GENERALIZED EMPIRICAL LIKELIHOOD TESTS IN TIME SERIES MODELS WITH POTENTIAL IDENTIFICATION FAILURE

Patrik Guggenberger
Richard J. Smith

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP01/05

Generalized Empirical Likelihood Tests in Time Series Models With Potential Identification Failure*

Patrik Guggenberger**
Department of Economics
U.C.L.A.

Richard J. Smith
*c*emmap
U.C.L and I.F.S.
and
Department of Economics
University of Warwick

April 7, 2005

Abstract

We introduce test statistics based on generalized empirical likelihood methods that can be used to test simple hypotheses involving the unknown parameter vector in moment condition time series models. The test statistics generalize those in Guggenberger and Smith (2005) from the i.i.d. to the time series context and are alternatives to those in Kleibergen (2001) and Otsu (2003). The main feature of these tests is that their empirical null rejection probabilities are not affected much by the strength or weakness of identification. More precisely, we show that the statistics are asymptotically distributed as chi–square under both classical asymptotic theory and weak instrument asymptotics of Stock and Wright (2000). A Monte Carlo study reveals that the finite–sample performance of the suggested tests is very competitive.

JEL Classification: C12, C31.

Keywords: Generalized Empirical Likelihood, Nonlinear Moment Conditions, Similar Tests, Size Distortion, Weak Identification.

* This paper is an extension of the first chapter of Guggenberger's dissertation at Yale University and was written in August 2003. Guggenberger thanks his advisor D.W.K. Andrews who played a very important role in the writing process. We would also like to thank J. Hahn for detailed comments and seminar participants at Brown, Harvard/MIT, Maryland and NYU in 2003 and at Irvine, Laval, Riverside and San Diego in 2004. Smith gratefully acknowledges financial support through a 2002 Leverhulme Major Research Fellowship.

** Corresponding author: Patrik Guggenberger, Bunche Hall 8385, Department of Economics, U.C.L.A., Box 951477, Los Angeles, CA 90095-1477. Email: guggenbe@econ.ucla.edu

1 Introduction

There has recently been a lot of interest in robust inference in weakly identified models, see *inter alia* Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2001, 2002), Caner (2003), Dufour and Taamouti (2003), Moreira (2003), Otsu (2003), Andrews and Marmer (2004), Andrews et al. (2004), Chao and Swanson (2005) and Guggenberger and Smith (2005, GS henceforth). For a recent discussion of that literature, see Dufour (2003). This paper adds to this literature by introducing two types of test statistics that can be used to test simple hypotheses involving the unknown parameter vector in *nonlinear* moment condition *time series* models. The main feature of these statistics is that they lead to tests whose empirical rejection probabilities (ERP) under the null hypothesis do not depend much on the strength or weakness of identification of the model. More precisely, we show that the statistics are asymptotically distributed as chi–square under both classical and the weak instrument asymptotic theory of Stock and Wright (2000). This is in contrast to many of the classical test statistics, like for example a Wald statistics, that have a chi–square under the former but a nonstandard asymptotic distribution under the latter theory.

The first test statistic is given as the renormalized criterion function of the generalized empirical likelihood (GEL) estimator, see Newey and Smith (2004), and the second one as a quadratic form in the first order condition (FOC) of the GEL estimator; both statistics are evaluated at the hypothesized parameter vector. The statistics generalize those in GS from the i.i.d. and martingale difference sequence (m.d.s.) setup to the time series case. One advantage of the second statistic over the first one is that the degrees of freedom parameter of its asymptotic chi–square distribution equals p, the dimension of the unknown parameter vector, while for the first statistic the degrees of freedom parameter equals k, the number of moment conditions. This negatively affects power properties of tests based on the first statistic in overidentified situations. To adapt the statistics to the time series context, we work with smoothed counterparts of the moment indicator functions based on a kernel function $k(\cdot)$ and a bandwidth parameter S_n, an approach which was originally used in Kitamura and Stutzer (1997) and Smith (1997, 2001). This method for the construction of test statistics in the weakly identified framework was suggested by Guggenberger (2003, Introduction of the first chapter). See also Otsu (2003).

While most of the papers on robust testing with weak identification are written for the linear i.i.d. instrumental variables model, there are two closely related procedures for robust inference in nonlinear time series models available in the literature. First, Kleibergen (2001) introduces a test statistic that is given as a quadratic form in the FOC of the generalized method of moments (GMM, Hansen (1982)) continuous updating estimator (CUE). The statistic includes consistent estimators for the long–run covariance matrix of the sums of the renormalized moment indicators and derivatives thereof. Kleibergen (2001) suggests the use of heteroskedasticity and autocorrelation consistent (HAC) estimators as given in Andrews (1991). Secondly, Otsu's (2003) procedure is based on the criterion function of the GEL estimator. An asymptotic chi–square distribution with p degrees of freedom is obtained by evaluating the GEL criterion function at transformed moment indicators of dimension p rather than at the original moment indicators that are k–dimensional. In section 2.3. below we give a detailed comparison of the various approaches. There we also introduce modifications to Otsu's (2003) statistic that are computationally more attractive.

Besides technicalities, the main assumptions needed to establish the asymptotic chi–square distribution of the new test statistics introduced in this paper are that 1) an appropriate HAC estimator of the long–run covariance matrix of the sums of the moment indicators is consistent and that 2) a central limit theorem (CLT) holds for the moment indicators and derivatives thereof with respect to the weakly identified parameters. These assumptions are very similar to the ones used in Kleibergen (2001). They are stated and discussed in the Appendix.

The tests in this paper are for simple hypotheses on the *full* parameter vector. They could straightforwardly be generalized to subvector tests under the assumption that the parameters not under test are strongly identified, see e.g. Kleibergen (2001, 2004), Otsu (2003) and GS. The idea is to replace the parameters not under test by consistently estimated counterparts in the test statistics. We omit this generalization here to avoid complicating the presentation. In any case, simulations in Guggenberger and Wolf (2004) indicate that the use of subsampling techniques for subvector tests may prove advantageous, especially in scenarios where the assumption of strong identification of the parameters not under test is questionable.

To investigate the finite–sample performance of the new tests, we compare them to those in Kleibergen (2001) and Otsu (2003) in a Monte Carlo study that focuses on a time series linear model with AR(1) or MA(1) variables. We find that both in terms of size and power the new tests compare very favorably to the alternative procedures. Even though the tests are first–order equivalent, there can be huge power differences between Otsu's (2003) and the tests in this paper.

To implement the tests here and those in Kleibergen (2001) and Otsu (2003) a bandwidth S_n has to be chosen. Andrews (1991) and Newey and West (1994) provide theory of how to choose the bandwidth, if the goal is to minimize the mean squared error of a (HAC) covariance matrix estimator. However, in the testing context here, we are really interested in size and power properties of the tests and it is unclear of how to develop a theory of bandwidth choice. One could still follow the procedures in Andrews (1991) or Newey and West (1994) but very likely this would not lead to any optimality result. The bandwidth choice is an important problem that is beyond the scope of this paper. Future research has to tackle this challenging question.

The remainder of the paper is organized as follows. In section 2 the model is introduced and the test statistics and their asymptotic theory are discussed. The tests are compared to Kleibergen's (2001) and Otsu's (2003) approaches. Section 3 contains the Monte Carlo study. All technical assumptions and proofs are relegated to the Appendix.

The symbols " $\rightarrow_d$ " and " $\rightarrow_p$ " denote convergence in distribution and convergence in probability, respectively. Convergence "almost surely" is written as "a.s." and "with probability approaching 1" is replaced by "w.p.a.1". The space $C^i(S)$ contains all functions that are i–times continuously differentiable on the set S. Furthermore, $vec(M)$ stands for the column vectorization of the $k \times p$ matrix M, i.e. if $M = (m_1, ..., m_p)$ then $vec(M) = (m_1', ..., m_p')'$, "$M'$" denotes the transpose matrix of M, $(M)_{i,j}$ the element in the i–th row and j–th column, "$M > 0$" means that M is positive definite and $||M||$ stands for the square root of the largest eigenvalue of $M'M$. By I_p we denote the p–dimensional identity matrix.

2 Robust Testing

2.1 Model and Notation

The paper considers models specified by a finite number of moment restrictions. More precisely, let $\{z_i : i = 1, ..., n\}$ be $\mathbb{R}^l$–valued time series data, where $n \in \mathbb{N}$ denotes the sample size. Let $g_n : H \times \Theta \to \mathbb{R}^k$, where $H \subset \mathbb{R}^l$ and $\Theta \subset \mathbb{R}^p$ denotes the parameter space. The model has a true parameter θ_0 for which the moment condition

$$Eg_n(z_i, \theta_0) = 0 \tag{2.1}$$

is satisfied. For $g_n(z_i, \theta)$, usually the shorter $g_i(\theta)$ is used.[1] Interest focuses on testing a simple hypothesis

$$H_0 : \theta_0 = \theta \text{ versus the alternative } H_1 : \theta_0 \neq \theta. \tag{2.2}$$

Define

$$\begin{aligned}&\widehat{g}(\theta) : = n^{-1} \sum_{i=1}^{n} g_i(\theta),\ \Psi_n(\theta) := n^{1/2}(\widehat{g}(\theta) - E\widehat{g}(\theta)) \text{ and}\\ &\Delta(\theta) : = \lim_{n\to\infty} E\Psi_n(\theta)\Psi_n(\theta)' \in \mathbb{R}^{k\times k},\end{aligned}$$

the long–run covariance matrix of $g_i(\theta)$. Let $\theta = (\alpha', \beta')'$, where $\alpha \in A$, $A \subset \mathbb{R}^{p_A}$, $\beta \in B$, $B \subset \mathbb{R}^{p_B}$ for $\Theta = A \times B$ and $p_A + p_B = p$. In the following, we adopt Assumption C from Stock and Wright (2000) in which α_0 and β_0 are modelled respectively as weakly and strongly identified parameter vectors. For a detailed discussion of this assumption, see Stock and Wright (2000, pp. 1060–1). Let $\mathcal{N} \subset B$ denote an open neighborhood β_0.

Assumption ID: The true parameter $\theta_0 = (\alpha_0', \beta_0')'$ is in the interior of the compact space $\Theta = A \times B$ and **(i)** $E\widehat{g}(\theta) = n^{-1/2}m_{1n}(\theta) + m_2(\beta)$, where $m_{1n}, m_1 : \Theta \to \mathbb{R}^k$ and $m_2 : B \to \mathbb{R}^k$ are continuous functions such that $m_{1n}(\theta) \to m_1(\theta)$ uniformly on Θ, $m_1(\theta_0) = 0$ and $m_2(\beta) = 0$ if and only if $\beta = \beta_0$; **(ii)** $m_2 \in C^1(\mathcal{N})$; **(iii)** let $M_2(\beta) := (\partial m_2/\partial\beta)(\beta) \in \mathbb{R}^{k\times p_B}$. $M_2(\beta_0)$ has full column rank p_B.

Following the suggestion in Guggenberger (2003), we work with smoothed counterparts of the moment indicators $g_i(\theta)$ to handle the general time series setup considered here as in Kitamura and Stutzer (1997)

[1] The function g is allowed to depend on the sample size n to model weak identification, see Assumption ID below. For example, consider the i.i.d. linear instrumental variable (IV) model given by the structural and reduced form equations $y = Y\theta_0 + u$, $Y = Z\Pi + V$, where $y, u \in \mathbb{R}^n$, $Y, V \in \mathbb{R}^{n\times p}$, $Z \in \mathbb{R}^{n\times k}$ and $\Pi \in \mathbb{R}^{k\times p}$. The matrices Y and Z contain the endogenous and instrumental variables, respectively. Denote by Y_i, V_i, Z_i, ... $(i = 1, ..., n)$ the i^{th} row of the matrix Y, V, Z, ... written as a column vector. Assume $EZ_iu_i = 0$ and $EZ_iV_i' = 0$. The first condition implies that $Eg_i(\theta_0) = 0$, where for each $i = 1, ..., n$, $g_i(\theta) := Z_i(y_i - Y_i'\theta)$. Note that in this example $g_i(\theta)$ depends on n if the reduced form coefficient matrix Π is modeled to depend on n, see Stock and Wright (2000), where $\Pi = \Pi_n = (n^{-1/2}\Pi_A, \Pi_B)$ and Π_A and Π_B are fixed matrices with p_A and p_B columns, $p = p_A + p_B$ and Π_B has full column rank.

and Smith (1997, 2001). See also Otsu (2003) and Smith (2000, 2005).[2] For $i = 1, ..., n$, define

$$g_{in}(\theta) := S_n^{-1} \sum_{j=i-n}^{i-1} k(j/S_n) g_{i-j}(\theta),$$

where S_n is a bandwidth parameter ($S_n \to \infty$ as $n \to \infty$) and $k(\cdot)$ is a kernel. For simplicity, from now on the truncated kernel is used given by

$$k(x) = 1 \text{ if } |x| \leq 1 \text{ and } k(x) = 0 \text{ otherwise}$$

and thus $g_{in}(\theta) = S_n^{-1} \sum_{j=\max\{-S_n, i-n\}}^{\min\{S_n, i-1\}} g_{i-j}(\theta)$.[3] Define

$$\widehat{g}_n(\theta) := n^{-1} \sum_{i=1}^{n} g_{in}(\theta) \text{ and } \widehat{\Delta}(\theta) := S_n \sum_{i=1}^{n} g_{in}(\theta) g_{in}(\theta)'/n. \tag{2.3}$$

Under assumptions given in Lemma 2 below, the estimator $\widehat{\Delta}(\theta_0)$ is shown to be consistent for $2\Delta(\theta_0)$ whereas the "unsmoothed" version of the estimator $\sum_{i=1}^{n} g_i(\theta_0) g_i(\theta_0)'/n$ used in GS, while being consistent in an i.i.d. or m.d.s. setup, would not be consistent in the general time series context considered here. See GS's discussion of their assumption M_θ(ii). The consistency of $\widehat{\Delta}(\theta)$ is crucial for the testing procedures suggested in the next section.

The statistics below are based on the GEL estimator. In what follows, a brief definition of the GEL estimator is given. For a more comprehensive discussion see Smith (1997, 2001), Newey and Smith (2004) and GS. Let ρ be a real–valued function $Q \to \mathbb{R}$, where Q is an open interval of the real line that contains 0 and $\widehat{\Lambda}_n(\theta) := \{\lambda \in \mathbb{R}^k : \lambda' g_{in}(\theta) \in Q \text{ for } i = 1, ..., n\}$. If defined, let $\rho_j(v) := (\partial^j \rho / \partial v^j)(v)$ and $\rho_j := \rho_j(0)$ for nonnegative integers j.

The GEL estimator is the solution to a saddle point problem

$$\widehat{\theta}_\rho : = \arg \min_{\theta \in \Theta} \sup_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}_\rho(\theta, \lambda), \text{ where} \tag{2.4}$$

$$\widehat{P}_\rho(\theta, \lambda) : = 2 \sum_{i=1}^{n} (\rho(\lambda' g_{in}(\theta)) - \rho_0)/n. \tag{2.5}$$

Assumption ρ**: (i)** ρ is concave on Q; **(ii)** ρ is C^2 in a neighborhood of 0 and $\rho_1 = \rho_2 = -1$.

Examples of GEL estimators include the CUE, see Hansen, Heaton and Yaron (1996), empirical likelihood (EL, see Imbens (1997) and Qin and Lawless (1994)) and exponential tilting (ET, see Kitamura and Stutzer (1997) and Imbens, Spady and Johnson (1998)) which correspond to $\rho(v) = -(1+v)^2/2$, $\rho(v) = \ln(1-v)$ and $\rho(v) = -\exp v$, respectively.

[2] An alternative procedure would be to work with a blocking method as in Kitamura (1997).

[3] In general, one could employ kernels in the class $\mathcal{K}_1$ of Andrews (1991, p.821) taking into account technical modifications in Jansson (2002); see for example Smith (2001) and Otsu (2003). Here we focus on the truncated kernel because it significantly simplifies the proofs and notation. In addition, for the testing purpose in this paper, it is not clear on what basis a kernel should be chosen and Monte Carlo simulations reveal that the finite sample performance is not very sensitive to the kernel choice, see also Newey and West (1994) for similar findings in the HAC literature.

2.2 Test Statistics

Here statistics are introduced that can be used to test (2.2) in the time series model given by (2.1). It is established that they are asymptotically pivotal quantities and have limiting chi–square null distributions under Assumption ID. Therefore these statistics lead to tests whose level properties in finite samples should not be affected much by the strength or weakness of identification. There are other statistics that share this property in the general time series set–up considered here, namely Kleibergen's (2001) GMM–based and Otsu's (2003) GEL–based statistic.[4] These statistics are compared to the approach of this paper in more detail below.

Let ρ be any function satisfying Assumption ρ. The first statistic is given by[5]

$$\begin{aligned} GELR_\rho(\theta) &: = S_n^{-1} n \widehat{P}_\rho(\theta, \lambda(\theta))/2, \text{ where, if it exists,} \\ \lambda(\theta) &: = \arg \max_{\lambda \in \widehat{\Lambda}_n(\theta)} \widehat{P}_\rho(\theta, \lambda). \end{aligned} \tag{2.6}$$

The statistic $GELR_\rho(\theta)$ has a nonparametric likelihood ratio interpretation, see GS, where motivation is provided in the i.i.d. context.

The second set of statistics is based on the FOC with respect to θ of the GEL estimator $\widehat{\theta}$. If the minimum of the objective function $\widehat{P}(\theta, \lambda(\theta))$ is obtained in the interior of Θ, the score vector with respect to θ must equal 0 at $\widehat{\theta}$. Using the envelope theorem it can be shown that this results in

$$0' = \lambda(\widehat{\theta})' \sum_{i=1}^{n} \rho_1(\lambda(\widehat{\theta})' g_{in}(\widehat{\theta})) G_{in}(\widehat{\theta})/n, \text{ where if defined} \tag{2.7}$$

$$G_{in}(\theta) \ : = (\partial g_{in}/\partial \theta)(\theta) \in \mathbb{R}^{k \times p}, \tag{2.8}$$

see Newey and Smith (2004) and GS for a rigorous argument of this statement in the i.i.d. case. For $\theta \in \Theta$, define

$$D_\rho(\theta) := \sum_{i=1}^{n} \rho_1(\lambda(\theta)' g_{in}(\theta)) G_{in}(\theta)/n \in \mathbb{R}^{k \times p}. \tag{2.9}$$

Thus, (2.7) may be written as $\lambda(\widehat{\theta})' D_\rho(\widehat{\theta}) = 0'$. The test statistic is given as a quadratic form in the score vector $\lambda(\theta)' D_\rho(\theta)$ evaluated at the hypothesized parameter vector θ and renormalized by the appropriate rate

$$S_\rho(\theta) := S_n^{-2} n \lambda(\theta)' D_\rho(\theta) \left(D_\rho(\theta)' \widehat{\Delta}(\theta)^{-1} D_\rho(\theta) \right)^{-1} D_\rho(\theta)' \lambda(\theta)/2. \tag{2.10}$$

In addition, the following variant of $S_\rho(\theta)$

$$LM_\rho(\theta) := n \widehat{g}_n(\theta)' \widehat{\Delta}(\theta)^{-1} D_\rho(\theta) \left(D_\rho(\theta)' \widehat{\Delta}(\theta)^{-1} D_\rho(\theta) \right)^{-1} D_\rho(\theta)' \widehat{\Delta}(\theta)^{-1} \widehat{g}_n(\theta)/2 \tag{2.11}$$

is considered that substitutes $S_n^{-1}\lambda(\theta)$ in $S_\rho(\theta)$ by the asymptotically equivalent expression $-\widehat{\Delta}(\theta)^{-1}\widehat{g}_n(\theta)$, see eq. (A.6) below. The names $S_\rho(\theta)$ and $LM_\rho(\theta)$ of the statistics are taken from GS and are based

[4] There are various other robust tests introduced for i.i.d. models, e.g. Kleibergen (2002), Caner (2003) and Moreira (2003).

[5] The generalization of the $GELR_\rho$ statistic in GS to the time series context has now been independently introduced by Otsu (2003), see his $\widehat{S}_{GEL}$ statistic.

on the interpretation of the statistics as score and Lagrange multiplier statistics, respectively, see GS for more discussion.

The next theorem discusses the asymptotic distribution of these test statistics evaluated at θ_0. The technical assumptions M_{θ_0} and their interpretation are given in the Appendix.

Theorem 1 *Suppose ID, ρ and M_{θ_0}(i)–(iii) hold. Then for $S_n \to \infty$ as $n \to \infty$ and $S_n = o(n^{1/2})$ it follows that*

(i) $GELR_\rho(\theta_0) \to_d \chi^2(k)$.

If in addition M_θ (iv)–(vii) hold then

(ii) $S_\rho(\theta_0), LM_\rho(\theta_0) \to_d \chi^2(p)$.

Remarks: 1) Theorem 1 implies a straightforward method to construct confidence regions or hypothesis tests for θ_0. For example, a critical region for the test (2.2) at significance level r is given by $\{GELR_\rho(\theta_0) \geq \chi^2_r(k)\}$, where $\chi^2_r(k)$ denotes the $(1-r)$–critical value from the $\chi^2(k)$ distribution. Unlike classical test statistics such as a Wald statistic, the statistics $GELR_\rho(\theta_0)$, $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ are asymptotically pivotal statistics under Assumption ID. Therefore, the level of tests based on these statistics should not vary much with the strength or weakness of identification in finite samples. For the statistics $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ to be pivotal, it is crucial that $D_\rho(\theta_0)$ (appropriately renormalized) and $n^{1/2}\widehat{g}_n$ are asymptotically independent under both weak and strong identification, see the proof of the theorem.[6] Theorem 1 also shows that the asymptotic null distribution of the test statistics does not depend on the choice of ρ.

2) Theorem 1 provides an approach to full–vector inference for θ_0. There are various approaches to subvector inference for θ_{01}, where $\theta_0 = (\theta'_{01}, \theta'_{02})'$.

First, under the assumption that the parameters θ_{02} not under test are strongly identified, one can replace θ_{02} in the test statistics above by their consistently estimated counterparts $\widehat{\theta}_{02}$, where $\widehat{\theta}_{02}$ is a GEL estimator, say, calculated under the restriction that $\theta_1 = \theta_{01}$. This approach is investigated in Kleibergen (2001, 2004) for the GMM CUE, in Otsu (2003) and GS for GEL and could also be implemented here at the expense of more difficult notation and longer proofs.

Second, confidence intervals can be constructed by a projection argument, see Dufour (1997). However, this approach is conservative and in general computationally cumbersome. In a recent paper, Dufour and Taamouti (2003) show that the Anderson and Rubin (1949) statistic is an exception, in that a closed form solution is available.

Third, Guggenberger and Wolf (2004) suggest an alternative approach based on subsampling. Unlike the first and second procedures, subsampling leads to subvector tests whose type I error converges to the desired nominal level *without* additional identification assumptions for each fixed degree of identification. Guggenberger and Wolf's (2004) Monte Carlos suggest that for subvector inference subsampling seems to do better than Kleibergen (2001, 2004) and Dufour and Taamouti (2003). In their simulation study,

[6] Also see Smith (2001) which demonstrates this property for the strongly identified case.

the former procedure tends to underreject when θ_{02} is only weakly identified and the latter seems to underreject across all the scenarios. On the other hand, they find that for full–vector inference, subsampling is outperformed by Kleibergen (2001) and GS. Therefore, in this paper, subvector inference along the first approach, is not included because it would unnecessarily complicate the presentation without contributing much.

3) A drawback of $GELR_\rho(\theta_0)$ is that its limiting null distribution has degrees of freedom equal to k, the number of moment conditions rather than the dimension of the parameter vector p. In general, this has a negative impact on the power properties of hypothesis tests based on $GELR_\rho(\theta_0)$ in over–identified situations. On the other hand, the limiting null distribution of $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ has degrees of freedom equal to p. Therefore the power of tests based on these statistics should not be negatively affected by a high degree of over–identification.

4) Besides technicalities, assumption M_{θ_0} (given in the Appendix) essentially states that (i) the Bartlett HAC estimator consistently estimates the long–run variance matrix $\Delta(\theta_0)$ and (ii) that a CLT holds for the times series $(vecG'_{iA}(\theta_0), g'_i(\theta_0))'$ with full rank asymptotic covariance matrix $V(\theta_0)$, where $G_{iA}(\theta_0)$ is the submatrix of $G_{in}(\theta_0)$ corresponding to the weakly identified parameters, see the Appendix for a detailed discussion. Part (ii) is very closely related to Assumption 1 in Kleibergen (2001) that states a CLT for $(vecG'_i(\theta_0), g'_i(\theta_0))'$ with possibly singular covariance matrix. Assumptions (i) and (ii) are compatible with many time series models. Therefore, the approach taken in this paper generalizes the setup in GS whose applications were restricted to m.d.s..

5) The theorem does not give any guidelines on how to choose the bandwidth S_n in finite samples. In fact, just as for the choice of the kernel k, it is difficult to provide theory for its choice in the testing context considered here, where size and power properties matter. One could still follow Andrews (1991) and choose S_n such that the mean-squared error of the covariance matrix estimator is minimized after a time series model has been specified. However, it is unclear what effect this procedure would have on size and power of the test and it would be surprising if this procedure led to any optimality property.

2.3 Comparison with Kleibergen (2001) and Otsu (2003)

Here we compare our statistics to the $K-$ and $\widehat{K}_{GEL}$–statistics of Kleibergen (2001) and Otsu (2003). These statistics, S_ρ and LM_ρ, and the ones defined below have the same first–order theory under the null hypothesis; asymptotically they are all distributed as $\chi^2(p)$ under the null.

Kleibergen's K–statistic is defined as

$$K(\theta) := n\widehat{g}(\theta)'\widetilde{\Delta}(\theta)^{-1}D_\theta(D'_\theta\widetilde{\Delta}(\theta)^{-1}D_\theta)^{-1}D'_\theta\widetilde{\Delta}(\theta)^{-1}\widehat{g}(\theta), \tag{2.12}$$

where

$$\begin{aligned} G_i(\theta) &:= (\partial g_i/\partial\theta)(\theta),\ \widehat{G}(\theta) := n^{-1}\textstyle\sum_{i=1}^n G_i(\theta) \in \mathbb{R}^{k\times p}, \\ D_\theta &:= \widehat{G}(\theta) - \widetilde{\Omega}(\theta)[I_p \otimes (\widetilde{\Delta}(\theta)^{-1}\widehat{g}(\theta))] \in \mathbb{R}^{k\times p} \text{ and} \end{aligned} \tag{2.13}$$

$\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$ are consistent estimators for $\Delta(\theta)$ and the long–run covariance matrix $\lim_{n\to\infty} E\{n^{-1}\sum_{i,j=1}^n [G_i(\theta) - EG_i(\theta)][(I_p \otimes g_j(\theta)') - E(I_p \otimes g_j(\theta)')]\}$, respectively. Kleibergen (2001) suggests the use of HAC estimators for $\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$, see e.g. Andrews (1991). The statistics LM_ρ and the K–statistic are given as quadratic forms in the FOC of the GEL and the GMM CUE estimator, respectively. The intuition for tests based on these statistics is as follows: under strong identification, GEL and GMM estimators are consistent. In consequence, in large samples the FOC for the estimator also holds at the true parameter vector θ_0. Therefore, the statistics are quadratic forms which are expected to be small at the true vector θ_0. Even though the GMM CUE and GEL CUE are numerically identical (see Newey and Smith (2004, fn. 2)), their FOC are different and therefore LM_{CUE} and K will typically differ. For i.i.d. or m.d.s. scenarios GS specify for which estimators $\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$ in the K–statistic, K and LM_{CUE} are identical. These statements in GS cannot be generalized to the general time series setup, where K and LM_{CUE} are different. The reason is that in this latter statistic functions of the smoothed indicators g_{in} and G_{in} are used, e.g. $\widehat{g}_n$, while the former statistic uses functions of the unsmoothed indicators, e.g. $\widehat{g}$.

Otsu's (2003) statistic is given by

$$\begin{aligned}\widehat{K}_{GEL}(\theta) &:= S_n^{-1} n \sup_{\gamma\in\Gamma(\theta)} \widehat{P}_\rho(\theta, \widehat{\Delta}(\theta)^{-1} D_\rho(\theta)\gamma)/2, \text{ where} \\ \Gamma(\theta) &:= \{\gamma \in \mathbb{R}^p; \widehat{\Delta}(\theta)^{-1} D_\rho(\theta)\gamma \in \widehat{\Lambda}_n(\theta)\} \text{ and}\end{aligned} \tag{2.14}$$

$\widehat{\Delta}(\theta)$ and $D_\rho(\theta)$ are defined in (2.3) and (2.9), respectively. Otsu's (2003) statistic has been formulated here based on the truncated kernel but can of course be implemented using more general kernels, see Otsu (2003). Otsu's (2003) statistic is not given as a quadratic form in the FOC and the above intuition does not apply. Unlike the $GELR_\rho$ statistic, however, the asymptotic null distribution of $\widehat{K}_{GEL}$ does not depend on the number of moment conditions k. This is achieved by considering the transformed moment indicators $g_{in}'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)$ in (2.14) rather than g_{in}' as in (2.6). A drawback of Otsu's (2003) approach is that two maximizations are necessary to calculate the statistic, one to calculate $\lambda(\theta)$ in $D_\rho(\theta)$ of (2.9) and one in (2.14). The latter maximization may be simply avoided as follows. Let

$$\begin{aligned}\mu_\rho(\theta) &:= -S_n \widehat{\Delta}(\theta)^{-1} D_\rho(\theta) \left(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1} D_\rho(\theta)\right)^{-1} D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}\widehat{g}_n(\theta), \\ \tilde{\mu}_\rho(\theta) &:= \widehat{\Delta}(\theta)^{-1} D_\rho(\theta) \left(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1} D_\rho(\theta)\right)^{-1} D_\rho(\theta)'\lambda(\theta).\end{aligned}$$

Define the statistic

$$GELR_\rho(\theta, \mu) := S_n^{-1} n \hat{P}_\rho(\theta, \mu)/2.$$

Theorem 2 *Suppose ID, ρ and $M_{\theta_0}(i)$–(vii) hold. Then for $S_n \to \infty$ as $n \to \infty$ and $S_n = o(n^{1/2})$ it follows that*

$$GELR_\rho(\theta_0, \mu_\rho(\theta_0)), GELR_\rho(\theta_0, \tilde{\mu}_\rho(\theta_0)) \to_d \chi^2(p).$$

Remark: The function ρ used in obtaining $\mu_\rho(\theta)$ or $\tilde{\mu}_\rho(\theta)$ through $D_\rho(\theta)$ and $\lambda(\theta)$ may be allowed to differ from that defining $GELR_\rho(\theta, \mu)$ as long as both functions satisfy Assumption ρ.

3 Monte Carlo Study

In this section, the finite sample properties of the hypotheses tests in Theorem 1 are investigated in a Monte Carlo study and compared to the tests suggested in Kleibergen (2001) and Otsu (2003).

3.1 Design

The data generating process is given by the linear instrumental variables time series model

$$\begin{aligned} y &= Y\theta_0 + u, \\ Y &= Z\Pi + V. \end{aligned} \tag{3.1}$$

There is only a single right hand side endogenous variable Y and no included exogenous variables. Let $Z \in \mathbb{R}^{n\times k}$, where k is the number of instruments and n the sample size. The reduced form matrix $\Pi \in \mathbb{R}^k$ equals a vector of ones times a constant Π_1 that determines the strength or weakness of identification. Similar to the design in Otsu (2003), each column of Z and u are generated as AR(1) or MA(1) processes (with autoregressive and moving–average parameters ϕ and ν, respectively) with innovations distributed as independent $N(0,1)$ random variables and V has i.i.d. $N(0,1)$ components. The innovations of the process for u and the i–th component of V are correlated; their joint distribution is $N(0,\Sigma)$, where $\Sigma \in \mathbb{R}^{2\times 2}$ with diagonal elements equal to unity and off–diagonal elements ρ_{uV}.

Interest focuses on testing the scalar null hypothesis $H_0 : \theta_0 = 0$ versus the alternative hypothesis $H_1 : \theta_0 \neq 0$. Results are reported at nominal levels of 5% for sample size $n = 200$. The following 40 parameter combinations are considered and for each we simulate $10,000$ repetitions:

$$\begin{aligned} &k = 2, 10;\ \Pi_1 = .01, .5;\ \rho_{uV} = 0, .5; \\ &\phi = 0, .5, .9;\ \nu = .5, .9. \end{aligned}$$

We include the test statistics LM_{EL}, S_{EL}, $GELR_{EL}$, K and $\widehat{K}_{GEL}$ in the study. For the K–statistic we use a Bartlett kernel to calculate the covariance matrix estimators and for $\widehat{K}_{GEL}$ we use the EL specification. To implement the statistics, the bandwidth S_n has to be chosen. We consider fixed bandwidths $S_n = 1, ..., 10$ and also calculate the i.i.d. versions of the test statistics. Note that for the Bartlett kernel, $S_n = 1$ leads to numerically identical results for K as no smoothing. To solve the maximization problems in the GEL–based statistics, a Newton–Raphson algorithm is used. Size and power properties are investigated by considering $\theta_0 = 0, 1$ and -1.

3.2 Results

Tests based on the statistics S_{EL} and $GELR_{EL}$ have less desirable size properties in our study than tests based on LM_{EL} and therefore only results for LM_{EL}, K and $\widehat{K}_{GEL}$ are discussed in detail. Size problems of the i.i.d. versions of S_{EL} and $GELR_{EL}$ in finite–samples were also reported in GS.

Size distortion, if any, is generally smaller for the MA(1) than for the AR(1) cases and for size purposes we therefore restrict attention to the AR(1) cases. The Monte Carlo results show that dependence of the results on $\rho_{uV} = 0, .5$ is small, especially when $k = 2$. If anything, the rejection probabilities of the tests are slightly higher for higher endogeneity. In what follows, we therefore restrict attention to $\rho_{uV} = .5$. As to be expected from theory, the size results do generally not vary much with Π_1, the strength of the instruments. The exceptions are occasionally cases of high degree of overidentification and high endogeneity ($k = 10, \rho_{uV} = .5$), where the size performance is typically somewhat worse. Therefore, for size, we restrict attention to $\Pi_1 = .01$. In contrast to size, power properties do of course strongly depend on Π_1, with usually little power for small Π_1. Therefore, for power, we restrict attention to $\Pi_1 = .5$. While the size properties of the K–test appear to be better when $k = 10$, the effect of the number of instruments is mixed for the GEL–type tests. The power results for $\theta_0 = -1$ and 1 are qualitatively identical and therefore we restrict attention to the former. All results not reported here are available upon request.

Figures I.1–4 contain size and Figures II. 1–4, III.1–4 power curves of the LM_{EL}, K and $\widehat{K}_{GEL}$ (referred to as Khat$_{EL}$ in the Figures) tests as functions of the smoothing parameters S_n for the cases

$$\begin{aligned}
\text{Figure I} \quad &: \quad k = 2, 10,\ \Pi_1 = .01,\ \rho_{uV} = .5,\ \phi = .5, .9,\ \theta_0 = 0 \text{ (size)},\\
\text{Figure II} \quad &: \quad k = 2, 10,\ \Pi_1 = .5,\ \rho_{uV} = .5,\ \phi = .5, .9,\ \theta_0 = -1 \text{ (power, AR(1)–case)},\\
\text{Figure III} \quad &: \quad k = 2, 10,\ \Pi_1 = .5,\ \rho_{uV} = .5,\ \nu = .5, .9,\ \theta_0 = -1 \text{ (power, MA(1)–case)}.
\end{aligned}$$

For convenience, at $S_n = 0$ we report the results for the unsmoothed i.i.d. versions of the statistics.

We first discuss the **size** results. As to be expected, all tests are typically size–distorted in the time series models considered here when there is no smoothing. The higher the autoregressive coefficient ϕ the higher the size–distortion, e.g. compare Figures I.1–2 and I.3–4. ERPs are typically decreasing functions of S_n for all tests in the study and in most cases the maximum smoothing number $S_n = 10$ considered here is enough to reduce ERPs to about the nominal level or even less. However, for various scenarios with few instruments, Otsu's (2003) test continues to overreject even for $S_n = 10$, see Figures I.1 and especially I.3, where $k = 2$. Many times, the ERPs of the LM_{EL} test decrease fastest as a function of S_n and then seem to level out at about the nominal level, see Figures I.1 and 3. This is a desirable property because it makes the test least dependent on the choice of S_n.

Next the **power** results are summarized. It seems that increasing k has a negative impact on the power properties of $\widehat{K}_{GEL}$ and K (see, e.g., Figures II.1–2 and II.3–4). On the other hand, for LM_{EL}, the effect of k on power is mixed and seems to depend on the bandwidth S_n. See, e.g., Figures II.1–2, where for small bandwidths $S_n \leq 2$ power of LM_{EL} is smaller for $k = 10$ but bigger for larger bandwidths. While increasing the autoregressive coefficient ϕ generally seems to have a negative impact on power (Figure II), the impact of the moving average parameter ν seems to be minor (Figure III). While for $\widehat{K}_{GEL}$ and K an increase in S_n generally leads to a reduction in ERPs, the effect of the bandwidth on the power of LM_{EL} depends on the scenario. For example, for $k = 10$ and $\phi = .5$, power of LM_{EL} increases in S_n (Figure II.2). While for $k = 2$ the K test tends to have best power properties (for basically all S_n values considered here), the LM_{EL} test seems to be the winner in the many instruments case $k = 10$ with oftentimes huge power gains for bandwidths $S_n \geq 3$. GS found that the comparative advantage of GEL based tests in i.i.d. simulations occur in situations with thick tailed or asymmetric error distributions. Here we find that even with normal errors, GEL–based tests can outperform the K–test, depending on the scenario, most crucially the number of instruments.

In summary we find that both the finite–sample size and power properties of the tests based on the new statistic LM_{EL} are very competitive.

Appendix

Additional notation is given and then the assumptions for Theorem 1 are stated.

For the proof of Theorem 1, consistency of $\widehat{\Delta}(\theta_0)/2$ in (2.3) for the long–run variance matrix $\Delta(\theta_0)$ is essential. To show consistency of $\widehat{\Delta}(\theta_0)/2$, we assume consistency of the classical Bartlett kernel HAC estimator (which holds under appropriate assumptions given in Andrews (1991, Proposition 1)) and then show that the HAC estimator differs from $\widehat{\Delta}(\theta_0)/2$ by a $o_p(1)$ term only. The latter is similar to Lemmata 2.1 and A.3 in Smith (2001, 2005). The same procedure can be applied to other long–run variance expressions, such as $\Delta_A(\theta_0)$, defined in M_{θ_0}(vii) below and its corresponding estimator $\widehat{\Delta}_A(\theta_0)/2$, where

$$\widehat{\Delta}_A(\theta_0) := S_n \sum_{i=1}^{n} (vecG_{inA}(\theta_0))g_{in}'(\theta_0)/n,$$

$G_{inA}(\theta)$ is defined by $G_{in}(\theta) = (G_{inA}(\theta), G_{inB}(\theta))$ for $G_{inA}(\theta) \in \mathbb{R}^{k\times p_A}$ and $G_{inB}(\theta) \in \mathbb{R}^{k\times p_B}$, see eq. (2.8). We now give the details.

In (2.13), decompose $G_i(\theta)$ into $(G_{iA}(\theta), G_{iB}(\theta))$, where $G_{iA}(\theta) \in \mathbb{R}^{k\times p_A}$ and $G_{iB}(\theta) \in \mathbb{R}^{k\times p_B}$ and analogously, decompose $\widehat{G}(\theta)$ into $(\widehat{G}_A(\theta), \widehat{G}_B(\theta))$.

Denote by k^* the Bartlett kernel given by

$$k^*(x) := 1 - |x/2| \text{ if } |x| \leq 2 \text{ and } k^*(x) = 0 \text{ otherwise.}$$

The Bartlett kernel is essentially the convolution of the truncated kernel, in fact, $k^*(x) = \int k(x-y)k(y)dy/2$, see Smith (2001, Example 2.1). The Bartlett HAC estimator of the long–run covariance between sequences of mean zero random vectors $r = (r_i)_{i=1,...,n}$ and $s = (s_i)_{i=1,...,n}$, is given by

$$\widetilde{J}_n(r,s) := \sum_{j=-n+1}^{n-1} k^*(j/S_n)\widetilde{\Gamma}_j(r,s), \text{ where}$$

$$\widetilde{\Gamma}_j(r,s) := \begin{cases} \sum_{i=j+1}^{n} r_i s_{i-j}'/n & \text{for } j \geq 0, \\ \sum_{i=-j+1}^{n} r_{i+j} s_i'/n & \text{for } j < 0, \end{cases}$$

see Andrews (1991, eq. (3.2)). Under certain assumptions, that include stationarity, it can be shown that (see Andrews (1991, Assumption A, Proposition 1))

$$\widetilde{J}_n(g_i, g_i) \to_p \Delta,\ \widetilde{J}_n(vecG_{iA}, g_i) \to_p \Delta_A, \tag{A.1}$$

where the argument θ_0 was left out to simplify notation. Below it is shown that the Bartlett HAC estimator and $\widehat{\Delta}(\theta_0)/2$ have the same probability limit.[7] Therefore, assuming (A.1) and some technicalities,

[7] Note that the assumptions $\widetilde{J}_n(vecG_{iA}, g_i) \to_p \Delta_A$ and $\widetilde{J}_n(vec(G_{iA} - EG_{iA}), g_i) \to_p \Delta_A$ are equivalent under weak conditions, for example under stationarity. Therefore, for consistency of the HAC estimator the possibly non–zero mean of $vecG_{iA}$ does not matter as long as $Eg_i = 0$. More precisely, it can be shown that under stationarity

$$\widetilde{J}_n(vecG_{iA}, g_i) - \widetilde{J}_n(vec(G_{iA} - EG_{iA}), g_i) = \widetilde{J}_n(vecEG_{iA}, g_i) \to_p 0.$$

This can be shown by establishing that for any $s = 1, ..., p_A k$ and $t = 1, ..., k$ and for some $c < \infty$ it holds that $(E\widetilde{J}_n(vecEG_{iA}, g_i))_{s,t} = 0$ and $(n/S_n^2)E(\widetilde{J}_n(vecEG_{iA}, g_i))_{s,t}^2 \leq c$, see Hannan (1970, p.280) for similar calculations. Because by assumption $(n/S_n^2) \to \infty$, the latter implies consistency.

$\widehat{\Delta}(\theta_0)/2$ is consistent for the long–run variance $\Delta(\theta_0)$. The same statement is true for $\Delta_A(\theta_0)$ and its estimator.

A.1 Assumptions

The assumptions of Theorem 1 are now stated and discussed. For the asymptotic distribution of $GELR_\rho$ the following assumptions are made. Here Z denotes the set of integer numbers.

Assumption $\mathbf{M}_{\theta_0}$: Suppose **(i)** $\max_{1\leq i\leq n}||g_i(\theta_0)|| = o_p(S_n^{-1}n^{1/2})$; **(ii)** for $S_n \to \infty$ and $S_n = o(n^{1/2})$ we have $\widetilde{J}_n((g_i(\theta_0)),(g_i(\theta_0))) \to_p \Delta(\theta_0) > 0$; $\sup_{i,j\geq 1} E||g_i(\theta_0)g_j'(\theta_0)|| < \infty$, $\sup_{i\in Z} n^{-1}\sum_{j=1}^n E||g_{j+i}(\theta_0)g_i'(\theta_0)|| = o(1)$, $S_n n^{-1}\sum_{i=1}^n ||g_{in}(\theta_0)g_{in}(\theta_0)'|| = O_p(1)$; **(iii)** $\Psi_n(\theta_0) \to_d \Psi(\theta_0)$, where $\Psi(\theta_0) \equiv N(0,\Delta(\theta_0))$.

To describe the asymptotic distribution of the statistics $LM_\rho(\theta_0)$ and $S_\rho(\theta_0)$, we need the following additional assumptions. For notational simplicity, the argument θ_0 is left out in M$_\theta$(v)–(vii) and in the following discussion.

Assumption $\mathbf{M}_{\theta_0}$: (cont.)

$$\begin{aligned} \textbf{(iv) } & M_{1n}(\theta_0) \,:\, = (\partial m_{1n}/\partial\theta)|_{\theta=\theta_0} \to M_1(\theta_0) := (\partial m_1/\partial\theta)|_{\theta=\theta_0} \in \mathbb{R}^{k\times p}, \\ & E\widehat{G}(\theta_0) = n^{-1/2}M_{1n}(\theta_0) + (0, M_2(\beta_0)) \to (0, M_2(\beta_0)); \end{aligned} \tag{A.2}$$

(v) $\widetilde{J}_n((vecG_{iA}),(g_i)) \to_p \Delta_A$ (Δ_A is defined in (vii)), $\sup_{i,j\geq 1} E||vecG_{iA}g_j'|| < \infty$, $\sup_{i\in Z} n^{-1}\sum_{j=1}^n E||vecG_{j+iA}g_i'|| = o(1)$, $\widehat{G}_B := n^{-1}\sum_{i=1}^n G_{iB} \to_p E\widehat{G}_B$; **(vi)** $\max_{1\leq i\leq n}||G_{iA}|| = o_p(S_n^{-1}n^{1/2})$, $S_n n^{-1}\sum_{i=1}^n ||vecG_{inA}g_{in}'|| = O_p(1)$, $\max_{1\leq i\leq n}||G_{iB}|| = o_p(S_n^{-1}n)$, $S_n n^{-3/2}\sum_{i=1}^n ||vecG_{inB}g_{in}'|| = o_p(1)$; **(vii)** $n^{-1/2}\sum_{i=1}^n ((vec(G_{iA} - EG_{iA}))', g_i')' \to_d N(0,V)$, where

$$V := \lim_{n\to\infty} var(n^{-1/2}\sum_{i=1}^n (vecG_{iA}', g_i')') \in \mathbb{R}^{k(p_A+1)\times k(p_A+1)}$$

has full column rank. Decompose V into

$$V = \begin{pmatrix} \Delta_{AA} & \Delta_A \\ \Delta_A' & \Delta \end{pmatrix}, \text{ where } \Delta_{AA} \in \mathbb{R}^{p_Ak\times p_Ak}.$$

A discussion of Assumption M$_{\theta_0}$ now follows. Assuming $S_n = cn^\alpha$ for positive constants c and $\alpha < 1/2$, a sufficient condition for M$_{\theta_0}$(i) is given by the moment condition $\sup_{i\geq 1} E||g_i(\theta_0)||^\xi < \infty$ for some $\xi > 2/(1-2\alpha)$, see GS, eq. (2.4), for a similar statement and a proof. Analogous sufficient conditions can be formulated for M$_{\theta_0}$(vi).

The high level assumption $\widetilde{J}_n((g_i),(g_i)) \to_p \Delta$ in M$_{\theta_0}$(ii) is satisfied under sufficient conditions given in Andrews (1991, Proposition 1) which include stationarity. We prefer the high level assumption to the sufficient condition because it may hold even when the data are not stationary, e.g. in cases of non–identically distributed data. M_{θ_0}(ii) then guarantees that $\widehat{\Delta} \rightarrow_p 2\Delta$, see Lemma 2 below. The technical assumption $\sup_{i \in Z} n^{-1} \sum_{j=1}^{n} E||g_{j+i}g_i'|| = o(1)$ can be interpreted as some mild form of mixing, see also M_{θ_0}(v), and is needed in the proof of Lemma 2. The assumption $S_n n^{-1} \sum_{i=1}^{n} ||g_{in}g_{in}'|| = O_p(1)$ is needed in the proof of Theorem 1(i) to show that $S_n \sum_{i=1}^{n} (\rho_2(\widetilde{\lambda}' g_{in}) + 1) g_{in}g_{in}'/n$ is $o_p(1)$. A sufficient condition in terms of the moment functions g_i is $\sup_{i \in Z} n^{-1} \sum_{j=1}^{n} E||g_{j+i}(\theta_0)g_i'(\theta_0)|| = O(S_n^{-1})$, which is a stronger form of mixing condition.[8] Similar comments apply for the analogous assumptions in M_{θ_0}(v) and (vi), parts of which are needed in deriving (A.10).

M_{θ_0}(iii) is the "high level" assumption also used in Stock and Wright (2000).

A sufficient condition for M_{θ_0}(iv) is given by: For some open neighborhood $\mathcal{M} \subset \Theta$ of θ_0, $\widehat{g}(\cdot)$ is differentiable at $\overline{\theta}$ a.s. for each $\overline{\theta} \in \mathcal{M}$, $\widehat{g}(\overline{\theta})$ is integrable for all $\overline{\theta} \in \mathcal{M}$ (with respect to the probability measure), $\sup_{\overline{\theta} \in \mathcal{M}} ||\widehat{G}(\overline{\theta})||$ is integrable, $m_{1n} \in C^1(\Theta)$ and $M_{1n}(\cdot)$ converges uniformly on Θ to some function. These conditions allow the interchange of the order of integration and differentiation in Assumption ID, i.e. $(\partial E\widehat{g}/\partial\theta)|_{\theta=\theta_0} = E\widehat{G}(\theta_0)$. Note that by ID the limit matrix $(0, M_2(\beta_0))$ is singular of rank p_B.

Let $\widehat{G}_n(\theta) := n^{-1} \sum_{i=1}^{n} G_{in}(\theta)$ and decompose $\widehat{G}_n(\theta)$ as $(\widehat{G}_{nA}(\theta), \widehat{G}_{nB}(\theta))$, where $\widehat{G}_{nA}(\theta) \in \mathbb{R}^{k \times p_A}$ and $\widehat{G}_{nB}(\theta) \in \mathbb{R}^{k \times p_B}$. The assumption $\max_{1 \leq i \leq n} ||G_{iB}|| = o_p(S_n^{-1} n)$ in M_{θ_0}(vi) ensures that $\widehat{G}_{nB} - 2\widehat{G}_B = o_p(1)$. This can be shown along the lines of Lemma 1.

Besides technical assumptions, M_{θ_0} essentially states that the HAC estimator $\widetilde{J}_n$ is consistent (parts (ii) and (v)) and that a CLT holds for $((vec(G_{iA} - EG_{iA}))', g_i')'$, (parts (iii) and (vii)). For the latter, primitive sufficient conditions based on mixing properties can be stated along the lines of Wooldridge and White (1988). The CLT assumption is very closely related to Assumption 1 in Kleibergen (2001).

A.2 Proofs

The next lemmata are helpful in the proof of the main result. Note that the assumptions made in Lemma 1 are implied by M_{θ_0}(i), (iii), (vi) and (vii), e.g. $\widehat{G}_A(\theta_0) = O_p(n^{-1/2})$ follows from M_{θ_0}(vii) and eq. (A.2). Recall $\widehat{G}_{nA}(\theta) = n^{-1} \sum_{i=1}^{n} G_{inA}(\theta)$.

Lemma 1 *Suppose $S_n \rightarrow \infty$ and $S_n = o(n^{1/2})$.*

$$\text{If } \max_{1 \leq i \leq n} ||g_i|| = o_p(S_n^{-1} n^{1/2}), \widehat{g} = O_p(n^{-1/2}) \text{ then } n^{1/2}(\widehat{g}_n - 2\widehat{g}) = o_p(1).$$

$$\text{If } \max_{1 \leq i \leq n} ||G_{iA}|| = o_p(S_n^{-1} n^{1/2}), \widehat{G}_A = O_p(n^{-1/2}) \text{ then } n^{1/2}(\widehat{G}_{nA} - 2\widehat{G}_A) = o_p(1),$$

where again θ_0 is left out to simplify the notation.

[8] The tedious proof of this statement is along the exact same lines as the proof of Lemma 2.

Proof: For the first equation tedious but straightforward calculations imply that

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n} g_{in} &= n^{-1}\sum_{i=1}^{n} S_n^{-1}\sum_{j=i-n}^{i-1} k(j/S_n)g_{i-j} = n^{-1}\sum_{i=1}^{n} S_n^{-1}\sum_{j=\max(i-n,-S_n)}^{\min(i-1,S_n)} g_{i-j} \\
&= n^{-1}\sum_{i=S_n+1}^{n-S_n}\frac{2S_n+1}{S_n}g_i + n^{-1}\sum_{i=1}^{S_n}\frac{S_n+i}{S_n}g_i + n^{-1}\sum_{i=n-S_n+1}^{n}\frac{n-i+S_n+1}{S_n}g_i \\
&= 2n^{-1}\sum_{i=1}^{n} g_i + n^{-1}\sum_{i=S_n+1}^{n-S_n}\frac{1}{S_n}g_i + \\
&\quad n^{-1}\sum_{i=1}^{S_n}\frac{i-S_n}{S_n}g_i + n^{-1}\sum_{i=n-S_n+1}^{n}\frac{-S_n+n-i+1}{S_n}g_i \\
&= 2n^{-1}\sum_{i=1}^{n} g_i + o_p(n^{-1/2}),
\end{aligned}$$

where the last equation uses $\max_{1\leq i\leq n}||g_i|| = o_p(S_n^{-1}n^{1/2})$ and $\widehat{g} = O_p(n^{-1/2})$ to show that the remainder terms are $o_p(n^{-1/2})$. The proof of the second equation can be derived in exactly the same way. □

It is now shown that under M_{θ_0}, $\widehat{\Delta}/2$ and $\widehat{\Delta}_A/2$ are consistent for Δ and Δ_A. The first part of the following lemma is similar to Lemma A.3 in Smith (2001). Note that the assumptions in the Lemma are part of M_{θ_0}(ii) and (v).

Lemma 2 *For $S_n \to \infty$ assume $S_n = o(n^{1/2})$. If $\sup_{i,j\geq 1} E||g_i g_j'|| < \infty$ and $\sup_{i\in Z} n^{-1}\sum_{j=1}^{n} E||g_{j+i}g_i'|| = o(1)$ then*

$$\widehat{\Delta} - 2\widetilde{J}_n((g_i),(g_i)) = o_p(1).$$

If $\sup_{i,j\geq 1} E||vecG_{iA}g_j'|| < \infty$ and $\sup_{i\in Z} n^{-1}\sum_{j=1}^{n} E||vecG_{j+iA}g_i'|| = o(1)$ then

$$\widehat{\Delta}_A - 2\widetilde{J}_n((vecG_{iA}),(g_i)) = o_p(1), \tag{A.3}$$

where the argument θ_0 is left out to simplify the notation.

Proof: For the first statement easy calculations lead to

$$\begin{aligned}
2\widetilde{J}_n((g_i),(g_i)) - \widehat{\Delta} &= \sum_{i=-n+1}^{n-1} n^{-1}\sum_{j=\max(1,1-i)}^{\min(n,n-i)} k_{ij}g_{j+i}g_j' \text{ for} \\
k_{ij} &:= 2k^*(i/S_n) - S_n^{-1}\sum_{l=1-j}^{n-j} k((l-i)/S_n)k(l/S_n).
\end{aligned}$$

Using the definitions of k and k^* tedious calculations show that for $0 \leq i < S_n$

$$k_{ij} = \begin{cases} S_n^{-1}(S_n - i - j) & \text{for } 1 \leq j \leq S_n - i + 1 \\ -S_n^{-1} & \text{for } S_n - i + 1 < j \leq n - S_n \\ -S_n^{-1}(n - j - S_n + 1) & \text{for } n - S_n < j \leq n - i \end{cases}$$

that for $-S_n < i < 0$

$$k_{ij} = \begin{cases} S_n^{-1}(S_n - j) & \text{for } 1 - i \leq j \leq S_n + 1 \\ -S_n^{-1} & \text{for } S_n + 1 < j < n - S_n - i \\ S_n^{-1}(S_n + i - n + j - 1) & \text{for } n - S_n - i \leq j \leq n \end{cases}$$

that $k_{ij} = -S_n^{-1}$ if $S_n \leq |i| \leq 2S_n$ and that $k_{ij} = 0$ otherwise. Using the moment assumptions, it then follows that $2\widetilde{J}_n((g_i),(g_i)) - \widehat{\Delta}$ reduces to $o_p(1)$ expressions, for example, by Markov's inequality $\Pr(|| -\sum_{i=S_n}^{2S_n} n^{-1}S_n^{-1}\sum_{j=1}^{n-i} g_{j+i}g_j'|| > \varepsilon)$ can be bounded by

$$\varepsilon^{-1}S_n^{-1}\sum_{i=S_n}^{2S_n} n^{-1}\sum_{j=1}^{n-i} E||g_{j+i}g_j'|| \leq \varepsilon^{-1}S_n^{-1}\sum_{i=S_n}^{2S_n} \sup_{i\in Z} n^{-1}\sum_{j=1}^{n} E||g_{j+i}g_j'|| = o(1)$$

and similarly for the other summands. The proof of the second claim is completely analogous and therefore omitted. □

Given the results in Lemma 1 and consistency of $\widehat{\Delta}/2$ and $\widehat{\Delta}_A/2$, the proof of Theorem 1 is along the same lines as the proofs of Theorems 3 and 4 in GS.

As in GS, the proof hinges on the following two lemmas. Let $c_n := S_n n^{-1/2}\max_{1\leq i\leq n}||g_{in}(\theta_0)||$. Let $\Lambda_n := \{\lambda \in \mathbb{R}^k : ||\lambda|| \leq S_n n^{-1/2}c_n^{-1/2}\}$ if $c_n \neq 0$ and $\Lambda_n = \mathbb{R}^k$ otherwise.

Lemma 3 *Assume* $\max_{1\leq i\leq n}||g_i(\theta_0)|| = o_p(S_n^{-1}n^{1/2})$. *Then* $\sup_{\lambda\in\Lambda_n, 1\leq i\leq n}|\lambda' g_{in}(\theta_0)| \to_p 0$ *and* $\Lambda_n \subset \widehat{\Lambda}_n(\theta_0)$ *w.p.a.1.*

Proof: The case $c_n = 0$ is trivial and thus w.l.o.g. $c_n \neq 0$ can be assumed. Note that $||g_{in}(\theta_0)|| \leq S_n^{-1}\sum\limits_{j=i-n}^{i-1} k(j/S_n)||g_{i-j}(\theta_0)||$ and thus by the definition of $k(\cdot)$

$$\begin{aligned}\max_{1\leq i\leq n}||g_{in}(\theta_0)|| &\leq \max_{1\leq i\leq n} S_n^{-1}\sum_{j=\max(-S_n,i-n)}^{\min(S_n,i-1)}||g_{i-j}(\theta_0)|| \\ &\leq (2S_n+1)S_n^{-1}\max_{1\leq i\leq n}||g_i(\theta_0)|| = o_p(S_n^{-1}n^{1/2}).\end{aligned}$$

Therefore, $c_n = o_p(1)$ and the first part of the statement follows from

$$\sup_{\lambda\in\Lambda_n, 1\leq i\leq n}|\lambda' g_{in}(\theta_0)| \leq S_n n^{-1/2}c_n^{-1/2}\max_{1\leq i\leq n}||g_{in}(\theta_0)|| =$$
$$n^{-1/2}S_n c_n^{-1/2}n^{1/2}S_n^{-1}c_n = c_n^{1/2} = o_p(1),$$

which also immediately implies the second part. □

In the next lemma $\lambda_{\min}(A)$ denotes the smallest eigenvalue in absolute value of the matrix A.

Lemma 4 *Suppose* $\max_{1\leq i\leq n}||g_i(\theta_0)|| = o_p(S_n^{-1}n^{1/2})$, $\lambda_{\min}(\widehat{\Delta}(\theta_0)) \geq \varepsilon$ *w.p.a.1 for some* $\varepsilon > 0$, $\widehat{g}_n(\theta_0) = O_p(n^{-1/2})$ *and Assumption* ρ *holds.*

Then $\lambda(\theta_0) \in \widehat{\Lambda}_n(\theta_0)$ *satisfying* $\widehat{P}_\rho(\theta_0, \lambda(\theta_0)) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda)$ *exists w.p.a.1,* $\lambda(\theta_0) = O_p(S_n n^{-1/2})$ *and* $\sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda) = O_p(S_n n^{-1})$.

Proof: W.l.o.g. $c_n \neq 0$ and thus Λ_n can be assumed compact. Let $\lambda_{\theta_0} \in \Lambda_n$ be such that $\widehat{P}_\rho(\theta_0,\lambda_{\theta_0}) = \max_{\lambda\in\Lambda_n}\widehat{P}_\rho(\theta_0,\lambda)$. Such a $\lambda_{\theta_0} \in \Lambda_n$ exists w.p.a.1 because a continuous function takes on its maximum on a compact set and by Lemma 3 and Assumption ρ, $\widehat{P}_\rho(\theta_0,\lambda)$ (as a function in λ for fixed θ_0) is C^2 on some open neighborhood of Λ_n w.p.a.1. It is now shown that actually $\widehat{P}_\rho(\theta_0,\lambda_{\theta_0}) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda)$

w.p.a.1 which then proves the first part of the lemma. By a second order Taylor expansion around $\lambda = 0$, there is a $\lambda^*_{\theta_0}$ on the line segment joining 0 and λ_{θ_0} such that for some positive constants C_1 and C_2

$$\begin{aligned} 0 &= S_n \widehat{P}_\rho(\theta_0, 0) \leq S_n \widehat{P}_\rho(\theta_0, \lambda_{\theta_0}) \\ &= -2S_n \lambda'_{\theta_0} \widehat{g}_n(\theta_0) + \lambda'_{\theta_0}[S_n \sum_{i=1}^n \rho_2(\lambda^{*\prime}_{\theta_0} g_{in}(\theta_0)) g_{in}(\theta_0) g_{in}(\theta_0)'/n]\lambda_{\theta_0} \\ &\leq -2S_n \lambda'_{\theta_0} \widehat{g}_n(\theta_0) - C_1 \lambda'_{\theta_0} \widehat{\Delta}(\theta_0)\lambda_{\theta_0} \leq 2S_n ||\lambda_{\theta_0}||\ ||\widehat{g}_n(\theta_0)|| - C_2 ||\lambda_{\theta_0}||^2 \end{aligned} \tag{A.4}$$

w.p.a.1, where the second inequality follows as $\max_{1 \leq i \leq n} \rho_2(\lambda^{*\prime}_{\theta_0} g_{in}(\theta_0)) < -1/2$ w.p.a.1 from Lemma 3, continuity of $\rho_2(\cdot)$ at zero and $\rho_2 = -1$. The last inequality follows from $\lambda_{\min}(\widehat{\Delta}(\theta_0)) \geq \varepsilon > 0$ w.p.a.1. Now, (A.4) implies that $(C_2/2)||\lambda_{\theta_0}|| \leq S_n ||\widehat{g}_n(\theta_0)||$ w.p.a.1, the latter being $O_p(S_n n^{-1/2})$ by assumption. It follows that $\lambda_{\theta_0} \in int(\Lambda_n)$ w.p.a.1. To prove this, let $\epsilon > 0$. Because $\lambda_{\theta_0} = O_p(S_n n^{-1/2})$ and $c_n = o_p(1)$, there exists $M_\epsilon < \infty$ and $n_\epsilon \in \mathbb{N}$ such that $\Pr(||S_n^{-1} n^{1/2} \lambda_{\theta_0}|| \leq M_\epsilon) > 1 - \epsilon/2$ and $\Pr(c_n^{-1/2} > M_\epsilon) > 1 - \epsilon/2$ for all $n \geq n_\epsilon$. Then $\Pr(\lambda_{\theta_0} \in int(\Lambda_n)) = \Pr(||S_n^{-1} n^{1/2} \lambda_{\theta_0}|| < c_n^{-1/2}) \geq \Pr((||S_n^{-1} n^{1/2} \lambda_{\theta_0}|| \leq M_\epsilon) \wedge (c_n^{-1/2} > M_\epsilon)) > 1 - \epsilon$ for $n \geq n_\epsilon$.

Hence, the FOC for an interior maximum $(\partial \widehat{P}_\rho / \partial \lambda)(\theta_0, \lambda) = 0$ hold at $\lambda = \lambda_{\theta_0}$ w.p.a.1. By Lemma 3, $\lambda_{\theta_0} \in \widehat{\Lambda}_n(\theta_0)$ w.p.a.1 and thus by concavity of $\widehat{P}_\rho(\theta_0, \lambda)$ (as a function in λ for fixed θ_0) and convexity of $\widehat{\Lambda}_n(\theta_0)$ it follows that $\widehat{P}_\rho(\theta_0, \lambda_{\theta_0}) = \sup_{\lambda \in \widehat{\Lambda}_n(\theta_0)} \widehat{P}_\rho(\theta_0, \lambda)$ w.p.a.1 which implies the first part of the lemma. From above $\lambda_{\theta_0} = O_p(S_n n^{-1/2})$. Thus the second and by (A.4) the third parts of the lemma follow. □

Proof of Theorem 1 (i): Lemma 4 implies that the FOC

$$n^{-1} \sum_{i=1}^n \rho_1(\lambda' g_{in}(\theta)) g_{in}(\theta) = 0 \tag{A.5}$$

have to hold at $(\theta_0, \lambda_0 := \lambda(\theta_0))$ w.p.a.1. Expanding the FOC in λ around 0, there exists a mean value $\widetilde{\lambda}$ between 0 and λ_0 (that may be different for each row) such that

$$0 = -\widehat{g}_n(\theta_0) + [S_n \sum_{i=1}^n \rho_2(\widetilde{\lambda}' g_{in}(\theta_0)) g_{in}(\theta_0) g_{in}(\theta_0)'/n] S_n^{-1} \lambda_0 = -\widehat{g}_n(\theta_0) - \widehat{\Delta}_{\widetilde{\lambda}} S_n^{-1} \lambda_0,$$

where the matrix $\widehat{\Delta}_{\widetilde{\lambda}}$ has been implicitly defined. Because $\lambda_0 = O_p(S_n n^{-1/2})$, Lemma 3 and Assumption ρ imply that $\max_{1 \leq i \leq n} |\rho_2(\widetilde{\lambda}' g_{in}(\theta_0)) + 1| \to_p 0$. By Assumption M_{θ_0}(ii) and Lemma 2 it follows that $\widehat{\Delta}_{\widetilde{\lambda}} \to_p 2\Delta(\theta_0) > 0$ and thus $\widehat{\Delta}_{\widetilde{\lambda}}$ is invertible w.p.a.1 and $(\widehat{\Delta}_{\widetilde{\lambda}})^{-1} \to_p \Delta(\theta_0)^{-1}/2$. Therefore

$$S_n^{-1} \lambda_0 = -(\widehat{\Delta}_{\widetilde{\lambda}})^{-1} \widehat{g}_n(\theta_0) \tag{A.6}$$

w.p.a.1. Inserting this into a second order Taylor expansion for $\widehat{P}(\theta, \lambda)$ (with mean value λ^* as in (A.4) above) it follows that w.p.a.1

$$S_n^{-1} n \widehat{P}_\rho(\theta_0, \lambda_0) = 2n \widehat{g}_n(\theta_0)' \widehat{\Delta}^{-1}_{\widetilde{\lambda}} \widehat{g}_n(\theta_0) - n \widehat{g}_n(\theta_0)' \widehat{\Delta}^{-1}_{\widetilde{\lambda}} \widehat{\Delta}_{\lambda^*} \widehat{\Delta}^{-1}_{\widetilde{\lambda}} \widehat{g}_n(\theta_0). \tag{A.7}$$

By Lemma 1 and M_{θ_0}(iii) $n^{1/2} \widehat{g}_n(\theta_0) = 2n^{1/2} \widehat{g}(\theta_0) + o_p(1) \to_d 2N(0, \Delta(\theta_0))$ and therefore $S_n^{-1} n \widehat{P}_\rho(\theta_0, \lambda_0)/2 \to_d \chi^2(k)$. □

Proof of Theorem 1 (ii): Define $D^* := D_\rho(\theta_0)\Lambda$ where the $p \times p$ diagonal matrix $\Lambda := diag(n^{1/2}, ..., n^{1/2}, 1, ..., 1)$ has first p_A diagonal elements equal to $n^{1/2}$ and the remainder equal to unity. Then, (in the remainder of the proof the argument θ_0 is left out for notational simplicity) it follows that

$$LM_\rho = n\widehat{g}_n'\widehat{\Delta}^{-1}D^*(D^{*\prime}\widehat{\Delta}^{-1}D^*)^{-1}D^{*\prime}\widehat{\Delta}^{-1}\widehat{g}_n/2. \tag{A.8}$$

It follows from (A.6) and $n^{1/2}\widehat{g}_n = O_p(1)$ that

$$S_n^{-1}n^{1/2}\lambda_0 = -\Delta^{-1}n^{1/2}\widehat{g}_n/2 + o_p(1) \tag{A.9}$$

and therefore the statement of the theorem involving S_ρ follows immediately from the one for LM_ρ. Therefore, only the statistic LM_ρ is dealt with using its representation in eq. (A.8).

First, it is shown that the matrix D^* is asymptotically independent of $n^{1/2}\widehat{g}_n$. By a mean value expansion about 0 it follows that $\rho_1(\lambda_0'g_{in}) = -1 + \rho_2(\xi_i)g_{in}'\lambda_0$ for a mean value ξ_i between 0 and $\lambda_0'g_{in}$ and thus by (2.9), (A.9) and the definition of Λ it follows that (modulo $o_p(1)$ terms)

$$\begin{aligned} D^* &= -n^{-1}\sum_{i=1}^n (n^{1/2}G_{inA}, G_{inB}) - S_n n^{-3/2}\sum_{i=1}^n [\rho_2(\xi_i)(n^{1/2}G_{inA}, G_{inB})g_{in}'\Delta^{-1}n^{1/2}\widehat{g}_n]/2 \\ &= -(n^{-1/2}\sum_{i=1}^n G_{inA} - S_n n^{-1}\sum_{i=1}^n G_{inA}g_{in}'\Delta^{-1}n^{1/2}\widehat{g}_n/2, 2M_2(\beta_0)), \end{aligned} \tag{A.10}$$

where for the last equality we use (A.2) and Assumptions M$_{\theta_0}$(v)-(vi). By Assumption M$_\theta$(v) and eq. (A.3) it follows that $\widehat{\Delta}_A = S_n n^{-1}\sum_{i=1}^n vec(G_{inA})g_{in}'/2 \to_p \Delta_A$ and thus

$$vec(D^*, n^{1/2}\widehat{g}_n) = w_1 + Mv + o_p(1),$$

where $w_1 := vec(0, -2M_2(\beta_0), 0) \in \mathbb{R}^{kp_A + kp_B + k}$ and

$$M := \begin{pmatrix} -I_{kp_A} & \Delta_A\Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix},\ v := n^{-1/2}\sum_{i=1}^n \begin{pmatrix} vecG_{inA} \\ g_{in} \end{pmatrix};$$

M and v have dimensions $(kp_A + kp_B + k) \times (kp_A + k)$ and $(kp_A + k) \times 1$, respectively. By Assumption ID, M$_{\theta_0}$(vii), Lemma 1 and (A.2) it follows that $v \to_d 2N(w_2, V)$, where $w_2 := ((vecM_{1A})', 0)'$ and M_{1A} are the first p_A columns of M_1. Therefore

$$vec(D^*, n^{1/2}\widehat{g}_n) \to_d N(w_1 + 2Mw_2, 4\begin{pmatrix} \Psi & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta \end{pmatrix}), \tag{A.11}$$

where $\Psi := \Delta_{AA} - \Delta_A\Delta^{-1}\Delta_A'$ has full column rank. Eq. (A.11) proves that D^* and $n^{1/2}\widehat{g}_n$ are asymptotically independent.

The asymptotic distribution of LM_ρ is derived next. Denote by $\overline{D}$ and $\overline{g}$ the limiting normal random matrices corresponding to D^* and $n^{1/2}\widehat{g}_n$, respectively, see (A.11). Below it is shown that the function $h : \mathbb{R}^{k\times p} \to \mathbb{R}^{p\times k}$ defined by $h(D) := (D'\Delta^{-1}D)^{-1/2}D'$ for $D \in \mathbb{R}^{k\times p}$ is continuous on a set $C \subset \mathbb{R}^{k\times p}$ with $\Pr(\overline{D} \in C) = 1$. By the Continuous Mapping Theorem and M$_{\theta_0}$(v) it follows that

$$2^{-1/2}(D^{*\prime}\widehat{\Delta}^{-1}D^*)^{-1/2}D^{*\prime}\widehat{\Delta}^{-1}n^{1/2}\widehat{g}_n \to_d (\overline{D}'\Delta^{-1}\overline{D})^{-1/2}\overline{D}'\Delta^{-1}\overline{g}/2. \tag{A.12}$$

By the independence of $\overline{D}$ and $\overline{g}$, the latter random variable is distributed as ζ, where $\zeta \sim N(0, I_p)$.

Finally, the continuity claim for h is dealt with. Note that h is continuous at each D that has full column rank. It is therefore sufficient to show that $\overline{D}$ has full column rank a.s.. From (A.11) it follows that the last p_B columns of $\overline{D}$ equal $-2M_2(\beta_0)$ which has full column rank by assumption. Define $O := \{o \in \mathbb{R}^{kp_A} : \exists \widetilde{o} \in \mathbb{R}^{k \times p_A}$, s.t. $o = vec(\widetilde{o})$ and the $k \times p$–matrix $(\widetilde{o}, -2M_2(\beta_0))$ has linearly dependent columns$\}$. Clearly, O is closed and therefore Lebesgue–measurable. Furthermore, O has empty interior and thus has Lebesgue–measure 0. For the first p_A columns of $\overline{D}$, $\overline{D}_{p_A}$ say, it has been shown that $vec\overline{D}_{p_A}$ is normally distributed with full rank covariance matrix Ψ. This implies that for any measurable set $O^* \subset \mathbb{R}^{kp_A}$ with Lebesgue–measure 0, $\Pr(vec(\overline{D}_{p_A}) \in O^*) = 0$, in particular, for $O^* = O$. This proves the continuity claim for h. $\square$

Proof of Theorem 2: Let $\mu_0 := \mu(\theta_0)$. Inserting this into a second order Taylor expansion for $\widehat{P}_\rho(\theta, \mu)$ around $\mu = 0$ with mean value $\widetilde{\mu}$, cf. eq. (A.4) above,

$$\begin{aligned} S_n\widehat{P}_\rho(\theta_0, \mu_0) &= -2S_n\mu_0'\widehat{g}_n(\theta_0) + \mu_0'[S_n \sum_{i=1}^n \rho_2(\widetilde{\mu}' g_{in}(\theta_0))g_{in}(\theta_0)g_{in}(\theta_0)'/n]\mu_0 \\ &= -2S_n\mu_0'\widehat{g}_n(\theta_0) + \mu_0'\widehat{\Delta}_{\widetilde{\mu}}\mu_0, \end{aligned}$$

where $\widehat{\Delta}_{\widetilde{\mu}}$ has been implicitly defined. As in the proof of Theorem 1(ii) define $D^* := D_\rho(\theta_0)\Lambda$. Hence, we may write $\mu_0 = -S_n\widehat{\Delta}(\theta_0)^{-1}D^* \left(D^{*\prime}\widehat{\Delta}(\theta_0)^{-1}D^*\right)^{-1} D^{*\prime}\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0)$. From Assumption M_{θ_0}(ii) and Lemma 2, $\lambda_{\min}(\widehat{\Delta}(\theta_0)), \lambda_{\min}(\widehat{\Delta}(\theta_0)^{-1}) \geq \varepsilon > 0$ w.p.a.1. Therefore, as the expression in (A.12) and D^* are $O_p(1)$, it follows that $\mu_0 = O_p(S_n n^{-1/2})$. By an analogous argument to that in the proof of Lemma 4, $\mu_0 \in int(\Lambda_n)$ w.p.a.1. Therefore, Lemma 3 and Assumption ρ imply that $\max_{1 \leq i \leq n} |\rho_2(\widetilde{\mu}' g_{in}(\theta_0)) + 1| \rightarrow_p 0$ and, thus from the last part of Assumption M_{θ_0}(ii), $\widehat{\Delta}_{\widetilde{\mu}} \rightarrow_p -2\Delta(\theta_0)$. Consequently, substituting for μ_0,

$$\begin{aligned} S_n^{-1}n\widehat{P}_\rho(\theta_0, \mu_0) &= n\widehat{g}_n(\theta_0)'\widehat{\Delta}(\theta_0)^{-1}D^* \left(D^{*\prime}\widehat{\Delta}(\theta_0)^{-1}D^*\right)^{-1} D^{*\prime}\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0) + o_p(1) \\ &= 2LM_\rho(\theta_0) + o_p(1) \rightarrow_d 2\chi^2(p) \end{aligned}$$

from the proof of Theorem 1(ii) as $\widehat{\Delta} \rightarrow_p 2\Delta(\theta_0)$ and by Lemma 1 and M_{θ_0}(iii) $n^{1/2}\widehat{g}_n(\theta_0) = 2n^{1/2}\widehat{g}(\theta_0) + o_p(1) \rightarrow_d 2N(0, \Delta(\theta_0))$. The result for $S_n^{-1}n\widehat{P}_\rho(\theta_0, \tilde{\mu}(\theta_0))/2$ then also follows immediately as $\lambda(\theta_0) = -S_n\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0) + o_p(S_n n^{-1/2})$. $\square$

References

Anderson, T.W. and H. Rubin (1949): "Estimators of the parameters of a single equation in a complete set of stochastic equations", *The Annals of Mathematical Statistics* 21, 570–582.

Andrews, D.W.K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica* 59(3), 817–858.

Andrews, D.W.K. and V. Marmer (2004): "Exactly Distribution–Free Inference in Instrumental Variables Regression with Possibly Weak Instruments", unpublished manuscript.

Andrews, D.W.K., M. Moreira, and J.H. Stock (2004): "Optimal invariant similar tests for instrumental variables regression", unpublished manuscript.

Caner, M. (2003): "Exponential Tilting with Weak Instruments: Estimation and Testing", unpublished manuscript.

Chao, J.C. and N.R. Swanson (2005): "Consistent Estimation With a Large Number of Weak Instruments", forthcoming in *Econometrica.*

Dufour, J. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models", *Econometrica* 65(6), 1365–1387.

——— (2003): "Identification, Weak Instruments and Statistical Inference in Econometrics. Presidential Address to the Canadian Economics Association", *Canadian Journal of Economics* 36(4), 767-808.

Dufour, J. and M. Taamouti (2003): "Projection–based statistical inference in linear structural models with possibly weak instruments", forthcoming in *Econometrica.*

Guggenberger, P. (2003): "Econometric Essays on Generalized Empirical Likelihood, Long–memory Time Series, and Volatility", Ph.D. thesis, Yale University.

Guggenberger, P. and R.J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification", *Econometric Theory* 21(4), 667-709.

Guggenberger, P. and M. Wolf (2004): "Subsampling tests of parameter hypotheses and overidentifying restrictions with possible failure of identification", unpublished manuscript.

Hannan, E.J. (1970): "Multiple Time Series", New York: Wiley.

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 & Economic Statistics* 14(3), 262–280.

Imbens, G. (1997): "One–step estimators for over–identified Generalized Method of Moments models", *Review of Economic Studies* 64, 359–383.

Imbens, G., R.H. Spady and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models", *Econometrica* 66(2), 333–357.

Jansson, M. (2002): "Consistent Covariance Matrix Estimation for Linear Processes", *Econometric Theory* 18, 1449-1459.

Kitamura, Y. (1997): "Empirical likelihood methods with weakly dependent processes", *Annals of Statistics* 25(5), 2084–2102.

Kitamura, Y. and M. Stutzer (1997): "An information–theoretic alternative to Generalized Method of Moments estimation", *Econometrica* 65(4), 861–874.

Kleibergen, F. (2001): "Testing parameters in GMM without assuming that they are identified", forthcoming in *Econometrica.*

——— (2002): "Pivotal statistics for testing structural parameters in instrumental variables regression", *Econometrica* 70(5), 1781-1805.

——— (2004): "Testing subsets of structural parameters in the instrumental variables regression model", *Review of Economics and Statistics* 86, 418–423.

Moreira, M.J. (2003): "A Conditional Likelihood Ratio Test for Structural Models", *Econometrica* 71(4), 1027-1048.

Newey, W.K. and R.J. Smith (2004): "Higher order properties of GMM and Generalized Empirical Likelihood estimators", *Econometrica* 72(1), 219-255.

Newey, W.K. and K.D. West (1994): "Automatic Lag Selection in Covariance Matrix Estimation", *Review of Economic Studies* 61, 631-653.

Otsu, T. (2003): "Generalized Empirical Likelihood Inference under Weak Identification", unpublished manuscript.

Qin J. and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations", *Annals of Statistics* 22(1), 300-325.

Smith, R.J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation", *Economic Journal* 107(441), 503-519.

——— (2000): "Empirical Likelihood Estimation and Inference", Chapter 4 in *Applications of Differential Geometry to Econometrics*, eds. M. Salmon and P. Marriott, 119-150. Cambridge University Press: Cambridge.

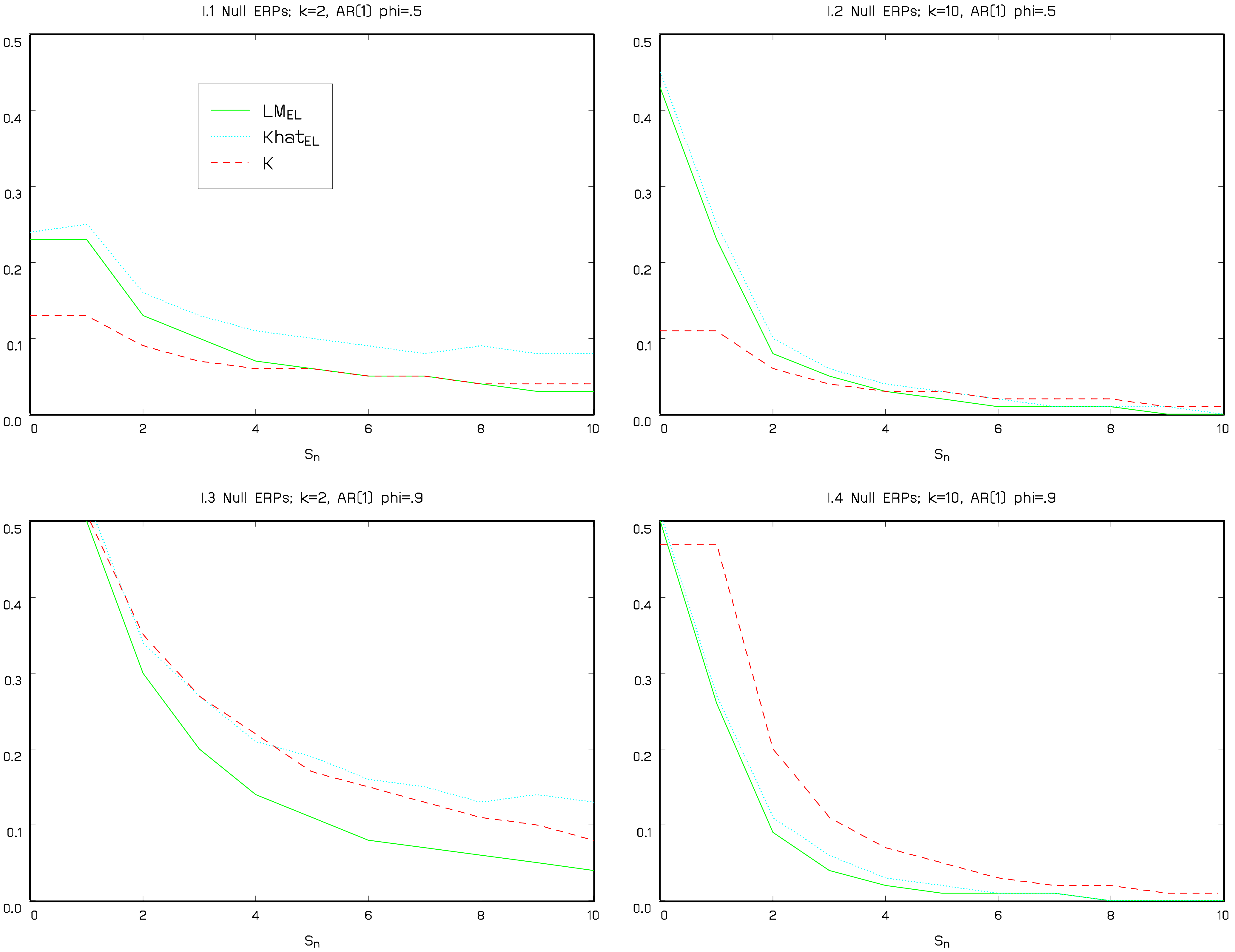
——— (2001): "GEL Criteria for Moment Condition Models", mimeo, University of Bristol. Revised version CWP 19/04, *c*emmap, I.F.S. and U.C.L. http://cemmap.ifs.org.uk/wps/cwp0419.pdf

——— (2005): "Automatic Positive Semi–Definite HAC Covariance Matrix and GMM Estimation", *Econometric Theory* 21(1), 158–170.

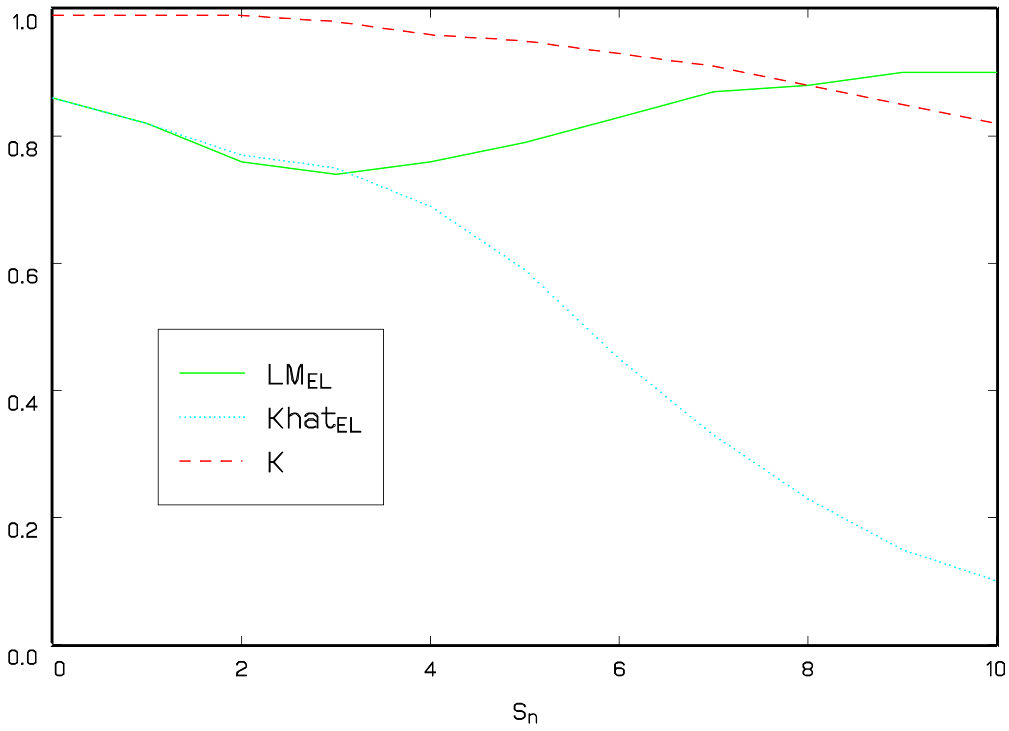
Staiger, D. and J.H. Stock (1997): "Instrumental Variables Regression With Weak Instruments", *Econometrica* 65(3), 557-586.

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

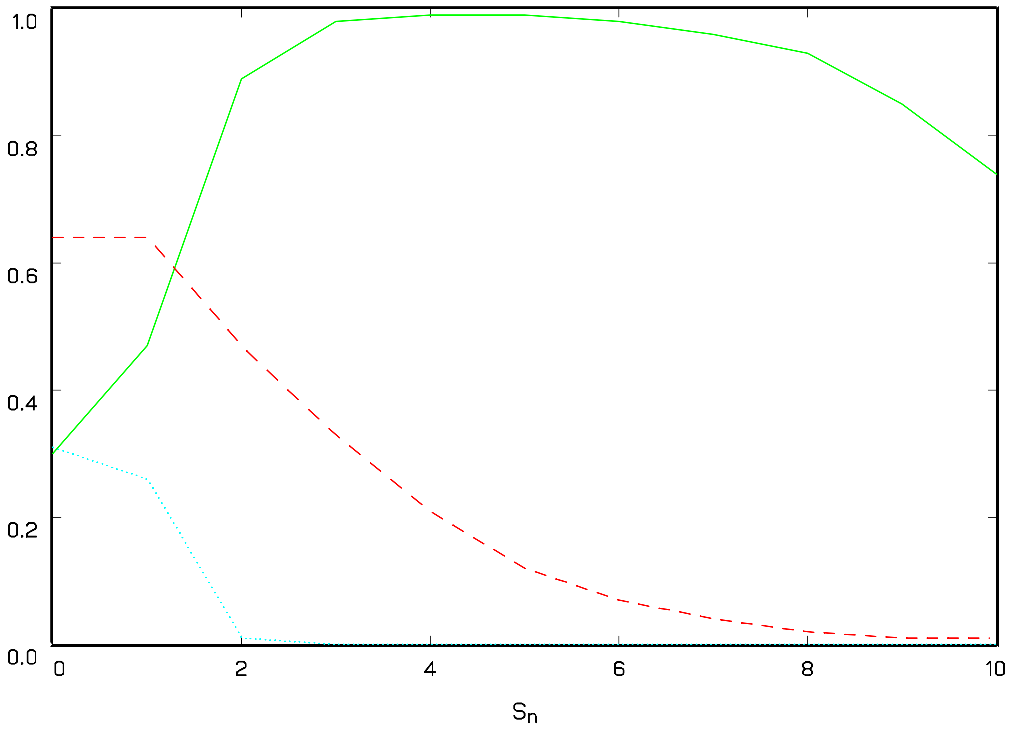
Wooldridge, J.M. and H. White (1988): "Some invariance principles and central limit theorems for dependent heterogeneous processes", *Econometric Theory* 4, 210–230.


I.1 Null ERPs; k=2, AR(1) phi=.5
I.2 Null ERPs; k=10, AR(1) phi=.5
I.3 Null ERPs; k=2, AR(1) phi=.9
I.4 Null ERPs; k=10, AR(1) phi=.9
LM_EL
Khat_EL
K
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
S_n

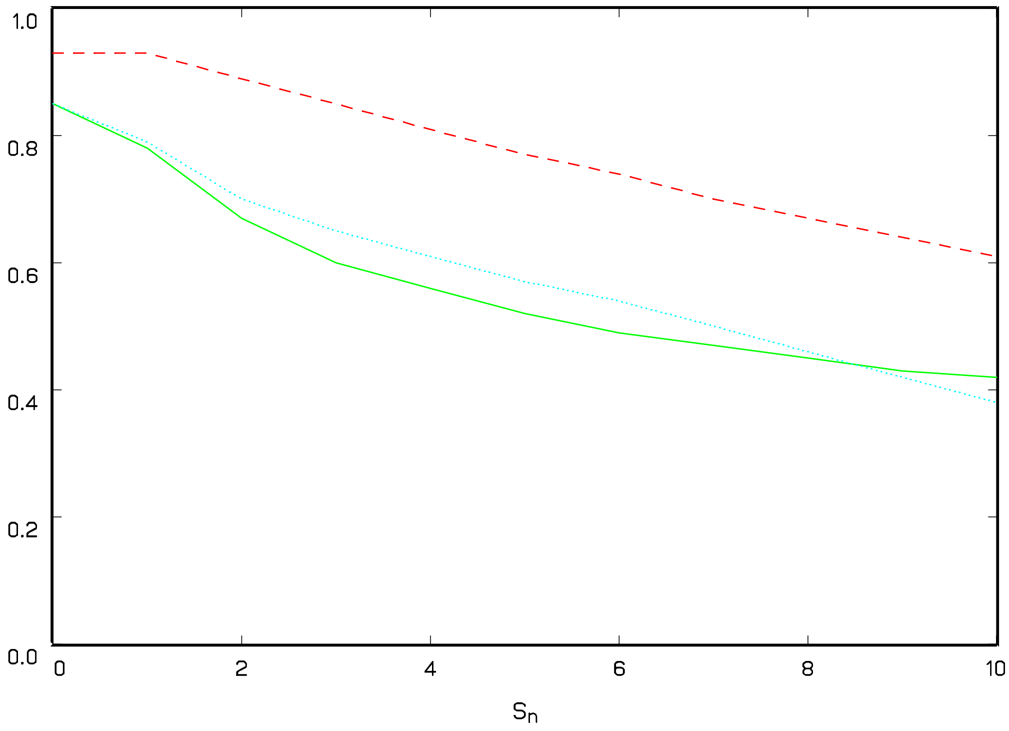
II.1 Power against -1; k=2, AR(1) phi=.5



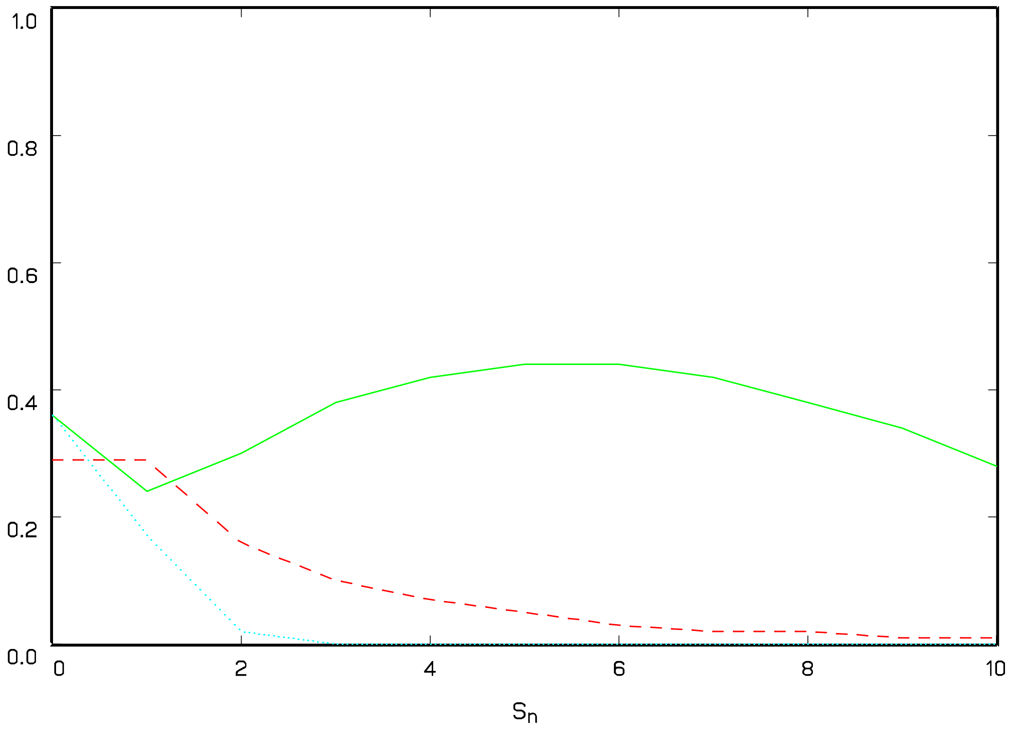
II.2 Power against -1; k=10, AR(1) phi=.5



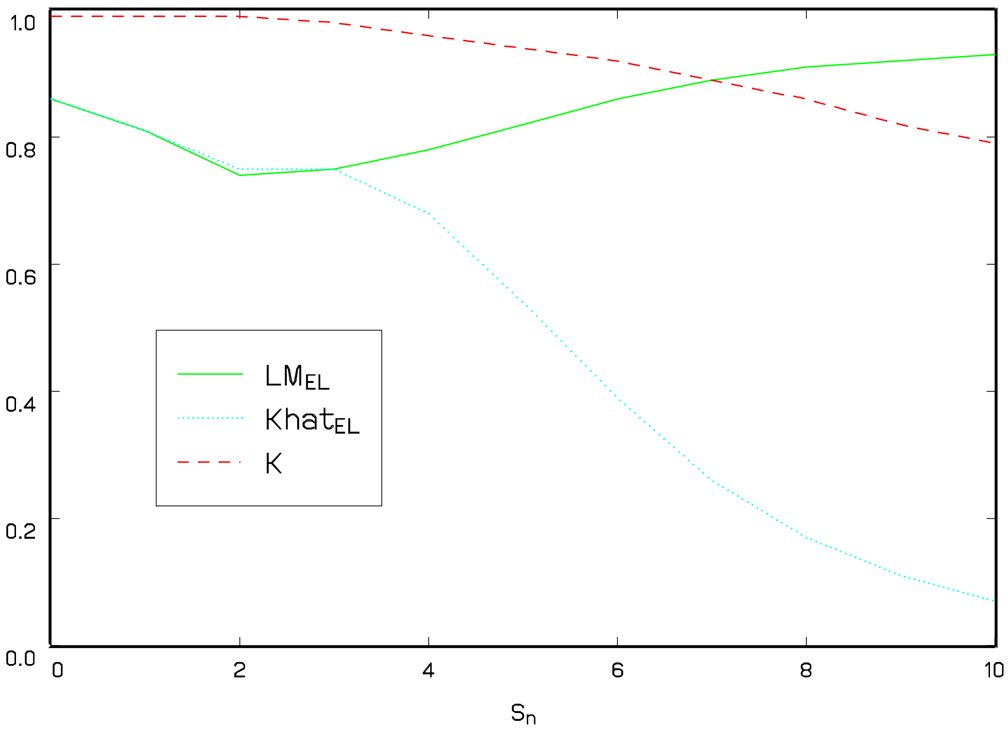
II.3 Power against -1; k=2, AR(1) phi=.9



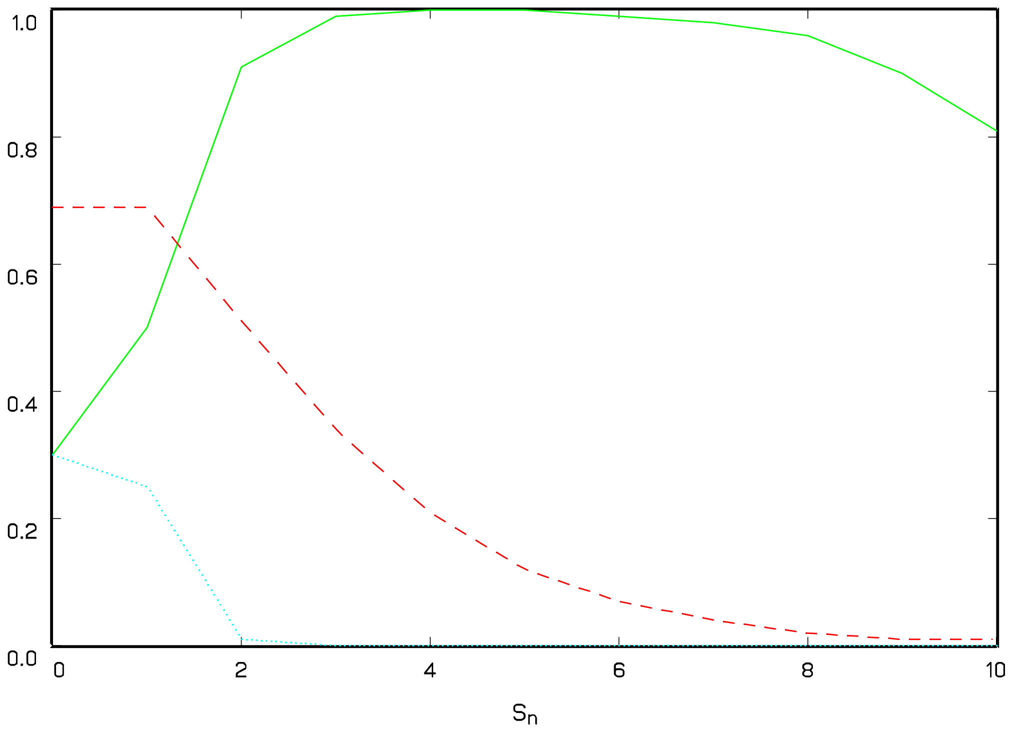
II.4 Power against -1; k=10, AR(1) phi=.9



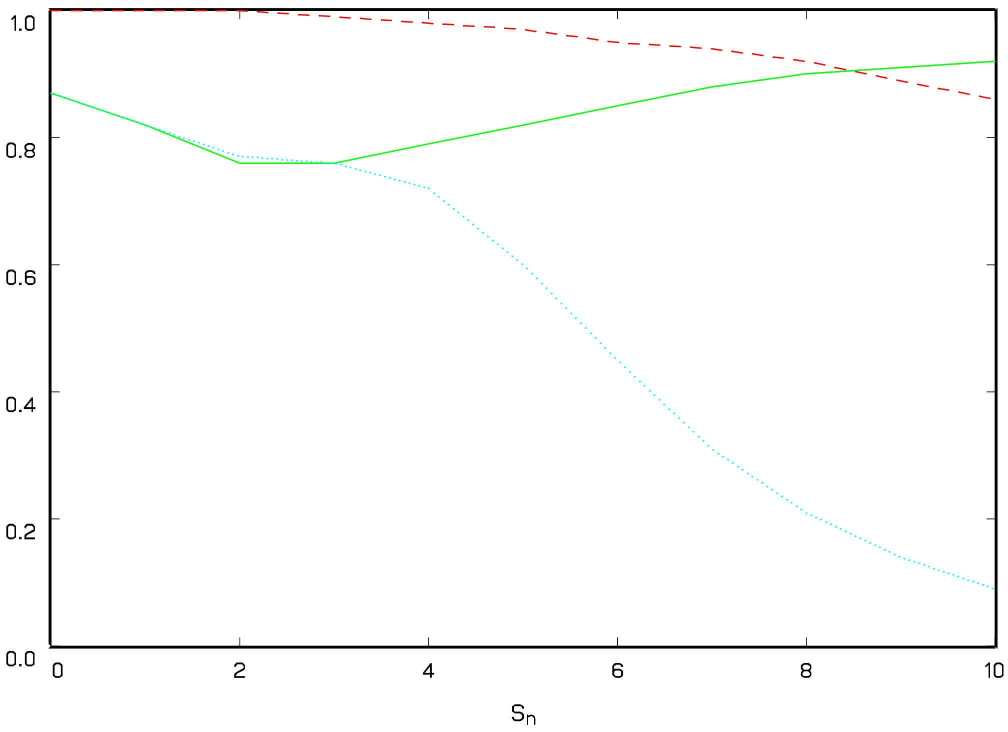
III.1 Power against −1; k=2, MA(1) v=.5



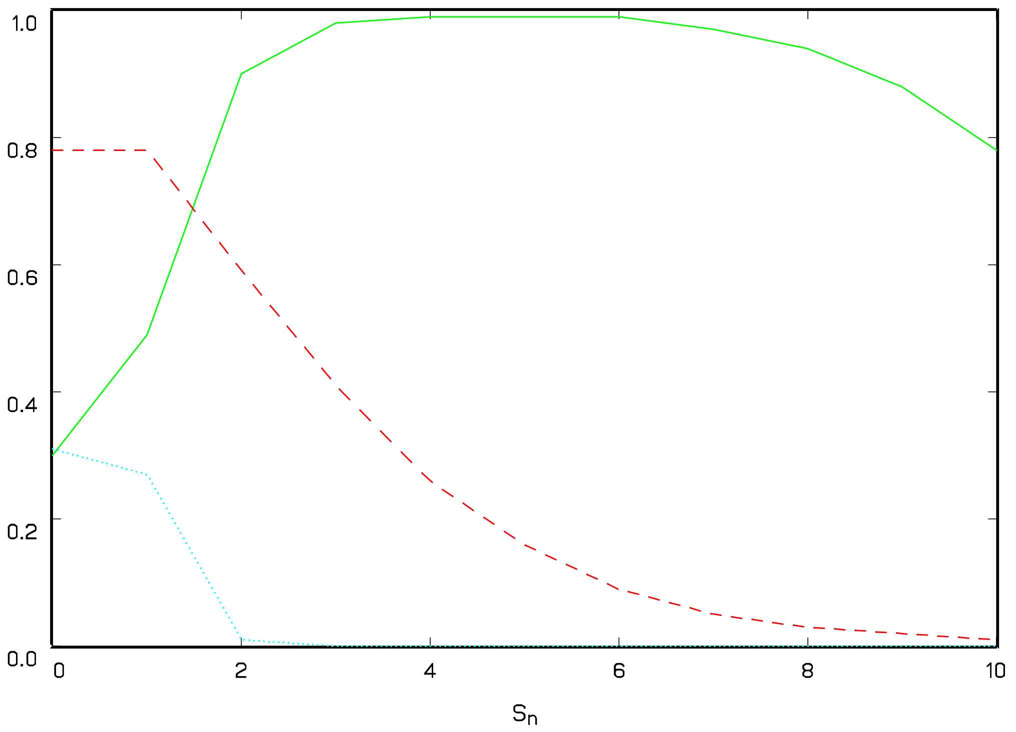
III.2 Power against −1; k=10, MA(1) v=.5



III.3 Power against −1; k=2, MA(1) v=.9



III.4 Power against −1; k=10, MA(1) v=.9


1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
S_n