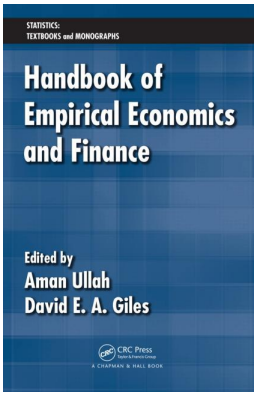

Handbook of Empirical Economics and Finance

Ullah Aman, E. A. Giles David

Efficient Inference with Poor Instruments

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b10440-3
Bertille Antoine, Eric Renault
Published online on: 20 Dec 2010

2

Efficient Inference with Poor Instruments: A General Framework

Bertille Antoine and Eric Renault

CONTENTS

2.1 Introduction

The generalized method of moments (GMM) provides a computationally convenient method for inference on the structural parameters of economic models. The method has been applied in many areas of economics but it was in empirical finance that the power of the method was first illustrated. Hansen (1982) introduced GMM and presented its fundamental statistical theory. Hansen and Hodrick (1980) and Hansen and Singleton (1982) showed the potential of the GMM approach to testing economic theories through their empirical analyzes of, respectively, foreign exchange markets and asset pricing. In such contexts, the cornerstone of GMM inference is a set of conditional moment restrictions. More generally, GMM is well suited for the test of an economic theory every time the theory can be encapsulated in the postulated unpredictability of some error term $u(Y_t, \theta)$ given as a known function of p

unknown parameters $\theta \in \Theta \subseteq \mathbb{R}^p$ and a vector of observed random variables Y_t. Then, the testability of the theory of interest is akin to the testability of a set of conditional moment restrictions,

$$E_t[u(Y_{t+1}, \theta)] = 0, \tag{2.1}$$

where the operator $E_t[.]$ denotes the conditional expectation given available information at time t. Moreover, under the null hypothesis that the theory summarized by the restrictions (Equation 2.1) is true, these restrictions are supposed to uniquely identify the true unknown value θ^0 of the parameters. Then, GMM considers a set of H instruments z_t assumed to belong to the available information at time t and to summarize the testable implications of Equation 2.1 by the implied unconditional moment restrictions:

$$E[\phi_t(\theta)] = 0 \quad \text{where} \quad \phi_t(\theta) = z_t \otimes u(Y_{t+1}, \theta). \tag{2.2}$$

The recent literature on weak instruments (see the seminal work by Stock and Wright 2000) has stressed that the standard asymptotic theory of GMM inference may be misleading because of the insufficient correlation between some instruments z_t and some components of the local explanatory variables of $[\partial u(Y_{t+1}, \theta)/\partial\theta]$. In this case, some of the moment conditions (Equation 2.2) are not only zero at θ^0 but rather flat and close to zero in a neighborhood of θ^0.

Many asset pricing applications of GMM focus on the study of a pricing kernel as provided by some financial theory. This pricing kernel is typically either a linear function of the parameters of interest, as in linear-beta pricing models, or a log-linear one as in most of the equilibrium based pricing models where parameters of interest are preference parameters. In all these examples, the weak instruments' problem simply relates to some lack of predictability of some asset returns from some lagged variables.

Since the seminal work of Stock and Wright (2000), it is common to capture the impact of the weakness of instruments by a drifting data generating process (hereafter DGP) such that the informational content of estimating equations $\rho_T(\theta) = E[\phi_t(\theta)]$ about structural parameters of interest is impaired by the fact that $\rho_T(\theta)$ becomes zero for all θ when the sample size goes to infinity. The initial goal of this so-called "weak instruments asymptotics" approach was to devise inference procedures robust to weak identification in the worst case scenario, as made formal by Stock and Wright (2000):

$$\rho_T(\theta) = \frac{\rho_{1T}(\theta)}{\sqrt{T}} + \rho_2(\theta_1) \quad \text{with} \quad \theta = [\theta_1' \ \theta_2']' \quad \text{and} \quad \rho_2(\theta_1) = 0 \Leftrightarrow \theta_1 = \theta_1^0. \tag{2.3}$$

The rationale for Equation 2.3 is the following. While some components θ_1 of θ would be identified in a standard way if the other components θ_2 were known, the latter ones are so weakly identified that for sample sizes typically available in practice, no significant increase of accuracy of estimators can be noticed when the sample size increases: the typical root-T consistency is

completely erased by the DGP drifting at the same rate through the term $\rho_{1T}(\theta)/\sqrt{T}$. It is then clear that this drifting rate is a worst case scenario, sensible when robustness to weak identification is the main concern, as it is the case for popular micro-econometric applications: for instance the study of Angrist and Krueger (1991) on returns to education.

The purpose of this chapter is somewhat different: taking for granted that some instruments may be poor, we nevertheless do not give up the efficiency goal of statistical inference. Even fragile information must be processed optimally, for the purpose of both efficient estimation and powerful testing. This point of view leads us to a couple of modifications with respect to the traditional weak instruments asymptotics.

First, we consider that the worst case scenario is a possibility but not the general rule. Typically, we revisit the drifting DGP (Equation 2.3) with a more general framework like:

$$\rho_T(\theta) = \frac{\rho_{1T}(\theta)}{T^{\lambda}} + \rho_2(\theta_1) \quad \text{with} \quad 0 \leq \lambda \leq 1/2.$$

The case $\lambda = 1/2$ has been the main focus of interest of the weak instruments literature so far because it accommodates the observed lack of consistency of some GMM estimators (typically estimators of θ_2 in the framework of Equation 2.3) and the implied lack of asymptotic normality of the consistent estimators (estimators of θ_1 in the framework of Equation 2.3). We rather set the focus on an intermediate case, $0 < \lambda < 1/2$, which has been dubbed nearly weak identification by Hahn and Kuersteiner (2002) in the linear case and Caner (2010) for nonlinear GMM. Standard (strong) identification would take $\lambda = 0$. Note also that nearly weak identification is implicitly studied by several authors who introduce infinitely many instruments: the large number of instruments partially compensates for the genuine weakness of each of them individually (see Han and Phillips 2006; Hansen, Hausman, and Newey 2008; Newey and Windmeijer 2009).

However, following our former work in Antoine and Renault (2009, 2010a), our main contribution is above all to consider that several patterns of identification may show up simultaneously. This point of view appears especially relevant for the asset pricing applications described above. Nobody would pretend that the constant instrument is weak. Therefore, the moment condition, $E[u(Y_{t+1}, \theta)] = 0$, should not display any drifting feature (as it actually corresponds to $\lambda = 0$). Even more interestingly, Epstein and Zin (1991) stress that the pricing equation for the market return is poorly informative about the difference between the risk aversion coefficient and the inverse of the elasticity of substitution. Individual asset returns should be more informative.

This paves the way for two additional extensions in the framework (Equation 2.3). First, one may consider, depending on the moment conditions, different values of the parameter λ of drifting DGP. Large values of λ would be assigned to components $[z_{it} \times u_j(Y_{t+1}, \theta)]$ for which either the pricing of asset j or the lagged value of return i are especially poorly informative. Second,

there is no such thing as a parameter θ_2 always poorly identified or parameter θ_1 which would be strongly identified if the other parameters θ_2 were known. Instead, one must define directions in the parameter space (like the difference between risk aversion and inverse of elasticity of substitution) that may be poorly identified by some particular moment restrictions.

This heterogeneity of identification patterns clearly paves the way for the device of optimal strategies for inferential use of fragile (or poor) information. In this chapter, we focus on a case where asymptotic efficiency of estimators is well-defined through the variance of asymptotically normal distributions. The price to pay for this maintained tool is to assume that the set of moment conditions that are not genuinely weak ($\lambda < 1/2$) is sufficient to identify the true unknown value θ^0 of the parameters. In this case, normality must be reconsidered at heterogeneous rates smaller than the standard root-T in different directions of the parameter space (depending on the strength of identification about these directions). At least, non-normal asymptotic distributions introduced by situations of partial identification as in Phillips (1989) and Choi and Phillips (1992) are avoided in our setting. It seems to us that, by considering the large sample sizes typically available in financial econometrics, working with the maintained assumption of asymptotic normality of estimators is reasonable; hence, the study of efficiency put forward in this chapter. However, there is no doubt that some instruments are poorer and that some directions of the parameter space are less strongly identified. Last but not least: even though we are less obsessed by robustness to weak identification in the worst case scenario, we do not want to require from the practitioner a prior knowledge of the identification schemes. Efficient inference procedures must be feasible without requiring any prior knowledge neither of the different rates λ of nearly weak identification, nor of the heterogeneity of identification patterns in different directions in the parameter space.

To delimit the focus of this chapter, we put an emphasis on efficient inference. There are actually already a number of surveys that cover the earlier literature on inference robust to weak instruments. For example, Stock, Wright, and Yogo (2002) set the emphasis on procedures available for detecting and handling weak instruments in the linear instrumental variables model. More recently, Andrews and Stock (2007) wrote an excellent review, discussing many issues involved in testing and building confidence sets robust to the weak instrumental variables problem. Smith (2007) revisited this review, with a special focus on empirical likelihood-based approaches. This chapter is organized as follows. Section 2.2 introduces framework and identification procedure with poor instruments; the consistency of all GMM estimators is deduced from an empirical process approach. Section 2.3 is concerned with asymptotic theory and inference. Section 2.4 compares our approach to others: we specifically discuss the linear instrumental variables regression model, the (non)equivalence between efficient two-step GMM and continuously updated GMM and the GMM-score test of Kleibergen (2005). Section 2.5 concludes. All the proofs are gathered in the appendix.

2.2 Identification with Poor Instruments

2.2.1 Framework

We consider the true unknown value θ^0 of the parameter $\theta \in \Theta \subset \mathbb{R}^p$ defined as the solution of the moment conditions $E[\phi_t(\theta)] = 0$ for some known function $\phi_t(.)$ of size K. Since the seminal work of Stock and Wright (2000), the weakness of the moment conditions (or instrumental variables) is usually captured through a drifting DGP such that the informational content of the estimating equations shrinks toward zero (for all θ) while the sample size T grows to infinity.

More precisely, the population moment conditions obtained from a set of *poor* instruments are modeled as a function $\rho_T(\theta)$ that depends on the sample size T and becomes zero when it goes to infinity. The statistical information about the estimating equations $\rho_T(\theta)$ is given by the sample mean $\bar{\phi}_T(\theta) = (1/T)\sum_{t=1}^{T}\phi_t(\theta)$ and the asymptotic behavior of the empirical process $\sqrt{T}[\bar{\phi}_T(\theta) - \rho_T(\theta)]$.

Assumption 2.1 *(Functional CLT)*
(i) There exists a sequence of deterministic functions ρ_T such that the empirical process $\sqrt{T}\left[\bar{\phi}_T(\theta) - \rho_T(\theta)\right]$, for $\theta \in \Theta$, weakly converges (for the sup-norm on Θ) toward a Gaussian process on Θ with mean zero and covariance $S(\theta)$.
(ii) There exists a sequence A_T of deterministic nonsingular matrices of size K and a bounded deterministic function c such that

$$\lim_{T\to\infty}\sup_{\theta\in\Theta}\|c(\theta) - A_T\rho_T(\theta)\| = 0.$$

The rate of convergence of coefficients of the matrix A_T toward infinity characterizes the degree of global identification weakness. Note that we may not be able to replace $\rho_T(\theta)$ by the function $A_T^{-1}c(\theta)$ in the convergence of the empirical process since

$$\sqrt{T}\left[\rho_T(\theta) - A_T^{-1}c(\theta)\right] = \left(\frac{A_T}{\sqrt{T}}\right)^{-1}[A_T\rho_T(\theta) - c(\theta)],$$

may not converge toward zero. While genuine weak identification like Stock and Wright (2000) means that $A_T = \sqrt{T}Id_K$ (with Id_K identity matrix of size K), we rather consider nearly weak identification where some rows of the matrix A_T may go to infinity strictly slower than $\sqrt{T}$. Standard GMM asymptotic theory based on strong identification would assume $A_T = Id_K$ and $\rho_T(\theta) = c(\theta)$ for all T. In this case, it would be sufficient to assume asymptotic normality of $\sqrt{T}\bar{\phi}_T(\theta^0)$ at the true value θ^0 of the parameters (while $\rho_T(\theta^0) = c(\theta^0) = 0$). By contrast, as already pointed out by Stock and

Wright (2000), the asymptotic theory with (nearly) weak identification is more involved since it assumes a functional central limit theorem uniform on Θ. However, this uniformity is not required in the linear case,[1] as now illustrated.

Example 2.1 (Linear IV regression)
We consider a structural linear equation: $y_t = x_t'\theta + u_t$ for $t = 1, \cdots, T$, where the p explanatory variables x_t may be endogenous. The true unknown value θ^0 of the structural parameters is defined through $K \geq p$ instrumental variables z_t uncorrelated with $(y_t - x_t'\theta^0)$. In other words, the estimating equations for standard IV estimation are

$$\bar{\Phi}_T(\hat{\theta}_T) = \frac{1}{T} Z'(y - X\hat{\theta}_T) = 0, \tag{2.4}$$

where X (respectively Z) is the (T, p) (respectively (T, K)) matrix which contains the available observations of the p explanatory variables (respectively the K instrumental variables) and $\hat{\theta}_T$ denotes the standard IV estimator of θ. Inference with poor instruments typically means that the required rank condition is not fulfilled, even asymptotically:

$$\texttt{Plim}\left[\frac{Z'X}{T}\right] \quad \textit{may not be of full rank.}$$

Weak identification means that only $\texttt{Plim}[\frac{Z'X}{\sqrt{T}}]$ has full rank, while intermediate cases with nearly weak identification have been studied by Hahn and Kuersteiner (2002). The following assumption conveniently nests all the above cases.

Assumption L1 *There exists a sequence A_T of deterministic nonsingular matrices of size K such that $\texttt{Plim}[A_T \frac{Z'X}{T}] = \Pi$ is full column rank.*

While standard strong identification asymptotics assume that the largest absolute value of all coefficients of the matrix A_T, $\|A_T\|$, is of order $\mathcal{O}(1)$, weak identification means that $\|A_T\|$ grows at rate $\sqrt{T}$. The following assumption focuses on nearly weak identification, which ensures consistent IV estimation under standard regularity conditions as explained below.

Assumption L2 *The largest absolute value of all coefficients of the matrix A_T is $o(\sqrt{T})$.*

To deduce the consistency of the estimator $\hat{\theta}_T$, we rewrite Equation (2.4) as follows and pre-multiply it by A_T:

$$\frac{Z'X}{T}(\hat{\theta}_T - \theta^0) + \frac{Z'u}{T} = 0 \;\Rightarrow\; A_T \frac{Z'X}{T}(\hat{\theta}_T - \theta^0) + A_T \frac{Z'u}{T} = 0. \tag{2.5}$$

After assuming a central limit theorem for $(Z'u/\sqrt{T})$ and after considering (for simplicity) that the unknown parameter vector θ evolves in a bounded subset of $\mathbb{R}^p$,

[1] Note also that uniformity is not required in the linear-in-variable case.

we get

$$\Pi(\hat{\theta}_T - \theta^0) = o_P(1).$$

Then, the consistency of $\hat{\theta}_T$ *directly follows from the full column rank assumption on* Π. *Note that uniformity with respect to* θ *does not play any role in the required central limit theorem since we have*

$$\sqrt{T}[\bar{\phi}_T(\theta) - \rho_T(\theta)] = \frac{Z'u}{\sqrt{T}} + \sqrt{T}\left[\frac{Z'X}{T} - E[z_t x_t']\right](\theta^0 - \theta)$$

with

$$\rho_T(\theta) = E[z_t x_t'](\theta^0 - \theta).$$

Linearity of the moment conditions with respect to unknown parameters allows us to factorize them out and uniformity is not an issue.

It is worth noting that in the linear example, the central limit theorem has been used to prove consistency of the IV estimator and not to derive its asymptotic normal distribution. This nonstandard proof of consistency will be generalized for the nonlinear case in the next subsection, precisely thanks to the uniformity of the central limit theorem over the parameter space. As far as asymptotic normality of the estimator is concerned, the key issue is to take advantage of the asymptotic normality of $\sqrt{T}\bar{\phi}_T(\theta^0)$ at the true value θ^0 of the parameters (while $\rho_T(\theta^0) = c(\theta^0) = 0$). The linear example again shows that, in general, doing so involves additional assumptions about the structure of the matrix A_T. More precisely, we want to stress that when several degrees of identification (weak, nearly weak, strong) are considered simultaneously, the above assumptions are not sufficient to derive a meaningful asymptotic distributional theory. In our setting, it means that the matrix A_T is not simply a scalar matrix $\lambda_T A$ with the scalar sequence λ_T possibly going to infinity but not faster than $\sqrt{T}$. This setting is in contrast with most of the literature on weak instruments (see Kleibergen 2005; Caner 2010 among others).

Example 2.1 (Linear IV regression – continued)
To derive the asymptotic distribution of the estimator $\hat{\theta}_T$, *pre-multiplying the estimating equations by the matrix* A_T *may not work. However, for any sequence of deterministic nonsingular matrices* $\tilde{A}_T$ *of size* p, *we have*

$$\frac{Z'X}{T}(\hat{\theta}_T - \theta^0) + \frac{Z'u}{T} = 0 \;\Rightarrow\; \frac{Z'X}{T}\tilde{A}_T\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0) = -\frac{Z'u}{\sqrt{T}}. \qquad (2.6)$$

If $[\frac{Z'X}{T}\tilde{A}_T]$ *converges toward a well-defined matrix with full column rank, a central limit theorem for* $(Z'u/\sqrt{T})$ *ensures the asymptotic normality of* $\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$. *In general, this condition cannot be deduced from Assumption L1 unless the matrix* A_T *appropriately commutes with* $[\frac{Z'X}{T}]$. *Clearly, this is not an issue if* A_T *is simply a scalar matrix* $\lambda_T Id_K$. *In case of nearly weak identification* ($\lambda_T = o(\sqrt{T})$), *it delivers*

asymptotic normality of the estimator at slow rate $\sqrt{T}/\lambda_T$ while, in case of genuine weak identification ($\lambda_T = \sqrt{T}$), consistency is not ensured and asymptotic Cauchy distributions show up.

In the general case, the key issue is to justify the existence of a sequence of deterministic nonsingular matrices $\tilde{A}_T$ of size p such that $[\frac{Z'X}{T}\tilde{A}_T]$ converges toward a well-defined matrix with full column rank. In the just-identified case ($K = p$), it follows directly from Assumption L1 with $\tilde{A}_T = \Pi^{-1}A_T$:

$$\texttt{Plim}\left[\frac{Z'X}{T}\Pi^{-1}A_T\right] = \texttt{Plim}\left[\frac{Z'X}{T}\left(A_T\frac{Z'X}{T}\right)^{-1}A_T\right] = Id_p.$$

In the overidentified case ($K > p$), it is rather the structure of the matrix A_T (and not only its norm, or largest coefficient) that is relevant. Of course, by Equation 2.5, we know that

$$\frac{Z'X}{T}\sqrt{T}\left(\hat{\theta}_T - \theta^0\right) = -\frac{Z'u}{\sqrt{T}}$$

is asymptotically normal. However, in case of lack of strong identification, $(Z'X/T)$ is not asymptotically full rank and some linear combinations of $\sqrt{T}(\hat{\theta}_T - \theta^0)$ may blow up. To provide a meaningful asymptotic theory for the IV estimator $\hat{\theta}_T$, the following condition is required. In the general case, we explain why such a sequence $\tilde{A}_T$ always exists and how to construct it (see Theorem 2.3).

Assumption L3 *There exists a sequence $\tilde{A}_T$ of deterministic nonsingular matrices of size p such that $\texttt{Plim}[\frac{Z'X}{T}\tilde{A}_T]$ is full column rank.*

It is then straightforward to deduce that $\sqrt{T}\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ is asymptotically normal. Hansen, Hausman, and Newey (2008) provide a set of assumptions to derive similar results in the case of many weak instruments asymptotics. In their setting, considering a number of instruments growing to infinity can be seen as a way to ensure Assumption L2, even though weak identification (or $\|A_T\|$ of order $\sqrt{T}$) is assumed for any given finite set of instruments.

The above example shows that, in case of (nearly) weak identification, a relevant asymptotic distributional theory is not directly about the common sequence $\sqrt{T}(\hat{\theta}_T - \theta^0)$ but rather about a well-suited reparametrization $\tilde{A}_T^{-1}\sqrt{T}(\hat{\theta}_T - \theta^0)$. Moreover, lack of strong identification means that the matrix of reparametrization $\tilde{A}_T$ also involves a rescaling (going to infinity with the sample size) in order to characterize slower rates of convergence. For sake of structural interpretation, it is worth disentangling the two issues: first, the rotation in the parameter space, which is assumed well-defined at the limit (when $T \to \infty$); second, the rescaling. The convenient mathematical tool is the singular value decomposition of the matrix A_T (see Horn and Johnson 1985, pp.414–416, 425). We know that the nonsingular matrix A_T can always be written as: $A_T = M_T\Lambda_T N_T'$ with M_T, N_T, and Λ_T three square matrices of size K, M_T, and N_T orthogonal and Λ_T diagonal with nonzero entries. In our

context of rates of convergence, we want to see the singular values of the matrix A_T (that is the diagonal coefficients of Λ_T) as positive and, without loss of generality, ranked in increasing order. If we consider Assumption 2.1(ii) again, N_T' can intuitively be seen as selecting appropriate linear combinations of the moment conditions and Λ_T as rescaling appropriately these combinations. On the other hand, M_T is related to selecting linear combinations of the deterministic vector c.

Without loss of generality, we always consider the singular value decomposition $A_T = M_T \Lambda_T N_T'$ such that the diagonal matrix sequence Λ_T has positive diagonal coefficients bounded away from zero and the two sequences of orthogonal matrices M_T and N_T have well-defined limits[2] when $T \to \infty$, M and N, respectively, both orthogonal matrices.

2.2.2 Consistency

In this subsection, we set up a framework where consistency of a GMM estimator is warranted in spite of lack of strong identification. The key is to ensure that a sufficient subset of the moment conditions is not impaired by genuine weak identification: in other words, the corresponding rates of convergence of the singular values of A_T are slower than $\sqrt{T}$. As explained above, specific rates of convergence are actually assigned to appropriate linear combinations of the moment conditions:

$$d(\theta) = M^{-1} c(\theta) = \lim_T \left[\Lambda_T N_T' \rho_T(\theta) \right].$$

Our maintained identification assumption follows:

Assumption 2.2 *(Identification)*
(i) The sequence of nonsingular matrices A_T writes $A_T = M_T \Lambda_T N_T'$ with $\lim_T [M_T] = M$, $\lim_T [N_T] = N$, M, and N orthogonal matrices.

(ii) The sequence of matrices Λ_T is partitioned as $\Lambda_T = \begin{bmatrix} \tilde{\Lambda}_T & 0 \\ 0 & \check{\Lambda}_T \end{bmatrix}$, such that $\tilde{\Lambda}_T$ and $\check{\Lambda}_T$ are two diagonal matrices, respectively, of size $\tilde{K}$ and $(K - \tilde{K})$, with[3] $\|\tilde{\Lambda}_T\| = o(\sqrt{T})$, $\|\check{\Lambda}_T\| = \mathcal{O}(\sqrt{T})$ and $\check{\Lambda}_T^{-1} = o(\|\tilde{\Lambda}_T\|^{-1})$.
(iii) The vector d of moment conditions, with $d(\theta) = M^{-1} c(\theta) = \lim_T [\Lambda_T N_T' \rho_T(\theta)]$, is partitioned accordingly as $d = [\tilde{d}' \; \check{d}']'$ such that θ^0 is a well-separated zero of the vectorial function $\tilde{d}$ of size $\tilde{K} \leq p$:

$$\forall \epsilon > 0 \quad \inf_{\|\theta - \theta^0\| > \epsilon} \|\tilde{d}(\theta)\| > 0.$$

(iv) The first $\tilde{K}$ elements of $N_T \rho_T(\theta^0)$ are identically equal to zero for any T.

[2] It is well known that the group of real orthogonal matrices is compact (see Horn and Johnson 1985, p. 71). Hence, one can always define M and N for convergent subsequences, respectively M_{T_n} and N_{T_l}. To simplify the notations, we only refer to sequences and not subsequences.

[3] $\|M\|$ denotes the largest element (in absolute value) of any matrix M.

As announced, the above identification assumption ensures that the first $\tilde{K}$ moment conditions are only possibly nearly weak (and not genuinely weak), $\|\tilde{\Lambda}_T\| = o(\sqrt{T})$, and sufficient to identify the true unknown value θ^0:

$$\tilde{d}(\theta) = 0 \;\Leftrightarrow\; \theta = \theta^0.$$

The additional moment restrictions, as long as they are strictly weaker ($\check{\Lambda}_T^{-1} = o(\|\tilde{\Lambda}_T\|^{-1})$, may be arbitrarily weak and even misspecified, since we do not assume $\check{d}(\theta^0) = 0$. It is worth noting that the above identification concept is nonstandard, since all singular values of the matrix A_T may go to infinity. In such a case, we have

$$\mathtt{Plim}\left[\bar{\phi}_T(\theta)\right] = 0 \quad \forall\, \theta \in \Theta. \tag{2.7}$$

This explains why the following consistency result of a GMM estimator cannot be proved in a standard way. The key argument is actually tightly related to the uniform functional central limit theorem of Assumption 2.1.

Theorem 2.1 *(Consistency of $\hat{\theta}_T$)*
We define a GMM-estimator:

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta}\left[\bar{\phi}_T'(\theta)\Omega_T\bar{\phi}_T(\theta)\right] \tag{2.8}$$

with Ω_T a sequence of symmetric positive definite random matrices of size K which converges in probability toward a positive definite matrix Ω.

Under the Assumptions 2.1 and 2.2, any GMM estimator like Equation 2.8 is weakly consistent.

We now explain why the consistency result cannot be deduced from a standard argument based on a simple rescaling of the moment conditions to avoid asymptotic degeneracy of Equation 2.7. The GMM estimator (Equation 2.8) can be rewritten as

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta}\left\{[\Lambda_T N_T'\bar{\phi}_T(\theta)]' W_T [\Lambda_T N_T'\bar{\phi}_T(\theta)]\right\}$$

with a weighting matrix sequence, $W_T = \Lambda_T^{-1} N_T'\Omega_T N_T \Lambda_T^{-1}$, and rescaled moment conditions $[\Lambda_T N_T'\bar{\phi}_T(\theta)]$ such that

$$\mathtt{Plim}[\Lambda_T N_T'\bar{\phi}_T(\theta)] = \lim_T[\Lambda_T N_T'\rho_T(\theta)] = d(\theta) \neq 0 \;\text{ for } \theta \neq \theta^0.$$

However, when all singular values of A_T go to infinity, the weighting matrix sequence W_T is such that

$$\mathtt{Plim}\,[W_T] = \lim_T\left[\Lambda_T^{-1} N'\Omega N \Lambda_T^{-1}\right] = 0.$$

In addition, the limit of the GMM estimator in Theorem 2.1 is solely determined by the strongest moment conditions that identify θ^0. There is actually no need to assume that the last $(K - \tilde{K})$ coefficients in $[\Lambda_T N_T'\rho_T(\theta^0)]$, or even

their limits $\check{d}(\theta^0)$, are equal to zero. In other words, the additional estimating equations $\check{d}(\theta) = 0$ may be biased and this has no consequence on the limit value of the GMM estimator insofar as the additional moment restrictions are strictly weaker than the initial ones, $\check{\Lambda}_T^{-1} = o(\|\tilde{\Lambda}_T\|^{-1})$. They may even be genuinely weak with $\|\check{\Lambda}_T\| = \sqrt{T}$. This result has important consequences on the power of the overidentification test defined in the next section.

2.3 Asymptotic Distribution and Inference

2.3.1 Efficient Estimation

In our setting, rates of convergence slower than square-root T are produced because some coefficients of A_T may go to infinity while the asymptotically identifying equations are given by $\rho_T(\theta) \overset{a}{\sim} A_T^{-1}c(\theta)$. Since we do not want to introduce other causes for slower rates of convergence (like singularity of the Jacobian matrix of the moment conditions, as done in Sargan 1983), first-order local identification is maintained.

Assumption 2.3 *(Local identification)*

(i) $\theta \to c(\theta)$, $\theta \to d(\theta)$ *and* $\theta \to \rho_T(\theta)$ *are continuously differentiable on the interior of* Θ.

(ii) θ^0 *belongs to the interior of* Θ.

(iii) *The* $(\tilde{K}, p)$*-matrix* $[\partial\check{d}(\theta^0)/\partial\theta']$ *has full column rank* p.

(iv) $\Lambda_T N_T'[\partial\rho_T(\theta)/\partial\theta']$ *converges uniformly on the interior of* Θ *toward* $M^{-1}[\partial c(\theta)/\partial\theta'] = \partial d(\theta)/\partial\theta'$.

(v) *The last* $(K - \tilde{K})$ *elements of* $N_T\rho_T(\theta^0)$ *are either identically equal to zero for any* T*, or genuinely weak with the corresponding element of* $\check{\Lambda}_T$ *equal to* $\sqrt{T}$.

Assumption 2.3(iv) states that rates of convergence are maintained after differentiation with respect to the parameters. Contrary to the linear case, this does not follow automatically in the general case. Then, we are able to show that the structural parameters are identified at the slowest rate available from the set of identifying equations. Assumption 2.3(v) ensures that the additional moment restrictions (the ones not required for identification) are either well-specified or genuinely weak: this ensures that these conditions do not deteriorate the rate of convergence of the GMM estimator (see Theorem 2.2). Intuitively, a GMM estimator is always a linear combination of the moment conditions. Hence, if some moments are misspecified and do not *disappear* as fast as $\sqrt{T}$, they can only deteriorate the rate of convergence of the estimator.

Theorem 2.2 *(Rate of convergence)*
Under Assumptions 2.1 to 2.3, any GMM estimator $\hat{\theta}_T$ like Equation 2.8 is such that

$$\|\hat{\theta}_T - \theta^0\| = \mathcal{O}_p(\|\tilde{\Lambda}_T\|/\sqrt{T}).$$

The above result is quite poor, since it assigns the slowest possible rate to all components of the structural parameters. We now show how to identify faster directions in the parameter space. The first step consists in defining a matrix $\tilde{A}_T$ similar to the one introduced in the linear example. The following result justifies its existence: in the appendix, we also explain in details how to construct it.

Theorem 2.3 *Under Assumptions 2.1 to 2.3, there exists a sequence $\tilde{A}_T$ of deterministic nonsingular matrices of size p such that the smallest eigenvalue of $\tilde{A}_T' \tilde{A}_T$ is bounded away from zero and*

$$\lim_T \left[\Lambda_T^{-1} M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \tilde{A}_T \right] \quad \textit{exists and is full column rank with } \|\tilde{A}_T\| = \mathcal{O}(\|\tilde{\Lambda}_T\|).$$

Following the approach put forward in the linear example, Theorem 2.3 is used to derive the asymptotic theory of the estimator $\hat{\theta}_T$. Since,

$$\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} \sqrt{T}(\hat{\theta}_T - \theta^0) = \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} \tilde{A}_T \sqrt{T} \tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0),$$

a meaningful asymptotic distributional theory is not directly about the common sequence $\sqrt{T}(\hat{\theta}_T - \theta^0)$, but rather about a well-suited reparametrization $\tilde{A}_T^{-1}\sqrt{T}(\hat{\theta}_T - \theta^0)$. Similar to the structure of A_T, $\tilde{A}_T$ involves a reparametrization and a rescaling. In others words, specific rates of convergence are actually assigned to appropriate linear combinations of the structural parameters.

Assumption 2.4 *(Regularity)*

$$(i)\ \sqrt{T}\left[\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} - A_T^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \right] = \mathcal{O}_P(1)$$

$$(ii)\ \sqrt{T} \frac{\partial}{\partial \theta} \left[\frac{\partial \bar{\phi}_T(\theta)}{\partial \theta'} \right]_{k.} - \frac{\partial}{\partial \theta} \left[A_T^{-1} \frac{\partial c(\theta)}{\partial \theta'} \right]_{k.} = \mathcal{O}_P(1) \textit{ and } \frac{\partial}{\partial \theta} \left[\frac{\partial c(\theta)}{\partial \theta'} \right]_{k.} = \mathcal{O}_P(1)$$

for any $1 \leq k \leq K$, uniformly on the interior of Θ with $[M]_{k.}$ the kth row the matrix M.

With additional regularity Assumption 2.4(i), Corollary 2.1 extends Theorem 2.3 to rather consider the empirical counterparts of the moment conditions: it is the nonlinear analog of Assumption L3.

Corollary 2.1 *(Nonlinear extension of L3)*
Under Assumptions 2.1–2.3 and 2.4(i), we have

$$\Gamma(\theta^0) \equiv \texttt{Plim}\left[\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'}\tilde{A}_T\right] \text{ exists and is full column rank.}$$

In order to derive a standard asymptotic theory for the GMM estimator $\hat{\theta}_T$, we need to impose an assumption on the homogeneity of identification.

Assumption 2.5 *(Homogenous identification)*

$$\frac{\sqrt{T}}{\underline{\tilde{\Lambda}}_T} = o\left(\frac{\sqrt{T}}{\|\tilde{\Lambda}_T\|}\right)^2$$

where $\|M\|$ $\underline{M}$ *denote respectively the largest and the smallest absolute values of all nonzero coefficients of the matrix M.*

Intuitively, assumption 2.5 ensures that second-order terms in Taylor expansions remain negligible in front of the first-order central limit theorem terms. Note that a sufficient condition for homogenous identification is dubbed nearly-strong and writes: $\|\tilde{\Lambda}_T\|^2 = o(\sqrt{T})$. It corresponds to the above homogenous identification condition when some moment conditions are strong, that is $\underline{\tilde{\Lambda}}_T = 1$. Then we want to ensure that the slowest possible rate of convergence of parameter estimators is strictly faster than $T^{1/4}$. This nearly-strong condition is actually quite standard in semiparametric econometrics to control for the impact of infinite dimensional nuisance parameters (see Andrews' (1994) MINPIN estimators and Newey's (1994) linearization assumption).

The asymptotic distribution of the rescaled estimated parameters $\sqrt{T}\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ can now be characterized by seemingly standard GMM formulas:

Theorem 2.4 *(Asymptotic distribution of* $\hat{\theta}_T$*)*
Under Assumptions 2.1–2.5, any GMM estimator $\hat{\theta}_T$ *like Equation 2.8 is such that* $\sqrt{T}\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ *is asymptotically normal with mean zero and variance* $\Sigma(\theta^0)$ *given by*

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

where $S(\theta^0)$ *is the asymptotic variance of* $\sqrt{T}\bar{\phi}_T(\theta^0)$.

Theorem 2.4 paves the way for a concept of efficient estimation in presence of poor instruments. By a common argument, the unique limit weighting matrix Ω minimizing the above covariance matrix is clearly $\Omega = [S(\theta^0)]^{-1}$.

Theorem 2.5 *(Efficient GMM estimator)*
Under Assumptions 2.1–2.5, any GMM estimator $\hat{\theta}_T$ *like Equation 2.8 with a weighting matrix* $\Omega_T = S_T^{-1}$*, where* S_T *denotes a consistent estimator of* $S(\theta^0)$,

is such that $\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ *is asymptotically normal with mean zero and variance* $[\Gamma'(\theta^0)S^{-1}(\theta^0)\Gamma(\theta^0)]^{-1}$.

In our framework, the terminology "efficient GMM" and "standard formulas" for asymptotic covariance matrices must be carefully qualified. On the one hand, it is true that for all practical purposes, Theorem 2.5 states that, for T large enough, $\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ can be seen as a Gaussian vector with mean zero and variance consistently estimated by

$$\tilde{A}_T^{-1}\left[\frac{\partial\bar{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}\tilde{A}_T^{-1\prime}, \tag{2.9}$$

since $\Gamma(\theta^0) = \mathtt{Plim}[\frac{\partial\bar{\phi}_T(\theta^0)}{\partial\theta'}\tilde{A}_T]$. However, it is incorrect to deduce from Equation (2.9) that, after simplifications on both sides by $\tilde{A}_T^{-1}$, $\sqrt{T}(\hat{\theta}_T - \theta^0)$ can be seen (for T large enough) as a Gaussian vector with mean zero and variance consistently estimated by

$$\left[\frac{\partial\bar{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}. \tag{2.10}$$

This is wrong since the matrix $[\frac{\partial\bar{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}]$ is asymptotically singular. In this sense, a truly standard GMM theory does not apply and at least some components of $\sqrt{T}(\hat{\theta}_T - \theta^0)$ must blow up. Quite surprisingly, it turns out that the spurious feeling that Equation 2.10 estimates the asymptotic variance (as usual) is tremendously useful for inference as explained in Subsection 2.3.2. Intuitively, it explains why standard inference procedures work, albeit for nonstandard reasons. As a consequence, for all practical purposes related to inference about the structural parameters θ, the knowledge of the matrices A_T and $\tilde{A}_T$ is not required.

However, the fact that the correct understanding of the "efficient GMM" covariance matrix as estimated by Equation 2.9 involves the sequence of matrices $\tilde{A}_T$ is important for two reasons.

First, it is worth reminding that the construction of the matrix $\tilde{A}_T$ only involves the first $\tilde{K}$ components of the rescaled estimating equations $[N_T'\rho_T(\theta)]$. This is implicit in the rate of convergence of $\|\tilde{A}_T\|$ put forward in Theorem 2.3 and quite clear in its proof. In other words, when the total number of moment conditions K is strictly larger than $\tilde{K}$, the last $(K - \tilde{K})$ rows of the matrix $\Gamma(\theta^0) = \mathtt{Plim}[\frac{\partial\bar{\phi}_T(\theta^0)}{\partial\theta'}\tilde{A}_T]$ are equal to zero. Irrespective of the weighting matrix's choice for GMM estimation, the associated estimator does not depend asymptotically on these last moment conditions. Therefore, there is an obvious waste of information: the so-called efficient GMM estimator of Theorem 2.5 does not make use of all the available information. Moment conditions based on poorer instruments (redundant for the purpose of identification) should actually be used for improved accuracy of the estimator, as explicitly shown in Antoine and Renault (2010a).

Second, the interpretation of the matrix $\tilde{A}_T$ in terms of reparametrization is underpinned by the proof of Theorem 2.3 which shows that

$$\tilde{A}_T = [\ \lambda_{1T} R_1 \vdots \lambda_{2T} R_2 \vdots \cdots \vdots \lambda_{LT} R_L\ ] = R\Delta_T \text{ with } R = [\ R_1 \vdots R_2 \vdots \cdots \vdots R_L\ ].$$

R is a nonsingular matrix of size p with each submatrix R_i of size (p, s_i); Δ_T is a diagonal matrix with L diagonal blocks equal to $\lambda_{iT} Id_{s_i}$. It is worth reinterpreting Theorem 2.5 in terms of the asymptotic distribution of the estimator of a new parameter vector[4]:

$$\eta = R^{-1}\theta = [\eta_1'\ \eta_2'\ \cdots\ \eta_L']'.$$

Theorem 2.5 states that $(R^{-1}\hat{\theta}_T)$ is a consistent asymptotically normal estimator of the true unknown value $\eta^0 = R^{-1}\theta^0$, while each subvector η_i of size s_i is attached to a specific (slower) rate of convergence $\sqrt{T}/\lambda_{iT}$. It is clear in the appendix that this reparametrization is performed according to the directions which span the range of the Jacobian matrix of the rescaled "efficient" moment conditions $\tilde{d}(\theta)$, that is according to the columns of the matrix R. Even though the knowledge of the matrix R (and corresponding rates λ_{iT}) is immaterial for the practical implementation of inference procedures on structural parameters (as shown in Section 2.3.2), it may matter for a fair assessment of the accuracy of this inference. As an illustration, Subsection 2.4.3 studies the power of score-type tests against sequences of local alternatives in different directions.

In the context of the consumption-based capital asset pricing model (CCAPM) discussed in Stock and Wright (2000) and Antoine and Renault (2009), there are two structural parameters: θ_1, the subjective discount factor and θ_2, the coefficient of relative risk aversion of a representative investor. Antoine and Renault (2009) provide compelling evidence that a first parameter η_1, estimated at fast rate $\sqrt{T}$, is very close to θ_1 (the estimation results show that $\eta_1 = 0.999\theta_1 - 0.007\theta_2$), while any other direction in the parameter space, like for instance the risk aversion parameter θ_2, is estimated at a much slower rate. In other words, all parameters are consistently estimated as shown in Stock and Wright's (2000) empirical results (and contrary to their theoretical framework), but the directions with $\sqrt{T}$-consistent estimation are now inferred from data instead of being considered as a prior specification.

The practical way to consistently estimate the matrix R from the sample counterpart of the Jacobian matrix of the moment conditions is extensively discussed in Antoine and Renault (2010a). Of course, since this Jacobian matrix involves in general the unknown structural parameters θ, there is little hope to consistently estimate R at a rate faster than the slowest one, namely $\sqrt{T}/\|\tilde{\Lambda}_T\|$. Interestingly enough, this slower rate does not impair the faster rates involved in Theorem 2.5. When R is replaced by its consistent estimator

[4] Note that the structural parameter θ is such that $\theta = \sum_{i=1}^{L} R_i \eta_i$.

$\hat{R}$, in the context of Theorem 2.5,

$$\sqrt{T}\Delta_T^{-1}\left(\hat{R}^{-1}\hat{\theta}_T - \hat{R}^{-1}\theta^0\right)$$

is still asymptotically normal with mean zero and variance $[\Gamma'(\theta^0)S^{-1}(\theta^0) \times \Gamma(\theta^0)]^{-1}$. The key intuition comes from the following decomposition:

$$\hat{R}^{-1}\hat{\theta}_T - \hat{R}^{-1}\theta^0 = R^{-1}(\hat{\theta}_T - \theta^0) + (\hat{R}^{-1} - R^{-1})(\hat{\theta}_T - \theta^0).$$

The potentially slow rates of convergence in the second term of the right-hand side do not deteriorate the fast rates in the relevant directions of $R^{-1}(\hat{\theta}_T - \theta^0)$: these slow rates show up as $T/\|\tilde{\Lambda}_T\|^2$ at worst, which is still faster than the fastest rate $\sqrt{T}/\lambda_{1T}$ by our nearly strong identification Assumption 2.5.

2.3.2 Inference

As discussed in the previous section, inference procedures are actually more involved than one may believe at first sight from the apparent similarity with standard GMM formulas. Nonetheless, the seemingly standard "efficient" asymptotic distribution theory of Theorem 2.5 paves the way for two usual results: the overidentification test and the Wald test.

Theorem 2.6 *(J test)*
Let S_T^{-1} be a consistent estimator of $\lim_T[\mathtt{Var}(\sqrt{T}\bar{\phi}_T(\theta^0))]^{-1}$.
Under Assumptions 2.1–2.5, for any GMM estimator like Equation (2.8), we have

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

As already announced, Theorem 2.1 has important consequences for the practice of GMM inference. We expect the above overidentification test to have little power to detect the misspecification of moment conditions when this misspecification corresponds to a subset of moment conditions of heterogeneous strengths. The proofs of Theorems 2.1 and 2.3 actually show that

$$T\bar{\phi}_T(\hat{\theta}_T)S_T^{-1}\bar{\phi}_T(\hat{\theta}_T) = \mathcal{O}_P\left(\frac{T}{\|\breve{\Lambda}_T\|^2}\right).$$

In other words, the standard J-test statistic for overidentification will not diverge as fast as the standard rate T of divergence and will even not diverge at all if the misspecified moment restrictions are genuinely weak ($\|\breve{\Lambda}_T\| = \sqrt{T}$).

Second, we are interested in testing the null hypothesis, $H_0 : g(\theta) = 0$, where the function $g : \Theta \to \mathbb{R}^q$ is continuously differentiable on the interior of Θ. We focus on Wald testing since it avoids estimation under the null which

may affect the reparametrization[5] previously defined. The following example illustrates how the standard delta-theorem is affected in our framework.

Example 2.2
Consider the null hypothesis $H_0 : g(\theta) = 0$ *with* g *a vector of size* q *such that*

$$\left[\frac{\partial g_j(\theta^0)}{\partial \theta}\right] \notin \texttt{col}\left[\frac{\partial \bar{d}_1'(\theta^0)}{\partial \theta}\right] \ \forall\ j = 1, \cdots, q$$

and a diagonal matrix Λ_T,

$$\Lambda_T = \begin{bmatrix} \lambda_{1T} Id_{K_1} & O \\ O & \lambda_{2T} Id_{K-K_1} \end{bmatrix} \text{with } \lambda_{1T} = o(\lambda_{2T}),\ \lambda_{2T} \to \infty, \text{ and } \lambda_{2T} = o(\sqrt{T}).$$

Applying the standard argument to derive the Wald test, we have that, under the null,

$$\left[\frac{\sqrt{T}}{\lambda_{2T}} g(\hat{\theta}_T)\right] \overset{a}{\sim} \left[\frac{\partial g(\theta^0)}{\partial \theta'} \frac{\sqrt{T}}{\lambda_{2T}} (\hat{\theta}_T - \theta^0)\right].$$

In other words, for T *large enough,* $\left[\frac{\sqrt{T}}{\lambda_{2T}} g(\hat{\theta}_T)\right]$ *can be seen as a normal random variable with mean 0 and variance*

$$\frac{\partial g(\theta^0)}{\partial \theta'} \left[\frac{\partial \bar{\phi}_T'(\theta^0)}{\partial \theta} [S(\theta^0)]^{-1} \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'}\right]^{-1} \frac{\partial g'(\theta^0)}{\partial \theta}.$$

Suppose now that there exists a nonzero vector α *such that*

$$\left[\frac{\partial g'(\theta^0)}{\partial \theta} \alpha\right] \in \texttt{col}\left[\frac{\partial \bar{d}_1'(\theta^0)}{\partial \theta}\right].$$

Then, under the null, $\left[\frac{\sqrt{T}}{\lambda_{1T}} \alpha' g(\hat{\theta}_T)\right]$ *is asymptotically normal and thus*

$$\frac{\sqrt{T}}{\lambda_{2T}} \alpha' g(\hat{\theta}_T) = \frac{\lambda_{1T}}{\lambda_{2T}} \frac{\sqrt{T}}{\lambda_{1T}} \alpha' g(\hat{\theta}_T) \xrightarrow{P} 0.$$

This means that even when a full rank assumption is maintained for the constraints to be tested, $\left[\frac{\sqrt{T}}{\lambda_{2T}} g(\hat{\theta}_T)\right]$ *does not behave asymptotically like a normal with a nonsingular variance matrix. This explains why deriving the asymptotic distributional theory for the Wald test statistic is nonstandard.*

Surprisingly enough, the above asymptotic singularity issue is immaterial and the standard Wald-type inference holds without additional regularity

[5] Typically, with additional information, the linear combinations of θ estimated respectively at specific rates of convergence may be defined differently. Caner (2010) derives the standard asymptotic equivalence results for the trinity of tests because he only considers testing when all parameters converge at the same nearly weak rate.

assumption as stated in Theorem 2.7. The intuition is the following. Consider a fictitious situation where the range of $[\partial \tilde{d}_1'(\theta^0)/\partial\theta]$ is known. Then, one can always define a nonsingular matrix H of size q and the associated vector h, $h(\theta) = Hg(\theta)$, in order to avoid the asymptotic singularity issue portrayed in Example 2.2. More precisely, with a (simplified) matrix A_T as in the above example, we consider

$$\begin{aligned}
&\text{for } j = 1, \cdots, q_1 : \left[\partial h_j(\theta^0)/\partial\theta\right] \in \mathtt{col}\left[\partial \tilde{d}_1'(\theta^0)/\partial\theta\right];\\
&\text{for } j = q_1 + 1, \cdots, q : \left[\partial h_j(\theta^0)/\partial\theta\right] \notin \mathtt{col}\left[\partial \tilde{d}_1'(\theta^0)/\partial\theta\right]\\
&\text{and no linear combinations of } \left[\partial h_j(\theta^0)/\partial\theta\right] \text{ does.}
\end{aligned}$$

Note that the new restrictions $h(\theta) = 0$ should be interpreted as a nonlinear transformations of the initial ones $g(\theta) = 0$ (since the matrix H depends on θ). It turns out that, for all practical purposes, by treating H as known, the Wald-type test statistics written with $h(.)$ or $g(.)$ are numerically equal; see the proof of Theorem 2.7 in the appendix.

Theorem 2.7 *(Wald test)*
Under Assumptions 2.1–2.5, the Wald test statistic ξ_T, for testing $H_0 : g(\theta) = 0$ with g twice continuously differentiable,

$$\xi_T = Tg'(\hat{\theta}_T)\left\{\frac{\partial g(\hat{\theta}_T)}{\partial\theta'}\left[\frac{\partial \bar{\Phi}_T'(\hat{\theta}_T)}{\partial\theta} S_T^{-1} \frac{\partial \bar{\Phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial\theta}\right\}^{-1} g(\hat{\theta}_T)$$

is asymptotically distributed as a chi-square with q degrees of freedom under the null.

In our framework, the standard result holds with respect to the size of the Wald test. Of course, the power of the test heavily depends on the strength of identification of the various constraints to test as extensively discussed in Antoine and Renault (2010a). See also the discussion in Subsection 2.4.3.

2.4 Comparisons with Other Approaches

2.4.1 Linear IV Model

Following the discussion in Examples 2.1 and 2.1, several matrices Π_T may be considered in the linear model with poor instruments. We now show that this choice is not innocuous.

(i) Staiger and Stock (1997) consider a framework with the same genuine weak identification pattern for all the parameters: $\Pi_T = C/\sqrt{T}$. To maintain Assumption L2, we can consider it as the limit case of: $\Pi_T = C/T^{\lambda}$, for $0 < \lambda < 1/2$ and C full column rank. Then $A_T = T^{\lambda} Id_K$ fulfills Assumption L1. Similarly, $\tilde{A}_T = T^{\lambda} Id_p$ fulfills

Assumption L3. Note that in this simple example, $\|A_T\|$ and $\|\tilde{A}_T\|$ grow at the same rate, which corresponds to the unique degree of nearly weak identification.

(ii) Stock and Wright (2000) reinterpret the above framework to accommodate simultaneously strong and weak identification patterns. This distinction is done at the parameter level and the structural parameter θ is (a priori) partitioned: $\theta = [\theta_1' \vdots \theta_2']'$ with θ_1 of dimension p_1 strongly identified and θ_2 of dimension $p_2 = p - p_1$ weakly identified. Following their approach, while maintaining Assumption L2, we consider the matrix

$$\Pi_T = \begin{bmatrix} \pi_{11} & \pi_{12}/T^{\lambda} \\ \pi_{21} & \pi_{22}/T^{\lambda} \end{bmatrix} = \Pi D_T^{-1},$$

with $0 < \lambda < 1/2$ while $\lambda = 1/2$ in Stock and Wright (2000); $\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$ and D_T a (p, p)-diagonal matrix (with 1 as the first p_1 coefficients and T^{λ} as the remaining ones). $\tilde{A}_T = D_T$ directly fulfills Assumption L3. Note that the degree of identification of each parameter has to be known (assumed) a priori in Stock and Wright's (2000) specification.

(iii) Antoine and Renault (2009) choose to distinguish between strong and nearly weak identification at the instrument level (see in particular their Subsection 2.3.2). They suppose that the set of K instruments can be partitioned between K_1 strong ones and $(K - K_1)$ nearly weak ones, so that

$$\Pi_T = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21}/T^{\lambda} & \pi_{22}/T^{\lambda} \end{bmatrix} = \Lambda_T^{-1}\Pi,$$

with Λ_T a (K, K)-diagonal matrix (with 1 as the first K_1 coefficients and T^{λ} as the K_2 remaining ones). The limit case with $\lambda = 1/2$ is the framework of Hahn, Ham, and Moon (2009).

Interestingly enough, the above approaches (ii) and (iii) lead to the same concentration matrix, a well-known measure of the strength of the instruments. As a consequence, one concludes that both approaches capture similar patterns of weak identification. In Examples 2.1 and 2.1, the concentration matrix and its determinant are respectively equal to

$$\mu = \Sigma_V^{-1/2\prime}\Pi_T' Z' Z \Pi_T \Sigma_V^{1/2}$$

and

$$\det(\mu) = \frac{1}{T^{2\lambda}} \det(Z'Z) \det(\Sigma_V^{-1}) \det(\Pi)^2$$

with $\Sigma_V \equiv \mathtt{Var}[V]$. With standard weak asymptotics ($T^{\lambda} = \sqrt{T}$), the concentration matrix has a finite limit (see also Andrews and

Stock 2007). Nearly weak asymptotics allow an infinite limit for the determinant of the concentration matrix, but at a rate smaller than $\det[Z'Z] = \mathcal{O}(T)$. In this respect, there is no difference between the two approaches, only the rate of convergence to zero of respectively a row or a column of the matrix Π_T matters.

(iv) Phillips (1989) introduces partial identification where Π_T matrices that may not be of full rank are considered. Generalization to asymptotic rank condition failures (at rate T^{λ}) comes at the price of having to specify which row (or column) asymptotically goes to zero. At least, Antoine and Renault's (2009) approach (iii) works with "estimable functions" of the structural parameters, or functions that can be identified and square-root T consistently estimated. By contrast, the approach (ii) implies directly a partition of the structural parameters between strongly and weakly identified ones.

(v) Antoine and Renault (2010a) generalize the above approach (iii) to accommodate matrices of reduced form like $\Pi_T = \Lambda_T^{-1}\Pi$ with Λ_T a (K, K)-diagonal matrix such that $\|\Lambda_T\| = o(\sqrt{T})$. Then $A_T = \Lambda_T E_{zz}^{-1}$ fulfills Assumption L1. By contrast with the former examples, the case where instruments may not be mutually orthogonal and may display different levels of strength leads to a nondiagonal matrix A_T. However, in this case, it is easy to imagine a standardization of instruments such that A_T eventually becomes diagonal (i.e., $A_T = \Lambda_T$). Then, a sequence of matrices $\tilde{A}_T$ fulfilling Assumption L3 can be built according to the general result provided in Theorem 2.3. The detailed construction provided in the appendix shows that we can actually choose $\tilde{A}_T = R\tilde{\Lambda}_T$ with R nonsingular (p, p)-matrix whose columns provide a basis for the orthogonal of the null space of Π while $\tilde{\Lambda}_T$ is a diagonal (p, p)-matrix such that $\|\tilde{\Lambda}_T\| \leq \|\Lambda_T\|$. In other words, all parameters are estimated with a rate of convergence at least equal to $\sqrt{T}/\|\tilde{\Lambda}_T\|$ irrespective of the slowest rate $\sqrt{T}/\|\Lambda_T\|$. The key is that some instruments (among the weakest) may be irrelevant, depending on the range of Π'. This analysis actually provides primitive conditions for the high-level Assumption 2 in Hansen, Hausman, and Newey (2008) where they assume that $\Upsilon = \Pi_T' z_t$ (where z_t denotes the tth observation of the K instruments) can be rewriten as $\Upsilon = S_T \tilde{z}_T$ for some p-dimensional vector $\tilde{z}_T$. This transformation exactly corresponds to our transformation of A_T into $\tilde{A}_T$ which is made explicit in the above detailed discussion. As also done in Antoine and Renault (2009, 2010a), Hansen, Hausman, and Newey (2008) take advantage of the matrix S_T to characterize how some linear combinations of the parameters may be identified at different rates.

2.4.2 Continuously Updated GMM

We now show that the nearly strong identification Assumption 2.5 is exactly needed to ensure that any direction in the parameter space is equivalently

estimated by efficient two-step GMM and continuously updated GMM. This will also explain the equivalence between GMM score test and Kleibergen's modified score test discussed in the next section. Hansen, Heaton, and Yaron (1996) define the continuously updated GMM estimator $\hat{\theta}_T^{CU}$ as follow:

Definition 2.1 *Let $S_T(\theta)$ be a family of nonsingular random matrices such that*[6]

(i) $S_T(\theta^0)$ *is a (unfeasible) consistent estimator of* $S \equiv \lim_T[\mathtt{Var}(\sqrt{T}\bar{\phi}_T(\theta^0))]$.

(ii) $\|S_T^{-1}(\theta^0)\| = \mathcal{O}_P(1)$.

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

(iv) $\sup_{\|\theta-\theta^0\|<\delta_T} \|S_T^{-1}(\theta) - S^{-1}\| = o_p(1)$ *for some real sequence* δ_T.

The continuously updated GMM estimator $\hat{\theta}_T^{CU}$ of θ^0 is then defined as

$$\hat{\theta}_T^{CU} = \arg\min_{\theta\in\Theta}\left[\bar{\phi}_T'(\theta)S_T^{-1}(\theta)\bar{\phi}_T(\theta)\right]. \tag{2.11}$$

PROPOSITION 2.1 *(Equivalence between CU-GMM and efficient 2S-GMM) Under Assumptions 2.1–2.5, any direction in the parameter space is equivalently estimated by efficient two-step GMM and continuously updated GMM. That is,*

$$\sqrt{T}\,\tilde{A}_T^{-1}\big(\hat{\theta}_T^{CU} - \hat{\theta}_T\big) = o_p(1).$$

In the special case where the same degree of global identification weakness λ_T is assumed for all coefficients of $\tilde{A}_T$, CU-GMM and efficient 2S-GMM are equivalent without the homogenous identification Assumption 2.5 (insofar as $\lambda_T = o(\sqrt{T})$).

Several comments are in order.

First, since nondegenerate asymptotic normality is obtained for $\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \theta^0)$ (and not for $\sqrt{T}(\hat{\theta}_T - \theta^0)$), the relevant (nontrivial) equivalence result between two-step efficient GMM and continuously updated GMM relates to the suitably *rescaled* difference $\sqrt{T}\,\tilde{A}_T^{-1}(\hat{\theta}_T - \hat{\theta}_T^{CU})$.

Second, the case with nearly weak (and not homogenous) identification ($\|\tilde{\Lambda}_T\|^2/\sqrt{T} = o(1)$) breaks down the standard theory of efficient GMM: the proof shows that there is no reason to believe that continuously updated GMM may be an answer. Two-step GMM and continuously updated GMM, albeit no longer equivalent, are both perturbed by higher-order terms with ambiguous effects on asymptotic distributions. The intuition given by higher-order asymptotics in standard identification settings cannot be extended to the case of nearly weak identification. While the latter approach shows that continuously updated GMM is, in general, higher-order efficient (see Newey

[6] The following regularity assumptions are standard when defining the continuously updated GMM estimator. See Pakes and Pollard (1989, pp. 1044–1046).

and Smith 2004; Antoine, Bonnal, and Renault 2007), there is no clear ranking of asymptotic performances under weak identification.

Third, it is important to keep in mind that all these difficulties are due to the fact that we consider realistic circumstances where several degrees of global identification weakness are simultaneously involved. Standard results (equivalence, or rankings between different approaches) carry on when the same rate λ_T is assumed for all coefficients of $\bar{A}_T$.

2.4.3 GMM Score-Type Testing

As already explained, when the same degree of global identification weakness λ_T is assumed for all coefficients of the matrix Λ_T, standard procedures and results hold. One of the contribution of this paper is to characterize the heterogeneity of the informational content of moment conditions along different directions in the parameter space. We now illustrate how the power of tests is affected. More precisely, we are interested in testing the null hypothesis: $\underline{H}_0 : \theta = \theta_0$. To simplify the exposition, we focus here on a diagonal matrix A_T:

$$A_T = \begin{bmatrix} Id_{K_1} & O \\ O & \lambda_T Id_{K-K_1} \end{bmatrix} \quad \text{with} \quad \lambda_T \to \infty \quad \text{and} \quad \lambda_T = o(\sqrt{T}).$$

Assumption 2.3 is modified accordingly:

(simplified) Assumptions 2.3

$$\begin{bmatrix} Id_{K_1} & O \\ O & \lambda_T Id_{K-K_1} \end{bmatrix} \frac{\partial \bar{\phi}_T(\theta_0)}{\partial \theta'} = \begin{bmatrix} Id_{K_1} & O \\ O & \lambda_T Id_{K-K_1} \end{bmatrix} \begin{pmatrix} \frac{\partial \bar{\phi}_{1T}(\theta^0)}{\partial \theta'} \\ \frac{\partial \bar{\phi}_{2T}(\theta^0)}{\partial \theta'} \end{pmatrix}$$

$$\to \frac{\partial \bar{d}(\theta^0)}{\partial \theta'} \equiv \begin{pmatrix} \frac{\partial \bar{d}_1(\theta^0)}{\partial \theta'} \\ \frac{\partial \bar{d}_2(\theta^0)}{\partial \theta'} \end{pmatrix}$$

with the (K, p)*-matrix* $[\partial \bar{d}(\theta^0)/\partial \theta']$ *full column rank.*

The following (simple) example illustrates our focus of interest.

Example 2.3
Consider the functions ϕ_{1t} *and* ϕ_{2t} *defined as*

$$\phi_{1t}(\theta) = Y_{1t} - g(\theta) \quad \textit{and} \quad \phi_{2t}(\theta) = -Z_t \otimes (Y_{2t} - X_{2t}\theta),$$

and associated moment conditions

$$E[Y_{1t}] = g(\theta^0) \quad \textit{and} \quad E\left[Z_t \otimes (Y_{2t} - X_{2t}\theta^0)\right] = 0.$$

The instruments Z_t introduced in ϕ_{2t} are only nearly weak instruments since

$$E[Z_t \otimes X_{2t}] = \frac{1}{\lambda_T}\frac{\partial \tilde{d}_2(\theta^0)}{\partial \theta'} \quad with \quad \lambda_T \xrightarrow{T} \infty, \quad and \quad \frac{\lambda_T}{\sqrt{T}} \xrightarrow{T} 0.$$

Then the associated Jacobian matrices are

$$\begin{aligned}
\mathtt{Plim}\left[\frac{\partial \bar{\phi}_{1T}(\theta^0)}{\partial \theta'}\right] &= \frac{\partial g(\theta^0)}{\partial \theta'} = \frac{\partial \tilde{d}_1(\theta^0)}{\partial \theta'} \\
\mathtt{Plim}\left[\lambda_T \frac{\partial \bar{\phi}_{2T}(\theta^0)}{\partial \theta'}\right] &= \mathtt{Plim}\left[\lambda_T \frac{1}{T}\sum_{t=1}^{T}(Z_t \otimes X_{2t})\right] \\
&= \lim_T [\lambda_T E\,(Z_t \otimes X_{2t})] = \frac{\partial \tilde{d}_2(\theta^0)}{\partial \theta'},
\end{aligned}$$

and we assume that $\left[\frac{\partial \tilde{d}_1'(\theta^0)}{\partial \theta} \vdots \frac{\partial \tilde{d}_2'(\theta^0)}{\partial \theta}\right]'$ has full column rank.

The GMM score-type testing approach wonders whether the test value θ_0 is close to fulfill the first-order conditions of the (efficient) two-step GMM minimization, that is whether the score vector is close to zero. The score vector is defined at the test value θ_0 as

$$V_T(\theta_0) = \frac{\partial \bar{\phi}_T'(\theta_0)}{\partial \theta} S_T^{-1}(\theta_0)\bar{\phi}_T(\theta_0).$$

The GMM score test statistic (Newey and West 1987) is then a suitable norm of $V_T(\theta_0)$:

$$\xi_T^{NW} = T V_T'(\theta_0)\left[\frac{\partial \bar{\phi}_T'(\theta_0)}{\partial \theta} S_T^{-1}(\theta_0)\frac{\partial \bar{\phi}_T(\theta_0)}{\partial \theta'}\right]^{-1} V_T(\theta_0).$$

Kleibergen's (2005) approach rather considers the first-order conditions of the CU-GMM minimization. The corresponding score vector is defined at the test value θ_0 as

$$V_T^{CU}(\theta_0) = \frac{\partial \phi_T^{CU\prime}(\theta_0)}{\partial \theta} S_T^{-1}(\theta_0)\bar{\phi}_T(\theta_0),$$

where each row of $\left[\frac{\partial \phi_T^{CU}(\theta_0)}{\partial \theta'}\right]$ is the residual of the long-term affine regression of $\left[\frac{\partial \bar{\phi}_T(\theta_0)}{\partial \theta'}\right]_{[i.]}$ on $\bar{\phi}_T(\theta_0)$:

$$\begin{aligned}
\left[\frac{\partial \phi_T^{CU}(\theta_0)}{\partial \theta'}\right]_{[i.]}' &= \left[\frac{\partial \bar{\phi}_T(\theta_0)}{\partial \theta'}\right]_{[i.]}' \\
&\quad -\mathrm{Cov}_{as}\left(\sqrt{T}\left[\frac{\partial \bar{\phi}_T(\theta_0)}{\partial \theta'}\right]_{[i.]}', \sqrt{T}\bar{\phi}_T(\theta_0)\right)\mathrm{Var}_{as}\left(\sqrt{T}\bar{\phi}_T(\theta_0)\right)^{-1}\bar{\phi}_T(\theta_0)
\end{aligned} \tag{2.12}$$

where $\text{Var}_{as}(\sqrt{T}\bar{\phi}_T(\theta_0)) = S^0$ is the long-term covariance matrix of the moment conditions $\phi_t(\theta_0)$ and $\text{Cov}_{as}\left(\sqrt{T}\left[\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta}\right]'_{[i.]}, \sqrt{T}\bar{\phi}_T(\theta_0)\right)$ is the long-term covariance between $\left[\frac{\partial\phi_t(\theta_0)}{\partial\theta}\right]_{[i.]}$ and $\phi_t(\theta_0)$ (which is assumed well-defined).

This characterization of the score of continuously updated GMM in terms of residual of an affine regression is extensively discussed in Antoine, Bonnal, and Renault (2007) through their Euclidean empirical likelihood approach. It explains the better finite sample performance of CU-GMM since the regression allows to remove the perverse correlation between the Jacobian matrix and the moment conditions. In finite sample, this perverse correlation implies that the first order conditions of standard (two-step) efficient GMM are biased. As clearly explained by Kleibergen (2005), this perverse correlation is even more detrimental with genuinely weak instruments since it does not even vanish asymptotically. This is the reason why Kleibergen (2005) puts forward a modified version of the Newey-West (1987) score test statistic:

$$\xi_T^K = T V_T^{CU\prime}(\theta_0)\left[\frac{\partial\phi_T^{CU\prime}(\theta_0)}{\partial\theta} S_T^{-1}(\theta_0)\frac{\partial\phi_T^{CU}(\theta_0)}{\partial\theta'}\right]^{-1} V_T^{CU}(\theta_0).$$

In contrast with Kleibergen (2005), we show that with nearly weak instruments, the aforementioned correlation does not matter asymptotically and that the standard GMM score test statistic ξ_T^{NW} works. It is actually asymptotically equivalent to the modified Kleibergen's score test statistic under the null:

PROPOSITION 2.2 *(Equivalence under the null)*
Under the null $\underline{H}_0 : \theta = \theta_0$, *we have:* $\texttt{Plim}\left[\xi_T^{NW} - \xi_T^K\right] = 0$. *Both* ξ_T^{NW} *and* ξ_T^K *converge in distribution toward a chi-square with* p *degrees of freedom.*

The following example illustrates how a proper characterization of the heterogeneity of the informational content of moment conditions matters when considering power of tests under sequences of local alternatives.

Example 2.3 (continued)
Consider a sequence of local alternatives defined by a given deterministic sequence $(\gamma_T)_{T\geq 0}$ *in* $\mathbb{R}^p$, *going to zero when* T *goes to infinity, and such that the true unknown value* θ_0 *is defined as:* $\theta_T = \theta_0 + \gamma_T$. *For* T *large enough,* $g(\theta_T)$ *can be seen as* $g(\theta_0) + [\partial g(\theta_0)/\partial\theta']\gamma_T$. *Therefore, the strongly identified moment restrictions* $E[Y_{1t} - g(\theta_T)] = 0$ *are informative with respect to the violation of the null* $(\theta_T \neq \theta_0)$ *if and only if:* $[\partial g(\theta_0)/\partial\theta']\gamma_T \neq 0$.

As a consequence, we expect GMM-based tests of $\underline{H}_0 : \theta = \theta_0$ *to have power against sequences of local alternatives converging at standard rate* $\sqrt{T}$, $\theta_T = \theta_0 + \gamma/\sqrt{T}$, *if and only if* $[\partial g(\theta_0)/\partial\theta']\gamma \neq 0$, *or, when* γ *does not belong to the null space of* $[\partial g(\theta_0)/\partial\theta'] = [\partial\bar{d}_1(\theta_0)/\partial\theta']$. *By contrast, if* $[\partial\bar{d}_1(\theta_0)/\partial\theta']\gamma = 0$, *violations of the null can only be built from the other identifying conditions:*

$$E[Z_t \otimes Y_t] = \frac{\partial\bar{d}_2(\theta_0)}{\partial\theta'}\frac{\theta_T}{\lambda_T}.$$

We show that the sequences of local alternatives relevant to characterize nontrivial power are necessarily such that $\theta_T = \theta_0 + \lambda_T \frac{\gamma}{\sqrt{T}}$.

In other words, the degree of weakness of the moment conditions λ_T *downplays the standard rate* $[\gamma/\sqrt{T}]$ *of sequences of local alternatives against which the tests have nontrivial local power. Under such a sequence of local alternatives,*

$$E\left[Z_t \otimes Y_t\right] = \frac{\partial \tilde{d}_2(\theta_0)}{\partial \theta'}\left[\frac{\theta_0}{\lambda_T} + \frac{\gamma}{\sqrt{T}}\right]$$

differs from its value under the null by the standard scale $1/\sqrt{T}$.

PROPOSITION 2.3 *(Local power of GMM score tests)*
(i) With a (drifted) true unknown value, $\theta_T = \theta_0 + \gamma/\sqrt{T}$, *for some* $\gamma \in \mathbb{R}^p$, *we have* $\texttt{Plim}[\xi_T^{NW} - \xi_T^K] = 0$, *and both* ξ_T^{NW} *and* ξ_T^K *converge in distribution toward a noncentral chi-square with* p *degrees of freedom and noncentrality parameter*

$$\mu = \left(\gamma' \frac{\partial \tilde{d}_1'(\theta_0)}{\partial \theta} \vdots 0\right)[S(\theta^0)]^{-1}\begin{pmatrix} \frac{\partial \tilde{d}_1(\theta_0)}{\partial \theta'}\gamma \\ 0 \end{pmatrix}.$$

(ii) In case of nearly strong identification $(\lambda_T^2 = o(\sqrt{T}))$, *with a (drifted) true unknown value* $\theta_T = \theta_0 + \gamma/\lambda_T$, *for some* $\gamma \in \mathbb{R}^p$ *such that* $\frac{\partial \tilde{d}_1(\theta_0)}{\partial \theta'}\gamma = 0$, *we have* $\texttt{Plim}[\xi_T^{NW} - \xi_T^K] = 0$, *and both* ξ_T^{NW} *and* ξ_T^K *converge in distribution toward a noncentral chi-square with* p *degrees of freedom and noncentrality parameter*

$$\mu = \left(0 \vdots \gamma' \frac{\partial \tilde{d}_2'(\theta_0)}{\partial \theta}\right)[S(\theta^0)]^{-1}\begin{pmatrix} 0 \\ \frac{\partial \tilde{d}_2(\theta_0)}{\partial \theta'}\gamma \end{pmatrix}.$$

Two additional conclusions follow from Proposition 2.3:

(i) First, if $\frac{\partial \tilde{d}_1(\theta_0)}{\partial \theta'}\gamma \neq 0$, the two GMM score tests behave more or less as usual against sequences of local alternatives in the direction γ. They are asymptotically equivalent and both consistent against sequences converging slower than $\sqrt{T}$. They both follow asymptotically a noncentral chi-square against sequences with the usual rate $\sqrt{T}$.

(ii) Second, if $\frac{\partial \tilde{d}_1(\theta_0)}{\partial \theta'}\gamma = 0$, the two GMM score tests have no power against sequences of local alternatives $\theta_T = \theta_0 + \gamma/\sqrt{T}$. They may have power against sequences $\theta_T = \theta_0 + \gamma\lambda_T/\sqrt{T}$ (or slower); their behavior is pretty much the standard one, but only in the homogenous identification case where $\lambda_T^2 = o(\sqrt{T})$.

We now explain why nonstandard asymptotic behavior of both score tests may arise when we consider sequences of local alternatives in the weak directions ($\theta_T = \theta_0 + \gamma\lambda_T/\sqrt{T}$ with $\frac{\partial \tilde{d}_1(\theta_0)}{\partial \theta'}\gamma = 0$) with severe nearly weak identification issues. Recall that the genuine weak identification usually considered in the literature ($\lambda_T = \sqrt{T}$) is a limit case, since we always maintain the nearly

weak identification condition $\lambda_T = o(\sqrt{T})$. Under such a sequence of local alternatives, while $\sqrt{T}\bar{\phi}_T(\theta_T)$ is asymptotically normal with zero mean, the key to get a standard noncentral chi-square for the asymptotic distribution of a score test statistic is to ensure that $\sqrt{T}\bar{\phi}_T(\theta_0)$ is asymptotically normal with nonzero mean if and only if γ is not zero. This result should follow from the Taylor approximation around θ_0 with θ_T^* between θ_0 and θ_T:

$$\sqrt{T}\bar{\phi}_T(\theta_T) \approx \sqrt{T}\bar{\phi}_T(\theta_0) + \sqrt{T}\frac{\partial\bar{\phi}_T(\theta_T^*)}{\partial\theta'}(\theta_T - \theta_0) \approx \sqrt{T}\bar{\phi}_T(\theta_0) + \begin{pmatrix} 0 \\ \frac{\partial\tilde{d}_2(\theta_0)}{\partial\theta'}\gamma \end{pmatrix}.$$

This approximation is justified by (simplified) Assumption 2.3 as long as

$$\frac{\partial\tilde{d}_1(\theta_0)}{\partial\theta'}\gamma = 0 \Rightarrow \texttt{Plim}\left[\lambda_T\frac{\partial\bar{\phi}_{1T}(\theta_T^*)}{\partial\theta'}\gamma\right] = 0. \tag{2.13}$$

This is not an issue if, as in Kleibergen (2005), the same degree of global identification weakness[7] is assumed for all coefficients of the matrix A_T. In other words, we can easily state that in the interesting case with mixture of strong and nearly weak identification (or nonempty subsets of components ϕ_1 and ϕ_2), Equation 2.13 should follow from

$$\frac{\partial\tilde{d}_1(\theta_0)}{\partial\theta'}\gamma = 0 \Rightarrow \texttt{Plim}\left[\lambda_T\frac{\partial\bar{\phi}_{1T}(\theta_T)}{\partial\theta'}\gamma\right] = 0. \tag{2.14}$$

Fragile identification may be wasted by Kleibergen's modification precisely because it comes with another piece of information which is stronger. To see this, the key is the aforementioned lack of logical implication from Equation 2.14 to Equation 2.13. As a result, the modified score statistic and the original one may have quite different asymptotic behaviors. It is quite evident from Equation 2.12 that, when $\sqrt{T}\bar{\phi}_T(\theta_0)$ is not $\mathcal{O}_P(1)$, the modified score statistic may have an arbitrarily nasty asymptotic behavior. However, it is worth noting that if we maintain Assumption 2.1 of a functional central limit theorem,

$$\sqrt{T}\bar{\phi}_T(\theta_0) - \sqrt{T}\rho_T(\theta_0) = \mathcal{O}_P(1),$$

a sufficient condition to ensure that Kleibergen's modified test statistic is well-behaved under the sequence of local alternatives $\theta_T = \theta_0 + \gamma\frac{\lambda_T}{\sqrt{T}}$ is

$$\sqrt{T}\rho_T\left(\theta_T - \gamma\frac{\lambda_T}{\sqrt{T}}\right) = \mathcal{O}(1).$$

The proof of Theorem 2.2 in the appendix allows us to think that this condition is plausible, since it precisely states that the rate of convergence of any GMM

[7] Smith (2007) already pointed out that the standard equivalence between tests holds when only one rate of convergence is considered.

estimator $\hat{\theta}_T$ ($\|\hat{\theta}_T - \theta_T\| = \mathcal{O}_P(\lambda_T/\sqrt{T})$) precisely comes from the fact (see Lemma 2.1 in the appendix) that:

$$\sqrt{T}\rho_T(\hat{\theta}_T) = \mathcal{O}_P(1).$$

To put it differently, Kleibergen's modified score test is well-behaved under a given sequence of local alternatives insofar as this sequence behaves as well as any GMM estimator. Such a result is not surprising. The novel feature introduced by nearly weak instruments asymptotics is that the rate of sequences of local alternatives must be assessed not only in the parameter space ($\|\theta_T - \theta_0\| = \mathcal{O}(\lambda_T/\sqrt{T})$) but also in the moments space ($\sqrt{T}\rho_T(\theta_0) = \mathcal{O}(1)$).

2.5 Conclusion

To conclude, we have proposed a general framework where weaker patterns of identification may arise without giving up the efficiency goal of statistical inference. We actually believe that even fragile information should be processed optimally for the purpose of both efficient estimation and powerful testing.

Our main contribution has been to consider that several patterns of identification may arise simultaneously. This heterogeneity of identification schemes paved the way for the device of optimal strategies for inferential use of information of poor quality. More precisely, we focus on a case where asymptotic efficiency of estimators is well-defined through the variance of asymptotically normal distributions. The price to pay for this maintained tool was to assume that the set of moment conditions that are not genuinely weak was sufficient to identify the true unknown value of the parameters. In this case, normality was characterized at heterogeneous rates smaller than the standard root-T in different directions of the parameter space. Finally, we were able to show that in such a case standard efficient estimation procedures still hold and are even feasible without requiring the prior knowledge of the identification schemes.

As emphasized in the survey of Andrews and Stock (2007), there are three main topics related to inference with weak identification: hypothesis tests and confidence intervals that are robust to weak instruments; point estimation; and pretesting for weak instruments. Andrews and Stock (2007) have focused on the first topic *"for which a solution is closer at hand than it is for estimation."* Our paper focuses on point estimation as well as power. This can only be done because we consider that the worst case scenario of genuine weak identification is not always warranted. As far as testing for strong/weak identification is concerned, the framework put forward in the present chapter allows us in a companion paper (Antoine and Renault 2010b) to add to the available literature that includes Hahn and Hausman (2002) and Hahn, Ham, and Moon (2009) among others.

Appendix

Notations

- For any vector v with element $(v_i)_{1\leq i\leq H}$, we define: $\|v\|^2 = \sum_{i=1}^{H} v_i^2$.
- For any matrix M with elements m_{ij} that is not a vector, we define: $\|M\| = \max_{i,j} |m_{ij}|$.
- Id_l denotes the identity matrix of size l.
- $[M]_{k.}$ denotes the kth row of the matrix M.
- $\texttt{col}[M]$ denotes the subspace generated by the columns of the matrix M.
- $\texttt{col}[M]^{\perp}$ denotes the subspace orthogonal to the one generated by $\texttt{col}[M]$.

We start with a preliminary result useful for the proofs of consistency and rates of convergence.

Lemma 2.1 *(i) Under Assumptions 2.1 and 2.2,*

$$\|\tilde{\rho}_T(\tilde{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right) \quad with \quad \tilde{\rho}_T(\theta) = [Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\rho_T(\theta)$$

where $\tilde{\theta}_T$ is the GMM-estimator deduced from the partial set of moment conditions as follows:

$$\tilde{\theta}_T = \arg\min_{\theta\in\Theta} \tilde{Q}_T(\theta) = \arg\min_{\theta\in\Theta} \left[\tilde{\phi}_T'(\theta)\tilde{\Omega}_T\tilde{\phi}_T(\theta)\right] \quad with$$

$$\tilde{\phi}_T(\theta) = [Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\bar{\phi}_T(\theta)$$

where $\tilde{\Omega}_T$ is a sequence of symmetric positive definite random matrices of size $\tilde{K}$ converging toward a positive definite matrix $\tilde{\Omega}$.
(ii) Under Assumptions 2.1, 2.2, and 2.3(v),

$$\|\rho_T(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right),$$

where $\hat{\theta}_T$ is the GMM-estimator defined in (2.8).

Proof of Lemma 2.1 First, we prove (ii); second, we show how (i) directly follows.

From Assumption 2.1(i), the objective function is written as follows:

$$TQ_T(\theta) \equiv T\phi_T'(\theta)\Omega_T\phi_T(\theta) = \left[\Psi_T(\theta) + \sqrt{T}\rho_T(\theta)\right]'\Omega_T\left[\Psi_T(\theta) + \sqrt{T}\rho_T(\theta)\right],$$

where the empirical process $(\Psi_T(\theta))_{\theta\in\Theta}$, is asymptotically Gaussian. Since $\hat{\theta}_T$ is the minimizer of Q_T, we have

$$\begin{aligned} Q_T(\hat{\theta}_T) \leq Q_T(\theta^0) \Leftrightarrow\ & T\rho_T'(\hat{\theta}_T)\Omega_T\rho_T(\hat{\theta}_T) + 2\sqrt{T}\rho_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) \\ & + \Psi_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) \\ & \leq T\rho_T'(\theta^0)\Omega_T\rho_T(\theta^0) + 2\sqrt{T}\rho_T'(\theta^0)\Omega_T\Psi_T(\theta^0) \\ & + \Psi_T'(\theta^0)\Omega_T\Psi_T(\theta^0). \end{aligned} \tag{2.15}$$

Following the notations introduced in Assumption 2.2, we define: $N_T\rho_T(\theta) = [\tilde{\rho}_T(\theta)'\ \check{\rho}_T(\theta)']'$.

From Assumption 2.2(iv), we have: $\tilde{\rho}_T(\theta^0) = 0$ for any T. From Assumptions 2.2(ii) and (iii), we have: $\|\check{\Lambda}_T\check{\rho}_T(\theta)\| = \mathcal{O}_P(1)$. Following Assumption 2.3(v), we distinguish two cases[8]:
(a) the additional restrictions are well-specified, $\check{\rho}_T(\theta^0) = 0$, and we have

$$(2.15) \Rightarrow T\rho_T'(\hat{\theta}_T)\Omega_T\rho_T(\hat{\theta}_T) + 2\sqrt{T}\rho_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) + h_T \leq 0, \tag{2.16}$$

with $h_T = \Psi_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) - \Psi_T'(\theta^0)\Omega_T\Psi_T(\theta^0)$.
(b) the additional restrictions are not well-specified, but genuinely weak, $\check{\Lambda}_T = \sqrt{T}Id_{K-\tilde{K}}$ which implies $\|\sqrt{T}\check{\rho}_T(\theta)\| = \mathcal{O}_P(1)$, and we have

$$(2.15) \Rightarrow T\rho_T'(\hat{\theta}_T)\Omega_T\rho_T(\hat{\theta}_T) + 2\sqrt{T}\rho_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) + h_T + \epsilon_T \leq 0, \tag{2.17}$$

with $\epsilon_T = \mathcal{O}_P(1)$.

Hence, we can always write:

$$T\rho_T'(\hat{\theta}_T)\Omega_T\rho_T(\hat{\theta}_T) + 2\sqrt{T}\rho_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) + h_T + \nu_T \leq 0, \tag{2.18}$$

with h_T defined above and $\nu_T = \mathcal{O}_P(1)$: actually, $\nu_T = 0$ in case (a) and $\nu_T = \epsilon_T$ in case (b).

Then, after defining μ_T as the smallest eigenvalue of Ω_T, it follows that

$$T\mu_T\|\rho_T(\hat{\theta}_T)\|^2 - 2\sqrt{T}\|\rho_T(\hat{\theta}_T)\| \times \|\Omega_T\Psi_T(\hat{\theta}_T)\| + h_T + \nu_T \leq 0.$$

In other words, $x_T \equiv \|\sqrt{T}\rho_T(\hat{\theta}_T)\|$ solves the inequality:

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

Therefore, we must have $\Delta_T \geq 0$

$$\text{with} \quad \Delta_T = \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|^2}{\mu_T^2} - \frac{h_T + \nu_T}{\mu_T},$$

[8] Note that a combination of these two cases also works similarly: by combination, we have in mind that some components of $\check{\rho}_T(\theta^0)$ are well-specified whereas some others are not well-specified but genuinely weak.

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

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

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \quad \text{and} \quad \frac{\|\Omega_T \Psi_T(\theta^0)\|}{\mu_T} = \mathcal{O}_P(1).$$

Denote by $\det(M)$ the determinant of any square matrix M. Since $\det(\Omega_T) \xrightarrow{P} \det(\Omega) > 0$, no subsequence of μ_T can converge in probability toward 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

$$\|\Omega_T \Psi_T(\hat{\theta}_T)\| = \mathcal{O}_P(1) \quad \text{and} \quad \|\Omega_T \Psi_T(\theta^0)\| = \mathcal{O}_P(1).$$

Denote by $\mathtt{tr}(M)$ the trace of any square matrix M. Since $\mathtt{tr}(\Omega_T) \xrightarrow{P} \mathtt{tr}(\Omega)$ and the sequence $\mathtt{tr}(\Omega_T)$ is upper bounded in probability, so are all the eigenvalues of Ω_T. Therefore the required boundedness in probability follows from the functional CLT in Assumption 2.1(i) which ensures

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

This completes the proof of (ii).
(i) easily follows after realizing that Assumption 2.3(v) is irrelevant since dealing with the additional moment restrictions and that an inequality similar to (2.18) can be obtained as follows:

$$\tilde{Q}_T(\tilde{\theta}_T) \leq \tilde{Q}_T(\theta^0) \Leftrightarrow T\tilde{\rho}_T'(\tilde{\theta}_T)\tilde{\Omega}_T\tilde{\rho}_T(\tilde{\theta}_T) + 2\sqrt{T}\tilde{\rho}_T'(\tilde{\theta}_T)\tilde{\Omega}_T\tilde{\Psi}_T(\tilde{\theta}_T) + \tilde{h}_T \leq 0,$$

with $\tilde{h}_T = \tilde{\Psi}_T'(\tilde{\theta}_T)\tilde{\Omega}_T\tilde{\Psi}_T(\tilde{\theta}_T) - \tilde{\Psi}_T'(\theta^0)\tilde{\Omega}_T\tilde{\Psi}_T(\theta^0)$.

Proof of Theorem 2.1 *(Consistency)*
Consider the GMM-estimators $\hat{\theta}_T$ defined in (2.8) and $\tilde{\theta}_T$ deduced from the partial set of moment conditions as follows:

$$\tilde{\theta}_T = \arg\min_{\theta \in \Theta} \tilde{Q}_T(\theta) = \arg\min_{\theta \in \Theta} \left[\tilde{\phi}_T'(\theta)\tilde{\Omega}_T\tilde{\phi}_T(\theta)\right] \text{ with}$$
$$\tilde{\phi}_T(\theta) = [Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\bar{\phi}_T(\theta),$$

where $\tilde{\Omega}_T$ is a sequence of symmetric positive definite random matrices of size $\tilde{K}$ converging toward a positive definite matrix $\tilde{\Omega}$. The proof of consistency of $\hat{\theta}_T$ is divided into two steps: (1) we show that $\tilde{\theta}_T$ is a consistent estimator of θ^0; (2) we show that $\text{Plim}[\hat{\theta}_T] = \text{Plim}[\tilde{\theta}_T]$.

(1) The weak consistency of $\tilde{\theta}_T$ follows from a contradiction argument. If $\tilde{\theta}_T$ were not consistent, there would exist some positive ϵ such that

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

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

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

From Assumption 2.2(ii), we have

$$\alpha \equiv \inf_{\|\theta-\theta^0\|>\epsilon} \|\tilde{d}(\theta)\| > 0.$$

Note that since c is bounded and the orthogonal matrix M_T is norm-preserving, $[\tilde{\Lambda}_T \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\rho_T(\theta)$ converges to $\tilde{d}(\theta)$ uniformly on Θ by Assumption 2.1. Then, by Assumption 2.2(ii), we have

$$\inf_{\|\theta-\theta^0\|>\epsilon} \|[\tilde{\Lambda}_T \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\rho_T(\theta)\| \geq \frac{\alpha}{2} \quad \text{for all } T \text{ sufficiently large.}$$

That is, for all T sufficiently large, we have

$$\inf_{\|\theta-\theta^0\|>\epsilon} \|\tilde{\Lambda}_T\tilde{\rho}_T(\theta)\| \geq \frac{\alpha}{2} \quad \text{where} \quad \tilde{\rho}_T(\theta) = [Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\rho_T(\theta).$$

Since $\|\tilde{\Lambda}_T\|/\sqrt{T} = o(1)$ by Assumption 2.2(ii) and $\sqrt{T}\tilde{\rho}(\tilde{\theta}_T) = \mathcal{O}_P(1)$ by Lemma 2.1, we get a contradiction when considering a subsequence T_n. We conclude that $\tilde{\theta}_T$ is a consistent estimator of θ^0.

(2) We now show that $\theta^0 = \texttt{Plim}[\hat{\theta}_T]$, by showing that it is true for any subsequence. If we could find a subsequence which does not converge toward θ^0, we could find a sub-subsequence with a limit in probability $\theta^1 \neq \theta^0$ (by assumption Θ is compact). To avoid cumbersome notations with sub-subsequences, it is sufficient to show that: $\texttt{Plim}[\hat{\theta}_T] = \theta^1 \Rightarrow \theta^1 = \theta^0$. Consider the criterion function: $Q_T(\theta) = \bar{\phi}_T'(\theta)\Omega_T\bar{\phi}_T(\theta)$. We show that

(i) $Q_T(\tilde{\theta}_T) = \mathcal{O}_P(1/\|\check{\Lambda}_T\|^2)$

(ii) $\theta^1 \neq \theta^0 \Rightarrow \|\check{\Lambda}_T\|^2 Q_T(\hat{\theta}_T) \xrightarrow{T} \infty$.

This would lead to a contradiction with the definition of GMM estimators: $Q_T(\hat{\theta}_T) \leq Q_T(\tilde{\theta}_T)$.To show (i) and (ii), we assume without loss of generality that the weighting matrices Ω_T, $\tilde{\Omega}_T$, Ω and $\tilde{\Omega}$ are all identity matrices; otherwise, this property would come with a convenient rescaling of the moment conditions.

$$
\begin{aligned}
T Q_T(\theta) &= T\bar{\phi}_T'(\theta)\bar{\phi}_T(\theta) \\
&= \|\sqrt{T}[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\bar{\phi}_T(\theta)\|^2 + \|\sqrt{T}[O_{K-\tilde{K},\tilde{K}} \vdots Id_{K-\tilde{K}}]\bar{\phi}_T(\theta)\|^2 \\
&= \|\sqrt{T}[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\bar{\phi}_T(\theta)\|^2 + \|[O_{K-\tilde{K},\tilde{K}} \vdots Id_{K-\tilde{K}}](\Psi(\theta) - \sqrt{T}\rho_T(\theta))\|^2
\end{aligned}
$$

From Lemma 2.1: $\|\sqrt{T}[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\bar{\phi}_T(\tilde{\theta}_T)\| = \mathcal{O}_P(1)$.

From Assumption 2.2(ii): $[O_{K-\tilde{K},\tilde{K}} \vdots \check{\Lambda}_T]N_T'\rho_T(\theta) \to \check{d}(\theta)$ uniformly.
Thus: $Q_T(\theta) = \mathcal{O}_P(1/(\|\check{\Lambda}_T\|^2))$.

$$
\begin{aligned}
&\|\check{\Lambda}_T\|^2 Q_T(\hat{\theta}_T) \\
&\geq \left\| \frac{\|\check{\Lambda}_T\|}{\sqrt{T}}[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\Psi(\hat{\theta}_T) + \|\check{\Lambda}_T\|[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\rho_T(\hat{\theta}_T) \right\|^2 \\
&\geq \left[\|\check{\Lambda}_T\|\|[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\rho_T(\hat{\theta}_T)\| - \frac{\|\check{\Lambda}_T\|}{\sqrt{T}}\|[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\Psi(\hat{\theta}_T)\| \right]^2
\end{aligned}
$$

From Assumption 2.2(ii): $\frac{\|\check{\Lambda}_T\|}{\sqrt{T}}\|[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\Psi(\hat{\theta}_T)\| = \mathcal{O}_P\left(\|\Psi(\hat{\theta}_T)\|\right) = \mathcal{O}_P(1)$, while

$$
\|\check{\Lambda}_T\|\|[Id_{\tilde{K}} \vdots O_{\tilde{K},K-\tilde{K}}]\rho_T(\hat{\theta}_T)\| \geq \frac{\|\check{\Lambda}_T\|\|[\tilde{\Lambda}_T \vdots O_{\tilde{K},K-\tilde{K}}]\rho_T(\hat{\theta}_T)\|}{\|\tilde{\Lambda}_T\|},
$$

with $\|[\tilde{\Lambda}_T \vdots O_{\tilde{K},K-\tilde{K}}]N_T'\rho_T(\hat{\theta}_T)\| \to \|\check{d}(\theta^1)\| \neq 0$. Thus,

$$
\frac{\|\check{\Lambda}_T\|\|[\tilde{\Lambda}_T \vdots O_{\tilde{K},K-\tilde{K}}]\rho_T(\hat{\theta}_T)\|}{\|\tilde{\Lambda}_T\|} \to +\infty,
$$

and we get the announced result.

Proof of Theorem 2.2 *(Rate of convergence)*
From Lemma 2.1, $\|\rho_T(\hat{\theta}_T)\| = \mathcal{O}_P(1/\sqrt{T})$. We know that N_T is bounded. Hence, we have: $\|N_T'\rho_T(\hat{\theta}_T)\| = \mathcal{O}_P(1/\sqrt{T})$. Recall also that from Assumption 2.2(iii), the first $\tilde{K}$ elements of $N_T\rho_T(\theta^0)$ are identically zero. The mean-value theorem, for some $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0 (component by component), yields to

$$
\left\| [\tilde{\Lambda}_T\ 0]N_T' \frac{\partial \rho_T(\tilde{\theta}_T)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0\right) \right\| = \mathcal{O}_P\left(\frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}}\right).
$$

Note that (by a common abuse of notation) we omit to stress that $\tilde{\theta}_T$ actually depends on the component of ρ_T. Define now z_T as follows:

$$z_T \equiv \frac{\partial \tilde{d}(\theta^0)}{\partial \theta'}(\hat{\theta}_T - \theta^0). \tag{2.19}$$

Since $[\partial \tilde{d}(\theta^0)/\partial \theta']$ is full column rank by Assumption 2.3(iii), we have

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

Hence, we only need to prove that $\|z_T\| = \mathcal{O}_P(\|\tilde{\Lambda}_T\|/\sqrt{T})$. By definition of z_T, we have

$$z_T = \left[\frac{\partial \tilde{d}(\theta^0)}{\partial \theta'} - [\tilde{\Lambda}_T\ 0]N_T'\frac{\partial \rho_T(\tilde{\theta}_T)}{\partial \theta'}\right]\left(\hat{\theta}_T - \theta^0\right) + w_T \ \text{ with}$$

$$\|w_T\| = \mathcal{O}_P\left(\frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}}\right). \tag{2.20}$$

However, since $\tilde{\theta}_T \xrightarrow{P} \theta^0$ and $[\tilde{\Lambda}_T\ 0]N_T'[\partial \rho_T(\theta)/\partial \theta']$ converges uniformly on the interior of Θ toward $[\partial \tilde{d}(\theta)/\partial \theta']$ by Assumption 2.3(iv), we have

$$\begin{aligned}
&\left[\frac{\partial \tilde{d}(\theta^0)}{\partial \theta'} - [\tilde{\Lambda}_T\ 0]N_T'\frac{\partial \rho_T(\tilde{\theta}_T)}{\partial \theta'}\right]\left(\hat{\theta}_T - \theta^0\right),\\
&\quad = \left[\frac{\partial \tilde{d}(\theta^0)}{\partial \theta'} - [\tilde{\Lambda}_T\ 0]N_T'\frac{\partial \rho_T(\tilde{\theta}_T)}{\partial \theta'}\right]\left[\frac{\partial \tilde{d}'(\theta^0)}{\partial \theta}\frac{\partial \tilde{d}(\theta^0)}{\partial \theta'}\right]^{-1}\frac{\partial \tilde{d}'(\theta^0)}{\partial \theta} z_T,\\
&\quad = D_T z_T,
\end{aligned}$$

for some matrix D_T such that $\|D_T\| \xrightarrow{P} 0$. Therefore: $\|z_T\| \leq \epsilon_T \|z_T\| + \|w_T\|$ with $\epsilon_T \to 0$. Hence, $\|z_T\| = \mathcal{O}_P(\|\tilde{\Lambda}_T\|/\sqrt{T})$ and we get: $\|\hat{\theta}_T - \theta^0\| = \mathcal{O}_P(\|\tilde{\Lambda}_T\|/\sqrt{T})$.

Proof of Theorem 2.3
Without loss of generality, we write the diagonal matrix Λ_T as:

$$\Lambda_T = \left(\begin{array}{ccc|ccc} \lambda_{1T} Id_{K_1} & & & & & \\ & \ddots & & & & \\ & & \lambda_{LT} Id_{K_L} & & & \\ \hline & & & \lambda_{L+1,T} Id_{K_{L+1}} & & \\ & & & & \ddots & \\ & & & & & \lambda_{\overline{L},T} Id_{K_{\overline{L}}} \end{array}\right) = \left(\begin{array}{c|c} \tilde{\Lambda}_T & O \\ \hline O & \check{\Lambda}_T \end{array}\right)$$

with $\overline{L} \leq K$, $\sum_{l=1}^{\overline{L}} K_l = K$ and $\lambda_{lT} = o(\lambda_{l+1,T})$. For convenience, we also rewrite the (p, K)-matrix $\left[\frac{\partial c'(\theta^0)}{\partial \theta} M^{-1\prime}\right]$ by stacking horizontally $\overline{L}$ blocks of

size (p, K_l) denoted J_l, $(l = 1, \cdots, \overline{L})$ as follows:

$$\frac{\partial c'(\theta^0)}{\partial \theta} M^{-1\prime} = \left(J_1 \cdots J_L \middle| J_{L+1} \cdots J_{\overline{L}} \right) = \left(\begin{array}{c|c} \frac{\partial \tilde{d}'(\theta^0)}{\partial \theta'} & \frac{\partial \check{d}'(\theta^0)}{\partial \theta'} \end{array} \right),$$

$$\text{with } J_1' \equiv \begin{pmatrix} \left[M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \right]_{[1.]} \\ \vdots \\ \left[M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \right]_{[K_1]} \end{pmatrix} \text{ and } J_l' \equiv \begin{pmatrix} \left[M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \right]_{[K_1+\cdots+K_{l-1}+1.]} \\ \vdots \\ \left[M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \right]_{[K_1+\cdots+K_l.]} \end{pmatrix}$$

$$\text{for } l = 2, \cdots, \overline{L}.$$

Recall also that by Assumption 2.3(iii), the columns of $\frac{\partial \tilde{d}'(\theta^0)}{\partial \theta}$ span the whole space $\mathbb{R}^p$. We now introduce the square matrix of size p, $R = [R_1 \; R_2 \; \cdots \; R_L]$ which spans $\mathbb{R}^p$. The idea is that each (p, s_l)-block R_l defined through $\texttt{col}[J_l]$ collects the directions associated with the specific rate λ_{lT}, $l = 1, \cdots, L$ and $\sum_{l=1}^{L} s_l = p$. Then, the matrix $\tilde{A}_T$ is built as

$$\tilde{A}_T = \left[\lambda_{1T} R_1 \vdots \lambda_{2T} R_2 \vdots \cdots \vdots \lambda_{LT} R_L \right].$$

By convention, $\lambda_{lT} = o(\lambda_{l+1,T})$ for any $1 \leq l \leq L-1$. We now explain how to construct the matrix R. The idea is to separate the parameter space into L subspaces. More specifically:

- R_L is defined such that $J_i' R_L = 0$ for $1 \leq i < L$ and $\texttt{rk}[R_L] = \texttt{rk}[J_L]$. In other words, R_L spans $\texttt{col}[J_1 \; J_2 \; \cdots \; J_{L-1}]^{\perp}$.
- R_{L-1} is defined such that $J_i' R_{L-1} = 0$ for $1 \leq i < L-1$ and $\texttt{rk}[R_{L-1} \; R_L] = \texttt{rk}[J_{L-1} \; J_L]$.
- And so on, until R_2 is defined such that $J_1' R_2 = 0$ and $\texttt{rk}[R_2 \; \cdots \; R_L] = \texttt{rk}[J_2 \; \cdots \; J_L]$.
- Finally, R_1 is defined such that $R = [R_1 \; R_2 \; \cdots \; R_L]$ is full rank.

We now check that $\lim_T [\Lambda_T^{-1} M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \tilde{A}_T]$ exists and is full column rank. First, recall that we have

$$\lim_T \left(\Lambda_T^{-1} M^{-1} \frac{\partial c(\theta^0)}{\partial \theta'} \tilde{A}_T \right) = \lim_T \left(\begin{bmatrix} \tilde{\Lambda}_T^{-1} & 0 \\ 0 & \check{\Lambda}_T^{-1} \end{bmatrix} \begin{bmatrix} \frac{\partial \tilde{d}(\theta^0)}{\partial \theta'} \\ \frac{\partial \check{d}(\theta^0)}{\partial \theta'} \end{bmatrix} \tilde{A}_T \right)$$

$$= \lim_T \begin{pmatrix} \tilde{\Lambda}_T^{-1} \frac{\partial \tilde{d}(\theta^0)}{\partial \theta'} \tilde{A}_T \\ 0 \end{pmatrix},$$

since $\check{\Lambda}_T^{-1} = o(\|\tilde{\Lambda}_T^{-1}\|)$ and $\|\tilde{A}_T\| = \mathcal{O}(\tilde{\Lambda}_T)$.

We now show that $[\tilde{\Lambda}_T^{-1}\frac{\partial \bar{d}(\theta^0)}{\partial \theta'}\tilde{A}_T]$ converges to a block diagonal matrix of rank p.

$$\tilde{\Lambda}_T^{-1}\frac{\partial \bar{d}(\theta^0)}{\partial \theta'}\tilde{A}_T = \begin{pmatrix} \lambda_{1T}^{-1} Id_{K_1} & & \\ & \ddots & \\ & & \lambda_{LT}^{-1} Id_{K_L} \end{pmatrix} \times \left[\lambda_{1T}\frac{\partial \bar{d}(\theta^0)}{\partial \theta'}R_1 \vdots \lambda_{2T}\frac{\partial \bar{d}(\theta^0)}{\partial \theta'}R_2 \vdots \cdots \vdots \lambda_{LT}\frac{\partial \bar{d}(\theta^0)}{\partial \theta'}R_L \right].$$

- The L diagonal blocks are equal to $J_l' R_l$; these (K_l, s_l)-blocks are full column rank s_l by construction of the matrices R_l with $\sum_{l=1}^{L} s_l = p$.
- The lower triangular blocks converge to zero since $\lambda_{jT} = o(\lambda_{lT})$ for any $1 \leq j < l \leq L$.
- The upper triangular blocks converge to zero by construction of the matrices R_l since $J_l' R_i = 0$ for any $1 \leq l < i \leq L$.

Proof of Corollary 2.1 *(Extended Theorem 2.3)*
From Assumption 2.4(i):

$$\begin{aligned} &\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} - A_T^{-1}\frac{\partial c(\theta^0)}{\partial \theta'} = \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right) \\ &\Leftrightarrow \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} - N_T^{-1}\Lambda_T^{-1}M_T^{-1}\frac{\partial c(\theta^0)}{\partial \theta'} = \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right) \\ &\Rightarrow \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'}\tilde{A}_T - N_T^{-1}\Lambda_T^{-1}M_T^{-1}\frac{\partial c(\theta^0)}{\partial \theta'}\tilde{A}_T = \mathcal{O}_P\left(\frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}}\right) \\ &\Rightarrow \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'}\tilde{A}_T - N^{-1}H = \mathcal{O}_P\left(\frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}}\right), \end{aligned}$$

with H full column rank matrix from Theorem 2.3.

Proof of Theorem 2.4 *(Asymptotic distribution)*
Mean-value expansion of the moment conditions around θ^0 for $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0,

$$\bar{\phi}_T(\hat{\theta}_T) = \bar{\phi}_T(\theta^0) + \frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'}(\hat{\theta}_T - \theta^0), \tag{2.21}$$

combined with the first-order conditions,

$$\frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \bar{\phi}_T(\hat{\theta}_T) = 0$$

yields to

$$\begin{aligned}
&\frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \bar{\phi}_T(\theta^0) + \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} (\hat{\theta}_T - \theta^0) = 0 \\
&\Leftrightarrow \tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \sqrt{T} \bar{\phi}_T(\theta^0) \\
&+ \tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \tilde{A}_T \tilde{A}_T^{-1} (\hat{\theta}_T - \theta^0) = 0 \\
&\Leftrightarrow \tilde{A}_T^{-1} \sqrt{T} (\hat{\theta}_T - \theta^0) = - \left[\tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \tilde{A}_T \right]^{-1} \\
&\times \tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \sqrt{T} \bar{\phi}_T(\theta^0) \\
&\Rightarrow \tilde{A}_T^{-1} \sqrt{T} (\hat{\theta}_T - \theta^0) = \mathcal{O}_P(1). \qquad (2.22)
\end{aligned}$$

We then get the expected result after justifying the invertibility of $\left[\tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} \Omega_T \frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \tilde{A}_T \right]$ for T large enough.

Lemma 2.2 *(Extension of Corollary 2.1)*
Under Assumptions 2.1–2.5, for any consistent estimator θ_T *s.t.* $\|\theta_T - \theta^0\| = \mathcal{O}_P(\|\tilde{\Lambda}_T\|/\sqrt{T})$,

$$\texttt{Plim} \left[\frac{\partial \bar{\phi}_T(\theta_T)}{\partial \theta'} \tilde{A}_T \right] \quad \textit{exists and is full column rank.}$$

Proof Mean-value expansion of the kth row of $\partial[\bar{\phi}_T(\theta_T)/\partial\theta']$ around θ^0 for $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0:

$$\begin{aligned}
&\left[\frac{\partial \bar{\phi}_T(\theta_T)}{\partial \theta'} \tilde{A}_T \right]_{k.} = \left[\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} \tilde{A}_T \right]_{k.} + (\theta_T - \theta^0)' \frac{\partial}{\partial \theta} \left[\frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \tilde{A}_T \right]_{k.}, \\
&\Leftrightarrow \left[\frac{\partial \bar{\phi}_T(\theta_T)}{\partial \theta'} \tilde{A}_T \right]_{k.} - \left[\frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} \tilde{A}_T \right]_{k.} = \frac{\sqrt{T}}{\|\tilde{\Lambda}_T\|} (\theta_T - \theta^0) \\
&\times \frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}} \frac{\partial}{\partial \theta} \left[\frac{\partial \bar{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \tilde{A}_T \right]_{k.}.
\end{aligned}$$

From Assumption 2.4(ii), the Hessian term is such that

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\frac{\partial\bar{\phi}_T(\tilde{\theta}_T)}{\partial\theta'}\right]_{k.} &= \frac{\partial}{\partial\theta}\left[A_T^{-1}\frac{\partial c(\tilde{\theta}_T)}{\partial\theta'}\right]_{k.} + \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right)\\ &= \frac{\partial}{\partial\theta}\left[N_T^{-1}\Lambda_T^{-1}M_T^{-1}\frac{\partial c(\tilde{\theta}_T)}{\partial\theta'}\right]_{k.} + \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right)\\ &= \mathcal{O}_P\left(\frac{1}{\lambda_{lT}}\right) + \mathcal{O}_P\left(\frac{1}{\sqrt{T}}\right) \quad \text{from Assumption 2.4(ii)}\\ &= \mathcal{O}_P\left(\frac{1}{\lambda_{lT}}\right), \end{aligned} \tag{2.23}$$

for any k such that $K_1 + \cdots + K_{l-1} < k \leq K_1 + \cdots + K_l$ and $l = 1, \cdots, L$.

Recall that $\tilde{A}_T = [\lambda_{1T}R_1 \vdots \cdots \vdots \lambda_{LT}R_L]$. To get the final result, we distinguish two cases to show that the RHS of the following equation is $o_p(1)$.

$$\begin{aligned}&\left(\left[\frac{\partial\bar{\phi}_T(\theta_T)}{\partial\theta'}\right]_{k.} - \left[\frac{\partial\bar{\phi}_T(\theta^0)}{\partial\theta'}\right]_{k.}\right)\lambda_{iT}R_i\\ &\quad= \frac{\sqrt{T}}{\|\tilde{\Lambda}_T\|}(\theta_T - \theta^0) \times \frac{\|\tilde{\Lambda}_T\|}{\sqrt{T}}\frac{\partial}{\partial\theta}\left[\frac{\partial\bar{\phi}_T(\tilde{\theta}_T)}{\partial\theta'}\right]_{k.}\lambda_{iT}R_i.\end{aligned}$$

- For $1 \leq i \leq l$, $\lambda_{iT} = o(\lambda_{lT})$ and the result directly follows from equation (2.23).
- For $i > l$, $\lambda_{lT} = o(\lambda_{iT})$ and the result follows from nearly-strong identification Assumption 2.5.

Note that when the same degree of global identification weakness is assumed, the asymptotic theory is available under Assumptions 2.1–2.4, since Lemma 2.2 holds without the nearly-strong identification Assumption 2.5.

Proof of Theorem 2.6 *(J test)*
Plugging (2.22) into (2.21), we get

$$\begin{aligned}&\sqrt{T}\bar{\phi}_T(\hat{\theta}_T) = \sqrt{T}\bar{\phi}_T(\theta^0) - \frac{\partial\bar{\phi}_T(\tilde{\theta}_T)}{\partial\theta'}\tilde{A}_T\left[\tilde{A}_T'\frac{\partial\bar{\phi}_T'(\hat{\theta}_T)}{\partial\theta}\Omega_T\frac{\partial\bar{\phi}_T(\tilde{\theta}_T)}{\partial\theta'}\tilde{A}_T\right]^{-1}\\ &\tilde{A}_T'\frac{\partial\bar{\phi}_T'(\hat{\theta}_T)}{\partial\theta}\Omega_T\sqrt{T}\bar{\phi}_T(\theta^0)\\ &\Rightarrow T\,Q_T(\hat{\theta}_T) = \left[\sqrt{T}\bar{\phi}_T(\theta^0)\right]'\Omega_T^{\prime 1/2}\left[Id_K - P_X\right]\Omega_T^{1/2}\left[\sqrt{T}\bar{\phi}_T(\theta^0)\right] + o_P(1),\end{aligned}$$

with $\Omega_T = \Omega_T^{\prime 1/2}\Omega_T^{1/2}$ and $P_X = X(X'X)^{-1}X'$ for $X = \Omega_T^{1/2}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\tilde{A}_T$.

Proof of Theorem 2.7 *(Wald test)*
To simplify the exposition, the proof is performed with Λ_T as defined in Example 2.2. In 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 identified at the

fast rate λ_{1T}, while the remaining ones are identified at the slow rate λ_{2T} without any linear combinations of the latter being identified at the fast rate. In step 2, we show that the Wald test statistic $\xi_T^W(h)$ on $H_0 : h(\theta) = 0$ asymptotically converges to the proper chi-square distribution with q degrees of freedom and that it is numerically equal to the Wald test statistic $\xi_T^W(g)$ on $H_0 : g(\theta) = 0$.

Step 1: The space of fast directions to be tested is

$$I^0(g) = \left[\text{col}\frac{\partial g'(\theta^0)}{\partial \theta}\right] \cap \left[\text{col}\frac{\partial \tilde{d}_1'(\theta^0)}{\partial \theta}\right].$$

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

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

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,

$$\texttt{Plim}\left[D_T^{-1}\frac{\partial h(\theta^0)}{\partial \theta'}\tilde{A}_T\right] = B^0 \quad \text{with} \quad D_T = \begin{bmatrix} \lambda_{1T} Id_{n^0(g)} & 0 \\ 0 & \lambda_{2T} Id_{q-n^0(g)} \end{bmatrix},$$

and B^0 a full column rank (q, p)-matrix.
Step 2: we show that the two induced Wald test statistics $\xi_T^W(g)$ and $\xi_T^W(h)$ are equal.

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

Then, we show $\xi_T^W(h)$ is asymptotically distributed as a chi-square with q degrees of freedom. First, a preliminary result naturally extends the above convergence toward B^0 when θ^0 is replaced by a λ_{2T}-consistent estimator θ_T^*:

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

The proof is very similar to Lemma 2.2 and is not reproduced here. The Wald test statistic $\xi_T^W(h)$ now writes:

$$\xi_T^W(h) = \left[\sqrt{T} D_T^{-1} h(\hat{\theta}_T)\right]' \left\{ D_T^{-1} \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} \tilde{A}_T^{-1} \left[\tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \tilde{A}_T \right]^{-1} \times \tilde{A}_T \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T^{-1} \right\}^{-1} \times \left[\sqrt{T} D_T^{-1} h(\hat{\theta}_T)\right].$$

From Lemma 2.2,

$$\left[\tilde{A}_T' \frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \tilde{A}_T \right] \xrightarrow{P} \Sigma \quad \text{nonsingular matrix.}$$

Now, from the mean-value theorem under H_0 we deduce

$$\sqrt{T} D_T^{-1} h(\hat{\theta}_T) = \sqrt{T} D_T^{-1} \frac{\partial h(\theta_T^*)}{\partial \theta'} (\hat{\theta}_T - \theta^0) = \left[D_T^{-1} \frac{\partial h(\theta_T^*)}{\partial \theta'} \tilde{A}_T \right] \sqrt{T} \tilde{A}_T^{-1} (\hat{\theta}_T - \theta^0)$$

with

$$\left[D_T^{-1} \frac{\partial h(\theta_T^*)}{\partial \theta'} \tilde{A}_T \right] \xrightarrow{P} B^0 \quad \text{and} \quad \sqrt{T} \tilde{A}_T^{-1} \left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}).$$

Finally we get

$$\xi_T^W(h) = \left[\sqrt{T} \tilde{A}_T^{-1} (\hat{\theta}_T - \theta^0)\right]' B_0' (B_0 \Sigma B_0')^{-1} B_0 \left[\sqrt{T} \tilde{A}_T^{-1} (\hat{\theta}_T - \theta^0)\right] + o_P(1).$$

Following the proof of Theorem 2.6 we get the expected result.

Proof of Proposition 2.1 *(Equivalence between CU-GMM and 2S-GMM)*
FOC of the CU-GMM optimization problem can be written as follows (see Antoine, Bonnal, and Renault 2007):

$$\sqrt{T} \frac{\partial \bar{\phi}_T'(\hat{\theta}_T^{CU})}{\partial \theta} S_T^{-1}(\hat{\theta}_T^{CU}) \sqrt{T} \bar{\phi}_T(\hat{\theta}_T^{CU}) - P \sqrt{T} \frac{\partial \bar{\phi}_T'(\hat{\theta}_T^{CU})}{\partial \theta} S_T^{-1}(\hat{\theta}_T^{CU}) \sqrt{T} \bar{\phi}_T(\hat{\theta}_T^{CU}) = 0,$$

where P is the projection matrix onto the moment conditions. Recall that

$$P \sqrt{T} \frac{\partial \bar{\phi}_T^{(j)}(\hat{\theta}_T^{CU})}{\partial \theta} = \text{Cov}\left(\frac{\partial \bar{\phi}_T^{(j)}(\hat{\theta}_T^{CU})}{\partial \theta}, \bar{\phi}_T(\hat{\theta}_T^{CU}) \right) S_T^{-1}(\hat{\theta}_T^{CU}) \sqrt{T} \bar{\phi}_T(\hat{\theta}_T^{CU}).$$

With a slight abuse of notation, we define conveniently the matrix of size (p, K^2) built by stacking horizontally the K matrices of size (p, K),

$\text{Cov} \times (\partial\bar{\phi}_{j,T}(\hat{\theta}_T^{CU})/\partial\theta, \bar{\phi}_T(\hat{\theta}_T^{CU}))$, as

$$\text{Cov}\left(\frac{\partial\bar{\phi}_T'(\hat{\theta}_T^{CU})}{\partial\theta}, \bar{\phi}_T(\hat{\theta}_T^{CU})\right)$$

$$\equiv \left[\text{Cov}\left(\frac{\partial\bar{\phi}_T^{(1)}(\hat{\theta}_T^{CU})}{\partial\theta}, \bar{\phi}_T(\hat{\theta}_T^{CU})\right) \ldots \text{Cov}\left(\frac{\partial\bar{\phi}_T^{(j)}(\hat{\theta}_T^{CU})}{\partial\theta}, \bar{\phi}_T(\hat{\theta}_T^{CU})\right) \ldots \times \text{Cov}\left(\frac{\partial\bar{\phi}_T^{(K)}(\hat{\theta}_T^{CU})}{\partial\theta}, \bar{\phi}_T(\hat{\theta}_T^{CU})\right)\right].$$

Then, we can write:

$$P\sqrt{T}\frac{\partial\bar{\phi}_T(\hat{\theta}_T^{CU})}{\partial\theta} = \text{Cov}\left(\frac{\partial\bar{\phi}_T'(\hat{\theta}_T^{CU})}{\partial\theta}, \bar{\phi}_T(\hat{\theta}_T^{CU})\right) \times \left(Id_K \otimes [S_T^{-1}(\hat{\theta}_T^{CU})\sqrt{T}\bar{\phi}_T(\hat{\theta}_T^{CU})]\right) \equiv H_T,$$

where $H_T = \mathcal{O}_P(1)$. Next, pre-multiply the above FOC by $\tilde{A}_T'/\sqrt{T}$ to get

$$\tilde{A}_T'\frac{\partial\bar{\phi}_T'(\hat{\theta}_T^{CU})}{\partial\theta}S_T^{-1}(\hat{\theta}_T^{CU})\sqrt{T}\bar{\phi}_T(\hat{\theta}_T^{CU}) - \frac{\tilde{A}_T'}{\sqrt{T}}H_T S_T^{-1}(\hat{\theta}_T^{CU})\sqrt{T}\bar{\phi}_T(\hat{\theta}_T^{CU}) = 0.$$

To get the equivalence between both estimators, we now show that the second element of the LHS is equal to $o_P(1)$.

From Assumption 2.1, we have $\sqrt{T}\bar{\phi}_T(\hat{\theta}_T^{CU}) = \sqrt{T}\rho_T(\hat{\theta}_T^{CU}) + \Psi_T(\hat{\theta}_T^{CU})$ with Ψ_T a Gaussian process. Hence, we have

$$\frac{\tilde{A}_T'}{\sqrt{T}}H_T S_T^{-1}(\hat{\theta}_T^{CU})\sqrt{T}\bar{\phi}_T(\hat{\theta}_T^{CU}) = \frac{\tilde{A}_T'}{\sqrt{T}}H_T S_T^{-1}(\hat{\theta}_T^{CU})\Psi_T(\hat{\theta}_T^{CU}) - \frac{\tilde{A}_T'}{\sqrt{T}}H_T S_T^{-1}(\hat{\theta}_T^{CU})\sqrt{T}\rho_T(\hat{\theta}_T^{CU}).$$

The first term of the RHS is obviously $o_p(1)$ since $\|\tilde{A}_T\| = o(\sqrt{T})$. The same remains to be shown for the second term,

$$\frac{\tilde{A}_T'}{\sqrt{T}}H_T\sqrt{T}S_T^{-1}(\hat{\theta}_T^{CU})\rho_T(\hat{\theta}_T^{CU}).$$

A result and proof similar to Lemma 2.1 for $\hat{\theta}_T^{CU}$ yield to: $\|\sqrt{T}\rho_T(\hat{\theta}_T^{CU})\| = \mathcal{O}_P(1)$.

Also, we already know that $H_T = \mathcal{O}_P(1)$ and $\|\tilde{A}_T\| = o(\sqrt{T})$. So, we combine these results to get

$$\frac{\tilde{A}_T'}{\sqrt{T}}H_T\sqrt{T}S_T^{-1}(\hat{\theta}_T^{CU})\rho_T(\hat{\theta}_T^{CU}) = o_p(1).$$

We conclude that both estimators are always defined by the same set of equations. To deduce that they are equivalent, we need Assumption 2.5 in order to get the same asymptotic theory. When the same degree of global identification weakness is assumed, the asymptotic theory holds without Assumption 2.5.

References

Andrews, D. W. K. 1994. Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity, *Econometrica* 62: 43–72.

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

Angrist, J. D., and A. B. Krueger. 1991. Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106: 979–1014.

Antoine, B., H. Bonnal, and E. Renault. 2007. On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood. *Journal of Econometrics* 138(2): 461–487.

Antoine, B., and E. Renault. 2009. Efficient GMM with Nearly-Weak Instruments. *Econometric Journal* 12: 135–171.

_____. 2010a. Efficient Minimum Distance Estimation with Multiple Rates of Convergence. *Journal of Econometrics*, forthcoming.

_____. 2010b. Specification Tests for Strong Identification. *Working Paper*, UNC-CH.

Caner, M. 2010. Testing, Estimation in GMM and CUE with Nearly-Weak Identification. *Econometric Reviews*, 29(3): 330–363.

Choi, I., and P. C. B. Phillips. 1992. Asymptotic and Finite Sample Distribution Theory for IV Estimators and Tests in Partially Identified Structural Equations. *Journal of Econometrics* 51: 113–150.

Epstein, L. G., and S. E. Zin. 1991. Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis. *Journal of Political Economy* 99(2): 263–286.

Hahn, J., J. Ham, and H. R. Moon. 2009. The Hausman Test and Weak Instruments. *Working Paper*, USC.

Hahn, J., and J. Hausman. 2002. A new Specification Test for the Validity of Instrumental Variables. *Econometrica* 70(1): 163–190.

Hahn, J., and G. Kuersteiner. 2002. Discontinuities of Weak Instruments Limiting Distributions. *Economics Letters* 75: 325–331.

Han, C., and P. C. B. Phillips. 2006. GMM with Many Moment Conditions. *Econometrica* 74(1): 147–192.

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

Hansen, L. P. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50(4): 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron. 1996. Finite Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics* 14: 262–280.

Hansen, L. P., and R. J. Hodrick. 1980. Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *Journal of Political Economy* 88: 829–853.

Hansen, L. P., and K. J. Singleton. 1982. Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica* 50: 1269–1286.

Horn, R. A., and C. R. Johnson. 1985. *Matrix Analysis*. Cambridge, U.K.: Cambridge University Press.

Kleibergen, F. 2005. Testing Parameters in GMM without Assuming that They Are Identified. *Econometrica* 73: 1103–1123.

Newey, W. K. 1994. The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62: 1349–1382.

Newey, W. K., and R. J. Smith. 2004. Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* 72: 219–255.

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

Newey, W. K., and F. Windmeijer. 2009. GMM with Many Weak Moment Conditions. *Econometrica* 77: 687–719.

Pakes, A., and D. Pollard. 1989. Simulation and the Asymptotics of Optimization Estimators. *Econometrica* 57(5): 1027–1057.

Phillips, P. C. B. 1989. Partially Identified Econometric Models. *Econometric Theory* 5: 181–240.

Sargan, J. D. 1983. Identification and Lack of Identification. *Econometrica* 51(6): 1605–1634.

Smith, R. J. 2007. *Weak Instruments and Empirical Likelihood: A discussion of the Papers by D. W. K. Andrews and J. H. Stock and Y. Kitamura*. Econometric Society Monograph Series, vol. 3, ch. 8 in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, pp. 238–260, Cambridge, U.K.: Cambridge University Press.

Staiger, D., and J. Stock. 1997. Instrumental Variables Regression with Weak instruments. *Econometrica* 65: 557–586.

Stock, J. H., and J. H. Wright. 2000. GMM with Weak Identification. *Econometrica* 68(5): 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo. 2002. A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics* 20: 518–529.