omega_T) \xrightarrow{P} det(\Omega) > 0$ no subsequence of Ψ_T can converge in probability towards zero and thus we can assume (for T sufficiently large) that μ_T remains lower bounded away from zero with asymptotic probability one. Therefore, we just have to show that:

$$\|\Omega_T \Psi_T(\hat{\theta}_T)\| = \mathcal{O}_P(1) \quad \text{and} \quad d_T = \mathcal{O}_P(1)$$

Since $trace(\Omega_T) \xrightarrow{P} trace(\Omega)$ and the sequence $trace(\Omega_T)$ is upper bounded in probability, so are all the eigenvalues of Ω_T. Therefore, the required boundedness in probability just results from our assumption 1(ii) ensuring that:

$$\sup_{\theta \in \Theta} \|\Psi_T(\theta)\| = \mathcal{O}_P(1)$$

The proof of lemma A is completed. Let us then deduce the weak consistency of $\hat{\theta}_T$ by a contradiction argument. If $\hat{\theta}_T$ is not consistent, there exists some positive ϵ such that:

$$P\left[\|\hat{\theta}_T - \theta^0\| > \epsilon\right]$$

does not converge to zero. Then we can define a subsequence $(\hat{\theta}_{T_n})_{n \in \mathbb{N}}$ such that, for some positive η:

$$P\left[\|\hat{\theta}_{T_n} - \theta^0\| > \epsilon\right] \geq \eta \quad \text{for } n \in \mathbb{N}$$

Let us denote

$$\alpha = \inf_{\|\theta - \theta^0\| > \epsilon} \|\rho(\theta)\| > 0 \quad \text{by assumption 1(i)}$$

Then for all $n \in \mathbb{N}$:

$$P\left[\|\rho(\hat{\theta}_{T_n})\| \geq \alpha\right] > 0$$

When considering the identification assumption 1(iii), this last inequality contradicts lemma *A*. This completes the proof of consistency. ■

Proof of Theorem 2.2 *(Score test)*:
The entire proof is written under the maintained null hypothesis that $\theta_0 = \theta^0$. The score statistic can be written as follows:

$$\begin{aligned} LM_T(\theta_0) &= T\overline{\phi}'_T(\theta_0)S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\overline{\phi}_T(\theta_0) \\ &= \left(S_T^{-1/2}\Psi_T(\theta_0)\right)' S_T^{-1/2}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1} \\ &\quad \times\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta'}S_T'^{-1/2}\left(S_T^{-1/2}\Psi_T(\theta_0)\right) \end{aligned}$$

From assumption 1(ii) $S_T^{-1/2}\Psi_T(\theta_0)$ is asymptotically distributed as a gaussian process with mean 0 and identity covariance matrix. To be able to conclude, we only need to find an invertible matrix D_T and a full column rank matrix B such that

$$\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}D_T \xrightarrow{P} B$$

This would ensure that

$$S_T^{-1/2}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta'}S_T'^{-1/2}$$

is a full rank p idempotent matrix and this leads to the desired result. Using assumption 2(iii), we call s_1 the rank of $[\partial\rho_1(\theta_0)/\partial\theta']$ and $(p-s_1)$ the one of $[\partial\rho_2(\theta_0)/\partial\theta']$. Define

$$D_T = \begin{bmatrix} D_1 & \frac{\sqrt{T}}{\lambda_T}D_2 \end{bmatrix}$$

where D_1 and D_2 are respectively (p, s_1) and $(p, p-s_1)$ full column rank matrices such that $D_2'D_1 = 0$ and the range of D_1 is the range of $[\partial\rho_1(\theta_0)/\partial\theta']$. This ensures that D_T is invertible for every fixed sample size T. We now have:

$$\begin{aligned} \frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}D_T &= \begin{bmatrix} \frac{\partial\overline{\phi}_{1T}(\theta_0)}{\partial\theta'}D_1 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{1T}(\theta_0)}{\partial\theta'}D_2 \\ \frac{\partial\overline{\phi}_{2T}(\theta_0)}{\partial\theta'}D_1 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_0)}{\partial\theta'}D_2 \end{bmatrix} \\ &\xrightarrow{P} \begin{bmatrix} \frac{\partial\rho_1(\theta_0)}{\partial\theta'}D_1 & 0 \\ 0 & \frac{\partial\rho_2(\theta_0)}{\partial\theta'}D_2 \end{bmatrix} \equiv B \quad \text{which is of full column rank } p \end{aligned}$$

where the zero south-west and north-east blocks of B are deduced respectively from assumptions 2(iv) and (v). ■

Proof of Theorem 2.3 *(Rate of convergence)*:
From lemma A $\|\rho(\hat{\theta}_T)\| = \|\rho(\hat{\theta}_T) - \rho(\theta^0)\| = \mathcal{O}_P(1/\lambda_T)$ and by application of the mean-value theorem, for some $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0 component by component, we get:

$$\left\| \frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0 \right) \right\| = \mathcal{O}_P \left(\frac{1}{\lambda_T} \right)$$

Note that, by a common abuse of notation, we omit to stress that $\tilde{\theta}_T$ actually depends on the component of $\rho(.)$. The key point is that since $\rho(.)$ is continuously differentiable and $\tilde{\theta}_T$, as $\hat{\theta}_T$, converges in probability towards θ^0, we have:

$$\frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \xrightarrow{P} \frac{\partial \rho(\theta^0)}{\partial \theta'}$$

and thus:

$$\frac{\partial \rho(\theta^0)}{\partial \theta'} \times \left(\hat{\theta}_T - \theta^0 \right) = z_T$$

with $\|z_T\| = \mathcal{O}_P(1/\lambda_T)$. Since $\partial \rho(\theta^0)/\partial \theta'$ is full column rank, we deduce that:

$$\left(\hat{\theta}_T - \theta^0 \right) = \left[\frac{\partial \rho'(\theta^0)}{\partial \theta} \frac{\partial \rho(\theta^0)}{\partial \theta'} \right]^{-1} \frac{\partial \rho'(\theta^0)}{\partial \theta} z_T$$

also fulfills:

$$\left\| \hat{\theta}_T - \theta^0 \right\| = \mathcal{O}_P \left(\frac{1}{\lambda_T} \right)$$

■

Proof of Theorem 3.1 *(Asymptotic Normality)*:
First we need a preliminary result which naturally extend the convergence towards J^0 in (3.2) when the true value θ^0 is replaced by some preliminary consistent estimator θ_T^*.

Lemma B *Under assumptions 1 to 3, if θ_T^* is such that $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_T)$, then*

$$\sqrt{T} \frac{\partial \overline{\phi}_T(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \xrightarrow{P} J^0 \quad \textit{when } T \to \infty$$

Proof of Lemma B:

First note that

$$\sqrt{T}\frac{\partial\overline{\phi}_T(\theta_T^*)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1} = \begin{bmatrix} \frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_1^0 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_2^0 \\ \frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_1^0 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_2^0 \end{bmatrix}$$

To get the results, we have to show the following:

$$\begin{aligned} &i) \quad \frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'} \xrightarrow{P} \frac{\partial\rho_1(\theta^0)}{\partial\theta'} \\ &ii) \quad \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} \xrightarrow{P} \frac{\partial\rho_2(\theta^0)}{\partial\theta'} \\ &iii) \quad \frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_1^0 \xrightarrow{P} 0 \\ &iv) \quad \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_2^0 \xrightarrow{P} 0 \end{aligned}$$

i) From assumption 2(iv), we have: $\frac{\partial\overline{\phi}_{1T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_1(\theta^0)}{\partial\theta'} = o_P(1)$. The mean-value theorem applies to the k^{th} component of $[\partial\overline{\phi}_{1T}/\partial\theta']$ for $1 \leq k \leq k_1$. For some $\tilde{\theta}$ between θ^0 and θ_T^*, we have:

$$\frac{\partial\overline{\phi}_{1T,k}(\theta_T^*)}{\partial\theta'} - \frac{\partial\overline{\phi}_{1T,k}(\theta^0)}{\partial\theta'} = (\theta_T^* - \theta^0)'\frac{\partial^2\overline{\phi}_{1T,k}(\tilde{\theta}_T)}{\partial\theta\partial\theta'} = o_P(1)$$

where the last equality follows from assumption 3(ii) and the assumption on θ_T^*.

ii) From assumption 2(iv), we have:

$$\sqrt{T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'} - \lambda_T\frac{\partial\rho_2(\theta_0)}{\partial\theta'} = \mathcal{O}_P(1) \quad \Rightarrow \quad \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_2(\theta^0)}{\partial\theta'} = o_P(1)$$

because $\lambda_T \xrightarrow{T} \infty$. The mean-value theorem applies to the kth component of $\partial\overline{\phi}_{2T}/\partial\theta'$ for $1 \leq k \leq k_2$. For some $\tilde{\theta}_T$ between θ^0 and θ_T^*, we have:

$$\frac{\sqrt{T}}{\lambda_T}\left(\frac{\partial\overline{\phi}_{2T,k}(\theta_T^*)}{\partial\theta'} - \frac{\partial\overline{\phi}_{2T,k}(\theta^0)}{\partial\theta'}\right) = (\theta_T^* - \theta^0)\frac{\sqrt{T}}{\lambda_T}\frac{\partial^2\overline{\phi}_{2T,k}(\tilde{\theta}_T)}{\partial\theta\partial\theta'} = o_P(1)$$

where the last equality follows from assumption 3(ii) and the assumption on θ_T^*.

iii)

$$\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} = \frac{\lambda_T}{\sqrt{T}} \times \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} = o_P(1)$$

because of (ii) and $\lambda_T = o(\sqrt{T})$.

iv) Recall the mean-value theorem from i). For $1 \leq k \leq k_1$ and $\tilde{\theta}_T$ between θ^0 and θ_T^*, we have:

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta_T^*)}{\partial \theta'} = \frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta^0)}{\partial \theta'} + \lambda_T(\theta_T^* - \theta^0)'\frac{1}{\lambda_T}\frac{\sqrt{T}}{\lambda_T}\frac{\partial^2 \overline{\phi}_{1T,k}(\tilde{\theta}_T)}{\partial \theta \partial \theta'}$$

The second member of the RHS is $o_P(1)$ because of assumptions 1(iii), 3(ii) and 3(iii) and the assumption on θ_T^*. Now we just need to show that the first member of the RHS is $o_P(1)$. Recall from assumption 2(v) that

$$\sqrt{T}\left[\frac{\partial \overline{\phi}_{1T}(\theta^0)}{\partial \theta'} - \frac{\partial \rho_1(\theta^0)}{\partial \theta'}\right] = \mathcal{O}_P(1) \Rightarrow \frac{\sqrt{T}}{\lambda_T}\left[\frac{\partial \overline{\phi}_{1T}(\theta^0)}{\partial \theta'} - \frac{\partial \rho_1(\theta^0)}{\partial \theta'}\right] R_2^0 = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

By definition R_2^0 is such that $\frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_2^0 = 0$. Hence we get

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta^0)}{\partial \theta'} R_2^0 = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right) = o_P(1)$$

This concludes the proof of lemma B. We now return to the proof of theorem 3.1. From the optimization problem (2.2), the first order conditions for $\hat{\theta}_T$ are written as:

$$\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \overline{\phi}_T(\hat{\theta}_T) = 0$$

A mean-value expansion yields to:

$$\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \overline{\phi}_T(\theta^0) + \frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \frac{\partial \overline{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \times \left(\hat{\theta}_T - \theta^0\right) = 0$$

where $\tilde{\theta}_T$ is between $\hat{\theta}_T$ and θ^0. Premultiplying the above equation by the non-singular matrix $T\tilde{\Lambda}_T^{-1}R^{0\prime}$ yields to an equivalent set of equations:

$$\hat{J}_T'\Omega\left[\sqrt{T}\overline{\phi}_T(\theta^0)\right] + \hat{J}_T'\Omega\tilde{J}_T \times \tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) = 0$$

after defining:

$$\hat{J}_T = \sqrt{T}\frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'}R^0\tilde{\Lambda}_T^{-1} \quad \text{and} \quad \tilde{J}_T = \sqrt{T}\frac{\partial \overline{\phi}_T(\tilde{\theta}_T)}{\partial \theta}R^0\tilde{\Lambda}_T^{-1}$$

From theorem 2.3 and lemma B, we can deduce that:

$$Plim\tilde{J}_T = J^0 \quad \text{and} \quad Plim\hat{J}_T = J^0$$

Hence,

$$\hat{J}_T'\Omega\tilde{J}_T \xrightarrow{P} J^{0\prime}\Omega J^0 \quad \text{nonsingular by assumption}$$

Recall now that by assumption 1ii), $\Psi_T(\theta^0) = \sqrt{T}\left[\ \overline{\phi}_T(\theta^0)\right]$ converges to a normal distribution with mean 0 and variance $S(\theta^0)$. We then get the announced result.■

Proof of Theorem 3.2 *(Overidentifying test)*:
A Taylor expansion of order 1 of the moment conditions gives:

$$\begin{aligned}\sqrt{T}\overline{\phi}_T(\hat{\theta}_T) &= \sqrt{T}\overline{\phi}_T(\theta^0) + \sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}(\hat{\theta}_T - \theta^0) + o_P(1)\\ &= \sqrt{T}\overline{\phi}_T(\theta^0) + \hat{J}_T\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) + o_P(1)\end{aligned}$$

with $\hat{J}_T = \sqrt{T}\partial\overline{\phi}_T(\hat{\theta}_T)/\partial\theta' R^0\tilde{\Lambda}_T^{-1}$.

A Taylor expansion of the FOC gives:

$$\begin{aligned}\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) &= -\left[\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)\right]^{-1}\\ &\quad\times\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta})}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

with S_T a consistent estimator of the asymptotic covariance matrix of the process $\Psi(\theta)$.

Combining the 2 above results leads to:

$$\sqrt{T}\overline{\phi}_T(\hat{\theta}_T) = \sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)$$

Use the previous result to rewrite the criterion function:

$$\begin{aligned}TQ_T(\hat{\theta}_T) &= \left[\sqrt{T}\overline{\phi}_T(\hat{\theta}_T)\right]' S_T^{-1}\sqrt{T}\overline{\phi}_T(\hat{\theta}_T)\\ &= \left[\sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0)\right]' S_T^{-1}\\ &\quad\times\left[\sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0)\right] + o_P(1)\\ &= \left[\sqrt{T}\overline{\phi}_T(\theta^0)\right]' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0)\\ &\quad-\sqrt{T}\overline{\phi}_T(\theta^0)S_T^{-1}\hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)\\ &= \sqrt{T}\overline{\phi}_T(\theta^0)'S_T'^{-1/2}\left[I - M\right]^{-1}S_T^{1/2}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

where $S_T^{1/2}$ is such that $S_T = S_T'^{-1/2} S_T^{-1/2}$ and $M = S_T^{-1/2} \hat{J}_T \left[\hat{J}_T' S_T^{-1} \hat{J}_T \right]^{-1} \hat{J}_T' S_T'^{-1/2}$ which is a projection matrix, hence idempotent and of rank $(K-p)$. The expected result follows. ■

Proof of Theorem 3.3 *(Orthogonalization)*:
Recall the inverse formulae:

$$\begin{aligned} S^{-1} &= \begin{bmatrix} [S_1^0]^{-1}(I + S_{12}^0 P^{-1} S_{21}^0 [S_1^0]^{-1}) & -[S_1^0]^{-1} S_{12}^0 P^{-1} \\ -P^{-1} S_{21} [S_1^0]^{-1} & P^{-1} \end{bmatrix} \\ &= \begin{bmatrix} Q^{-1} & -Q^{-1} S_{12}^0 [S_2^0]^{-1} \\ -[S_2^0]^{-1} S_{21}^0 Q^{-1} & S_2^{-1}(I + S_{21}^0 Q^{-1} S_{12}^0 [S_2^0]^{-1}) \end{bmatrix} \end{aligned}$$

with $Q = S_1^0 - S_{12}^0 [S_2^0]^{-1} S_{21}^0$ and $P = S_2^0 - S_{21}^0 [S_1^0]^{-1} S_{12}^0$.

Recall the block-diagonality of the matrix J^0 (see page 15):

$$J^0 = \begin{bmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_1^0 & 0 \\ 0 & \frac{\partial \rho_2(\theta^0)}{\partial \theta'} R_2^0 \end{bmatrix}$$

Recall

$$AVar(\hat{\eta}_T) = \left[J^{0\prime} S^{-1} J^0 \right]^{-1} \quad \text{and} \quad AVar(\tilde{\eta}_T) = \left[J^{0\prime} [\Sigma^0]^{-1} J^0 \right]^{-1}$$

We need to compare the north-west and the south-east blocks of the above matrices.

Straightforward calculations lead to:

$$\left[J^{0\prime} [\Sigma^0]^{-1} J^0 \right]^{-1} = \begin{bmatrix} \left[\tilde{R}_1' Q^{-1} \tilde{R}_1 \right]^{-1} & 0 \\ 0 & \left[\tilde{R}_2' S_2^{-1} \tilde{R}_2 \right]^{-1} \end{bmatrix}$$

with $\tilde{R}_i \equiv \frac{\partial \rho_i}{\partial \theta'} R_i$ for $i = 1, 2$ and $Q \equiv S_1 - S_{12} S_2^{-1} S_{21}$.

On the other end, we have:

$$J^{0\prime} S^{-1} J^0 = \begin{bmatrix} \tilde{R}_1' Q^{-1} \tilde{R}_1 & -\tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ -\tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 & \tilde{R}_2' \left[S_2^{-1} + S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \right] \tilde{R}_2 \end{bmatrix} \equiv \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with $B' = C$.

i) We have $AVar(\hat{\eta}_{1T}) = A^{-1} + A^{-1} B (D - CA^{-1}B)^{-1} C A^{-1}$ that needs to be compared to $AVar(\tilde{\eta}_{1T}) = \left(\tilde{R}_1' Q^{-1} \tilde{R}_1 \right)^{-1}$.

Note that $A^{-1} = \left(\tilde{R}_1' Q^{-1} \tilde{R}_1\right)^{-1}$. Hence it is enough to study $A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$. Recall that $AVar(\hat{\eta}_{2T}) = (D - CA^{-1}B)^{-1}$, hence it is a positive definite matrix (see also ii)). Also we have $B = C'$ and A symmetric. Then, we can deduce that $A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1}$ is positive semi-definite.

Finally, we can conclude: $AVar(\hat{\eta}_{1T}) \ \geq\geq \ AVar(\tilde{\eta}_{1T})$

ii) We have $AVar(\hat{\eta}_{2T}) = \left(D - CA^{-1}B\right)^{-1}$ that needs to be compared to $AVar(\tilde{\eta}_{2T}) = \left(\tilde{R}_2' S_2^{-1} \tilde{R}_2\right)^{-1}$.

It is enough to compare $D - CA^{-1}B$ with $\tilde{R}_2' S_2^{-1} \tilde{R}_2$.

$$\begin{aligned} D - CA^{-1}B &= \tilde{R}_2' S_2^{-1} \tilde{R}_2 + \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ &\quad - \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \end{aligned}$$

The last 2 terms of the RHS can be rewritten as follows:

$$\begin{aligned} & \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 - \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ = \ & \tilde{R}_2' S_2^{-1} S_{21} \left\{ Q^{-1} - Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} \right\} S_{12} S_2^{-1} \tilde{R}_2 \end{aligned}$$

It is enough to study the middle matrix that appears between the brackets:

$$\begin{aligned} & Q^{-1} - Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} \\ = \ & Q^{-1/2\prime} \left\{ I - Q^{-1/2} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1/2\prime} Q^{-1/2} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1/2\prime} \right\} Q^{-1/2} \\ = \ & Q^{-1/2\prime} \left\{ I - X(X'X)^{-1} X' \right\} Q^{-1/2} \\ = \ & Q^{-1/2\prime} M_X Q^{-1/2} \end{aligned}$$

with $Q^{-1} \equiv Q^{-1/2\prime} Q^{-1/2}$, $X \equiv Q^{-1/2} \tilde{R}_1$ and $M_X \equiv I - X(X'X)^{-1}X'$.

Finally, we have:

$$\begin{aligned} D - CA^{-1}B &= \tilde{R}_2' S_2^{-1} \tilde{R}_2 + \left(Q^{-1/2} S_{12} S_2^{-1} \tilde{R}_2\right)' M_X \left(Q^{-1/2} S_{12} S_2^{-1} \tilde{R}_2\right) \\ &\geq\geq \tilde{R}_2' S_2^{-1} \tilde{R}_2 \end{aligned}$$

because by definition, M_X is a projection matrix. Hence it is positive semi-definite as well as $H' M_X H$ for any matrix H.

We can then conclude: $AVar(\hat{\eta}_{2T}) \leq\leq AVar(\tilde{\eta}_{2T})$. ■

Proof of Proposition 3.4:

We consider the following set of moment conditions,

$$\begin{pmatrix} \overline{\phi}^H_{1T}(\theta^0) \\ \overline{\phi}^H_{2T}(\theta^0) \end{pmatrix} = \begin{pmatrix} \overline{\phi}_{1T}(\theta^0) + H\overline{\phi}_{2T}(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) \end{pmatrix}$$

such that

$$\sqrt{T}\begin{bmatrix} \phi^H_{1T}(\theta^0) - \rho_1(\theta^0) \\ \phi^H_{2T}(\theta^0) - \frac{\lambda_T}{\sqrt{T}}\rho_2(\theta^0) \end{bmatrix} \Rightarrow \tilde{\Psi}(\theta^0)$$

where $\tilde{\Psi}(\theta)$ is a Gaussian random variable with mean zero and variance

$$S^H = \begin{bmatrix} S^0_1 + HS^0_2H' + S^0_{12}H' + HS^0_{21} & S^0_{12} + HS^0_2 \\ S^0_{21} + S^0_2H' & S^0_2 \end{bmatrix}$$

The above set is orthogonalized as follows:

$$\begin{pmatrix} \tilde{\phi}^H_{1T}(\theta^0) \\ \tilde{\phi}^H_{2T}(\theta^0) \end{pmatrix} = \begin{pmatrix} \overline{\phi}^H_{1T}(\theta^0) - Cov\left(\sqrt{T}\overline{\phi}^H_{1T}(\theta^0), \sqrt{T}\overline{\phi}^H_{2T}(\theta^0)\right)\left[Var\left(\sqrt{T}\overline{\phi}^H_T(\theta^0)\right)\right]^{-1}\overline{\phi}^H_{2T}(\theta^0) \\ \overline{\phi}^H_{2T}(\theta^0) \end{pmatrix}$$

with

$$\begin{aligned} \tilde{\phi}^H_{1T}(\theta^0) &= \overline{\phi}_{1T}(\theta^0) + H\overline{\phi}_{2T}(\theta^0) - (S_{12} + HS_2)S_2^{-1}\overline{\phi}_{2T}(\theta^0) \\ &= \overline{\phi}_{1T}(\theta^0) + H\overline{\phi}_{2T}(\theta^0) - S_{12}S_2^{-1}\overline{\phi}_{2T}(\theta^0) - H\overline{\phi}_{2T}(\theta^0) \\ &= \tilde{\phi}_{1T}(\theta^0) \end{aligned}$$

■

Proof of Corollary 3.5:

The proof directly follows from the results of Theorem 3.3 and Property 3.4. ■

Proof of Lemma 3.6

In theorem 3.1, we have established that the following vector is asymptotically normally distributed:

$$\begin{bmatrix} \sqrt{T}A\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T B\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

We now show that the above convergence result is not altered when matrices A and B are replaced by some λ_T-consistent estimators, respectively $\hat{A}$ and $\hat{B}$.

(i) Convergence of the nearly-weak directions:

$$\lambda_T \hat{B}\left(\hat{\theta}_T - \theta^0\right) = \underbrace{\lambda_T B\left(\hat{\theta}_T - \theta^0\right)}_{(1)} + \underbrace{\lambda_T\left(\hat{B} - B\right)\left(\hat{\theta}_T - \theta^0\right)}_{(2)}$$

$(1) = \mathcal{O}_P(1)$. $\hat{B}$ is a λ_T-consistent estimator of B, so clearly (1) dominates (2): this is denoted as $(2) \prec (1)$.

(ii) Convergence of the standard directions:

$$\sqrt{T}\hat{A}\left(\hat{\theta} - \theta^0\right) = \underbrace{\sqrt{T}A\left(\hat{\theta} - \theta^0\right)}_{(1)} + \underbrace{\frac{\sqrt{T}}{\lambda_T}\left(\hat{A} - A\right)\lambda_T\left(\hat{\theta} - \theta^0\right)}_{(2)}$$

We have $(1) = \mathcal{O}_P(1)$ and $\lambda_T(\hat{\theta} - \theta^0) = \mathcal{O}_P(1)$. Hence,

$$(2) \prec (1) \iff \frac{\sqrt{T}}{\lambda_T}(\hat{A} - A) = o_P(1) \iff \hat{A} - A = o_P\left(\frac{\lambda_T}{\sqrt{T}}\right)$$

By assumption $\left\|\hat{A} - A\right\| = \mathcal{O}_P(\frac{1}{\lambda_T})$, so we get:

$$(2) \prec (1) \iff \frac{1}{\lambda_T} = o\left(\frac{\lambda_T}{\sqrt{T}}\right) \iff \sqrt{T} = o\left(\lambda_T^2\right)$$

which corresponds to assumption 3(i). ■

Proof of Theorem 4.1 *(Wald test)*:
The proof is divided into two steps:
- step 1: we define an algebraically equivalent formulation of $H_0 : g(\theta) = 0$ as $H_0 : h(\theta) = 0$ such that its first components are strongly identified while the remaining ones are nearly-weakly identified without any linear combinations of the latter being strongly identified.
- step 2: we show that the Wald test statistic on $H_0 : h(\theta) = 0$ asymptotically converges to the proper $\chi^2(q)$ distribution and that it is numerically equal to the Wald test statistic on $H_0 : g(\theta) = 0$.

- Step 1: The space of strongly identified directions to be tested is:

$$I^0(g) = \left[Im\frac{\partial g'(\theta^0)}{\partial\theta}\right] \cap \left[Im\frac{\partial \rho_1'(\theta^0)}{\partial\theta}\right]$$

Denote $n^0(g)$ the dimension of $I^0(g)$. Then, among the q restrictions to be tested, $n^0(g)$ are strongly identified and the $(q - n^0(g))$ remaining ones are nearly-weakly identified.
Define q vectors of $\mathbb{R}^q$ denoted as ϵ_j $(j = 1, \cdots, q)$ such that $\left[\partial g'(\theta^0)/\partial\theta \times \epsilon_j\right]_{j=1}^{q_1}$ is a basis of $I^0(g)$ and $\left[\partial g'(\theta^0)/\partial\theta \times \epsilon_j\right]_{j=q_1+1}^{q}$ is a basis of

$$\left[I^0(g)\right]^{\perp} \cap \left[Im \frac{\partial g'(\theta^0)}{\partial\theta}\right]$$

We can then define a new formulation of the null hypothesis $H_0 : g(\theta) = 0$ as: $H_0 : h(\theta) = 0$ where $h(\theta) = Hg(\theta)$ with H invertible matrix such that $H' = [\epsilon_1 \ \cdots \ \epsilon_q]$. The two formulations are algebraically equivalent since $h(\theta) = 0 \iff g(\theta) = 0$. Moreover,

$$\underset{T\to\infty}{Plim} \left[D_T \frac{\partial h(\theta^0)}{\partial\theta'} R^0 \left[\tilde{\Lambda}_T\right]^{-1}\right] = B^0$$

with D_T a (q, q) invertible diagonal matrix with its first $n^0(g)$ coefficients equal to $\sqrt{T}$ and the $(p - n^0(g))$ remaining ones equal to λ_T and B^0 a (q, p) matrix with full column rank.

- Step 2: first we show that the 2 induced Wald test statistics are numerically equal.

$$\begin{aligned}
\zeta_T^W(g) &= Tg'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial\theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial\theta} \right\}^{-1} g(\hat{\theta}_T) \\
&= TH'g'(\hat{\theta}_T) \left\{ H \frac{\partial g(\hat{\theta}_T)}{\partial\theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial\theta} H' \right\}^{-1} Hg(\hat{\theta}_T) \\
&= \zeta_T^W(h)
\end{aligned}$$

Then we show $\zeta_T^W(h) \xrightarrow{d} \chi^2(q)$. First we need a preliminary result which naturally extends the above convergence towards B^0 when θ^0 is replaced by a λ_T-consistent estimator θ_T^*:

$$Plim_T \left[D_T \frac{\partial h(\theta_T^*)}{\partial\eta'} [\tilde{\Lambda}_T]^{-1}\right] = B^0$$

The proof is very similar to lemma B in the proof of theorem 3.1 and is not reproduced here. Note that the fact that $g(.)$ is twice continuously differentiable is needed for this proof.

The Wald test statistic on $h(.)$ can be written as follows:

$$\begin{aligned}\zeta_T^W(h) &= T\left[D_T h(\hat{\theta}_T)\right]' \left\{ D_T \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T \right\}^{-1} \left[D_T h(\hat{\theta}_T)\right] \\ &= \left[D_T h(\hat{\theta}_T)\right]' \left\{ D_T \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \left[\hat{J}_T' S_T^{-1} \hat{J}_T\right]^{-1} \tilde{\Lambda}_T^{-1} R^{0\prime} \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T \right\}^{-1} \left[D_T h(\hat{\theta}_T)\right]\end{aligned}$$

where $\hat{J}_T \equiv \sqrt{T} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}$ with $\hat{J}_T \xrightarrow{P} J^0$ and $\hat{J}_T' S_T^{-1} \hat{J}_T \xrightarrow{P} J^{0\prime} \left[S(\theta^0)\right]^{-1} J^0 \equiv \Sigma$.
Now from the mean-value theorem under H_0 we deduce:

$$D_T h(\hat{\theta}_T) = D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0\right) = \left[D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}\right] \tilde{\Lambda}_T \left[R^0\right]^{-1} \left(\hat{\theta}_T - \theta^0\right)$$

$$\text{with} \quad \left[D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}\right] \xrightarrow{P} B^0 \quad \text{and} \quad \tilde{\Lambda}_T \left[R^0\right]^{-1} \left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

Finally we get

$$\xi_T^W(h) = \left[\tilde{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0)\right]' B_0' (B_0 \Sigma B_0')^{-1} B_0 \left[\tilde{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0)\right] + o_P(1)$$

Following the proof of theorem 3.2 we get the expected result. ■

T	$\mathrm{Var}(\hat{\eta}_{1T})$	$\mathrm{Var}(\hat{\eta}_{2T})$	$\mathrm{Var}(\hat{\theta}_{1T})$	$\mathrm{Var}(\hat{\theta}_{2T})$
50	0.0534	0.0264	0.0233	0.0081
100	0.0236	0.01799	0.0154	0.0056
200	0.0133	0.0103	0.0082	0.0037
300	0.0093	0.0080	0.0059	0.0030
400	0.0073	0.0058	0.0041	0.0024
500	0.0060	0.0049	0.0031	0.0021
600	0.0051	0.0044	0.0029	0.0019
700	0.0042	0.0042	0.0027	0.0018
800	0.0035	0.0039	0.0026	0.0017
900	0.0031	0.0035	0.0023	0.0015
1000	0.0028	0.0033	0.0022	0.0014
1500	0.0019	0.0023	0.0015	0.0010
2000	0.0014	0.0020	0.0012	0.0009
3000	0.0009	0.0015	0.0009	0.0007
5000	0.0005	0.0011	0.0006	0.0005
6000	0.0005	0.0010	0.0006	0.0005
7000	0.0004	0.0009	0.0005	0.0004
8000	0.0003	0.0008	0.0005	0.0004
9000	0.0003	0.0008	0.0004	0.0004
10000	0.0003	0.0007	0.0004	0.0003
11000	0.0002	0.0007	0.0004	0.0003
12000	0.0002	0.0006	0.0004	0.0003
13000	0.0002	0.0006	0.0003	0.0003

Table 1: Single-equation linear IV model: Estimation results for the variance of the Monte Carlo distributions of the new parameters $\hat{\eta}_T$ as well as the original one $\hat{\theta}_T$ for various sample sizes.

Entire specter of sample sizes				
	$\hat{\beta}$	95% Confidence Interval		Estimated rate
$\hat{\eta}_{1T}$	-0.9976	-1.0111	-0.9842	0.4988
$\hat{\eta}_{2T}$	-0.6806	-0.6950	-0.6663	0.3403
$\hat{\theta}_{1T}$	-0.7577	-0.7808	-0.7345	0.3788
$\hat{\theta}_{2T}$	-0.6130	-0.6221	-0.6039	0.3065
Large sample sizes (>5000)				
	$\hat{\beta}$	95% Confidence Interval		Estimated rate
$\hat{\eta}_{1T}$	-0.9903	-1.0042	-0.9764	0.4951
$\hat{\eta}_{2T}$	-0.6267	-0.6692	-0.5841	0.3133
$\hat{\theta}_{1T}$	-0.6667	-0.7088	-0.6246	0.3333
$\hat{\theta}_{2T}$	-0.6046	-0.6411	-0.5681	0.3023

Table 2: Single-equation linear IV model: Estimation of the β coefficients in the linear regression (5.4) and the rates of convergence of the variance series.

Entire specter of sample sizes			
	Estimated slope	95% Confidence Interval	
$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.3170	0.2905	0.3435
$Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$	0.1447	0.1209	0.1685
Large sample sizes (>5000)			
	Estimated slope	95% Confidence Interval	
$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.3636	0.3114	0.4158
$Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$	0.0621	0.0478	0.0764

Table 3: Single-equation linear IV model: Estimation of the β coefficients for the ratio series.

Large sample sizes (>10000)								
	$\hat{\beta}$	95% CI		Rate		Slope	95% CI	
$\hat{\eta}_{1T}$	-0.9862	-1.0069	-0.9654	0.4931	$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	-0.0011	-0.0068	0.0046
$\hat{\eta}_{2T}$	-0.9872	-1.0071	-0.9674	0.4936				
$\hat{\theta}_{1T}$	-0.9879	-1.0072	-0.9686	0.4940	$Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$	0.0006	0.0002	0.0010
$\hat{\theta}_{2T}$	-0.9885	-1.0077	-0.9693	0.4942				

Table 4: CCAPM for set 1: i) Estimation of the β coefficients in the linear regression (5.4) and the rates of convergence of the variance series; ii) Estimation of the β coefficient for the ratio series

Large sample sizes (>10000)								
	$\hat{\beta}$	95% CI		Rate		Slope	95% CI	
$\hat{\eta}_{1T}$	-0.9674	-0.9872	-0.9477	0.4837	$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.0018	0.0010	0.0027
$\hat{\eta}_{2T}$	-0.9656	-0.9854	-0.9458	0.4828				
$\hat{\theta}_{1T}$	-0.9633	-0.9831	-0.9436	0.4816	$Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$	-0.0002	-0.0052	0.0048
$\hat{\theta}_{2T}$	-0.9631	-0.9828	-0.9435	0.4815				

Table 5: CCAPM for set 2: i) Estimation of the β coefficient in the linear regression (5.4) and the rates of convergence of the variance series; ii) Estimation of the β coefficient for the ratio series

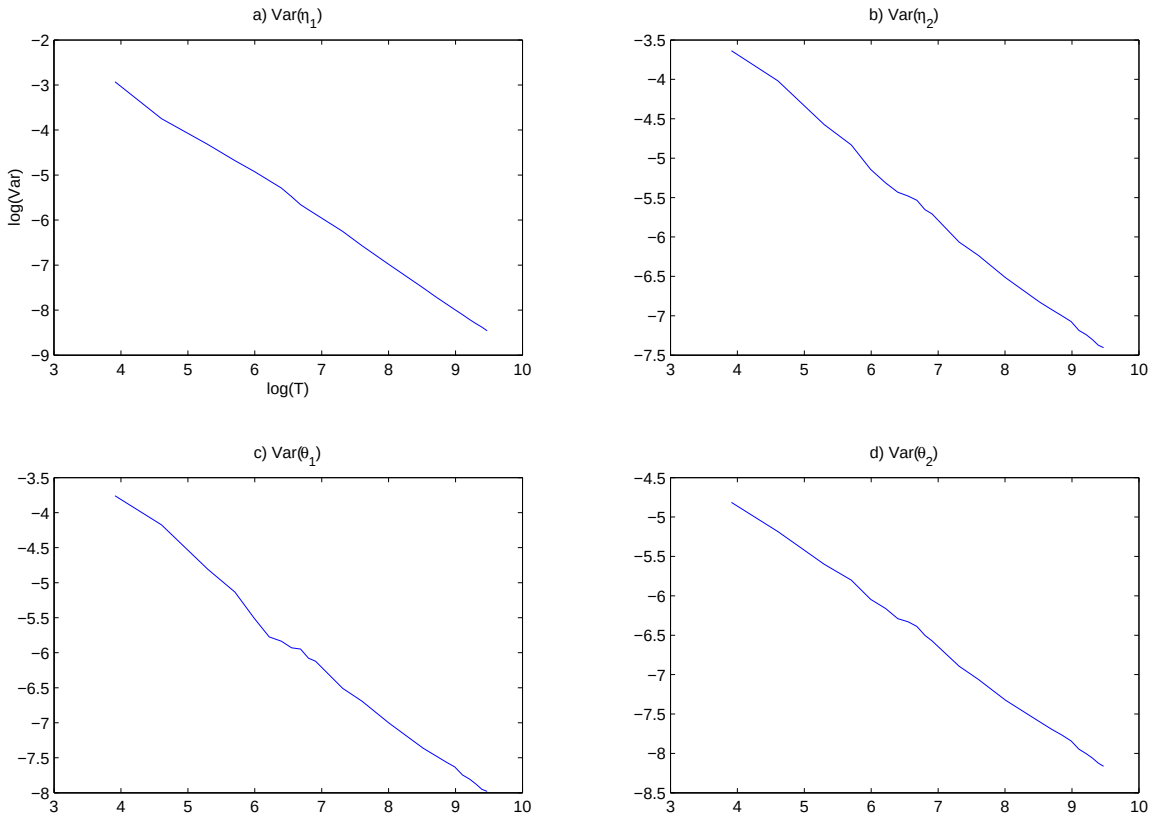

Figure 1: Single-equation linear IV model: Logarithm of the variance as a function of the log-sample size. Top figures for the new parameters with a) $\hat{\eta}_{1T}$; b) $\hat{\eta}_{2T}$; Bottom figures for the original parameters with c) $\hat{\theta}_{1T}$; d) $\hat{\theta}_{2T}$.

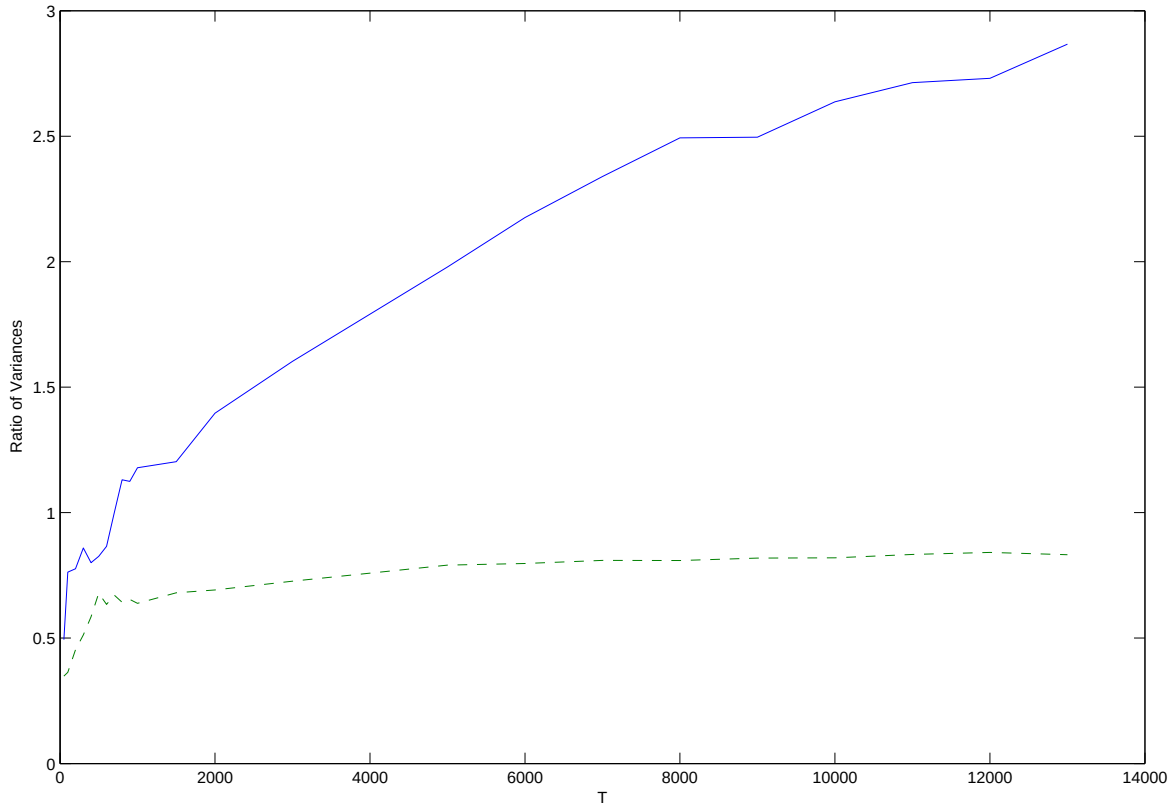

Figure 2: Single-equation linear IV model: Ratio of the variance of the parameters as a function of the sample size. Solid line for $Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$; Dashed line for $Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$.

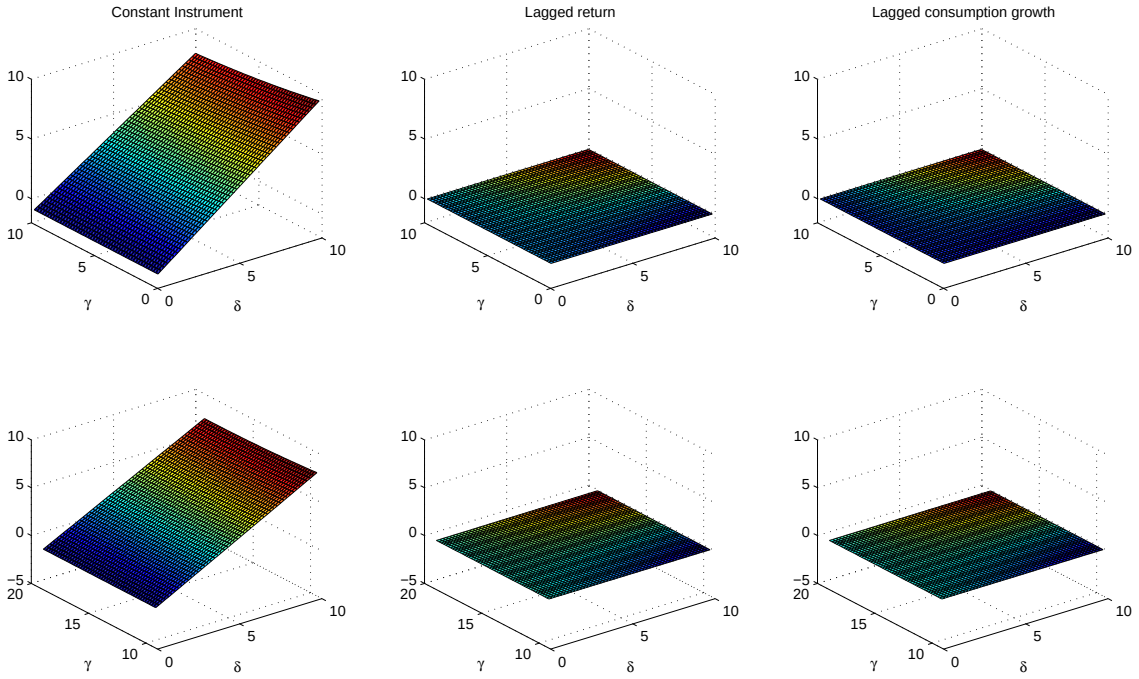

Figure 3: CCAPM: Moment restrictions as a function of the parameter values θ. Top figures for set 1 with a) constant instrument; b) lagged asset return; c) lagged consumption rate. Bottom figures for set 2. T=100 and M=2500.

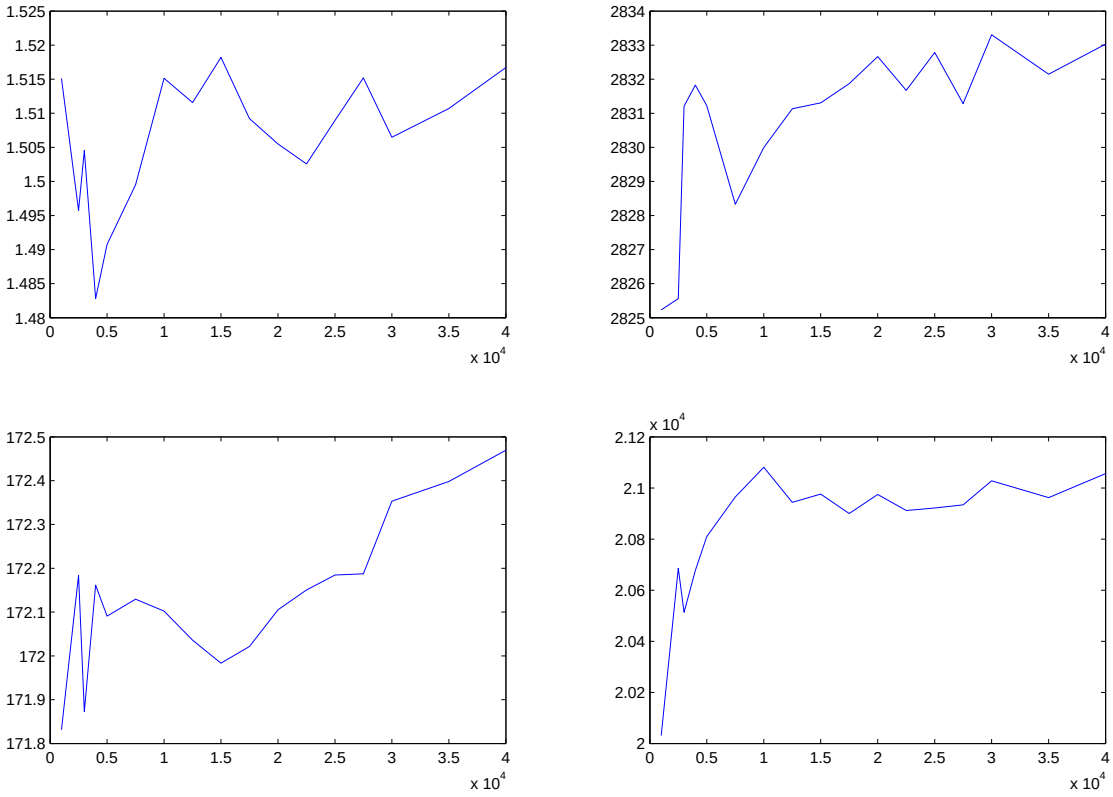

Figure 4: CCAPM: Ratio of the variances as a function of the sample size. Top for set 1, left for $Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$ and right for $Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$. Bottom for set 2.